Table of Contents
Manuals

SEER Program

Self Instructional Manual for Cancer Registrars

Book 7: Statistics and Epidemiology for Cancer Registries

777

NATIONAL INSTITUTES OF HEALTH
National Cancer Institute

SEER PROGRAM

SELF-INSTRUCTIONAL MANUAL FOR CANCER REGISTRARS

Book 7 - Statistics and Epidemiology for Cancer Registrars

SEER Program
Cancer Statistics Branch
National Cancer Institute

Authors

Evelyn M. Shambaugh, M.A., CTR
Cancer Statistics Branch
National Cancer Institute

John L. Young, Jr., Dr. P.H., CTR
Cancer Surveillance Section
California Department of Health Services

Calvin Zippin, Sc.D.
Department of Epidemiology and Biostatistics
University of California, San Francisco, CA

Diana Lum, B.S., CTR
Department of Epidemiology and Biostatistics
University of California, San Francisco, CA

Cheryl Akers, M.S.
East Lansing, Michigan

Mildred A. Weiss, B.A.
Los Angeles, CA

U.S. DEPARTMENT OF HEALTH and HUMAN SERVICES
Public Health Service
National Institutes of Health
NIH Publication No. 94-3766

Acknowledgements

We wish to extend special thanks to Pamela Derish for her work on Section F, Inferential Statistics, and Section G, Statistical Hypothesis Testing, and to Zarrin Navvab and Jennifer Seiffert for their help with reviewing all sections.

We also wish to thank Terry Swenson and Norma Guenterberg for their excellent technical assistance in setting up the equations and all of the tables, graphs, and special tables in the appendices in an eye-catching and understandable manner.

TABLE OF CONTENTS

Supplement

TABLES

FIGURES

SECTION A

OBJECTIVES AND CONTENT OF BOOK 7

SECTION A

OBJECTIVES AND CONTENT OF BOOK 7

I. GENERAL OBJECTIVES OF BOOK 7

A. Provide an introduction to basic biostatistical and epidemiological methodology.

B. Enhance the ability of the tumor registrar to prepare reports based on tumor registry data.

C. Increase the ability of the tumor registrar to understand statistical references in cancer literature.

D. Provide the requisite knowledge to meet the National Cancer Registrars Association (NCRA) Educational Standards.

E. Increase ability of the tumor registrar to provide assistance to the hospital cancer committee and other medical and research staff in the use of tumor registry data.

F. Provide definitions of statistical concepts and terms used in medical and epidemiologic literature.

Note:

If you wish to refresh your memory of arithmetic and algebra before you begin your study of statistics, turn to appendix 1, a refresher course on basic mathematics. Only high-school-level algebra is required.

II. CONTENT OF BOOK 7

If tumor registries are to be utilized to their fullest potential, the tumor registrars must be prepared to assemble and present statistical reports based on data contained in the registry system. The value of a tumor registry is determined primarily by use of the data. The success of the interaction between a tumor registry and its users depends upon the quality of the data and the facility with which the data can be retrieved and summarized.

Sections B-D are essential for all tumor registrars. These sections will be covered in the NCRA certification examination. Sections E-G are for tumor registrars who wish to increase their knowledge of the statistical methodologies and be able to carry out more complex analyses. These latter sections will not be included in the certification examination.

In the earlier manuals we learned how to collect and store data and maintain followup on cancer patients. In this manual we will learn how to assemble, summarize and analyze registry data.

"Statistics" is a branch of mathematics dealing with the collection, summarization, analysis, interpretation, and presentation of masses of numerical data. For cancer patients, statistics represent

counts or measurements of patient or disease factors. Statistical analysis is a means of summarizing the essential features and relationships of the data. Then, one can generalize to reveal the major characteristics of the patient group in order to determine broad patterns of behavior or tendencies.

Data can be prepared for presentation in the form of *tables* or *graphs*. Registry data may be presented by means that include:

- *Frequency distributions* (counts) or relative frequencies (percentages) which summarize the data according to variables such as primary site, stage, age, and sex
- *Measures of central tendency*, such as average or median age, or median survival time
- *Population-based measures* such as incidence and mortality rates
- *Survival curves* which can show trends in survival or make comparisons of survival for various groups of patients such as by sex, race, age, and histologic types

The interpretation of these analyses often requires that measures of *reliability* or *variability* be made for the results obtained.

The sections in this manual will discuss each of the aspects of preparing statistical reports. The scope and types of studies and reports will be determined by factors, such as:

- The registry setting, whether hospital-based or population-based
- The type of institution--community or teaching hospital, cancer center, central registry
- Data items collected by the registry
- Degree of dependability of selected data items
- The effect of coding changes over the years
- Target audience of the report
- Allocation of staff time for preparation of reports
- Ability of staff to conduct appropriate statistical tests and to interpret findings
- Geographic coverage of the registry
- Number of referrals to the hospital from outside the area
- Inclusion or exclusion of non-hospital cases i.e., outpatient cases

Some examples of the type of studies and reports that might be generated from tumor registry data are:

1. Hospital registry data

   a. Cancer control

      - Estimating patient accruals for treatment protocol studies
      - Assessing needs for screening programs
      - Assessing needs for community education programs
      - Assessing quality of patient care
      - Studying patterns of patient care in relation to short- and long-term outcome

   b. Physician education

      - Tumor conferences
      - Long-term follow-up reports

   c. Health care planning and administration

      - Studying patient's place of residence; defining service area, determining target population
      - Planning services and facilities: increase, reduce?
      - Studying utilization of services

   d. American College of Surgeons (ACoS) required reports

      - Hospital cancer program annual report
      - Ongoing patient care studies
      - Description of facility's cancer patient population
      - Monitoring quality of patient care in the facility

2. Central registry data

   - For a defined population, description of the kinds of cancers diagnosed and their importance in terms of incidence and survival
   - Study of cancer risks in a defined population
   - Study of cancer clusters
   - Identification of cases for research studies
   - Description of patterns of cancer patient care

The following references provide more detailed information on writing reports:

Fritz, A. Writing for Tumor Registrars. A Manual of Style, Elm Publications, Rockville, MD, 1987.

Cancer Program Manual. American College of Surgeons, Chicago.

Guidelines for Preparing a Hospital Cancer Program Annual Report. Tumor Registrars Association of California and American Cancer Society, California Division, 1986. (Distributed by the American College of Surgeons, Chicago.)

SECTION B

DESCRIPTIVE STATISTICS

SECTION B

DESCRIPTIVE STATISTICS

SELECTING, ASSEMBLING, PRESENTING, AND ANALYZING DATA

Defining the Problem

The first step in preparing a statistical report is to define the problem. The objectives and scope of the report must be defined at the outset. What information does the user want? What information is available in the registry? Are the data routinely collected by the registry, or will it require the collection of additional data?

Selecting the Cases

Once the objectives of the report are clearly identified, determine the cases to be included. For example, for a count of all cancer cases seen at a hospital for a given year, one would probably include both analytic and nonanalytic cases, alive and dead cases, and cases identified at autopsy only. For a count of all cancers occurring in a population covered by a state registry, one would limit the count to residents of that state who were first diagnosed during the period under study, including residents diagnosed out-of-state and those identified by death certificate only.

The criteria for inclusion may be limited to analytic cases, to cases of selected histologic types, to definitively-treated cases, to microscopically confirmed cases, to a certain age or ethnic group, or to residents of a defined geographic area. For example, the study group may be all patients under 15 years of age who had acute lymphocytic leukemia diagnosed between January 1, 1985 through December 31, 1989.

If all cases are not to be used, avoid bias[1] in the selection of cases. For example, if you report breast cancer patients by stage and omit those for whom there is no stage recorded, you are introducing a bias. Clearly define the population[2] to be studied. If only a sample[3] of cases is being used, be sure it is a random sample[4] to avoid bias.

If planning survival analyses, exclude those patients with cancers first identified at autopsy or identified by death certificate only. These are nonanalytic cases with no diagnosis date while alive, no treatment, and no survival. In-situ cases are generally excluded from survival reports. The excellent survival of in-situ cases camouflages the poorer survival of invasive cases.

---

[1]bias--The tendency of a statistical estimate to deviate in one direction from the true value.

[2]population--Any set of individuals (or objects) having some common observable characteristic that we are interested in studying.

[3]sample--A subset of the population under study.

[4]random sample--One in which every individual in the population has an equal and independent chance of being chosen for a sample.

For some studies it is desirable to select a random sample of cases. The most common aid to selecting a random sample is a table of random numbers (See appendix 2). A popular alternative to random selection is systematic selection, i.e., taking every fifth patient on a list. When a systematic selection is used, make sure that the number of items between successive selections does not correspond to some recurring cycle of cases. Sampling will be discussed in section F of this manual.

Determining the Data Items

Determine the types of information to be included and the availability and reliability of the data. If a data item is usually not available in the record, for example, occupation, then the information you collect will not be reliable. Select and define the variables[1] (usually the same as the data items) to be used, for example, age, race, sex, primary site, histologic type, stage, treatment modalities, and length of survival.

The data items or variables selected for a report will often vary from primary site to primary site. For example, of particular interest might be:

Histology (cell type) for leukemia, Hodgkin's disease, non-Hodgkin's lymphoma, brain, melanoma of skin

Sex distribution for lung, colon, bladder

Age distribution for leukemia, breast, kidney, brain, cervix/corpus

FIGO Stage for cervix, corpus, vagina, vulva

American Joint Committee on Cancer (AJCC) stage for breast, colon, bladder, melanoma of skin

Size of primary tumor for breast, oral cavity

Type of surgery for breast, colon/rectum, bone

Subsites for oral cavity, stomach, breast, colon/rectum

Assembling the Data

Before the process of assembling the data begins, it is a good idea to review previous studies and publications to get an idea of the expected results. This will help to set up categories, anticipate the range and concentration of values, and perhaps alert you to potential pitfalls.

The next step is to assemble the required information. This may involve going to computer file(s) or manually reviewing paper documents. In either case, some editing of the data may be necessary to ensure the quality of the recorded information.

---

[1]variable--A data item that can take on different values (vary).

Review preliminary tabulations for obvious errors or highly unusual cases. For example, male cervical cases, Wilms' tumor of the brain, or squamous cell carcinoma of the bone would all need to be reviewed, corrected, and the data retabulated.

Mutually Exclusive Categories

Summarizing the data involves setting up categories for the different variables and counting the number of cases that fall in each category, thereby creating a *frequency distribution.*

When grouping data into categories, the groupings should be mutually exclusive (each observation falls into one and only one category) and as a general rule should have between 6 and 15 classes.

In general, it is advisable to divide detailed data into a reasonable number of classes. If the number of classes is too few, important characteristics may be obscured. If there are too many classes with small frequencies, it may be difficult to see the underlying pattern, and some classes may contain no values. A proper balance must be struck so that the reader neither overlooks a relationship nor creates the effect of one by chance.

The values included in each class must be stated precisely to avoid ambiguity. Any of several methods of designating classes may be used depending in part on the nature of the data. The table below demonstrates four methods of designating classes of tumor size for breast cancer patients. Of the four methods, the one in column A is wrong, for it is ambiguous; it is not clear where a tumor of 2 cm should be counted. Column B clearly states the midpoint of each interval, but it is wrong because it is not clear what the limits of each class are. The class limits in column C are appropriate for *discrete* data only, that is, data that are recorded as whole numbers. The class limits in column D are the most suitable for *continuous* data when some values could include a decimal value.

Table 01. Examples of Classification

Classification for Tumor Size (in cm)			
WRONG	WRONG	CORRECT	CORRECT
A	B	C	D
0 - 2	1	0 - 1	<2.0
2 - 4	3	2 - 3	2.0 - 3.9
4 - 6	5	4 - 5	4.0 - 5.9
6 - 8	7	6 - 7	6.0 - 7.9
8 - 10	9	8 - 9	8.0 - 9.9
10+	11	10+	10.0+
Unknown	Unknown	Unknown	Unknown

It is highly desirable that all class intervals have the same width because equal intervals are easier to interpret. For example, it is preferable to have age categories of 5 years rather than unequal groupings such as 0-5, 6-15, 16-30, etc., although frequently childhood tumors are grouped together in the age group 0 - 14 followed by 10-year age groups thereafter, i.e., 15-24, 25-34, etc.

For some types of data, however, it may be desirable to use unequal intervals to summarize the data. For example, in a classification of breast cancers used to show a relationship between tumor size and prognosis, it may be more important to have narrower intervals for small tumors and wider intervals for the larger tumors such as:

<0.2 cm
0.2 - 0.4
0.5 - 0.9
1.0 - 1.9
2.0 - 2.9
3.0 - 3.9
4.0 - 4.9
5.0 - 9.9
10.0+

If the data have too many individual observations for easy analysis and presentation, the data may be grouped into broad intervals as below.

Table 02. Data Grouped into Broad Age Intervals for Survival Rates

Brain and Nervous System	Melanoma of Skin	Colon/Rectum
<15	<25	<45
15-24	25-34	45-54
25-34	35-44	55-64
35-44	45-54	65-74
45-54	55-64	≥75
55-64	65-74	
65-74	≥75	
≥75		

Because brain tumors arise in children, we have a <15 years of age group and then 10-year age groups. Melanomas are frequent in adults beginning at age 25, so we begin with 10-year age groups at age 25. Colon/Rectum cancers arise and become more frequent at about age 45.

Q1

What is the first step in preparing a statistical report? _____

_____.

Q2

Only after you define the problem and determine your variables, can you select the _____ to be included.

Q3

Match the terms on the left with the description on the right:

____1.	random sample	a.	Tendency of a statistical estimate to deviate from the true value
____2.	population	b.	Every individual has an equal and independent chance of being chosen
____3.	bias	c.	A subset of the population under study
____4.	sample	d.	A set of individuals having some common observable characteristic

Q4

Indicate which of the following categories are mutually exclusive (ME) and clearly defined.

A	B	C	D
0 - 15	<10	10	0 - 10
15 - 30	10.0 - 20.0	20	11 - 20
30 - 45	20.1 - 30.0	30	21 - 30
45 - 60	30.1 - 40.0	40	31 - 40
60+	40.1 - 50.0	50	41 - 50
	50.1+	51+	

Answer: Q1

You might have said the first step in defining a statistical report is defining the problem or defining the objectives.

Answer: Q2

Only after you define the problem and determine your variables, can you select the cases to be included.

Answer: Q3

b	1.	random sample	Every individual has an equal and independent chance of being chosen.
d	2.	population	A set of individuals having some common observable characteristic
a	3.	bias	Tendency of a statistical estimate to deviate from the true value
c	4.	sample	A subset of the population under study

Answer: Q4

No	A	(Not mutually exclusive)
Yes	B	
No	C	(Not clearly defined)
Yes	D	(However, D is not all inclusive because it contains no mention of any value greater than 50.)

It often happens that what is needed is not so much the count of patients which fall into each class but rather the *relative frequency* which is the percentage distribution. This is illustrated in the table below.

Table 03. Example of a Percentage Distribution

Percentage Distribution of Acute Lymphocytic Leukemia Patients by Age and Sex, Community Hospital, 1989		
Age in years	Male	Female
	Percent of Cases	
All Ages	100.0%	100.0%
0 - 14	55.3	53.4
15 - 24	14.5	13.1
25 - 34	4.9	4.0
35 - 44	4.4	4.8
45 - 54	2.0	3.6
55 - 64	5.1	5.2
65 - 74	5.4	7.0
75 - 84	6.0	6.3
85+	2.4	2.6

Relative frequency: The number in each subcategory divided by the total number in the class, then multiplied by 100. In our example, to arrive at 55.3%, you would have to know that there were 341 males in the subcategory age 0-14 out of a total number of 617. Then 341/617 = .553 x 100 = 55.3%. If data are to be compared with other series, the categories must be the same, for example, the same age groups, stage groupings, treatment categories.

There are different *kinds of data* which will influence the setting up of categories.

If a variable can have only a particular (limited) set of values, it is called *discrete*. For example, the number of children in a family is an example of discrete data. A family may contain two children or three children, but 2 1/4 or 3 1/2 children is impossible.

If a variable can have different or more precise values with successive refinements of the measuring scale, it is called *continuous*. For example, height is continuous data. You might say someone was approximately 6 feet tall, then refine it to 5 feet 10 inches, and then refine it further to 5 feet 10 1/2 inches tall.

Presenting the Data

The presentation of the data will depend on the purpose of the study. If the purpose requires only counts, percents, or relationships of the patient characteristics, the data may be presented in the form of a *table* or a *graph*.

Tables or Graphs: Advantages and Disadvantages

Ever since records were first kept, there has been the problem of understanding numerical data. Statistical tables were developed for summarizing data and graphs for presenting relationships in data in visual form.

Too often data are presented in an awkward or confusing format. By following certain simple rules, it should be possible to present the data with maximum effectiveness.

The question of whether to present data in the form of a table or a graph depends on the purpose and the audience.

Tables have the following advantages over graphs:

- More information can be presented.
- Exact values can be read from a table to retain precision.
- Less work and less cost are required in the preparation.
- Flexibility is maintained without distortion of data.

On the other hand, graphs have the advantage of:

- Attracting attention more readily
- Being more easily understood
- Showing trends or comparisons more vividly
- Being more easily remembered.

In short, one picture (graph) is worth a thousand words. However, in some studies it may be advantageous to give both the detailed table and a simple summary graph. Graphs can bring out hidden facts and relationships which stimulate analytic thinking, but tables provide the supportive details. Together they present a better balanced understanding.

Tables and graphs should not be presented alone, but should be accompanied by explanatory narrative. Significant results and relationships should be pointed out for the reader who does not have the time or opportunity to analyze the raw data.

Comparisons of groups of patients are often made in terms of measures of central tendency, such as the arithmetic average or mean, the median, or the mode. Variability in the data is described by measures such as the range or the standard deviation. These measures are discussed later in this section.

If the data in the registry cover all cases in a known population base, for example, a central registry which collects ALL cases within the population residing in a defined geographic area, it is then possible to compute incidence rates.[1] This type of summary description is a basic descriptive tool for epidemiologists--a stepping-stone to the study of possible causes of cancer, such as environmental factors, genetic differences, and host differences. The calculation of incidence rates is discussed in section C.

Often a hospital registry will want to calculate patient survival rates as a means of evaluating progress in treatment of patients. Survival rates are discussed in section D.

Analyzing the Data

Prerequisites to the analysis of registry data include:

1. Checking for completeness of casefinding
2. Editing of the data for abstracting and coding errors
3. Reviewing inconsistencies between fields
4. Determining omission of data or data items
5. Resolving questionable entries

Before the data are presented in final form, the process of analysis must take place. This will include deciding on the appropriate format for tables and graphs, e.g., age groups, time intervals, treatment categories. It may also involve the choice of appropriate statistical measures which will be discussed in other sections.

Most tumor registry reports provide a summary or description of cases collected by the registry, i.e., descriptive statistics.[2] For example, for all the cancers in the registry, a table providing the number of cases in each site by sex, race, and age would be called a summary table. If the analysis is to include inferential statistics[3] or the interpretation of population-based data, the tumor registrar will probably want to seek assistance from a statistician or an epidemiologist, for example, regarding inferences about the influence of the Mormon life-style on incidence of certain sites of cancer in Utah.

---

[1]incidence rates--Rate of occurrence of new cases that are diagnosed during a set time period in a defined population.

[2]descriptive statistics--Numerical summaries which describe an observed frequency distribution (i.e., mean, median, variance, range, etc.).

[3]inferential statistics--Sample statistics which estimate population statistics.

There are certain precautions which must be taken in analyzing registry data.

- Avoid faulty generalizations. Don't jump to conclusions or generalizations on the basis of too small a sample or a sample not typical of the whole population.

- Avoid comparison of dissimilar data, such as comparing observed survival rates for children under 15 years of age with those for persons over 65. Other causes of death must be taken into account in comparing different age groups.

- Provide clear definitions and a complete description of the demographic and disease characteristics of the cases included in the study. If any cases were excluded, be sure the exclusions are clearly documented.

- Follow the usual conventions for calculating survival, incidence, and mortality rates, and specify the methods used.

Q5

Indicate the proper order of work in preparing a statistical report by numbering the order of work for the following:

____ Selecting the cases

____ Defining the problem

____ Presenting the data

____ Analyzing the data

Q6

Match the purpose of the study on the left with the most appropriate method of presentation on the right:

_____	1.	Counts of breast cancer	a.	Survival rates of patients by age
_____	2.	Comparisons of characteristics of males and females with lung cancer	b.	Measures of central tendency such as average or median values
_____	3.	Comparison of successes of various treatment groups	c.	Frequency distributions in a table

Q7

Very complex detailed data can only be completely presented in a _____. Relationships in data can be emphasized more vividly by using a _____.

Q8

Indicate whether a table (T) or a graph (G) is the preferred method of presentation in the following situations:

_____ a. Frequency distribution by site, sex, race, and time period of all cancers in your institution

_____ b. Survival trends over time by sex for lung cancer

_____ c. Presentation by stage of disease of female breast cancer to illustrate a talk

_____ d. Detailed treatment distribution of cervical cancer for a doctor on the staff at your hospital

Q9

Precautions in use of registry data:

a. _____

b. _____

c. _____

d. _____

e. _____

Q10

If your sample is too small, your _____ may be faulty.

Answer: Q5

The proper order of work in preparing a statistical report is as follows:

1. Defining the problem
2. Selecting the cases
3. Analyzing the data
4. Presenting the data.

Answer: Q6

c 1. Counts of breast cancer patients: Frequency distributions in a table

b 2. Comparisons of males and females with lung cancer: Measures of central tendency such as the average or median values

a 3. Comparison of successes of various treatment groups: Survival rates

Answer: Q7

Very complex detailed data can only be completely presented in a table.

Relationships in data can be emphasized more vividly by using a graph.

Answer: Q8

Indicate whether a table (T) or a graph (G) is the preferred method of presentation in the following situations:

T a. Frequency distribution by site, sex, race, and time period of all cancers in your institution

G b. Survival trends over time by sex for lung cancer

G c. Presentation by stage of disease of female breast cancer to illustrate a talk

T d. Detailed treatment distribution of cervical cancer for a doctor on the staff of your hospital

Answer: Q9

Precautions in use of registry data:

a. Bias in selecting cases
b. Faulty generalizations
c. Comparison of noncomparable data
d. Unclear definitions
e. Improper use of survival, incidence, and mortality rates

Answer: Q10

If your sample is too small, your generalizations/conclusions may be faulty.

PREPARING TABLES

A table is an orderly arrangement of values which groups data into classes. Variables such as vital status, race, age, treatment, and stage of disease have a system of classification. Vital status has two classes while age can have any number depending on age groupings. The method of constructing a table depends to some extent on the manner in which the data are arranged. It may be useful to obtain the counts of cases with all possible values of the variable in logical order or in order of frequency. For example, if you are counting patients by age at diagnosis, then count the number of patients at each single year of age with age arranged in numerical order from youngest to oldest. You may then want to combine categories, e.g., patients 45-54 years old.

All table captions with the possible exception of brief text tables that are an integral part of the narrative should contain certain essentials.

Title

Number and Percent of Lung Cancer Patients by Age and Sex, Diagnosed At Community Hospital 1985-89						
Age (Stub Head)	Sex (Boxhead)					
	Total (Column Headings)		Male		Female	
	No.	%	No.	%	No.	%
All ages	<----------Row---------->					
<45 (Stub)	cell		↑ Column ↓			
45-54	cell					
55-64	cell					
65-74	cell					
75+	cell					

Footnote:

Source:

Essential Components of Tables

TITLE: The title must tell as simply as possible what is in the table. It should answer the questions:

- What are the data?--Counts; percentage distributions; rates
- Who?--White females with breast cancer; black males with lung cancer
- Where are the data from?--Your hospital; the entire state
- When?--A particular year; time period.

For example: Site Distribution by Age and Stage of Cancer Patients First Admitted to General Hospital in 1989

BOXHEAD: The boxhead contains the captions or column headings. The heading of each column should contain as few words as possible yet explain exactly what the data in the column represent.

STUB: The row captions are known as the stub. Items in the stub should be grouped to facilitate interpretation of the data. For example, group ages into 5-year age groups.

CELL: The box formed by the intersection of columns and rows

FOOTNOTE: Anything in a table that cannot be understood by the reader from the title, boxhead, or stub should be explained by footnotes. Footnotes contain information on missing numbers, preliminary or revised numbers, or explanations for any unusual numbers. Definitions, abbreviations, and/or qualifications for captions or cell names should be footnoted. A footnote usually applies to a specific cell(s) within the table, and a symbol, such as "*" or "#" may be used to key the cell to the footnote. If several footnotes are required, it is better to use small letters rather than symbols or numbers. Footnote numbers may be confused with the numbers within the table.

SOURCE: If data from a source outside the registry are used, the exact reference to the source should be given.

Denoting the source lends authenticity to the data and enables the reader to locate the source if further information is desired.

In the preceding diagram, sex is labeled horizontally in the BOXHEAD and age is labeled vertically in the STUB. The individual entries which are classified according to the row and column are called cells. The totals represent the distribution of age (or sex) alone, while the data in the cells represent the interaction of age with sex.

Tables usually are arranged so the length exceeds the width; it is generally better to use the longer wording in the stub. Important numbers to be compared should be placed in adjoining columns or rows. Time series are listed in chronological order, beginning usually with the earliest time period; classifications of numbers are usually listed from smallest to largest; traditional listings such as anatomical sites are usually listed in ICD-O order. For emphasis, the order may be changed to another order, such as the relative frequency of occurrence, e.g., the ten most common sites might be arranged in order by relative frequency, or for a non-technical audience, arranged alphabetically.

Cross-classified tables must always account completely for the data being classified. For this reason unimportant classes are put in a composite class labeled "Other." The "Other" categories are placed to the right or the bottom of the rows or columns, respectively.

Many analytic tables contain both numbers of cases and percentage distributions. Numbers provide information on magnitude; percentages facilitate comparisons.

Check the table to be sure that:

- It is a logical unit. (Separate analyses call for separate tables.)
- It is self-explanatory. (Can it stand alone if it is photocopied and removed from its context?)
- All sources and units are specified.
- Headings are specific and understandable for every column and row.
- Rows and columns add up to totals.
- No cell is left blank (Enter "0" or "-").
- Categories are mutually exclusive (do not overlap) and all inclusive.

Types of Tables

The registry may be called upon to prepare two types of tables:

REFERENCE TABLES are detailed so as to provide complete information (e.g., tabulation of all cancer cases seen at your hospital cross-classified by site, sex, race, and stage). They are not intended to be read through, but are presented so that source data are available.

SUMMARY TABLES are designed to present specific data for a particular use. In the process of preparing a summary table, it is often desirable to:

- Use only the most important categories (e.g., all stages and localized).
- Use grouped data instead of detailed (e.g., total colon instead of subsites of colon).
- Round off to whole numbers.
- Place the most important numbers (e.g., totals) at the top on the left for emphasis.
- Place data being compared in adjacent positions (e.g., male and female comparisons).

Often summary numbers other than frequencies or percents are presented to facilitate the interpretation of a table (e.g., average age or ratio of males to females).

Tables from which slides are made should be kept as simple as possible.

Construction of Tables

In table construction, good judgment is more important than blind adherence to rules. Present the data in a format to illustrate a specific idea. For complex tables, it is useful to construct the table in several different formats to see which one illustrates the idea the best. Do not be afraid to discard meaningless tables.

The simplest table is a one-way classification in which one variable, for example, sex, is presented either in terms of numbers of cases or a percentage distribution or both. Table 06 has both.

Table 04. A One-Way Classification--Numbers of Cases

Number by Sex of Children Age 0-14 With Acute Leukemia Diagnosed at Community Hospital, 1989	
Sex	Number of Cases
Total	50
Male	30
Female	20

Table 05. A One-Way Classification--Percentage Distribution

Percentage Distribution by Sex of Children Under Age 15 With Acute Leukemia Diagnosed at Community Hospital 1989	
Sex	Percent
Total	100%
Male	60
Female	40

Table 06. A One-Way Classification with Both Numbers of Cases and a Percentage Distribution

Number and Percentage Distribution by Sex of Children Age 0-14 With Acute Leukemia Diagnosed at Community Hospital, 1989		
Sex	Number of Cases	Percent
Total	50	100%
Male	30	60
Female	20	40

If a classification is desired according to two characteristics simultaneously, they are cross-classified in a *two-way* table. One classification will appear horizontally (sex) and the other vertically (histology) as shown in the table below:

Table 07. A Two-Way Classification

Number of Cases of Cancer of the Lung and Pleura by Histology and Sex First Admitted to General Hospital, 1989			
		Sex	
Histology	Total Cases	Male	Female
Total	261	159	102
Squamous Cell Carcinoma	54	37	17
Adenocarcinoma	94	58	36
Bronchiolar Carcinoma	17	9	8
Small/Oat Cell Carcinoma	40	22	18
Adenosquamous Carcinoma	12	5	7
Mesothelioma	6	4	2
Other	38	24	14

The following three tables (08 - 10) illustrate two-way classifications using a variety of variables.

Table 08. A Two-Way Classification

Number of Cases of Cancer of the Head and Neck By Site and Year of Admission, University Hospital, 1985-89						
		Year of Admission				
Primary Site	Total	1985	1986	1987	1988	1989
Total	823	148	177	185	161	152
Lip	13	2	1	2	2	6
Tongue	125	22	30	26	24	23
Major Salivary Glands	66	12	9	13	21	11
Gums	29	3	6	8	6	6
Floor of Mouth	55	9	14	13	6	13
Other Mouth	73	10	11	21	13	18
Oropharynx	60	13	14	12	14	7
Nasopharynx	30	5	6	9	4	6
Hypopharynx	50	6	11	12	11	10
Nasal Cavity/Sinuses	75	18	17	19	12	9
Larynx	223	42	51	46	46	38
Nonspecific Oral Cavity	24	6	7	4	2	5

Table 09. A Two-Way Classification

Number of Leukemia Cases by Histology and Age Memorial Hospital, 1989								
Histology	Total Cases	Age						
		<15	15-34	35-44	45-54	55-64	65-74	75+
All histologies	209	29	67	40	21	20	21	11
Acute Lymphocytic	61	24	22	11	0	1	1	2
Chronic Lymphocytic	25	0	1	1	6	8	6	3
Acute Granulocytic	52	2	16	16	5	6	5	2
Chronic Granulocytic	46	0	19	10	6	3	6	2
Monocytic	2	1	1	0	0	0	0	0
Myelomonocytic	7	0	4	0	1	1	1	0
Hairy Cell	10	0	1	2	3	1	2	1
Other and Unspecified	6	2	3	0	0	0	0	1

NOTE: Includes analytic cases only

Table 10. A Two-Way Classification

Number of Cases of Melanoma of the Skin by Histology and Stage Initially Diagnosed or Treated at Memorial Hospital 1985-89							
Histology	Total Cases	AJCC Stage					
		0	I	II	III	IV	NR
Total	373	18	184	75	5	6	85
Lentigo Maligna	36	8	14	5	0	0	9
Superficial Spreading	170	4	111	34	0	1	20
Nodular	52	0	23	15	2	1	11
Acral Lentiginous	11	0	2	4	2	0	3
Malig. Melanoma, NOS	90	5	30	14	1	4	36
Other Specified Melanoma	14	1	4	3	0	0	6

Note: NR = Not recorded

When three or more classifications of the data are desired, the problem becomes more difficult. This multidimensional relationship must be shown on a two-dimensional sheet of paper. Table 11 illustrates a *three-way* classification in which the row categories are subdivided by race.

Table 11. A Three-Way Classification

Number of Lung Cancer Patients by Sex and M/F Ratio
for Whites and Blacks by Age
General Hospital
1985-89

Race and Age	Total	Male	Female	M/F Ratio
White				
All ages	520	316	204	1.5
<45	40	20	20	1.0
45-54	66	36	30	1.2
55-64	158	92	66	1.4
65-74	162	112	50	2.2
75+	94	56	38	1.5
Black				
All ages	50	35	15	2.3
<45	2	2	-	0.0
45-54	10	6	4	1.5
55-64	20	14	6	2.3
65-74	14	10	4	2.5
75+	4	3	1	3.0

Often it is desirable to include summary information in a table. For example, in table 11 the ratio of males to females might be pertinent to the discussion and could be added to the caption entries as in the above example.

Table 12 illustrates a four-way classification of data where the rows are subdivided by histology and the columns by race, then by sex.

Table 12. A Four-Way Classification

Percentage Distribution of Leukemia Cases By Chronicity, Morphologic Classification, Race, and Sex, University Hospital, 1985-89						
Chronicity and Morphology	Black			White		
	Male + Female	Male	Female	Male + Female	Male	Female
Number of cases	571	320	251	7799	4510	3289
Percent:						
Total Acute	50 %	48 %	53 %	50 %	48 %	53 %
Acute Lymphocytic	14	14	15	14	14	14
Acute Myelocytic	16	14	17	15	14	17
Monocytic	10	9	12	11	10	12
Acute, NOS	10	11	9	10	10	10
Total Chronic	50	52	47	50	52	47
Chronic Lymphocytic	28	30	24	28	32	25
Chronic Myelocytic	21	21	21	19	18	19
Leukemia, NOS	1	1	2	3	2	3

Note: NOS = Not otherwise specified

Whenever further subdivision of data leads to tables which are too complex to be read easily, it is preferable to increase the number of tables. Reference tables, which may be even more complex, should be presented at the end of the report.

Q11

Why should there be complete documentation of tables and graphs?

_____.

Q12

What are the four essential components of the title of a table or graph, all of which begin with "W"? 1. _____

2. _____

3. _____

4. _____

Q13

The best medium for presenting data for quick visualization is:

[] A table

[] A graph

[] An abstract

[] The medical record

Q14

Indicate whether the following types of tables are reference (R) or summary (S) tables.

_____ a. Stage distribution for white females with breast cancer for your state in 1986

_____ b. Number and percent distribution of all cancer cases seen at your institution in 1986-87 by site, sex and age group

_____ c. Sex distribution of lung cancer from 1960-85 for your hospital

Q15

If you wish to classify data according to two variables simultaneously, such as sex and age, prepare a _____ -way table with one variable appearing _____ and the other variable appearing _____.

Q16

When a detailed cross-classification of more than two variables is to be presented in tabular form, list two possible methods of presentation. 1._____ and 2. _____.

Q17

In your own words what does it mean when you say "the classes should be mutually exclusive"?

_____

Q18

You may prefer to present the percent of patients which fall into each class rather than the count of patients. This is called a _____ _____.

Q19

Anything in a table that cannot be understood by the reader from the title, captions, and/or stub should be explained by a _____. Examples of such information are _____, _____ _____, and _____ _____.

Q20

If you use data from outside your institution for comparative purposes, always indicate the _____ of the data.

Answer: Q11

You might have said that there should be complete documentation of tables and graphs so they can stand alone, or if the tables and graphs are separated from the text, you know to what they refer.

Answer: Q12

The four essential components in the title of any table or graph are:

1. What
2. Who
3. Where
4. When

Answer: Q13

The best medium for presenting data for quick visualization is a graph.

Answer: Q14

Indicate whether the following table is a reference table (R) or a summary table (S).

S a. Stage distribution for white females with breast cancer for your state in 1986

R b. Number and percent distribution of all cancer cases seen at your institution in 1986-87 by site, sex, stage, treatment, and age group

S c. Sex distribution of lung cancer from 1960-85 for your hospital

Answer: Q15

If you wish to classify data according to two characteristics simultaneously, such as, sex and age, prepare a two-way table with one characteristic appearing horizontally and the other characteristic appearing vertically.

Answer: Q16

1. A three-way or four-way classification table.
2. More than one table.

Answer: Q17

To say that classes should be mutually exclusive means that each entry can appear in one and only one cell.

Answer: Q18

You may prefer to present the percent of patients which fall into each class rather than the count of patients. This is called a relative frequency.

Answer: Q19

Anything in a table that cannot be understood by the reader from the title, captions and/or stub should be explained by a footnote. Examples of such information are abbreviations, missing numbers, and revised numbers.

Answer: Q20

If you use data from outside your institution for comparison purposes, always indicate the source of the data.

TYPES OF GRAPHS AND THEIR CONSTRUCTION

A graph is the best medium for presenting data for quick visualization of relationships between various factors. Graphs effectively emphasize the main points in an analysis and clarify relationships which might otherwise remain elusive.

There are many types of graphs from which to choose: bar graphs, histograms, frequency polygons, line graphs, pie charts, scatter diagrams, and pictograms. The type of graph used will depend on the type of data.

Choosing the Right Graph

Selecting the most appropriate graph(s) to accompany your data will add a lot to the effectiveness of your presentation. On the other hand, an overabundance of graphs, or graphs which do not demonstrate anything in particular, should be avoided. The identification of specific relationships or trends inherent in the data by means of well designed graphs will have the greatest appeal for the reader. It is a good idea to lay out several versions of a graph and to use the one that turns out to be the most illuminating.

Computer Graphics

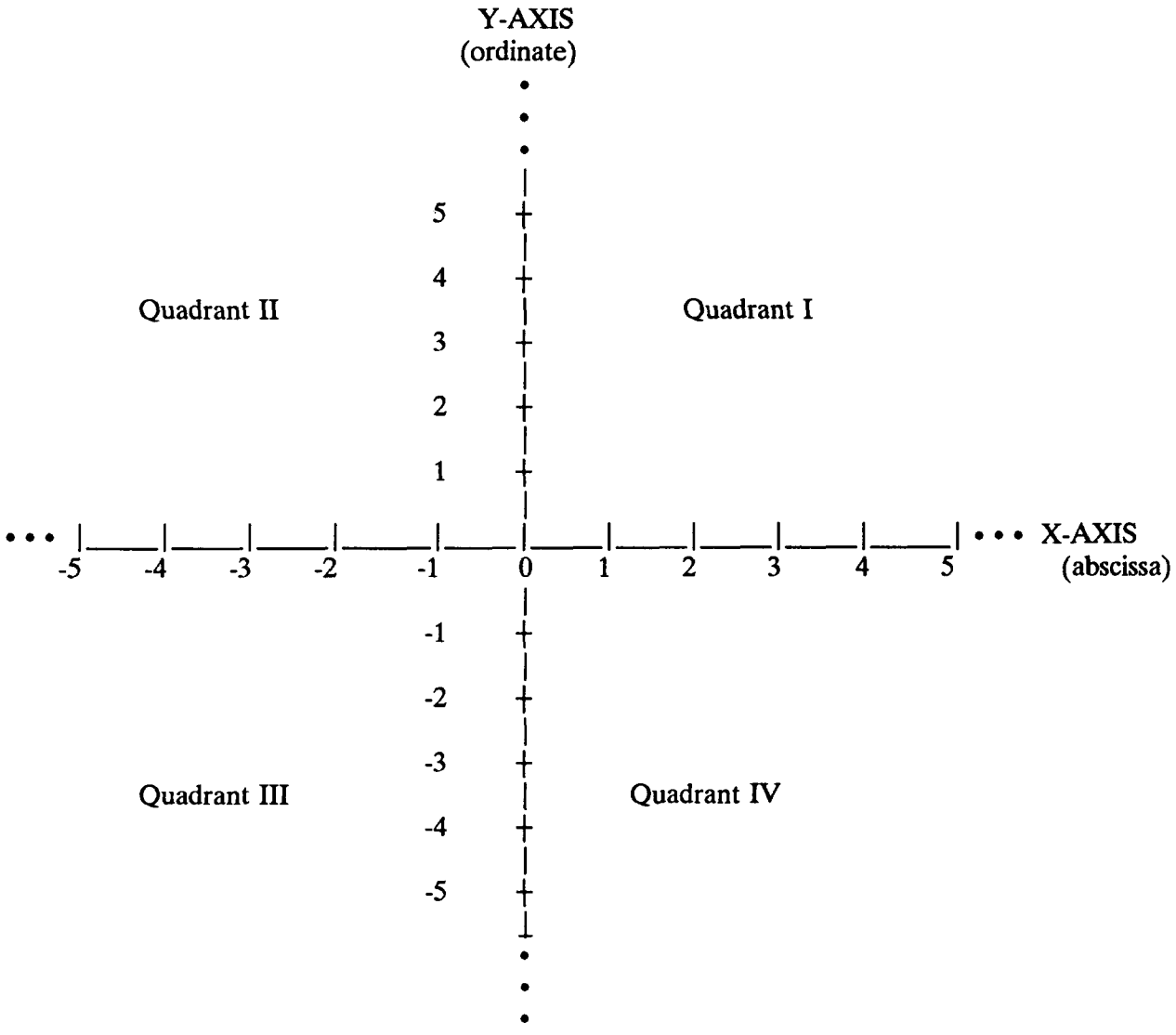
The availability of computer software tailored for tumor registry data enables computerized registries to produce attractive graphs quite easily. A choice of graphics packages is available on the market.

Whether the graphics used in reports are produced manually or by computer, the basic principles of design and construction are the same. For manual registries, a variety of drawing materials and graphic aids are to be found in artist supply stores and stationers.

Construction of Graphs

The basic form of a graph is usually constructed by plotting numbers in relation to two axes. A scale is arranged in both directions from a zero point at the intersection of the axes. The Y-axis (vertical) is called the ordinate and the X-axis (horizontal) is called the abscissa. Most graphs use positive values only, thus only the upper right-hand part of the grid (quadrant I) is usually shown. "Tic" marks are used to indicate the grid lines in the example below. The axes are marked off in equal units and may be extended as far as necessary in any direction.

Figure 01. Basic Graph Format



Essential Components

TITLE: The title must tell as simply as possible what the graph shows. It should answer the same questions as the title for a table.

- What are the data?--Counts; percentage distributions; rates
- Who--White females with breast cancer; black males with lung cancer
- Where are the data from?--One hospital; the entire state
- When?--A particular year; a time period.

LEGEND or KEY: When several variables are included on the same graph, it is necessary to identify each by using a key or legend. Place the legend in a clear space on the face of the graph and identify each line or bar on the graph as in the example below.

Line graph

White Males • • • • • • • •
Black Males - - - - - - - -
White Females •••-•••-•••-•••
Black Females ---•---•---•---•

Bar graph

Males  (Crosshatch)

Females  (Stripes)

Although different colors may be used for lines or bars, different patterns should still be used so that a photocopy will differentiate.

SCALE CAPTIONS: Scale captions are placed on both axes to identify the scale values clearly. It is essential that both the subject and the units used be identified. The caption for the horizontal scale is generally centered under the X-axis. The caption for the vertical axis is placed either at the top left of the Y-axis or along the Y-axis, whichever is the easier to read. The Y-axis is most often used for frequency or relative frequency; the X-axis for category.

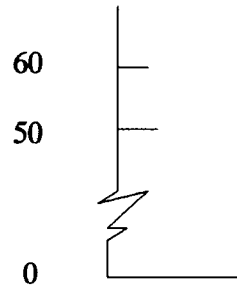
- The scale of values for the X-axis reads from the lowest value on the left to the highest value on the right.
- The scale of values for the Y-axis extends from the lowest value at the bottom to the highest value at the top of the graph.

FOOTNOTES: If the title, scale labels, and legend cannot explain everything in the graph, then footnotes should be used as in tables.

SOURCE: The exact reference to an outside source should be given just as for tables.

The scales should be set to fit the data. Comparisons can be magnified or minimized depending on the size of the scale. When setting up a graph, lay it out on graph paper allowing for a margin on all sides.

The zero point should appear on the vertical scale whenever possible. If this results in a large gap between the lowest value and 0, a scale break may be used. For example:



This technique is most often used in a line graph.

Like a table, a graph must be complete enough to stand alone when it is photocopied and read out of context.

A table sometimes accompanies a graph, or actual numbers are entered on the graph, so that the reader may see the numbers on which the graph is based.

Types of Graphs

BAR GRAPH

Frequencies, proportions, or percentages of categorized data are often displayed using bar graphs. They are easy to construct and can be readily interpreted. Bars are effective for showing the component parts of a whole and for making comparisons between groups such as number or percent of cancer patients by race or stage of disease.

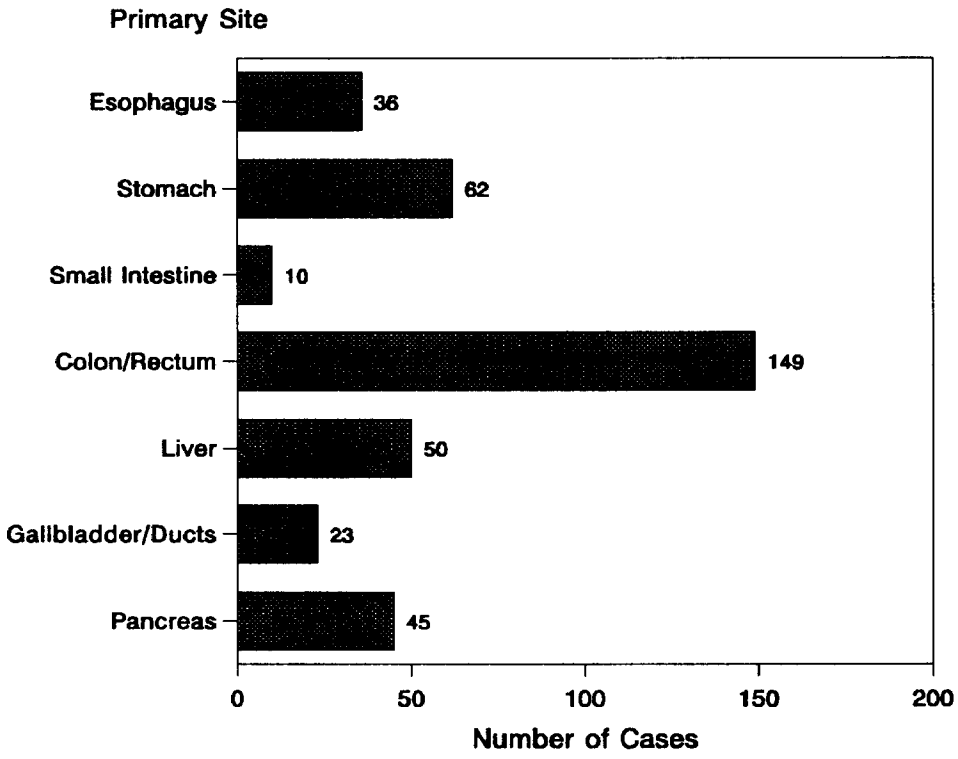
Bars may be either vertical (columns) or horizontal (columns turned sideways) and may show actual numbers or percentages. They are usually filled in with stripes, cross-hatching, dots or shadings to distinguish between categories. Because the bars represent magnitudes by their length, the zero line must be shown and the arithmetic scale[1] (numbers or percentages) must be used. In a simple bar graph, the spaces between the bars are usually about half of the width of each bar. Bar graphs are particularly effective when you want to compare values between categories.

---

[1]arithmetic scale--Scales in which the space between divisions are equal and measure absolute differences.

Figure 02. Simple Bar Graph (Horizontal)

Number of Cancer Cases Primary in the Digestive Tract
First Diagnosed at Community Hospital, 1990

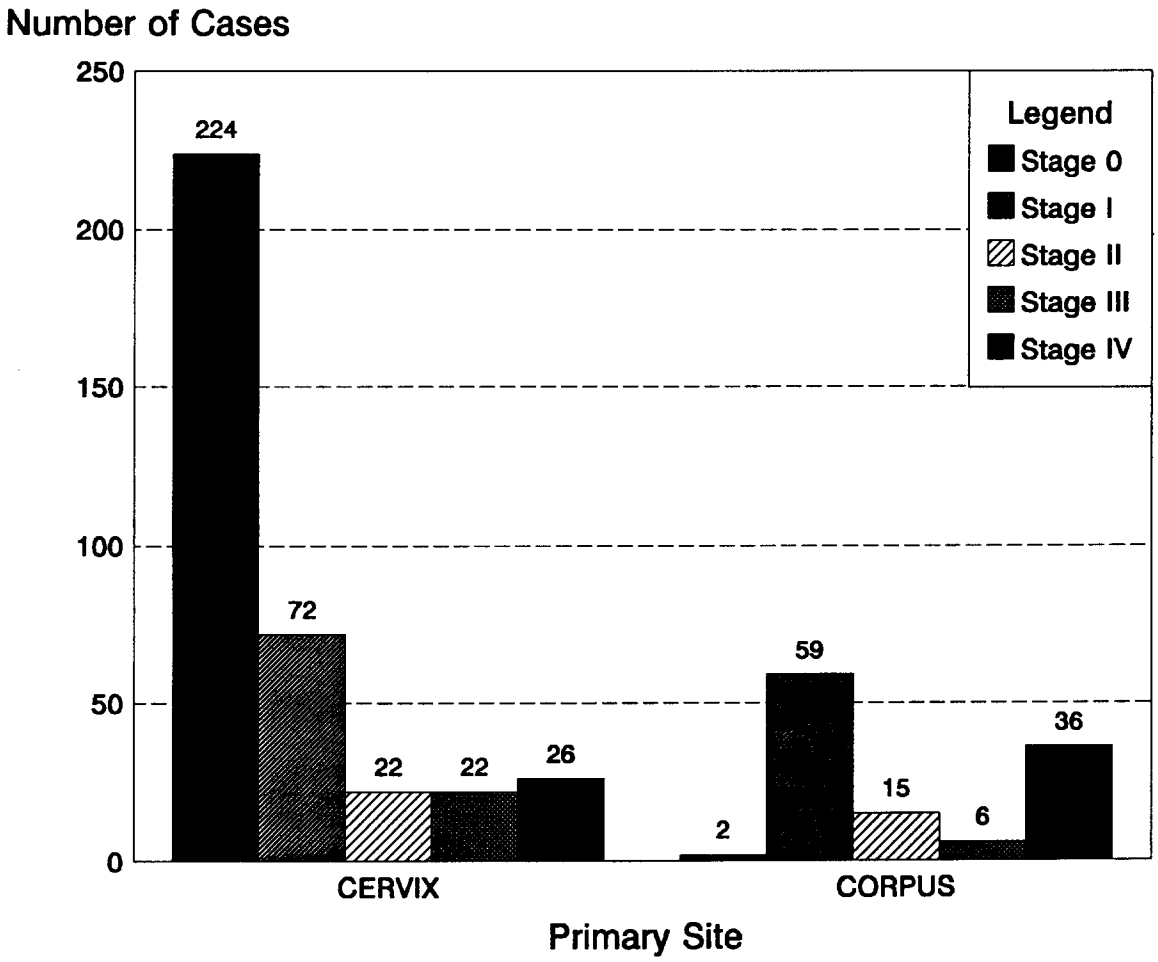


In the above graph, the width of the bars is the same, and the value of each bar is indicated on the Y-axis. The value of each bar is independent of the value of other bars.

Comparison of subdivisions within a group of cases may be illustrated by showing a series of adjacent bars. To compare more than one subdivided group, a space is made between each series of bars. Subdivisions are distinguished by the texture or shading of bars representing comparable categories within the different groups.

Figure 03. Bar Graph with Subdivisions (Vertical)

Number of Uterine Cancer Cases by Primary Site and Stage
First Diagnosed at Community Hospital, 1986-90

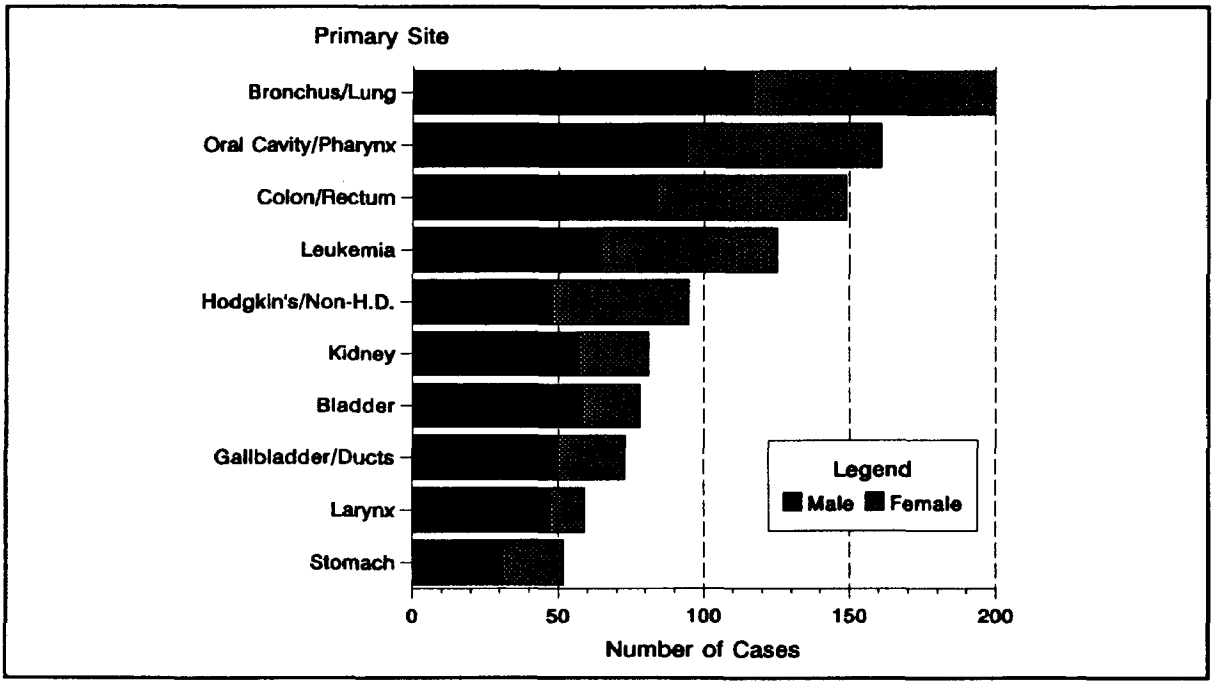


It is also possible to construct a *stacked* bar chart where one variable such as sex is subdivided within the bar. For example, in figure 04, the X-axis measures horizontally the number of cases in each site and the segment of that site group which are males or females.

Table 13. Data for Stacked Bar Graph

Number and Percentage of Cancer Cases for Leading Non-Sex-Specific Sites by Sex First Diagnosed at City Hospital, 1990					
Primary Site	Total Cases	Male		Female	
		No.	%	No.	%
Oral Cavity/Pharynx	161	94	58.4	67	41.6
Larynx	59	47	79.7	12	20.3
Bronchus/Lung	200	117	58.5	83	41.5
Stomach	52	31	59.6	21	40.4
Colon/Rectum	149	84	56.4	65	43.6
Gallbladder/Ducts	73	50	68.5	23	31.5
Bladder	78	58	74.4	20	25.6
Kidney	81	57	70.4	24	29.6
Hodgkin's/Non-H. Disease	95	48	50.5	47	49.5
Leukemia	125	65	52.0	60	48.0

Figure 04. Stacked Bar Graph (Numbers)



COMPONENT BAND GRAPH

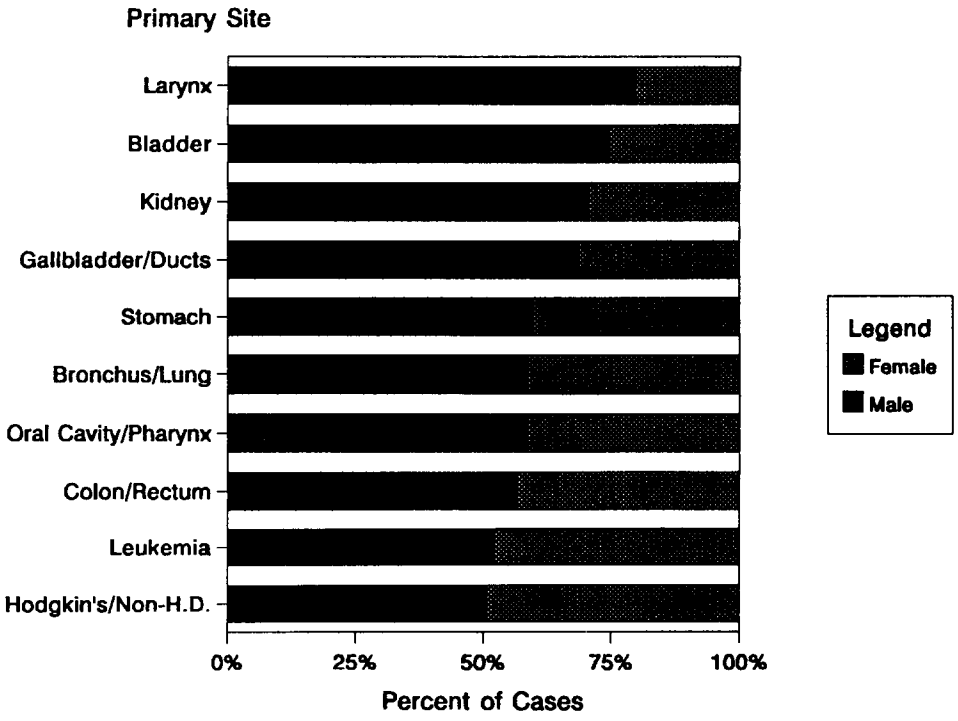
The component band graph is used to compare the relative sizes of various categories within two or more different groups. Like a bar graph, it presents frequencies of categorized data, but instead of individual bars it is subdivided into bands. It can be either vertical or horizontal, whichever is easier to read.

The length of the band and its component parts represent percentages in each category; each band is subdivided into categories. The different categories are arranged in the same order, either horizontally or vertically, in all of the groups.

The same data utilized to construct figure 04 can be presented in terms of percentages. In this instance, the length of each band represents 100 percent of the cases in each site. The segments of the band represent the percentage of the total in each category, i.e., males or females. The sites have been arranged in ascending order of percent males to emphasize the male/female differences.

Variables with more than two categories may be used, of course, but the number of subdivisions should be kept to a minimum to be visually effective.

Figure 05. Component Band Graph (Percentages)



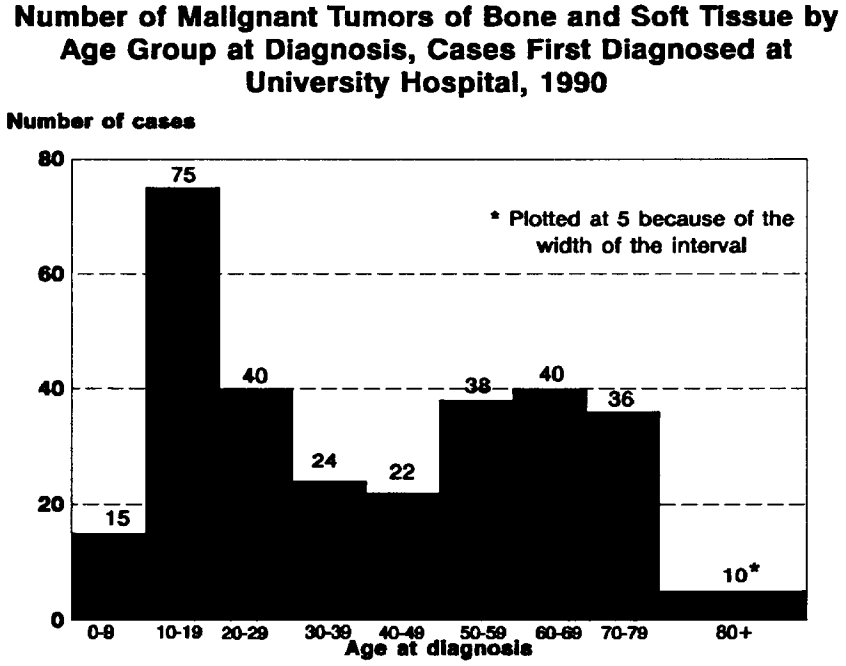
HISTOGRAM

A histogram is useful when the observations for one continuous variable are being presented. It is a distribution expressed either in terms of numbers or percentages. A histogram consists of a series of columns each having as its base one class interval and as its height the number or percent of cases in that class. In this type of graph there are no spaces between the columns. The sum of the heights of the columns represents the total number or 100 percent of the cases.

In other words, a histogram is a frequency distribution in bar graph form; the total area covered by the graph represents the whole. A histogram is most effective when only one distribution is shown. It is used when the distribution of the data needs to be emphasized more than the actual values.

In actual practice it is customary to represent the histogram in outline form, rather than show the sides of each column.

Figure 06. Histogram



Width of Intervals

In working with histograms it is a good idea, if possible, to use intervals of the same width, e.g., all 50 mm size intervals or all 10-year age groups. If the intervals are not equal, but have varied interval sizes, the frequency value on the vertical scale should be adjusted for differences in interval width. If all the intervals were for five years except one that was ten years, the 10-year interval would have to be converted by dividing its number or percentage in half. In figure 06 the age-group 80+ most likely represents cases diagnosed over a 20-year age span. Thus, the plot of cases is half as tall as the actual number, but the width of the bar is doubled. Area, not the height of a column, represents frequency. Each column MUST represent the same size group if the height of the column is to be used to represent frequency.

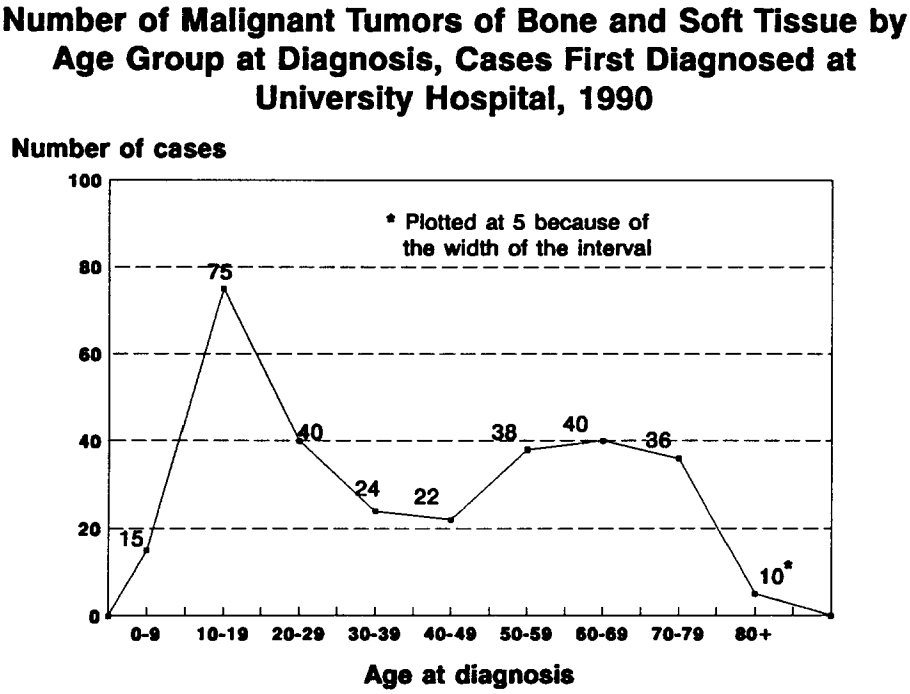
FREQUENCY POLYGON

A frequency polygon may be used as an alternative to the histogram. Simply join the midpoints at the top of each bar in the histogram as shown in the figure below. The advantage of the frequency polygon over the histogram is that several frequency polygons can easily be plotted on the same graph for purposes of comparison. It is also easy to interpret.

In constructing the graph of the frequency polygon, the X-axis should be longer than the Y-axis; a graph should be basically square. It is important not to distort data. The frequencies of observations are always placed on the Y-axis and the scale of values under study on the X-axis. Frequency values are plotted at the midpoint of each class interval.

Figure 07 shows the same data used in figure 06 plotted in the form of a frequency polygon. As with the histogram, the frequencies are placed on the Y-axis and the scale of values on the X-axis. Actual numbers or percentages may be used on the Y-axis. Since the X-axis represents the total distribution, the line always starts and ends with zero.

Figure 07. Frequency Polygon - Numbers



If more than one frequency polygon is to be shown on a single graph for comparison and the numbers in the different groups vary widely, it may be practical to convert the numbers into percentages.

Using the following data, the age distributions for three different histologies are compared in Figure 08.

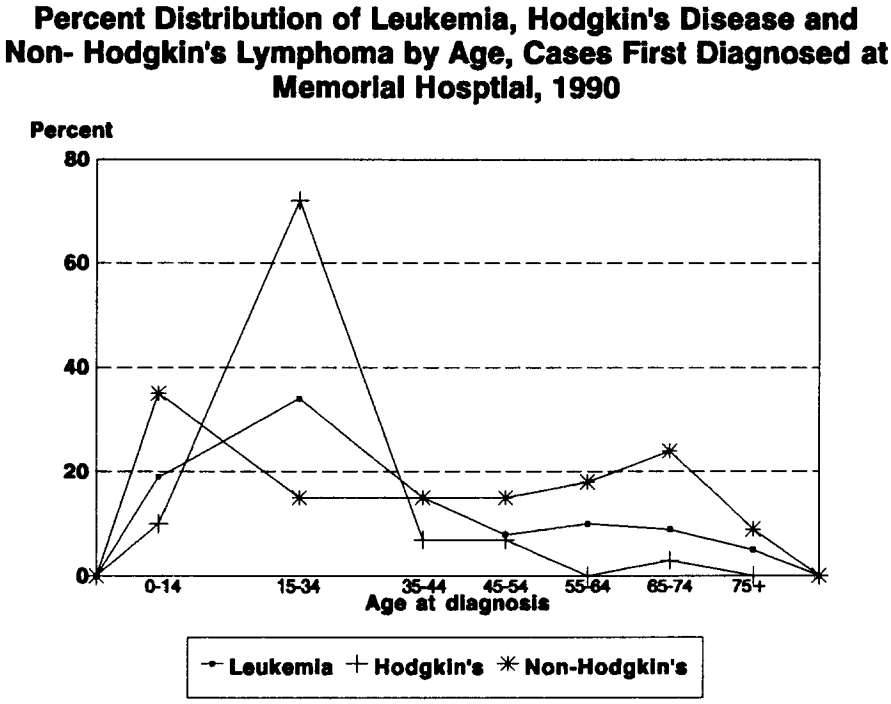
Table 14. Data for Frequency Polygon--Percentages

Age Distribution of Leukemia, Hodgkin's Disease and Non-Hodgkin's Lymphoma Cases First Diagnosed at Memorial Hospital, 1990						
Age Group	Leukemia		Hodgkin's		Non-Hodgkin's	
	No.	%	No.	%	No.	%
Total	125	100	29	100	66	100
<15	24	19	3	10	2	3
15-34	42	34	21	72	10	15
35-44	19	15	2	7	10	15
45-54	10	8	2	7	10	15
55-64	13	10	0	0	12	18
65-74	11	9	1	3	16	24
75+	6	5	0	0	6	9

Note: Percentages will *not* always add up to 100% because of the methodology used in rounding.

Frequency polygons are easy to understand. For example, they are useful for showing differences in age distributions of various forms of cancer as in figure 08 which indicates that Non-Hodgkin's lymphomas occur at all ages with the highest frequency between 65-74 years of age. On the other hand, Hodgkin's disease occurs primarily in adolescents and young adults ages 15-34 and occurs rarely after age 55. The age group with the highest frequency (15-34 for Hodgkin's disease) is called the modal interval. (See measures of central tendency.)

Figure 08. Frequency Polygon - Percentages



CUMULATIVE FREQUENCY POLYGON

A further step in the analysis of the frequency distribution might be the use of a cumulative frequency polygon, also known as an **ogive**. The cumulative frequency for any interval on the scale of values (Y-axis) is the total of the frequencies for that interval and for all lower intervals. It can be used to demonstrate graphically the number or percent of cases "less than" a certain value.

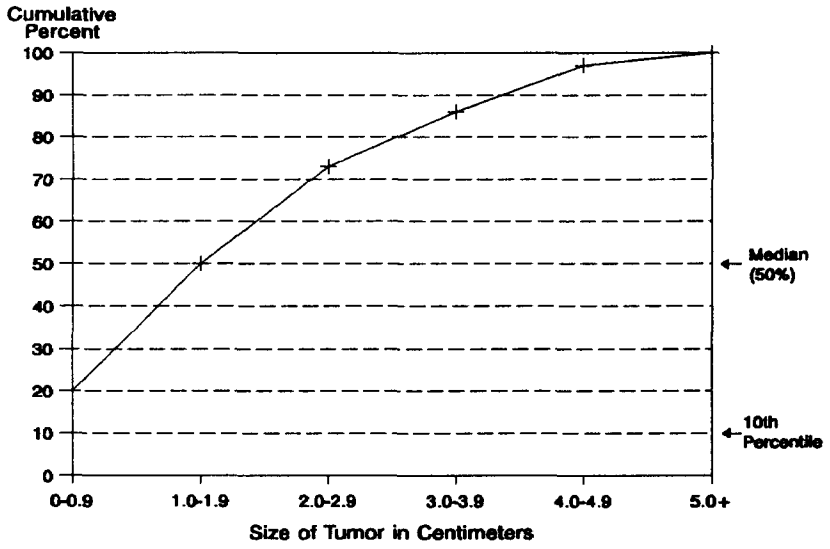
The cumulative frequency polygon is usually expressed in terms of percentages or percentiles[1] of the total. However, the shape of the polygon is the same whether actual figures, percentages, or percentiles are used on the Y-axis. The X-axis may be used, for example, to represent continuous variables such as age, weight, size of tumor or number of lymph nodes.

Plot the number or the cumulative percent on the Y-axis and the values of the continuous variable on the X-axis. Always plot the number or the cumulative percent at the upper limit of each interval.

In the following example, it appears that 50 percent of the tumors were under 2.0 cm. in size and 75 percent were under 3.0 cm.

Figure 09. Cumulative Frequency Polygon

Cumulative Percent of Female Breast Cancer by Size of Primary Tumor
Cases First Diagnosed at Community Hospital, 1990



NOTE: Excludes cases with microscopic foci only or size of primary unknown.

1percentiles--Numbers that divide a distribution into 100 equal parts, e.g., the 10th percentile includes the first 10 percent of the cases; the 50th percentile is the median.

LINE GRAPHS

The line graph is most often used to display time trends and survival curves. The X-axis shows the units of time from left to right, and the Y-axis measures the values of the variable being shown.

Sometimes the scale of values is so broad that it is difficult to include on a graph. A break in the vertical scale, indicated by a jagged line, may then be used. This will permit the value of zero to be included on the graph without unduly compressing the scale.

There are two ways of constructing the vertical scale. The most common is the arithmetic scale which illustrates absolute numerical differences and the other is the semilogarithmic scale which shows relative differences. The arithmetic scale is like an automobile odometer which indicates "how far," while the semilogarithmetic scale is like the speedometer which indicates "how fast."

Arithmetic Line Graph

An arithmetic line graph consists of a line connecting a series of points on an arithmetic scale. It should be designed to be easily read without too much information on any one graph. The selection of proper scales, complete and accurate titles, and informative legends is important. If a graph is too long and narrow, either vertically or horizontally, it has an awkward appearance and unduly exaggerates one aspect of the data.

The line graph is especially useful when there are a large number of values to be plotted, i.e., a continuous variable with an unlimited number of possible points. It also allows the presentation of several sets of data on one graph.

Actual numbers or percentages may be used on the Y-axis. A percentage distribution is particularly useful if more than one set of data is to be shown. It permits comparison of groups of patients with different totals on a common basis of 100 percent.

If more than one set of data is plotted on the same graph, different types of lines (solid or broken) should be used to distinguish between the lines. The number of lines should be kept to a minimum; a line graph can soon become too cluttered. Each line must be identified in a key or legend if not on the graph itself.

There are two kinds of time-trend data:

- Point data which are taken at a specified instant of time
- Period data which cover an average or total over a specified period of time, such as a year or a 5-year time interval.

In point data, the scale marker on the X-axis indicates a particular point in time, such as 1, 2, 3, 4, 5, etc., years of survival.

On the other hand, when plotting period data, the horizontal scale lines are used to indicate the interval limits, and the values are plotted at the midpoint of each interval. For example:

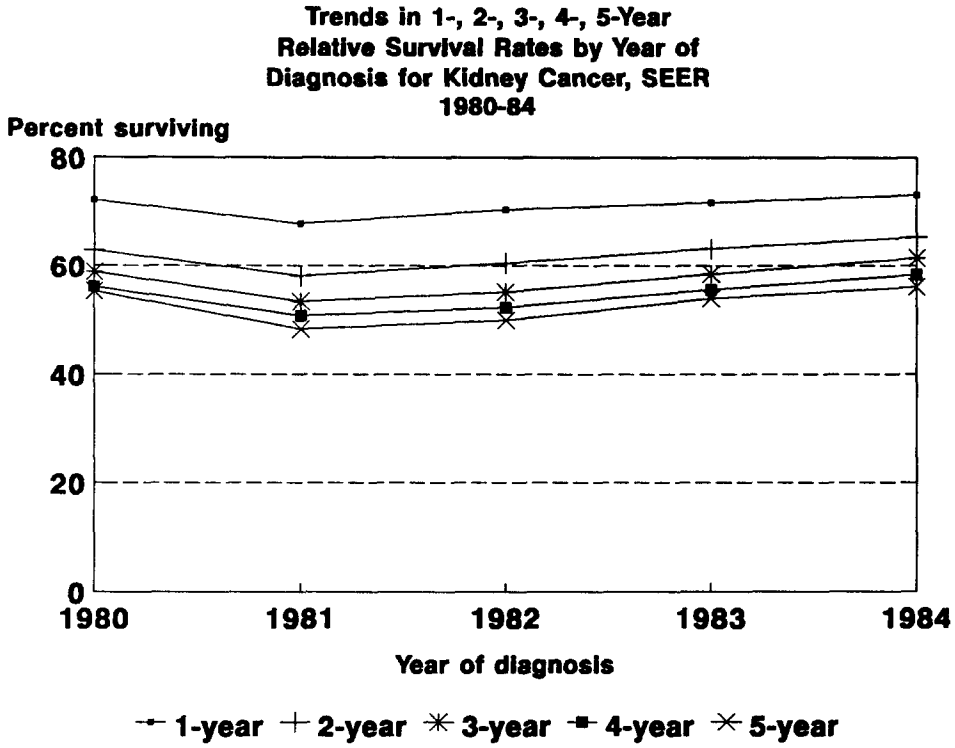
Year of Diagnosis	Midpoint of Interval
1980-1984	1982
1985-1989	1987
1990-1991	1990.5

Table 15. Example for Point Data

Relative Survival Rates by Year of Diagnosis for Kidney Cancer, SEER, 1980-84					
Years of Survival	Year of Diagnosis				
	1980	1981	1982	1983	1984
1-year	72.3	67.9	70.5	71.8	73.2
2-year	62.9	58.2	60.5	63.2	65.4
3-year	58.9	53.4	55.2	58.5	61.5
4-year	56.1	50.8	52.3	55.6	58.4
5-year	55.3	48.3	50.0	54.0	56.1

Source: Cancer Statistics Review, 1973-1989, National Cancer Institute, 1992.

Figure 10. Line Graph for Point Data



Source: Cancer Statistics Review,
1973-89, National Cancer Institute, 1992

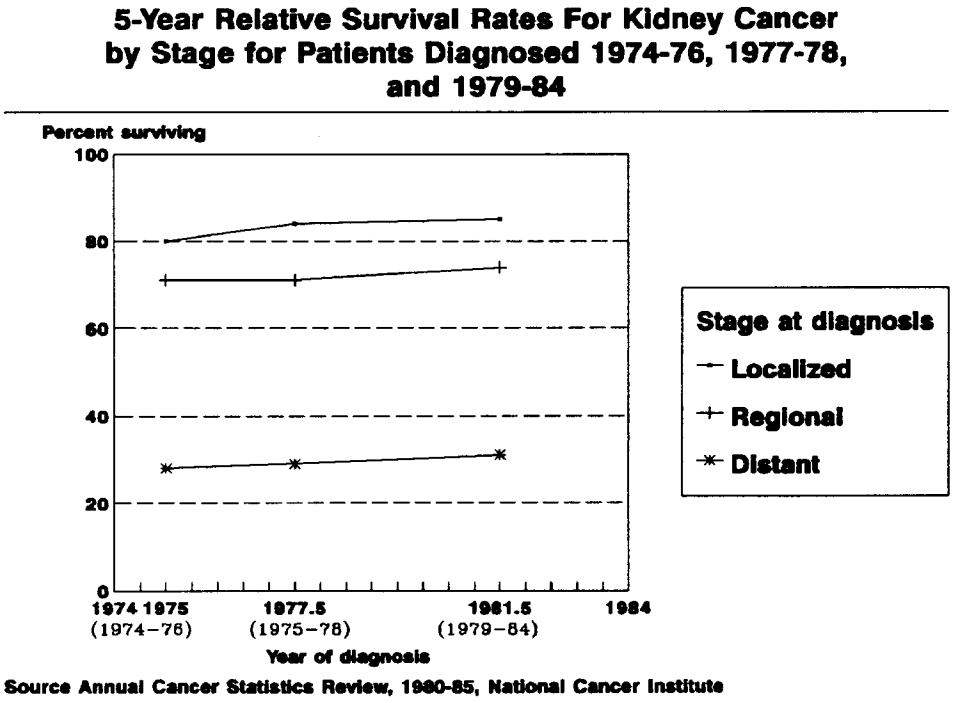
Table 16. Example for Period Data

5-Year Relative Survival Rates for Kidney Cancer by Stage for Patients Diagnosed 1974-76, 1977-78, and 1979-84				
		Survival Rate		
Year of Diagnosis	Midpoint of Interval	Localized	Regional	Distant
1974-76	1975	80	71	28
1977-78	1977.5	84	71	29
1979-84	1981.5	85	74	31

Source: Annual Cancer Statistics Review, 1980-85, National Cancer Institute, 1988.

The graph for the *period* data in table 16 is illustrated in figure 11 below. Since this is plotted on an arithmetic scale, the lines represent absolute changes in the survival values.

Figure 11. Line Graph for Period Data



When plotting a summary statistic such as 1-year and 5-year survival rates for several time periods, plot the values at the midpoint value of the time periods.

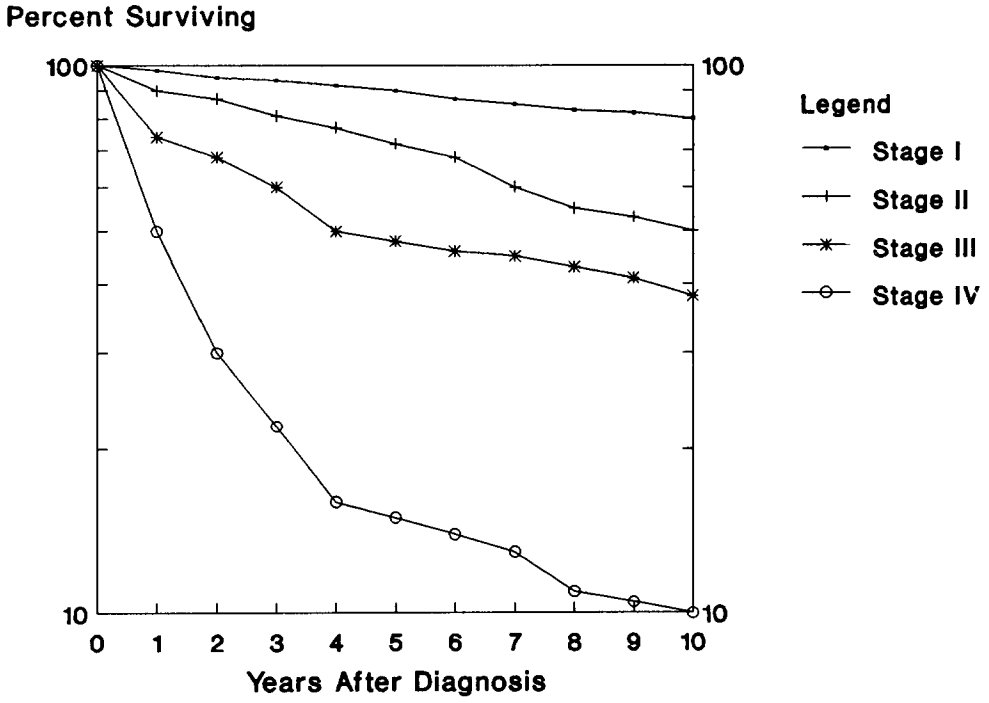
Semilog Line Graph

Lines plotted on semilogarithmic (or semilog) graph paper show the relative changes (rate of change) by the slope of the lines. The steeper the line, the greater the rate of change. The X-axis usually shows time and is plotted on the usual arithmetic scale. The values of the variable, usually rates such as survival or incidence rates, measured at each interval of time, are plotted on the Y-axis, which is a logarithmic scale. Logarithmic scales are scales in which the space between division marks are *not constant*, but vary according to the logarithms of the numbers that are represented on the scales (instead of the numbers themselves). The log scale is a *multiplicitous* scale unlike the arithmetic scale, which is additive. When values of the variable range in value between 10 and 100, a single-cycle log scale is used (See figure 12). Values which range between 1 and 100 must be plotted on a two-cycle scale. (See figure 13.)

When plotting the percent of the patients surviving to the end of each interval, plot the values and then connect each point by a straight line as in figure 12. In this example, the value at diagnosis (year = 0) is understood be be 100 percent.

Figure 12. Semilog Line Graph (One Cycle)

Observed Survival by Stage at Diagnosis for Cases of
Cutaneous Malignant Melanoma
Diagnosed at University Hospital, 1980-89



The following illustrates the assignment of possible values on a semilog scale and whether the range of values will cover one or two cycles:

Range of Values

One Cycle	Two Cycles
0.1-1.0	0.1-10
or	or
1-10	1-100
or	or
10-100	10-1,000
or	or
100-1,000	100-10,000

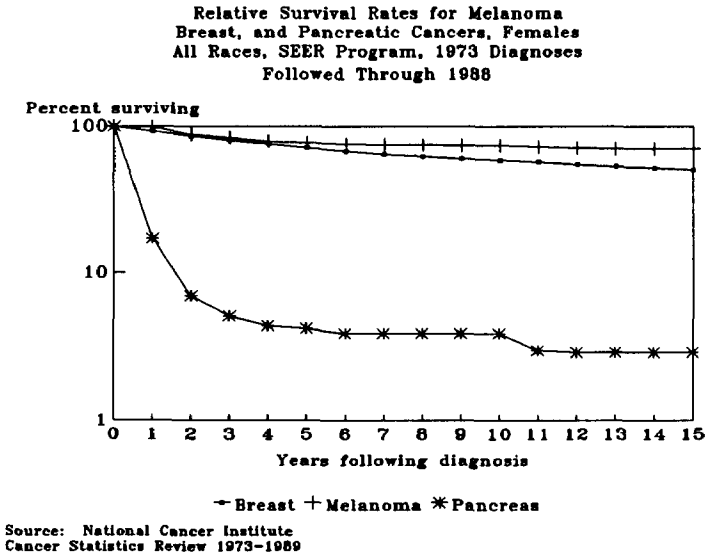
The logarithm of zero is minus infinity and, therefore, zero cannot be located on the scale. Each cycle begins with a power of 10, i.e., 0.1, 1, 10, 100, 1,000. Distances between 2 and 4, 4 and 8, 8 and 16 (100 percent increases) will be the same, and distances between 2 and 3, 8 and 12, 16 and 24 (50-percent increases) will also be constant. A scale brake can never be used on a semilog graph.

The slope of the line on a semilog graph indicates the percentage change between two points in time. The steeper the slope, the greater the percentage change. A line curving downward means a decreasing rate, while a line curving upward means an increasing rate. A rate of change which is constant over all years of observation would plot as a straight line.

Graphs plotted on semilog scale are useful for plotting survival curves when you want to emphasize rates of change or to compare patterns of survival for more than one group.

If data are plotted on a semilog scale, it should be explained in the accompanying narrative that the graph demonstrates the rate of change for each successive time period as opposed to the absolute (arithmetic) change.

Figure 13. Semilog Line Graph (Two Cycles)



PIE CHART

Another method of showing the component parts of the whole is to plot them on a circle (360 degrees) called a pie chart. Each part is expressed as a percent of the total and is plotted with a protractor (1 percent = 3.6 degrees) as a sector around a circle whose total circumference represents the whole or 100 percent.

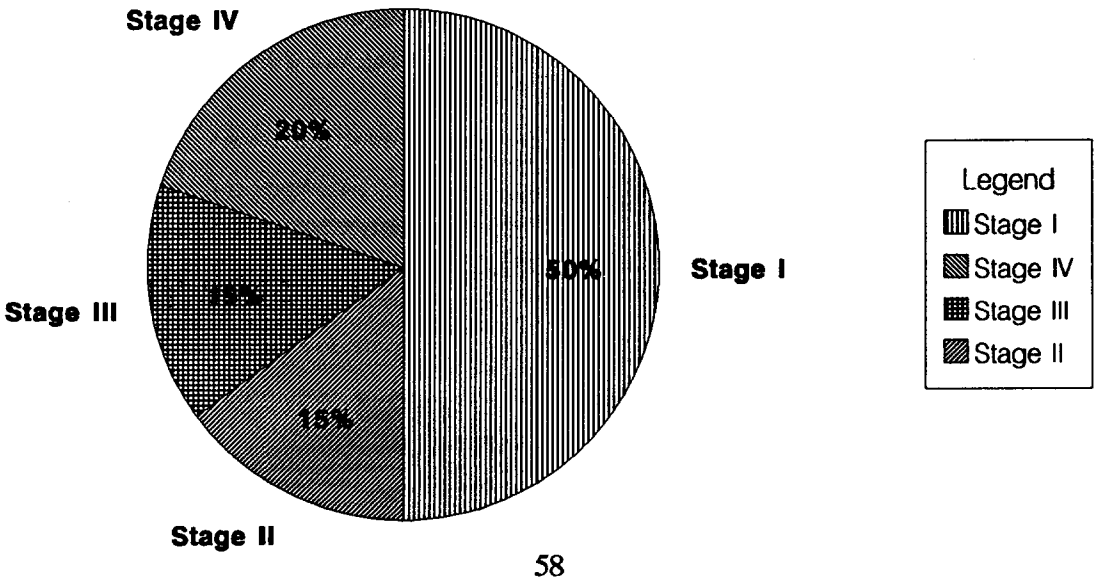
Pie charts are constructed as follows:

- Convert percents to degrees in a circle (100% = 360°). Multiply each percent by 3.6.
- Cumulate degrees for each succesive segment of the pie.
- Start at the 12 o'clock point and plot clockwise. (Many computerized graphics packages begin at 3 or 9 o'clock.)
- If there is a logical order to the values, use that order, otherwise plot in order of size of wedge.
- Label each segment on a horizontal plane, either within the circle or outside.

Never use two pie charts to compare distributions. Pie charts are not as appropriate as are component band graphs for such comparisons. A pie chart should only be used to illustrate how the whole is divided into segments, for example, stage of disease for a particular site is divided into stage groupings. Stage is an example where logical or conventional order is preferred to magnitude.

Figure 14. Pie Chart

Percentage Distribution of Invasive Cervical Cancer Cases by Stage
Women's Hospital, 1990-91

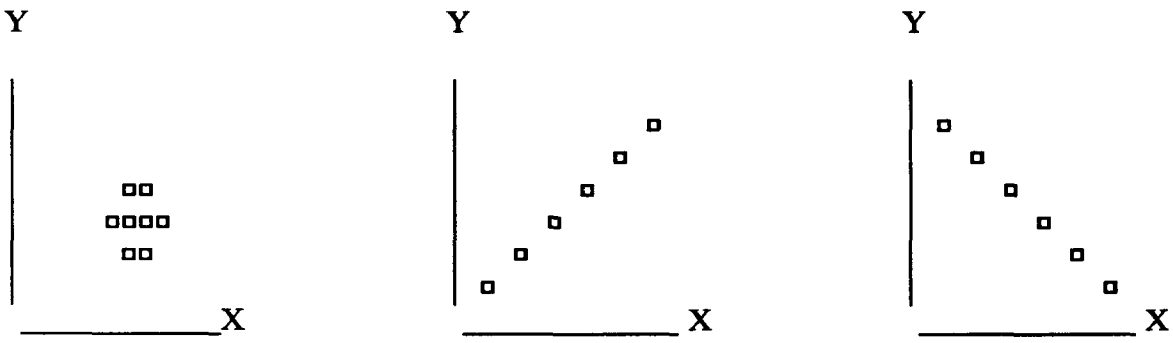


SCATTER DIAGRAM

A scatter diagram is a means of presenting relationships between two variables. One variable is plotted on the X-axis and the second variable on the Y-axis. Individual observations are plotted at the point of intersection of the values of the two variables.

If the points tend to form a line at an angle to the axes, there may exist either a positive or an inverse relationship. If the points are randomly distributed, there would appear to be no relationship.

Figure 15. Three Scatter Diagrams



Random (no) Relationship | Positive Relationship | Inverse Relationship

In analyzing tumor regsitry data, for example, one might want to assess the relationship between size of tumor and depth of invasion, or number of positive lymph nodes and length of survival.

PICTOGRAPH

A pictograph may be used as a dramatic way to catch the reader's attention. In constructing a pictograph, symbols are used to represent numbers. The number of symbols indicate the frequency of an occurrence. While pictographs are easy to understand, they are by nature imprecise in displaying numerical information.

Figure 16. Pictograph

OF EVERY FIVE DEATHS, ONE IS FROM CANCER

UNITED STATES, 1990

GEOGRAPHIC MAP

A map of an area is used as a reference, and certain statistical information is superimposed upon it. Two commonly used graphs of this type are dot maps and shaded maps.

• Dot Maps. Dots or colored pins are placed in their proper locations on a map to indicate the occurrence of a particular observation at that location and, thus, give the general effect of density. Each dot represents a certain number of cases. In some areas the dots may be too close to be counted, but an impression of density can be clearly brought out. The dots may represent the number of cases for a geographic area. For a large central registry, a better value would be the number of cases per 100,000 population. Such maps would be useful in pinpointing areas of excessive incidence which need to be investigated.

For an individual tumor registry, the place of residence of its patient population might be of interest in determining referral patterns and developing outreach programs.

Variations in quantities may be indicated also by varying the size, shape, and/or color of the dot or pin.

The construction of dot maps can be difficult because of the care that must be exercised in selection of the size of the dot and the quantity it is to represent. On the other hand, the pin map is flexible, quick, and easy to change.

• Shaded Maps.

These maps are most often used, instead of dots, for incidence or mortality rates. In designing a shaded map, the lightest shading should indicate the lowest rate, and the shading should increase with the darkest shading indicating the highest rate (See figure 17).

Maps may represent political divisions such as cities, counties, or states; metropolitan areas, census tracts or other defined population areas. The variable being illustrated should have geographic relevance and the number of classifications should be kept to a minimum. Areas should be sufficiently large to recognize boundaries. For instance, you would not divide the entire United States into census tracts.

Because of the differences in population density, rates obviously are more appropriate than actual numbers in constructing shaded maps.

No matter how well designed, graphs should not be used as a substitute for a narrative analysis of the data. The relevance of each graph to the presentation should be made clear to the reader, preferably on the same page as the graph. Appropriate background information and adequate interpretation of the graphics should be a part of the analysis.

Figure 17. Shaded Map

Age–Adjusted (1970 standard) Breast Cancer Mortality,
Females, All Races, Continental United States, 1983–87

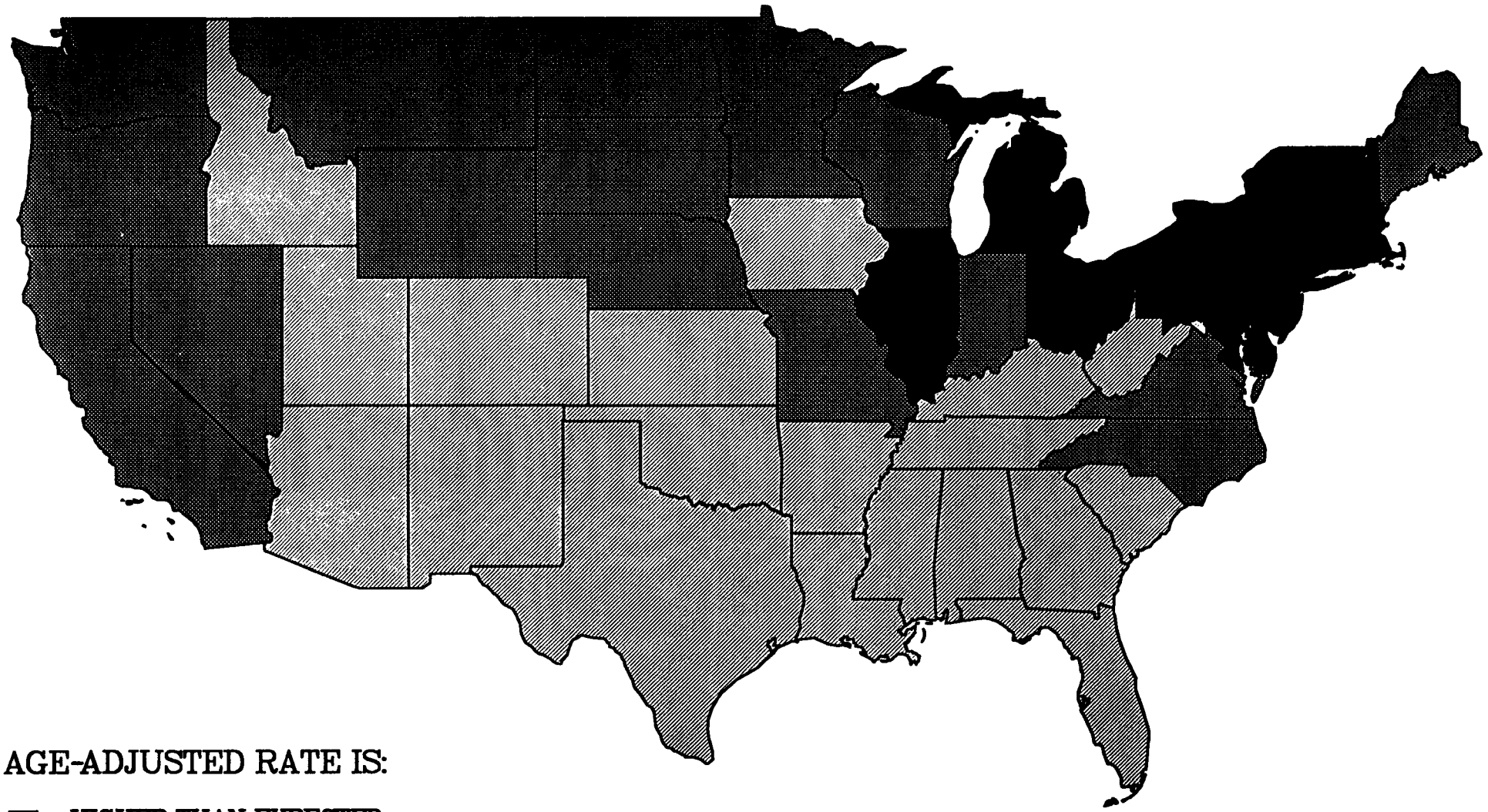


Table of Contents
Manuals

Q21

If there are too many values in your data item for easy analysis, you may wish to group your data into _____.

Q22

When data are grouped into intervals, we call these intervals_____.

Q23

A general rule for dividing detailed data is to have between 6 and ___ _____, and they must be stated precisely to avoid _____.

Q24

Which of the following methods of designating intervals for age groups is the best and why?

A	B	C	D
0-10	0-09	0-04	0-05
10-20	10-19	05-09	6-15
20-30	20-29	10-14	16-25
30-40	30-39	15-19	26-35
40-50	40-49	20-24	36-45
50-60	50-59	25-29	46-65
60-70	60-69	30-34	66-85
70-80	70-79	35-39	86+
80+	80+	40-45	
		46-49	
		50-54	
		55-59	
		60-64	
		65-69	
		70-74	
		75-79	
		80-84	
		85+	

A _____

_____

B _____

_____

C _____

_____

D _____

_____

Q25

If it is more important to you to know the relative number of patients in each class then it is to know the actual number of cases, use a _____ distribution.

Answer: Q21

If there are too many values in your array for easy analysis, you may wish to group your data into intervals.

Answer: Q22

When data are grouped into intervals, we call these intervals classes.

Answer: Q23

A general rule for dividing detailed data is to have between 6 and 15 classes, and they must be stated precisely to avoid ambiguity.

Answer: Q24

Group A: The classes are ambiguous because they overlap. Does age 20 go into group 10-20 or 20-30? You can't tell.

Group B: The classes are clear and unambiguous. There is no overlapping and the classes are all of the same size--10 years each. However, B is grouped by decades. Children and retirees (i.e., 65+) cannot be readily identified.

Group C: The classes are clear and unambiguous. There is no overlapping and the classes are all of the same size--5 years each.

Group D: The grouping is clear; there is no overlapping of classes; however, the age groups vary making it difficult to interpret.

Note: Group C is the best method for designating intervals for age groups.

Answer: Q25

If it is more important for you to know the relative number of patients in each class than it is to know the actual number of patients, use a percentage distribution.

Q26

_____ graphs emphasize individual amounts, while _____ graphs emphasize general trends.

Q27

A frequency distribution shown in bar graph form is called a _____.

Q28

Match the type of graph on the left with the description on the right.

_____	1. Bar graph	a.	The sum of the heights of the bars represents all the cases so no space is left between bars.
_____	2. Pie chart	b.	Shows proportional parts of the whole in terms of degrees
_____	3. Histogram	c.	Dots give the location and create the effect of density.
_____	4. Map	d.	The individual heights of each bar represent a whole, so space is usually left between the bars.
_____	5. Frequency polygon	e.	A line graph which represents all cases

Q29

The value of the frequency polygon over the histogram is:

a. Component parts of the whole can be shown.

b. It shows the distribution of all cases according to some variable.

c. Several sets of data can be presented simultaneously.

d. It shows trends over time.

Q30

If you have a cumulative frequency polygon of patients by age groups, you can:

a. Determine what percent of the patients are in each age group.

b. Determine what number of the patients are in each age group.

c. Determine what number of the patients are below a particular age.

Q31

A frequency polygon will tell you:

a. The total number of observations in each interval.

b. The total number of observations in a particular interval and for all lower intervals.

c. The percent of observations in each interval.

d. The percent of observations less than a given value.

Answer: Q26

Bar graphs emphasize individual amounts, while line graphs emphasize general trends.

Answer: Q27

A frequency distribution shown in bar graph form is called a histogram.

Answer: Q28

Match the type of graph on the left with the description on the right.

d	1. Bar graph--	The individual heights of each bar represent a whole, so space is usually left between bars.
b	2. Pie chart--	Shows proportional parts of the whole in terms of degrees
a	3. Histogram--	The sum of the heights of the bars represents all the cases so no space is left between bars.
c	4. Map--	Dots give the location and create the effect of density.
e	5. Frequency polygon--	A line graph which represents all cases

Answer: Q29

c. The value of the frequency polygon over the histogram is that several sets of data can be presented simultaneously.

Answer: Q30

c. If you have a cumulative frequency polygon of patients by age groups, you can determine what number of the patients are below a particular age.

Answer: Q31

a. A frequency polygon will tell you the total number of observations in each interval.

Q32

Pie charts are used:

a. To compare two distributions.

b. To illustrate how the whole is divided into segments.

c. To create an impression of density.

d. To emphasize general survival trends.

Q33

Match the type of scale on the left with effect on the right.

_____	1. Arithmetic scale	a.	Rate of change
_____	2. Semilog scale	b.	Absolute change

Q34

_____ graph paper has equal units while _____ graph paper has equal units on its _____ scale, but unequal units on its _____ scale.

Answer: Q32

b. Pie charts are used to illustrate how the whole is divided into segments. They are not appropriate for a, c, and d.

Answer: Q33

_b_ 1. Arithmetic scale: Absolute change

_a_ 2. Semilog scale: Rate of change

Answer: Q34

Arithmetic graph paper has equal intervals in contrast to semilog graph paper which has equal units on its horizontal scale, but unequal units on its vertical scale.

MEASURES OF CENTRAL TENDENCY AND VARIATION

If measurable characteristics, such as age, weight, stage of disease or response to treatment did not vary from individual to individual, describing a set of data would be completed after the first observation. However, biological differences and disease characteristics in which we are interested take on a range of values distributed among the subjects under study. In order to describe these variations we need to summarize.

How do we summarize a set of data? Let's gain command of some of the most widely used measures which we derive from a set of observations. We characterize a set of data in terms of:

1. *Central values* about which the data tend to cluster. These are called measures of *central tendency*. These measures could be described as "typical" values, e.g., the average age at diagnosis.

2. The amount of *spread* or the *variability* or *dispersion* of the observations. The measures we use here are called *measures of variation*, e.g., the average fluctuation of ages.

First let's introduce some shorthand notation which is in general usage:

1. Let X be the value of a measurement or observation.

2. Σ (the capital Greek letter sigma) tells us to carry out the process of summation (sum of the values of X).

3. Let "n" represent the number of observations (values) in our group.

4. $\bar{X}$ (spoken "X bar") is used to denote the mean, the average value, or a measure of central tendency.

5. SD is used to represent standard deviation, a measure of variability.

Measures of Central Tendency

Widely used measures of central tendency are the mean, median, and mode.

Example 01: Assume that the numbers of positive nodes seen in three female breast cancer patients were 2, 8, and 5, respectively. So the three values of X are $X_1 = 2$, $X_2 = 8$, and $X_3 = 5$ and $n = 3$.

- MEAN: The arithmetic average is the sum of all values, divided by the number of values. Using our notation, the sample mean is denoted by

$$\bar{X} = \frac{\Sigma X}{n}$$

OR

$$\bar{X} = \frac{\Sigma X}{n} = \frac{(X_1 + X_2 + X_3)}{n} = \frac{(2 + 8 + 5)}{3} = \frac{15}{3} = 5$$

The mean, $\bar{X}$, will be extremely valuable in drawing statistical inferences (predictions) about the mean of a larger population.

We will be using $\bar{X}$ in relation to the so-called NORMAL CURVE, about which we will learn more in succeeding sections.

- MEDIAN: The median is the middle value in terms of magnitude. Sorting the observations in order from smallest to largest, the median is the 50th percentile, i.e., half the values are smaller and half are larger.

In the above example, the values in order of magnitude are 2, 5, 8. Therefore, the middle value or MEDIAN is also 5 nodes (50th percentile).

The median is easy to calculate and easy to understand; it divides the series of observations such that half are smaller and half are larger than the median. Furthermore, the median is a quite stable measure, i.e., adding an extreme value to a series of observations tends to cause only a limited change in the value of the median. Thus, if a female breast cancer patient with 18 nodes was added to our series, the median would only increase from 5 to 6.5 (halfway between the two middle values of 5 and 8). The mean, $\bar{X}$, would be influenced more and would increase from 5 to 8.25, i.e., $\Sigma X/n = 33/4 = 8.25$.

- MODE: The mode is the most frequently seen value.

There is no most frequent value in the previous example so there is no modal value. With a frequency distribution, there is usually an interval with more observations than any other one. This is the *modal interval.* At times there may be more than one value that occurs most frequently.

Example 02: The weights (in pounds) of twenty white males with adenocarcinoma of the rectum were as follows:

198	189	148	170
158	142	175	175
200	155	173	151
165	185	155	193
164	186	183	175

The computations of the mean, median and mode for the above group of patients are as follows:

$$\text{MEAN:} \quad = \quad \bar{X} = \frac{\Sigma X}{n} = \frac{(198 + 189 + \cdots + 175)}{20} = \frac{3440}{20} = 172$$

MEDIAN: Put the values in order from smallest to largest: 142 148 151 155 155 158 164 165 170 173 175 175 175 183 185 186 189 193 198 200

This group has two middle values (173 and 175), therefore, the median is found by averaging the two middle values.

The *median* is the middle-most value $\frac{173 + 175}{2} = 174$

A general way for finding how far to count to find the middle value is to calculate $\frac{n + 1}{2}$. In the first example of patient lymph nodes, $\frac{(n + 1)}{2} = \frac{(3 + 1)}{2} = 2.$ This means that the second value is the median. In the second example of weights:

$\frac{(n + 1)}{2} = \frac{(20 + 1)}{2} = 10.5$, thus the median value is halfway between the 10th and 11th value. Count the ordered values to the 10th and 11th values, and average them.

$$\frac{(173 + 175)}{2} = 174.$$

MODE: The most frequently occurring value is 175. It occurs three times.

Measures of Variation

The most common measures of variation applicable to tumor registry data are the range and the standard deviation.

- RANGE: The easiest measure of variation is the range, which is the difference between the highest and the lowest values.

  In Example 01 above the range is 6 nodes (8 - 2).

  In Example 02 above the range is 58 pounds (200 - 142).

The problem with using the range is that it uses only the end points and therefore is greatly influenced by extreme values.

- STANDARD DEVIATION: Another approach is to look at measures of variation dealing with how far observations tend to vary from the mean.

The formula for calculating the standard deviation is:

$$SD = \sqrt{\frac{\Sigma\ (X - \bar{X})^2}{(n-1)}}$$

In Example 01, the calculation is as follows:

X	$\bar{X}$	$(X - \bar{X})$	$(X - \bar{X})^2$	
2	5	-3	9	n-1 = 2
5	5	0	0	
8	5	+3	9	
		0	18	$= \Sigma(X - \bar{X})^2$

$$SD = \sqrt{\frac{18}{2}} = \sqrt{9} = 3$$

Why do we square the deviations from the mean?

As you can see, the sum of X - $\bar{X}$ will always equal 0. Therefore, by squaring the differences from the mean, the difficulty of signs (+ or -) is eliminated since when squared, negative as well as positive values become positive.

The explanation of why we use (n-1) in the denominator instead of the actual number of observations is explained in appendix 1, page 5.

In example 02, the standard deviation of weights for the twenty white males with adenocarcinoma of the rectum is calculated below:

Range--Lowest to Highest Numbers

X	$(X - \bar{X})$	$(X - \bar{X})^2$	
142	-30	900	
148	-24	576	
151	-21	441	
155	-17	289	
155	-17	289	
158	-14	196	
164	-8	64	
165	-7	49	
170	-2	4	
173	1	1	
175	3	9	
175	3	9	
175	3	9	
183	11	121	*Mean* $(\bar{X})$ = 172
185	13	169	
186	14	196	n-1 = 19
189	17	289	
193	21	441	
198	26	676	
200	28	784	
	0	5512	= $\Sigma(X - \bar{X})^2$

$$SD = \frac{\sqrt{\Sigma(X - \bar{X})^2}}{(n-1)} = \sqrt{5512/19} = \sqrt{290.1} = 17.0$$

The significance of the standard deviation will be seen when we study the normal distribution curve (section F.)

Q35

The survival time from diagnosis until death of seven cancer patients was as follows: 0, 2, 3, 5, 5, 7 and 34 months.

a. What was the mean survival time?

b. What was the median survival time?

c. What is the modal survival time?

d. What was the range of survival times?

Q36

Which of the above measures of central tendency best describes the distribution of survival times?

Answer: Q35

a. mean survival = 56/7 = 8 months

b. median = 5 months (middle value = 5)

c. mode = 5 months (two patients survived 5 months)

d. range = 34 months-0 months = 34 months.

Answer: Q36

The median is the best descriptor of central tendency in this case since it is not affected by the extreme value of 34 months. Six of the seven patients survived 7 months or less, yet the average survival was 8 months due to the one patient who survived for 34 months, an "extreme" for this group of patients.

SECTION C

DESCRIPTIVE EPIDEMIOLOGY

SECTION C

DESCRIPTIVE EPIDEMIOLOGY

INTRODUCTION TO EPIDEMIOLOGY

Epidemiology is a branch of medical science concerned with the study of the distribution of disease in a population (descriptive epidemiology) and the search for determinants of disease (analytic epidemiology). In this section we will describe methods used in descriptive epidemiology. Analytic epidemiology will be considered in section E. The following several paragraphs will introduce in general terms the thinking behind standard epidemiologic methods and some of the tools which are employed. Following this introduction, the methods will be developed in more detail.

It is possible to study the distribution of cancer in human populations in terms of variables such as age, race, sex, place of residence, marital status, and socioeconomic status in order to identify high and low risk subgroups within a population.

In the field of cancer, the public wants to know if the risk of developing or dying from cancer is increasing or decreasing. If there are changes in cancer risks, they want to know which cancers are changing and by how much. This information is obtained and used by epidemiologists as well as by public health planners and administrators.

RATES AS MEASURES OF RISK

Our primary tool for the measurement of risk is called a rate. Rates of morbidity (illness) may be expressed in terms of either incidence rates (disease occurrence) or prevalence rates (disease presence). The risk of mortality (or death) is called a mortality rate.

Morbidity and mortality statistics are essential to public health agencies for comparison of disease risk among communities and for the study of time trends. The direction of cancer control efforts may be determined by these findings. The cancer registrar and those in allied fields should be familiar with techniques presented here since it will often be to the registrar that physicians, administrators, and researchers will turn for assistance.

What We Need To Know To Calculate a Rate

There are two primary components in either a morbidity rate or a mortality (death) rate. The first component is a count of the number of events we wish to measure. The second is the size of the group or population of interest which is subject to the risk of the event. A large number of disease occurrences in a small population sounds an alarm which is more likely to attract attention than a large number of diagnoses in a sizeable population. Hence we need to take both components into account in assessing the frequency of reports of disease. We employ similar considerations in evaluating mortality.

Do We Count Occurrences (diagnoses) or Individuals?

In determining mortality figures for a disease, an individual can be counted only once, since death is experienced only once. However, when we report morbidity from a disease such as the common cold, the event of interest (diagnosis of the disease) can occur more than once to the same individual even within a relatively brief time period. Sometimes we wish to record the total number of occurrences of the disease. At other times we only wish to note the number of different people afflicted either one or more times within a certain time interval.

Our tumor registry record keeping system should be set up so that we can keep track of multiple diagnoses such as cancers of two or more body sites in the same individual, i.e., multiple primaries. By so doing, we will also be able to determine the total number of diagnoses of cancer of a particular site as well as the total number of individuals with cancer.

In calculating site-specific mortality rates for persons with multiple primaries, the death should be attributed to the cancer site which led to the death of the patient if this can be determined. If not, the death will be considered as due to cancer of "Unknown Primary Site" and the case excluded from calculation of the site-specific rate.

Q1

Descriptive cancer epidemiology is the study of the _____ of cancer in man.

Q2

In studying the distribution of cancer in man one measures the _____ of getting cancer or dying from it.

Q3

What measures of risk do we associate with the study of cancer?

1. _____
2. _____

Q4

How is the measure of risk expressed?

_____.

Q5

What two components are required in order to calculate a rate?

1. _____
2. _____

Q6

Three measures of risk, two of which deal with morbidity and one of which deals with mortality, are:

1. _____
2. _____
3. _____

Answer: Q1

Descriptive cancer epidemiology is the study of the distribution of cancer in man.

Answer: Q2

In studying the distribution of cancer in man one measures the risk of getting cancer or dying from it.

Answer: Q3

Two measures of risk associated with the study of cancer are:

1. Morbidity
2. Mortality

Answer: Q4

The measure of risk is expressed in the form of a rate.

Answer: Q5

The two components which are required in order to express a rate are:

1) number of disease occurrences or deaths

2) number of people at risk of getting the disease

Answer: Q6

Three measures of risk, two of which deal with morbidity and one of which deals with mortality, are:

1. Incidence rates
2. Prevalence rates
3. Mortality rates

CRUDE RATES

Since descriptive cancer epidemiology employs rates--incidence, prevalence, or mortality rates--as measures of the risk of developing, having, or dying from cancer, these measures will now be discussed in greater detail.

1. Incidence Rates

An incidence rate is the rate of occurrence of NEW cases diagnosed in a defined population in a given time period.

Incidence data may originate either from a special survey of the population, such as the Third National Cancer Survey (1969-71), or from a routine population-based cancer reporting system, such as the Surveillance, Epidemiology, and End Results (SEER) Program or other cancer programs which cover a defined population.

2. Prevalence Rates

The purpose of a prevalence rate is to quantify the TOTAL amount of active disease present in a defined population at a particular point in time. For a disease such as cancer, prevalence is difficult to measure since it is not always possible to determine whether a person with a prior diagnosis of cancer still has active disease. Usually, a cancer prevalence rate is based on the TOTAL number of living cases, both new and previously diagnosed.

3. Mortality Rates

A mortality rate measures the risk of DEATH for the cause under study in a defined population during a given time period.

The National Center for Health Statistics collects data on all deaths occurring within the United States. These deaths can be classified by sex, age, race, and cancer site so that cancer mortality for a given time period can be determined for the entire United States or for selected areas.

Calculation of Crude Morbidity and Mortality Rates

As described earlier, a rate is based on two components:

1. Number of disease occurrences or deaths (numerator)
2. Number of people at risk of getting the disease (denominator)

With morbidity and mortality rates the time interval during which events occurred as counted in the numerator must be specified. For chronic disease such as cancer, this is generally one year.

A rate may be defined as the ratio of two related quantities per 100, 1,000, 10,000, 100,000, or 1,000,000 population as a base for a given period of time:

$$\frac{Numbers\ of\ events}{Population\ at\ risk}\ X\ Base$$

The numerator is a count of the number of diagnoses or deaths from the disease reported during a specific time period, usually a calendar year. The denominator is a mid-year estimate of the population at risk of having the disease during that time period. The *base* is a number sufficiently large to report the rate in whole numbers. For cancer morbidity and mortality rates, the convention is to speak of rates per 100,000 among adults. Childhood cancer rates are generally reported per 1,000,000 since the risk among children is quite low.

Examples:

1. Cancer Incidence Rates

An incidence rate is calculated as follows:

$$\frac{Number\ of\ new\ cancers\ diagnosed\ during\ a\ given\ time\ period}{Total\ number\ in\ population\ at\ risk}\ X\ 100{,}000$$

In 1987 there were 87,304 cases of cancer diagnosed among the 22,425,893 residents of the SEER areas.

Using the above formula, the cancer incidence rate for 1987 per 100,000 was:

$$\frac{87{,}304}{22{,}425{,}893}\ X\ 100{,}000$$

$$= 0.003893\ X\ 100{,}000$$

$$= 389.3 \text{ diagnoses per } 100{,}000 \text{ population in } 1987$$

This is called a CRUDE cancer incidence rate because it is based on the entire population, that is, it encompasses cancers of all sites for all persons irrespective of age, race, or sex. We will consider rates of a more specific nature shortly.

2. Cancer Prevalence Rates

Sometimes we wish to calculate a rate based on the number of persons in a community who have active cancer at some point in time. This is measured by the prevalence rate, which is calculated as follows:

$$\frac{\textit{Number of active (existing) cancer cases at a given point in time}}{\textit{Total number in population at risk}} \times 100{,}000$$

Since it is often difficult to know whether cancer is still active following diagnosis and treatment, one usually includes the total number of cases ever diagnosed and still alive at a given point in time. This could be thought of as "historical" prevalence.

Cancer Facts and Figures (1990) states that "there are over 6 million Americans alive today who have a history of cancer." Assuming the population of the United States to be 250,000,000, the cancer prevalence rate per 100,000 population is found to be:

$$(6{,}000{,}000/250{,}000{,}000) \times 100{,}000 = 2{,}400 \text{ per } 100{,}000$$

Many of these people have no evidence of active disease so that the true prevalence of active disease would be much lower, but much harder to accurately assess. Also, the prevalence rate will tend to be higher for older registries with more historical data.

3. Cancer Mortality Rates.

A cancer mortality rate is calculated as follows:

$$\frac{\textit{Number of cancer deaths during a given period of time}}{\textit{Total number in population at risk}} \; X \; 100{,}000$$

There were 476,927 cancer deaths in the United States in 1987. The mid-year population in the United States in 1987 was estimated to be 243,394,693. Using the formula above, the cancer mortality (death) rate per 100,000 for the United States in 1987 was:

$$\frac{\textit{Number of cancer deaths in 1987}}{\textit{Population at risk}} \; X \; 100{,}000$$

$$= \frac{476{,}927}{243{,}394{,}693} \; X \; 100{,}000$$

$$= 0.00195947 \; X \; 100{,}000$$

= 195.9 deaths per 100,000 population

This is a CRUDE death rate because it encompasses deaths from all forms of cancer for persons of all ages and races and of both sexes, that is, it is based on the entire population of the United States.

Q7

The rate of occurrence of NEW cases diagnosed in a defined population in a given time period is called an _____ rate.

Q8

The rate of occurrence of the TOTAL number of alive cases, both new and previously diagnosed, in a defined population at a particular point in time is called a _____ rate.

Q9

The rate of dying in a defined population during a given time period is called either a DEATH rate or a _____ rate.

Q10

If you knew that in your state there were 64,133 people alive with cancer and that 23,457 new cases would be diagnosed this year, what other information would you need to compute rates, and what kinds of rates could you compute? _____

_____

Q11

When a rate is based on the entire population and includes cancer of all sites for persons of all ages, races, and both sexes, it is called a _____ rate.

Answer: Q7

The rate of occurrence of NEW cases diagnosed in a defined population in a given time period is called an incidence rate.

Answer: Q8

The rate of occurrence of the TOTAL number of alive cases, both new and previously diagnosed, in a defined population at a particular point in time is called a prevalence rate.

Answer: Q9

The rate of dying in a defined population during a given time period is called either a DEATH rate or a mortality rate.

Answer: Q10

If you knew that there were 64,133 people in your state with cancer and that 23,457 new cases would be diagnosed this year, you would need only the population of your state for that year to compute:

1) a Cancer Prevalence Rate

$$\frac{23{,}457 + 64{,}133}{\textit{Total number of population}} \; X \; 100{,}000$$

2) a Cancer Incidence Rate

$$\frac{23{,}457}{\textit{Total number of population at risk}} \; X \; 100{,}000$$

Answer: Q11

When a rate is based on the entire population encompassing cancer of all sites for persons of all ages and races and both sexes, it is called a crude rate.

CRUDE RATES VS. SPECIFIC RATES

Up to this point we have only considered crude cancer rates, i.e., rates of all cancers combined and of entire populations without consideration of any subgroupings by characteristics such as age, race, or sex. Next we will consider rates which describe risks for specific cancers in entire populations or in specific subgroups of a population.

An *age-specific rate* is similar to a crude rate except that it is specific for persons within a given age group. In general, cancer rates increase with age.

We can even be specific for several factors such as age, sex, and cancer site in the same rate. For example, let us consider the lung cancer incidence rate for women between the ages of 60 and 64 in Iowa for the years 1973 and 1980. The required data are presented below:

	Lung Cancer	
Women, Ages 60-64 in Iowa	1973	1980
Number of Cases	33	51
Population	67,974	69,958
Rate per 100,000	48.54	72.91

33/67,974 X 100,000 = 48.54 51/69,958 X 100,000 = 72.91

Source: "Cancer in Iowa 1973-82, State Health Registry of Iowa, Iowa State Department of Health, 1985.

This is called an *age-sex-site-specific* incidence rate. By analogy we can also calculate rates specific for additional factors such as race, marital status, and histologic type of cancer.

It should be noted that when age is not considered, e.g., a lung cancer rate is calculated for all females, that rate is sometimes referred to as a crude rate even though it is specific for other characteristics (sex, site); hence, one can speak of the crude female lung cancer rate.

Where age distributions are dissimilar, the most meaningful approach is to compare rates in individual age groups in the study populations. Although this provides the most logical comparison, it may be cumbersome in some instances. Later in this chapter, we will discuss an approach which is widely used to summarize two sets of age-specific rates from populations whose age structures differ by calculating what has been called *age-adjusted* or *age-standardized* rates. Age adjustment makes possible comparison of the risks in two populations using a single summary measure which attempts to take into account (or adjust for) the differing age compositions of the two populations.

Before describing how age-adjusted rates are calculated, let us first see how crude rates depend on the age structure of populations under study.

A Comparison of Crude and Age-Specific Rates

Table 17 shows the crude and age-specific cancer incidence rates for males and females in the state of California in 1988. These data are shown graphically in figure 18. When looking at either table 17 or figure 18 with large differences in risk after age 60 one feels intuitively that the risk of cancer is greater among males. Yet, the crude rate (all ages) is exactly equal for the two groups since out of the total population of males, 13,966,886, there were 50,949 cases for a crude rate of 50,949/13,966,886 = 365 per 100,000, and out of the total population of females, 14,356,389, there were 52,465 cases for a crude rate of 52,465/14,356,389 = 365 per 100,000. However, when one thinks about the population of males versus females, females have a greater life expectancy, and therefore, the age distribution of females is probably different (older) than that for males, which may contribute to the apparent contradiction between the age-specific and crude rates.

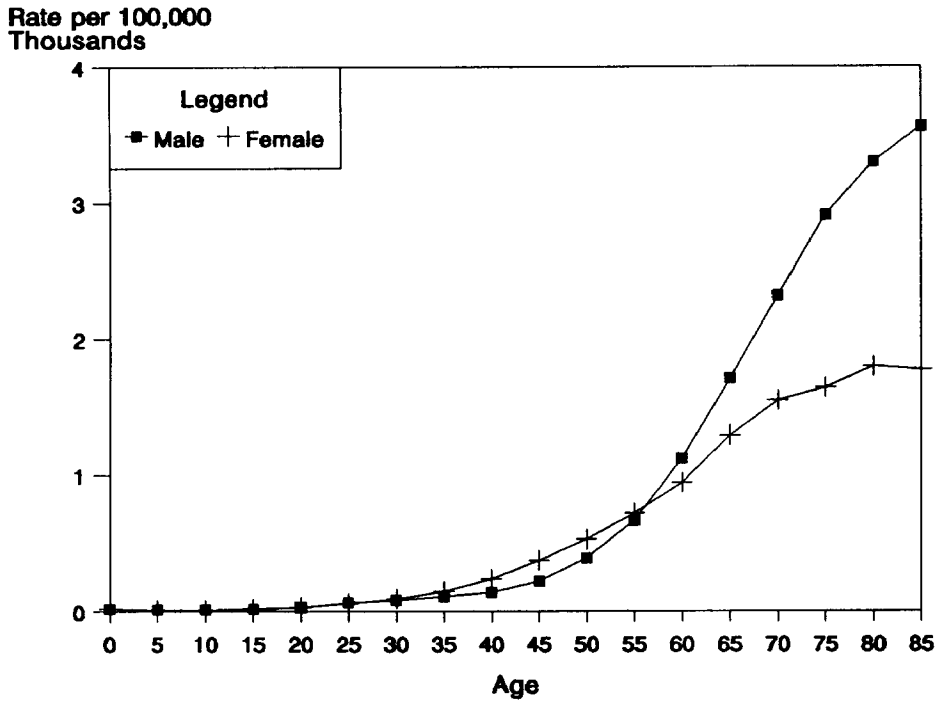
Table 17. Example of Equal Crude Rates and Differing Age-Specific Rates

Age-specific Cancer Incidence Rates per 100,000 Population by Sex, All Races, California, 1988

Age Group	Males	Females
All Ages	365	365
0-4	20	21
5-9	12	9
10-14	13	13
15-19	20	18
20-24	31	32
25-29	66	57
30-34	83	92
35-39	111	147
40-44	141	239
45-49	222	378
50-54	393	530
55-59	668	720
60-64	1121	942
65-69	1710	1289
70-74	2316	1548
75-79	2907	1642
80-84	3298	1797
85+	3556	1776

Source: California Cancer Registry, 1/91.

Figure 18. Age-Specific Cancer Incidence Rates per 100,000, All Sites, All Races, by Sex, California, 1988



Source: California Cancer Registry, 1/91.

How Crude Rates Depend on Age Composition of Population

Consider the simple case of two small communities (as shown in table 18) of 200 persons each. In community A, 50 people or one-fourth of the population, are under age 45 (col. 2) and 150 people or three-fourths are age 45 or above. The age composition in community B is different, overall younger, with 100 people or one-half of the population in the under 45 age group and 100 people or one-half in the older age group. In column 1 the number of diagnoses during a recent year is shown for each community by age group.

It should be noted that in the younger group the incidence rate (col. 3) is the same in both communities, i.e., 4 cases per 100 population. Similarly the rate in the older age interval is the same in community A and community B. The bottom line in the table gives the experience in communities A and B for both age groups combined from which to calculate the crude rate.

In contrast with the identical results in the two communities for the younger and older segments of the population, we find that the crude incidence rate (all ages) in community A is considerably higher (31 per 100) compared with community B (22 per 100). Since the age-specific rates are the same for each age group in community A and community B, we must conclude that the difference in the crude rates is attributed to the difference in the age composition of the two communities. Thus, the higher crude rate for community A reflects its heavier concentration in the older age group as compared with community B.

Table 18. Cancer Incidence in Communities A and B

Age	Community A			Community B		
	(1) No. Cases	(2) Population	(3) Rate per 100	(1) No. of Cases	(2) Population	(3) Rate per 100
Under 45	2	50	4	4	100	4
45+	60	150	40	40	100	40
All ages	62	200	31	44	200	22

The Crude Rate as an Average Measure of Risk

The crude rate of 31 per 100 for community A (table 18) may be viewed as an average of the risks to which the 200 people in community A are subject, i.e., the fifty persons under age 45 (one-fourth of the population) are subject to a risk of 0.04 (4/100) while the remaining 150 over age 45 (three-fourths of the population) have a risk of 0.40 (40/100). To get the average risk for all 200 persons in the population we would add up the 50 values of 0.04 and the 150 values of 0.40 and divide by 200 to obtain an average. This is equivalent to taking

$$\frac{50\ X\ (0.04)\ +150\ X\ (0.40)}{200} = 0.31\ or\ 31\ per\ 100$$

If, instead of using the actual number of persons, we divide the age groups into the proportion they comprise of the population with the total adding up to one, we obtain the same result by calculating

$$\frac{0.25\ X\ (0.04)\ +\ 0.75\ X\ (0.40)}{1} = 0.31 \text{ for community A (table 18)}$$

The proportions (0.25) and (0.75) by which the age-specific rates are multiplied are called *weights*. The calculations above show that in community A the crude rate is based on giving three times as much *weight* (0.75 vs. 0.25) to the rate for the older age group compared to that given to the younger age group. Thus, the crude rate of 0.31 is nearer to that for the over 45 age group (0.40) than that for the younger group (0.04).

We may proceed in a similar manner with the data from community B (table 18) in which half the population is in each of the two age groups, i.e., 100 in the younger and 100 in the older age categories. If we use population proportions as *weights* for the age-specific rates which are identical to those in community A, we find that the crude rate is

$$\frac{0.50\ X\ (0.04)\ +\ 0.50\ X\ (0.40)}{1} = 0.22\ or\ 22\ per\ 100 \text{ for community B (table 18).}$$

The crude rate for community B (22) is half way between the rates for the two age groups (0.04 and 0.40) and is lower than the crude rate for community A (31) since the weight assigned to the rate for the younger age group is greater than in community A (0.50 vs. 0.25).

Thus, a crude rate is regarded as a *weighted average* of the age-specific rates with the weights assigned to each reflecting the age structure of the population. In this example, even though the age-specific rates were identical in community A and community B, the crude rates were very different because of the differing weights of the age-specific rates. Thus, if only the crude rates were considered, we would conclude that the risk in community B is much less than that in community A when in fact the risk is identical when the effect of age is taken into account. An adjustment for the widely different proportions (weights) in each age group in community A versus community B would help to prevent drawing the erroneous conclusion that the risks were different in the two communities.

In our real life example from the state of California, we would have concluded on the basis of the crude rates (table 17) that the risk of cancer was equal among males and females even though our intuition told us that the risk must be higher among males.

In essence, populations with a high proportion of older persons will have a higher crude death rate than a population consisting of predominantly young persons. Therefore, to meaningfully compare cancer risk in the United States and developing countries in the world, account must be taken of the younger age structure of most developing nations as contrasted with that of this country. Additionally, America's population has been aging during this century as life expectancy has increased.

Q12

What is the difference between a crude incidence rate and an age-specific incidence rate?

Q13

If you wish your age-specific incidence rate to be even more specific, name two other factors which you might have considered.

Q14

How can two populations have the same age-specific rates but different crude rates?

Q15

When do you use age-adjusted rates?

Q16

What procedure is employed to correct for age differences in two or more populations?

Answer: Q12

A crude incidence rate encompasses all newly diagnosed cases within a given time period regardless of age. An age-specific incidence rate is specific for persons of a given age group.

Answer: Q13

If you wish your age-specific incidence rate to be even more specific, you might consider two of the following:

sex

race/ethnicity

primary site

geographic area

marital status

histologic type

Answer: Q14

Two populations with the same age-specific rates will have different crude rates if they have *different age distributions*.

Answer: Q15

You use age-adjusted rates when you wish to compare risks in two or more populations with differing age compositions.

Answer: Q16

The procedure employed to correct for age differences in two or more populations is called age adjustment or age standardization.

AGE-ADJUSTED RATES (DIRECT METHOD)

STANDARD SET of WEIGHTS

Age-adjusted rates are averages of the age-specific rates just as crude rates are. However, when we calculate age-adjusted rates for two or more communities (or countries or racial or sex groups or time periods), we operate as if the age compositions of each of the communities are identical by applying identical weights to the age-specific rates for each population under study. The weights we use are the proportions in each age interval of some so-called *standard population*, such as:

1. the age distribution of one of the populations under study
2. the age distribution of the combined study population
3. the population of the United States for a specific year (usually a census year such as 1970 or 1980)
4. the population of the world

Once a standard set of weights is chosen, it must be applied to all populations under study to arrive at comparable age-adjusted rates. These adjusted rates are actually fictitious rates, but they are comparable. Rates which have been adjusted to different standards (i.e., using different sets of weights) CANNOT be compared to one another. If, for example, different standards have been used for males than for females, rates among males can only be compared to each other; male rates CANNOT be compared to female rates.

The method of correcting for differences in the population age distributions of two or more communities by applying a standard set of weights to the age-specific rates of each community is known as the *direct method of age adjustment*. A second method, known as the *indirect method*, will be discussed later in this section.

Using Age Distribution of Combined Study Populations as Standard for Age-Adjustment

Now, let us obtain age-adjusted incidence rates for communities A and B, using the age distribution of the combined populations to arrive at a standard.

Table 19. Components of Age-Adjusted Rates

Age	Population		Combined	Proportion
	Community A	Community B	(Standard)	(Standard)
<45	50	100	150	.375
45+	150	100	250	.625
Total	200	200	400	1.000

As seen in table 19, 37.5 percent of the population in the combined communities was under the age of 45 and 62.5 percent was age 45 or older. We will therefore assign weights of .375 to the younger age group and 0.625 to the older age group and multiply these weights by the age-specific rates previously observed in communities A and B (shown in table 18). By utilizing these weights to obtain a new weighted average of the age-specific rates, we will have a new measure of risk in communities A and B which adjusts for the difference in their age compositions, hence an age-adjusted rate (shown in table 20 A).

Table 20 A. Calculation of Age-Adjusted Rates Utilizing Proportions

Age	Community A			Community B		
	Weight	Age-Specific Rate per 100	Weight x Rate per 100	Weight	Age-Specific Rate per 100	Weight x Rate per 100
<45	0.375	4	1.5	0.375	4	1.5
45+	0.625	40	25.0	0.625	40	25.0
Total	1.000		26.5	1.000		26.5
Age-adjusted rate (0.375 x 4 per 100) + (0.625 x 40 per 100) = 1.5 per 100 + 25 per 100 = 26.5 per 100*				(0.375 x 4 per 100) + (0.625 x 40 per 100) = 1.5 per 100 + 25 per 100 = 26.5 per 100*		

*This is equivalent to adding up 375 values of 4 per 100 and 625 values of 40 per 100 and dividing by 1,000 to get an average value of 26.5 per 100.

From table 20A we find that the age-adjusted rates for communities A and B are identical, 26.5 per 100. Note that the age-adjusted "rate" is different from both of the crude rates (31 and 22 in table 18). As previously noted, this adjusted rate is not a "real" rate but is an index of comparison between the two communities. It cannot be used as an indicator of the actual level of risk in either community A or B or to predict the risk in any other community. Its only use is in comparison of data adjusted using this same standard population.

The same age-adjusted rates for communities A and B can also be obtained by thinking in terms of the *number of persons* in each of the two age intervals of the standard population (table 19), rather than in terms of the *proportions* in each. Thus, for our standard population of size 400, 150 are in the younger age group and the remaining 250 in the older age group.

We may set up a table (table 20B) in a way that allows us to calculate and enter the number of "expected" cases in an age interval. To accomplish this we assume that those in that age interval in the standard population are subject to the age-specific rate of the community under study. For example, consider the younger age group in community A for whom the rate is 4 per 100. By multiplying this rate by the number of persons of this age group in the standard population, we find our "expected" number of cases to be 6, i.e., 4 per 100 X 150 = 6.

Table 20 B. Calculation of Age-Adjusted Rates Utilizing Expected Cases

Age	Community A			Community B		
	Age-Specific Rate	Standard Population	Expected Cases in Standard	Age-Specific Rate	Standard Population	Expected Cases in Standard
< 45	4/100	150	6	4/100	150	6
45+	40/100	250	100	40/100	250	100
Total		400	106		400	106
Rate A Age-adjusted = 106/400 = 26.5/100				Rate B Age-Adjusted = 106/400 = 26.5/100		

The rate in the older age group in community A is 40 per 100. With 250 persons assumed to be in this age group in the standard population, we find the *expected* number of cases to be 100. The total of *expected* cases adds to 106 in a total assumed population of 400. One hundred and six cases per 400 in the total standard population is equivalent to our previously obtained age-adjusted rate of 26.5 cases per 100. The same approach, using age-specific rates for community B yields the same result as using *weights* for each age interval (see table 20A).

The two methods discussed above use the combined study population as a standard to adjust for differences in age distributions. It is possible to perform adjustments for other differences in populations as well, for example, race or sex, using similar techniques. In the study of cancer it is most important to adjust for age differences since cancer risk is highly dependent on age. Age-adjustment of rates is widely practiced. In most instances there are more than two age intervals employed. However, the simple example used above demonstrates the procedures followed, whatever the number of intervals.

Comparing Two Populations Using Age-Adjustment

The following example uses data from two population-based registries divided into 18 5-year age groups.

Table 21. Age-specific, Crude, and Age-adjusted (1970 standard) Breast Cancer Incidence Rates per 100,000 White Females, Iowa and Atlanta, 1976

Age Group	Rate per 100,000 White-females	
	IOWA	ATLANTA
<5	-	-
5-9	-	-
10-14	-	-
15-19	-	-
20-24	1.6	-
25-29	13.3	8.5
30-34	22.0	35.1
35-39	47.2	48.5
40-44	76.7	119.1
45-49	180.8	177.5
50-54	154.7	238.2
55-59	193.1	251.5
60-64	221.7	279.5
65-69	237.0	281.0
70-74	318.0	276.8
75-79	242.7	368.3
80-84	346.7	237.4
85+	410.9	291.9
All Ages: Crude Rate	91.0	84.8
Age-Adjusted Rate	75.7	88.9

From this table we can see that if we consider only the rate for all ages, the risk of developing breast cancer appears to be higher in Iowa than in Atlanta, 91.0 per 100,000 versus 84.8 per 100,000. (Note, these rates are still referred to as *crude* rates because even though they are specific for race, sex, geographic area and cancer site, they have not been adjusted for age.) However, when one examines the age-specific rates, one notes that in 8 of the 14 age categories in which any cases occurred (there were no cases occurring before age 20), the rates were higher in Atlanta. Also, one might anticipate that the age structure might be different in Iowa versus Atlanta. Therefore, one would like to know what the risk would be if there were no difference in the age structure.

A technique used to adjust for age involves multiplying the standard population as calculated for each age group and expressed as a standard million (see table 22A) by the age-specific rate for each corresponding age group, and then dividing by 1,000,000. The calculations for our example are in table 22B. The resulting age-adjusted *rates* using this technique reveal that the risk is actually lower in Iowa compared to Atlanta, 75.7 versus 88.9 per 100,000.

Table 22 A. Developing a Standard Using the 1970 Population of the United States
All Races, Both Sexes

Age	Population	Percent	Standard Million
<5	17,154,337	8.4416	84,416
5-9	19,956,247	9.8204	98,204
10-14	20,789,468	10.2304	102,304
15-19	19,070,348	9.3845	93,845
20-24	16,371,021	8.0561	80,561
25-29	13,476,993	6.6320*	66,320
30-34*	11,430,436	5.6249	56,249
35-39	11,106,851	5.4656	54,656
40-44	11,980,954	5.8958	58,958
45-49	12,115,939	5.9622	59,622
50-54	11,104,018	5.4643	54,643
55-59	9,973,028	4.9077	49,077
60-64	8,616,784	4.2403	42,403
65-69	6,991,625	3.4406	34,406
70-74	5,443,831	2.6789	26,789
75-79	3,834,834	1.8871	18,871
80-84	2,284,311	1.1241	11,241
85+	1,510,901	.7435	7,435
All Ages	203,211,926	100.0000	1,000,000

*Median age group

In our example from California (table 17) when age adjustment is carried out using as a set of weights the 1970 population of the United States, the resulting age-adjusted rates are 386 per 100,000 for males versus 316 per 100,000 for females. Thus, we can conclude, as we had felt all along, that the overall cancer risk is higher among males. Again, these age-adjusted rates are an index for comparison and not real rates. The "real" rates are the crude rates. Calculations for establishing these rates are not included here.

Table 22 B. Age-Adjusting Using United States Population

Age Group	Standard* (Weight)	Iowa		Atlanta	
		Age-specific Rate		Age-specific Rate	
		Actual	Weighted	Actual	Weighted
<5	0.084	-	-	-	-
5-9	0.098	-	-	-	-
10-14	0.102	-	-	-	-
15-19	0.094	-	-	-	-
20-24	0.081	1.6	0.13	-	-
25-29	0.066	13.3	0.88	8.5	0.56
30-34	0.056	22.0	1.23	35.1	1.96
35-39	0.055	47.2	2.60	48.5	2.67
40-44	0.059	76.7	4.53	119.1	7.03
45-49	0.060	180.8	10.85	177.5	10.65
50-54	0.055	154.7	8.51	238.2	13.10
55-59	0.049	193.1	9.46	251.5	12.32
60-64	0.043	221.7	9.53	279.5	12.02
65-69	0.034	237.0	8.06	281.0	9.55
70-74	0.027	318.0	8.59	276.8	7.47
75-79	0.019	242.7	4.61	368.3	7.00
80-84	0.011	346.7	3.81	237.4	2.61
85+	0.007	410.9	2.88	291.9	2.04
Age-Adjusted Rate			75.67		88.98

*Standard has been divided by 1,000,000 for ease of computation.

Q17

What is the difference between a crude rate and an age-adjusted rate?

_____

_____

_____

Q18

Give three example(s) of weights you might use as a standard for age-adjusting two or more sets of rates.

1) _____

2) _____

3) _____

Q19

If two communities have equal age-specific rates but different crude rates, what will result when the rates are age-adjusted?

_____

_____

_____

Q20

If two communities have different age-adjusted rates, what does this mean?

_____

_____

Answer: Q17

A crude rate is based on the distribution of the actual total population at risk. An age-adjusted rate has been "adjusted" or "corrected" to take into account the difference in age distribution between two population groups.

Answer: Q18

Three weights that you might use as a standard in age-adjusting two or more sets of rates are:

1) The proportion in each age interval of the U. S. population for the year 1970 or 1980

   (The 1970 standard is the standard currently used to age adjust cancer incidence rates. There is no plan to change to a different standard in the foreseeable future.)

2) The proportion in each age interval of one of the populations under study

3) The proportion in each age group for both study populations combined.

Answer: Q19

If two communities have equal age-specific rates but different crude rates, age adjustment will result in *equal age-adjusted rates*.

Answer: Q20

If two communities have different age-adjusted rates, it means that the *age-specific rates* in the two communities are *different*.

AGE-ADJUSTED RATES (INDIRECT METHOD)

In the *indirect* method, instead of using a standard set of weights to adjust for differences in the age distribution of two (or more) populations, we initially select a standard set of age-specific rates observed in either one of the study populations or some other population, for example, the whole United States. We use these rates to compare what would have happened in our study populations if they had experienced the same risks as the standard population. That is, we ask the question "How many cases would we EXPECT to see if our study population were at the same risk of cancer as our standard population?" We then calculate the "expected" cases as described below and compare that number to the number of cases we actually observed.

STANDARDIZED RATIOS

The ratio of observed to expected (O/E) is known as a *standardized ratio*. The ratio is generally multiplied by 100 to convert it to a whole number. If we are comparing mortality data, i.e., observed deaths to expected deaths, the ratio is called a standardized mortality ratio or SMR. If we are comparing incidence data, i.e., observed new cases to expected new cases, the ratio is called a standardized incidence ratio or SIR. If the SMR or SIR is greater than 100, this means that the risk in the study population was greater than that in the standard population. Conversely, if the ratio is lower than 100, we conclude that the risk in the study population was lower than that in the standard population. One additional step converts our SMR or SIR into an indirect (age)-adjusted rate and will be discussed after the following example.

Calculation of Standardized Incidence Ratio

By taking as our standard the combined experience of communities A and B (table 23A), we can illustrate the calculation of a *standardized ratio*. Let us consider another community with a population of 200, community C (table 23B). In this community, 120 out of 200 or 60 percent of the population is under the age of 45 while 40 percent (80 out of 200) is age 45 or older. In the younger age group 10 cases were observed and 40 cases were observed in the older age group. We now decide that we would like to see how different the risk of disease in community C is compared to the risk in communities A and B.

Table 23 A. Combined Experience of Communities A and B

	Cases		Population		Communities A and B		
Age	Comm. A	Comm. B	Comm. A	Comm. B	Cases	Pop	Rate
<45	2	4	50	100	6	150	4/100
45+	60	40	150	100	100	250	40/100
Total	62	44	200	200	106	400	26.5/100

Based on the combined rates in A and B (table 23B) we would expect 4.8 cases to occur in the younger age group (4/100 x 120 pop. = 4.8) and 32 cases to occur in the older age group (40/100 x 80 people = 32) for a total expected of 36.8. However, we actually observed 50 cases.

Table 23 B. Expected Cases in Community C

Age	Observed Cases in Community C	Population of Community C	Rate in Standard (Combined A & B)	Expected Cases in Community C
<45	10	120	4/100	4.8
45+	40	80	40/100	32.0
Total	50	200		36.8
SIR = OBS/EXP x 100 = 50/36.8 x 100 = 135.9				

Thus, the standardized incidence ratio of 135.9 obtained by comparing the number of cases we expected in community C (based on the combined experience of communities A and B) with what actually happened (50 cases observed versus 36.8 expected) was large. Actually, persons in community C were at a higher risk than those in communities A and B. In fact, one could state that the risk in community C is 36 percent higher than expected based on the combined experience of communities A and B, i.e., 35.9 cases over 100.

In this example notice that the expected number of cases in a given subgroup does not have to be a whole number. This seems confusing since it is difficult to think of expecting 4.8 cases as in the age group <45 in the example above. However, since we are putting ourselves in the hypothetical situation of "what if community C were like communities A and B," these fractional numbers were calculated only for comparison and age-adjusting purposes and are thus "fictitious" numbers. Thus, the standardized ratios, like the direct age-adjusted rate, are index numbers for comparison purposes only and are not real numbers. Also, SMRs and SIRs developed using different sets of standardized rates CANNOT be compared.

Using the crude rate in the standard population as a baseline value for comparison, we can easily arrive at an age-adjusted rate for community C. We simply multiply the crude rate of 26.5 cases per 100 population by the standardized ratio of 1.36 which gives us an indirect age-adjusted rate of 36.0 per 100 population. As a cautionary note, the measure of relative risk resulting from the use of the indirect method versus the direct method may vary widely, and depending on the standard selected, may even give opposite results.

In our example of breast cancer among white females in Iowa and Atlanta, if we decide to calculate what would have happened in Atlanta if those women had experienced the same rates as those which occurred in Iowa (i.e., we select Iowa as our standard), we obtain the results shown in Table 24.

Table 24. Breast Cancer Cases Expected Among White Females in Atlanta Based on Rates Occurring Among White Females in Iowa, 1976

Age	Observed Cases Atlanta	Female Population of Atlanta	Rate in Iowa/ 100,000	Expected Atlanta Cases Based on Iowa's Rate
<5	—	43,304	—	—
5-9	—	43,688	—	—
10-14	—	50,667	—	—
15-19	—	51,373	—	—
20-24	1	55,274	1.6	0.9
25-29	5	58,814	13.3	7.8
30-34	17	48,451	22.0	10.7
35-39	18	37,134	47.2	17.5
40-44	40	33,588	76.7	25.8
45-49	59	33,242	180.8	60.1
50-54	79	33,170	154.7	51.3
55-59	64	25,450	193.1	49.1
60-64	58	20,751	221.7	46.0
65-69	51	18,150	237.0	43.0
70-74	39	14,092	318.0	44.8
75-79	38	10,317	242.7	25.0
80-84	17	7,160	346.7	24.8
85+	15	5,138	410.9	21.1
All ages	500			427.9

Thus, based on Iowa's experience, we would have expected 427.9 cases to have occurred, but we actually observed 500 cases. The resulting SIR of 500/427.9 X 100 = 117 tells us that the risk of developing breast cancer in Atlanta was 17 percent higher than expected based on Iowa's experience. Hence, we conclude that the risk of developing breast cancer was higher in Atlanta than in Iowa, which was the same conclusion we drew based on the rates age-adjusted by the direct method. If we go the extra step of calculating the indirect age-adjusted rate for Atlanta by multiplying Iowa's crude rate of 91.0 (see table 21) by 1.17, the resulting rate of 106.5 leads us to the same conclusion which we had reached before, that is, the risk of developing breast cancer is greater in Atlanta than in Iowa.

It is appropriate to use the indirect method of age-adjustment rather than the direct method of age-adjustment when the population in individual age groups is small with few observed study events recorded. In this situation rates derived from these few observations may be too unreliable for use in the direct adjustment procedure. The indirect adjustment method, on the other hand, employs the more stable rates from a larger standard population to estimate expected numbers of events within each age group of the study population for comparison with the observed numbers of cases or deaths.

The indirect method of adjustment is used widely in special studies to compare incidence or mortality ratios of individuals such as smokers or industrial workers at potential excess risk compared with other study or population groups. The SIR and SMR measures based on indirect adjustment are convenient and generally easily interpreted measures of relative risk.

CUMULATIVE RATES

With either the direct or the indirect method of adjusting rates for variables such as age, care must be taken in selecting either a standard population or a standard set of rates for the adjustment procedure. Further, data sets adjusted using different standards may not be compared to one another. Thus, in comparing risks in two or more populations, it would be desirable if there were a method of adjusting for a characteristic such as age without having to choose some arbitrary standard.

One alternative to age adjustment is to compare so-called cumulative rates, i.e., to look at the accumulated risk over a certain age span such as age 0-14 or 0-64 or 0-74 in the two (or more) populations under study. The concept behind cumulative rates can be explained as follows. If we have the risk of disease during the first year of life (the incidence rate at age zero) and the risk during the second year of life for those alive at age one (the incidence rate at age one), we can calculate the risk of disease at any point between birth and age two by adding the rates for the two individual years. Similarly, if we want the cumulative risk between ages 0-14 or 0-64 or 35-64, we would add the individual yearly risks (rates) over the time interval of interest.

Ordinarily, tables with cancer incidence or mortality rates are not given by individual years of age but are usually given by 5-year age intervals. However, the rate is considered to apply to each year of age contained within the interval. So using our example of breast cancer incidence among women in Iowa, the age-specific rate for the age group 20-24 is 1.6 per 100,000. This implies that women age 20 have this annual risk as do women of age 21 or 22 or 23 or 24. Thus, by analogy with the example above for the first two years of life, the cumulative risk between the ages of 20-24 is (1.6 + 1.6 + 1.6 + 1.6 + 1.6) per 100,000 or a total of 8.0 per 100,000. Hence, if we wish to consider the rates between 0-74, we simply add up the age-specific rates for ages 0-4 and 5-9, and 10-14 all the way up to 70-74 and then multiply this number by 5 since each rate covers a 5-year age span. To convert our calculation to a percent or a rate per 100, we must divide our sum by 1,000, since our age-specific rates are expressed as rates per 100,000.

In our example of Iowa women with breast cancer (table 21) we see that the cumulative risk of a woman developing breast cancer between the ages of 0 and 74 is 1,466.1 x 5/1,000 = 7.3 per 100. This can be compared to a similar calculation for Atlanta women which reveals a cumulative risk of 1,715.7 x 5/1,000 = 8.6 per 100. Therefore, we conclude, based on our cumulative rates, that the risk of developing breast cancer in Iowa versus Atlanta in women between ages 0 and 74 is greater in Atlanta. This is the same conclusion which we drew based on our age-adjusted rates and the SIRs.

POPULATION AT RISK

The computation of rates for both incidence and mortality requires reliable estimates of the population at risk by age, sex, and race/ethnicity for each group or time period being studied. In the United States, population estimates are periodically available for every county in the United States from the U.S. Census Bureau.

In recent years concern has been raised regarding the undercounting of various population subgroups, especially minorities in certain geographic areas of the United States. Note, if our rate calculations use population estimates that are too low, i.e., underestimate the population at risk, our disease rates will be too high, i.e., will overestimate the actual risk of disease.

Q21

Answer the questions below using table 25 on the next page.

a. What is the average annual crude breast cancer incidence rate for white females? for black females?

b. Calculate the average annual age-adjusted rates per 100,000 for white and black women, standardized to the 1970 U. S. population distribution from table 22. How do you explain the difference between these and the crude rates? Which measure (crude or adjusted) is more appropriate for interpopulation comparisons and why?

Q22

Using the California data given in table 26, calculate the cumulative rate for males and for females between the ages of 0-74. What is your conclusion?

Q23

If the experience in community A is used to calculate SIRs for communities B and C with the result that community B has a SIR of 160 and community C has a SIR of 94, what conclusions can be drawn?

Q24

How would you calculate the cumulative rate of dying from lung cancer between the ages of 30 and 64?

Table 25. Real Data for Q21

AVERAGE ANNUAL BREAST CANCER INCIDENCE, San Francisco-Oakland SMSA
THIRD NATIONAL CANCER SURVEY, 1969-71

Age	White Females			Black Females		
	Number of Cases	Number of Population	Rate/100,000 per year	Number of Cases	Number of Population	Rate/100,000 per year
0-19	0	404,117	-	0	70,439	-
20-24	3	119,004	0.8	2	15,885	4.2
25-29	22	101,843	7.2	4	12,886	-
30-34	45	77,597	19.3	8	10,705	24.9
35-39	137	70,504	64.8	22	9,580	76.5
40-44	288	80,154	-	28	9,862	-
45-49	503	88,875	188.7	36	10,341	116.0
50-54	495	79,843	206.7	43	8,691	164.9
55-59	519	71,819	-	25	6,850	-
60-64	495	61,479	268.4	29	5,017	192.7
65-69	386	50,187	256.4	21	3,806	183.9
70-74	389	42,505	-	11	2,264	-
75-79	288	32,076	299.3	9	1,403	213.8
80-84	179	20,697	288.3	4	765	174.3
85+	147	14,817	330.7	3	629	159.0
Total	3,896	1,315,517		245	169,123	

Crude Rates _____ _____

Age-Adjusted Rates/100,000
(Standardized to the 1970
U.S. Population) _____ _____

Table 26. More Real Data for Q22

Age-Specific Cancer Incidence Rates per 100,000 Population by Sex, All Races, California, 1988		
Age Group	Males	Females
All Ages	365	365
0-4	20.3	21.0
5-9	12.4	8.6
10-14	12.6	12.9
15-19	19.5	17.9
20-24	31.1	32.2
25-29	65.5	57.1
30-34	83.2	92.1
35-39	111	147
40-44	141	239
45-49	222	378
50-54	393	530
55-59	668	720
60-64	1121	942
65-69	1710	1289
70-74	2316	1548
75-79	2907	1642
80-84	3298	1797
85+	3556	1776

Source: California Cancer Registry

Answer: Q21

The average annual crude breast incidence rates are:

a. white females = 3,896/(1,315,517 x 3) X 100,000 = 98.7

black females = 245/(169,123 x 3) X 100,000 = 48.3

Remember, the cases cover a 3-year period, so to get an average annual rate, the cases must be divided by 3 or the population multiplied by 3.

The average annual age-adjusted rates are:

b. white females = 86.3 and black females = 60.1.

The age-adjusted rates account for the difference in the age structure of the white and black populations bringing the rates closer together, although the risk was still substantially higher among white females. To compare blacks versus whites, the age-adjusted rates are more appropriate than crude rates.

Answer: Q22

The cumulative rate is obtained by adding up the 15 (5-year) age-specific rates for the ages 0-4, 5-9, ...70-74 multiplying by 5 and dividing by 1,000.

For males the rate is 5 (20.3 + 12.4 + 12.6 + 19.5 + 31.1 + 65.5 + 83.2 + 111 + 141 + 222 + 393 + 668 + 1,121 + 1,710 + 2,316)/1,000 = 5 (6,927)/1,000 = 34.6 per 100 or 34.6 percent.

The rate for females is 30.2 per 100.

The conclusion is that considering the age range of 0 - 74, males have a higher risk of developing cancer than do females.

Answer: Q23

Compared with community A, community B has a higher risk of disease, in fact, a 60 percent higher risk (SIR = 160) while community C has a somewhat lower risk, in fact, 6 percent lower (SIR = 94).

Answer: Q24

You would calculate the cumulative rate of dying from lung cancer between the ages 30 and 64 by adding up the age-specific mortality rates for each 5-year age group between 30 and 64 (i.e., 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, and 60-64) multiplying by 5 and dividing by 1,000 to get a rate per 100.

SECTION D

SURVIVAL ANALYSIS

SECTION D

SURVIVAL ANALYSIS

INTRODUCTION

The use of some measure of survival is necessary for evaluating patient care. Unfortunately survival measures are usually the least understood of all the basic statistical measures used in a hospital cancer registry. This is because it is a specialized topic that is not usually covered in basic statistics courses or text books.

In this section we will present the methods used for doing survival analysis in a step by step fashion. We will also give some guidelines for choosing which patient group to use, which method of analysis to use, and how to present the results.

There are several different types of measures that can be used:

Survival time: Average (mean) or median survival time for a group of patients

Survival rate: Observed survival rates measure the proportion of persons surviving (survival) regardless of cause of death (basically the proportion of patients surviving for a certain amount of time). This can be calculated using the direct method, the actuarial method, or the Kaplan-Meier Method.

Adjusted and relative survival rates: account for deaths from causes other than cancer.

Recurrence rate: Measured from the time of complete remission until time of recurrence.

Before you begin any survival analysis, first decide on the purpose of the study. In some cases you may be participating in a study designed by others (e.g., the American College of Surgeons) in which case the criteria for patient selection will be specified for you. Or, you may be asked to carry out a study suggested by the cancer committee or an epidemiologist, in which case they will help you determine which cases to select and how to group the cases. The cases that you will analyze and the method of analysis will depend on the site that you choose and prognostic factors relevant to that particular site, such as, age, race, histology, and treatment options. Look for sources of comparison data. This will determine what patients you will select, how you will group them, and what measure you will use for survival. Before the study begins, the following must be determined.

1. Selection of Cases

If you are located in a hospital-based registry, you will want to limit your study to analytic cases. These are the cases for which your doctors took part in the primary care of the patient when the cancer was first diagnosed and/or treated. If the purpose of the study is to evaluate treatment given at your hospital, you will want to exclude those patients who were diagnosed at your hospital but whose full first course of therapy was done elsewhere. Cases first diagnosed at autopsy and cases for which the death certificate is the only indication of a cancer diagnosis (death certificate only cases) are always excluded from a survival study.

You need to decide what years of diagnosis to use. If you are going to look at a 5-year survival rate, then you will need to include patients diagnosed during a 5-year period. If you have a very old registry, you may want to limit yourself to the more recent cases (when diagnostic procedures and coding schemes are more current) or you might want to include cases diagnosed over a longer period of time and group them by decade (or other time grouping) of diagnosis.

Generally, analysis is done separately for each stage of disease (e.g., localized, regional, distant or stage I, II, III, IV), in which case you will exclude patients with unknown stage. In-situ or stage 0 cases are excluded from survival analysis since their survival is expected to be near 100 percent.

You may want to exclude cases without a microscopic diagnosis, as there may be some doubt as to the primary site and histologic type of cancer. For some sites of cancer (e.g., eye), there may be a high proportion of cases not microscopically confirmed. In that case you could include cases with a clinical diagnosis.

You may want to exclude cases with multiple primaries to avoid the problem of "from which cancer did the patient die (or survive)?"

Depending on the purpose of your study, you may want to exclude cases occurring among children or the extreme elderly.

All cases that meet your documented criteria must be included--all inclusions and exclusions must be accounted for.

2. Followup

Make sure you have at least 90 percent (closer to 100 percent is better) successful followup for the patient group that you will use. Every case that is lost before the cutoff date of your study is a source of potential bias because those lost to followup are likely to have different characteristics than those you have successfully followed. This might mean doing a special followup for those cases that you will be using in your study.

3. Grouping of Cases

Run some preliminary tabulations on your patient group. Then you can see if you have enough patients with similar characteristics to group them by prognostic factors such as stage or age groups. Use this as an opportunity to do quality control. Investigate any cases with suspicious characteristics (e.g., the diagnosis of liver cases that are not hepatomas) to make sure they were not misclassified metastatic disease.

Ideally you want to group your cases so that each group contains cases similar for all prognostic factors. Practically, since you like to have at least 30 cases in each group, you may have to combine groups. If you plan to compare your survival results with those reported by others, you will want to select and group your cases in the same way as the comparison group.

Some factors that you may want to group cases by are:

Primary site (or a group of related sites, such as colorectal)

Stage at diagnosis or treatment time (You can use broad groupings such as localized, regional, and distant or more detailed stage groupings such as AJCC stage I, IIA, IIB, IIIA, IIIB, IV). If you do not have enough cases to include all the stage categories separately, you may wish to group cases as early (localized) versus late (regional + distant) stages.

Histology: some histologies have a different prognosis than others (e.g., islet cell of pancreas vs. other histologies, squamous cell carcinoma of the lung vs. other histologies).

Calendar year of diagnosis

Sex

Age at diagnosis

Race or ethnicity

Lab markers (such as estrogen receptors)

Socioeconomic status

Never group nonanalytic cases with analytic cases. This will introduce a serious and unpredictable bias to your analysis. Only a select group of cases may live long enough to be readmitted. Conversely, good survivors may not be readmitted at all.

4. Choosing the Starting Point

Choose the starting point for your calculation (i.e., survival from when). Usually you will use the date of first diagnosis or the date of first treatment, depending on the purpose of your study. Other starting times may be date of first symptoms, or, for a recurrence study, date of first remission. If you are looking at survival for nonanalytic cases, you might want to use date of admission to your hospital. If you are comparing your survival rates to someone else's, make sure you choose the same starting point. Various reference dates are commonly used as starting times for evaluating the effects of therapy. These include (1) date of first diagnosis, (2) date of first visit to physician or clinic, (3) date of hospital admission, and (4) date of treatment initiation. The SEER program uses date of diagnosis as the starting point for their survival figures. For evaluating therapy, the American College of Surgeons uses date of first treatment. Survival measured from appearance of first symptoms will appear longer than survival measured from diagnosis or from the beginning of treatment because

there is a lag time between these events. Include which starting point you used in your report.

5. Choosing the Ending Point

It is natural to think of survival time as survival until death. For most studies, there will also be a study cutoff date. This may be based on the date of last complete follow-up information for the patient group or another date chosen to match the purpose of the study. In the absence of a study cut-off date, the ending date is generally date of death or date of last contact (for patients still alive). If there is a study cutoff date, information on survival beyond that date is not used in calculating the survival experience of the study group. If you are doing a recurrence rate study, you will have an ending point of date of first recurrence.

6. Calculating Survival Times

Survival time is calculated by subtracting the date of diagnosis (or whatever starting time you decide to use) from the date of last contact (or death). The time intervals can be measured in terms of years, months, or even weeks and days depending on the purpose of your study. For example, for patient #4 on the colon cancer listing in table 27, the follow-up date is 02/81 and the date of diagnosis is 09/79. Notice that the follow-up month is smaller than the diagnosis month, so the patient had only one complete year of survival plus 5 months, or a survival time of 1 year and 5 months. Patient #34 has an unknown month of diagnosis. When computing years surviving, if you have an unknown month and you can make no closer estimate, use the month of July (the 7th month) instead of unknown (unless this would make the date of diagnosis later than the date of last contact). Note, if a patient died after the study cutoff date, remember to change vital status to alive when doing survival calculations.

The listing on the next page (table 27) of localized colon cancer cases diagnosed during 1978-1987 will be used to illustrate the computations of some of the survival measures described on page 122. For our examples, the date of diagnosis will be used as the starting point. The study cutoff date is 12/88.

Table 27. CASES OF LOCALIZED COLON CANCER DIAGNOSED AT MY HOSPITAL, 1978-87

Obs	Sex	Age	Race/Ethnicity	DX Date	FUP Date	Status	Survival Time
1	Female	61	White	04/78	10/88	Alive	10 Y 06 M
2	Female	78	White	11/78	07/79	Dead	0 Y 08 M
3	Female	69	White	01/79	08/88	Alive	9 Y 07 M
4	Male	62	White	09/79	02/81	Dead	1 Y 05 M
5	Male	77	White	09/79	06/88	Alive	8 Y 09 M
6	Male	81	White	09/79	02/84	Dead	4 Y 05 M
7	Male	81	White	11/79	12/79	Dead	0 Y 01 M
8	Female	83	White	12/79	09/86	Dead	6 Y 09 M
9	Male	72	White	03/79	04/79	Dead	0 Y 01 M
10	Male	85	White	05/80	02/82	Dead	1 Y 09 M
11	Female	58	White	08/80	09/80	Alive	0 Y 01 M
12	Female	89	White	10/80	10/83	Dead	3 Y 00 M
13	Female	75	White	12/80	12/88	Dead	8 Y 00 M
14	Male	84	White	03/82	03/85	Dead	3 Y 00 M
15	Female	64	White	04/82	01/88	Alive	5 Y 09 M
16	Male	72	White	05/82	02/89	Alive	6 Y 06 M*
17	Male	67	White	07/82	07/87	Alive	5 Y 00 M
18	Female	60	White	08/82	09/88	Alive	6 Y 01 M
19	Female	70	White	06/82	07/88	Alive	6 Y 01 M
20	Female	76	White	11/82	02/88	Alive	5 Y 03 M
21	Male	86	White	12/82	12/88	Alive	6 Y 00 M
22	Female	66	Hispanic	04/83	03/88	Alive	4 Y 11 M
23	Female	64	Hispanic	06/83	01/87	Alive	3 Y 07 M
24	Female	69	Black	08/83	03/87	Alive	3 Y 07 M
25	Male	68	White	08/83	08/88	Alive	5 Y 00 M
26	Female	85	White	08/83	09/83	Dead	0 Y 01 M
27	Female	79	White	03/84	09/88	Dead	4 Y 06 M
28	Male	76	White	06/84	07/84	Dead	0 Y 01 M
29	Female	75	White	02/85	08/88	Alive	3 Y 06 M
30	Female	64	White	02/85	10/88	Alive	3 Y 08 M
31	Female	78	Korean	06/85	10/88	Alive	3 Y 04 M
32	Female	65	Hispanic	06/85	08/88	Alive	3 Y 02 M
33	Male	67	Chinese	09/85	07/88	Alive	2 Y 10 M
34	Female	71	Black	XX/85	08/88	Alive	3 Y 01 M
35	Female	97	White	02/86	10/87	Alive	1 Y 08 M
36	Female	72	White	03/87	12/87	Alive	0 Y 09 M
37	Female	72	White	04/87	12/87	Alive	0 Y 08 M
38	Female	91	White	07/87	10/87	Alive	0 Y 03 M
39	Female	84	White	07/87	11/87	Alive	0 Y 04 M
40	Male	59	White	10/87	09/88	Alive	0 Y 11 M
41	Female	66	White	12/87	12/88	Alive	1 Y 00 M

*Calculated to study cut-off date 12/88 - not to FUP date 02/89

SURVIVAL TIME

Typically, survival time is used to give an idea of how long patients tend to live after diagnosis with a certain type of cancer. It is a more easily understood measure than a survival rate. However, most comparison survival data are published as rates as opposed to survival times.

1. Average (Mean) Survival Time

To look at a measure of "typical" time surviving, our first instinct might be to use the average survival time (e.g., on the average, a person with lung cancer survives 6 months after diagnosis). There are two problems with using this measure. The first problem is that when we talk about average survival time, we are really thinking about an average time until death. If we knew the time until death for each one of our patients (i.e., all our patients have to be dead) then we could add up all the survival times and divide by the number of patients and get an average survival time. Fortunately, we are rarely in the situation where all of our patients are dead. The other disadvantage for using this measure is that the average is very sensitive to extreme values. Therefore, a patient who lives a lot longer (or a lot shorter) time than the others, will affect the average survival time inordinately.

2. Median Survival Time

To overcome the disadvantages of the average survival time, we turn to the median survival time. Although the median is not as commonly used in statistical tests, this measure has the advantage that extreme values do not much affect it. If you have a group of patients that were all diagnosed (or treated) at the same time, you can calculate a median if at least half of the patients are dead. Sort your patients in order from shortest to longest survival and choose the middle value to get the median survival time.

For patients who were not diagnosed or treated at the same time, the median survival time can be found at the 50 percent survival point on a graph of survival rates. (For an example see figure 20).

OBSERVED SURVIVAL RATE

An observed survival rate is a measure of survival of a patient group for a specific period of time after diagnosis (or treatment). This is interpreted as the proportion (or percent) of patients surviving a specified amount of time after cancer diagnosis or treatment. In computing the observed survival rate, deaths from other causes are treated just like deaths from cancer. Therefore, the observed survival rate should be interpreted as the likelihood of surviving all causes of death (i.e., being alive) for a certain time after cancer diagnosis, not the likelihood of surviving that cancer.

Most of us are familiar with seeing the 5-year survival rate reported. A 5-year rate has sometimes been considered the *cure rate*. However, 5 years is not an appropriate cut off time for all cancers. For some cancers such as breast cancer, it is more effective to calculate a 10- or even 20-year survival rate. For other cancers such as pancreas, we might be more interested in the 1-year or 2-year survival rate. For simplicity, the 5-year survival rate will be used in the discussion to follow.

1. Direct Method for Calculating an Observed Survival Rate

It is not recommended that you use the direct method for calculating survival, but you should know what it means when it is reported elsewhere, and understanding it will help you understand other methods of survival analysis. The direct method is the most intuitive approach for calculating a survival rate. Like other rates, it is the proportion of events that occur in a certain amount of time. From descriptive statistics, you know that a proportion is a part of the total divided by the total. In this case, the part of the total is the number surviving, and the total is the number at risk of dying, and usually the time period is 5 years. The calculation would look like:

$$\frac{\textit{Number surviving for 5 years}}{\textit{Number at risk for 5 years}}$$

- The number at risk for 5 years would be those patients for which you have at least 5 years of complete followup. To find those patients:

  Select a cohort of cases that have had a chance to survive 5 years, i.e., their date of diagnosis (or treatment) was at least 5 years prior to the study cutoff date.

  In our example in table 27, we have completed followup through February of 1989; however, our study cutoff date is December, 1988. Therefore, all patients diagnosed December 1983 or earlier are eligible for inclusion in the study group. Since the list is sorted by diagnosis date, we can look at the listing to see that all patients through patient number 26 can be included in the study. Patients 27 through 41 must be excluded because they were not diagnosed at least 5 years prior to the study cutoff date. It is helpful to have patients sorted by year of diagnosis (or treatment, if this is your starting point) for this type of calculation.

  If any of the qualified patients were lost to followup (vital status alive, survival time less than 5 years) they also must be excluded, because we don't have 5 years of information on them. In our example, patients 11, 22, 23 and 24 must be excluded. We now have 26 - 4 = 22 patients that can be included.

- After counting the number at risk for 5 years we need to count the number surviving for 5 years. Remember that patients who are known to have died after the cutoff date were still alive as of that date. These will be the patients that have a survival time of 5 years or greater. (It doesn't matter if they lived or died after that point). In our example, patients 1, 3, 5, 8, 13, 15, 16, 17, 18, 19, 20, 21, and 25 survived at least 5 years. Thus, we had 13 patients surviving 5 years.

- Divide the number of survivors by the number at risk to get the proportion surviving for 5 years. This is our 5-year observed survival rate calculated by the direct method. For our example we have: 13 divided by 22 = 0.59. If you would rather work with percentages, you can multiply the result by 100 and express the 5-year survival rate as a percent, 0.59 X 100 = 59 percent.

2. Actuarial (Life Table) Method for Calculating an Observed Survival Rate

Although both the actuarial and Kaplan-Meier methods are life table methods, many people use the term "life table" synonymously with the actuarial method.

This method applies a statistical "trick" to use information from patients who were diagnosed (or treated) less than 5 years ago in the calculation of a 5-year survival rate. To do this we calculate a 1-year survival rate for all our patients; then for those patients that survived 1 year, we calculate another 1-year survival rate for those who survived the second year and so on, ending with calculating the rate at which 4-year survivors lasted that fifth year. Then we multiply the rate for each interval by the rate for the succeeding interval to calculate the overall 5-year survival rate. We can multiply because of a rule in statistical probability theory that says if we want to get an overall estimate of the likelihood of two independent events both happening, we multiply the individual probabilities. Annual survival rates satisfy this rule. As an added bonus, we also get a picture of the pattern of survival, starting at 1 year.

Q1

Survival can be measured in terms of survival _____ (average (mean) or median) or in terms of a survival _____.

Q2

Before beginning your survival study name six things to be considered:

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____

Q3

What is the advantage of median survival time over average (mean) survival time?

_____

Q4

When you measure the survival of a patient group for a specific period of time after diagnosis, it is called an _____ _____ _____.

Q5

A 5-year survival rate is sometimes called a cure rate, but this term is not appropriate for all cancer sites. Why not?

_____

Q6

How is a 5-year observed survival rate calculated using the direct method?

_____

Q7

When you calculate a 1-year survival rate for all patients, then another 1-year survival rate for those who survived the second year and so on, this is called the _____ _____ or sometimes the _____ _____ _____ for calculating an observed survival rate.

Answer: Q1

Survival can be measured in terms of survival time (average (mean) or median) or in terms of a survival rate.

Answer: Q2

Before beginning your survival study, six things to be considered are:

1. Selection of cases (determine inclusions and exclusions)
2. Followup for at least 90 percent of the study group
3. At least 30 to each grouping; determine if you have enough cases
4. Choosing the starting point such as date of first diagnosis, date of first treatment, or date of first remission.
5. Choosing the ending point, that is, a study cutoff point
6. Calculating the survival in terms of years, months, weeks or days depending on the purpose of your study.

Answer: Q3

Median survival has the advantage that extreme values do not have as much effect as they do in an average (mean) survival time.

Answer: Q4

When you measure the survival of a patient group for a specific period of time after diagnosis, it is called an observed survival rate.

Answer: Q5

A 5-year "cure" rate is a term sometimes used but it is not appropriate for all cancer sites because cancers of many sites, such as breast, may recur as many as 15 or 20 years after treatment.

Answer: Q6

A 5-year observed survival rate using the direct method is calculated by dividing the number surviving 5 years by the number at risk for 5 years (those with 5 years of complete follow-up).

Answer: Q7

When you calculate a 1-year survival rate for all patients, then another 1-year survival rate for those who survived the second year and so on, this is called the actuarial method or sometimes the life table method for calculating an observed survival rate.

The easiest way to calculate survival using the actuarial method is to fill out a life table (hence the name, life table method). A blank life table is shown in table 28 below.

Table 28. Blank Life Table for Calculating Survival Rates by Actuarial Method

A (i)	B (l)	C (d)	D (w)	E (l')	F (q)	G (p)	H (P or CP)
Interval of Observation (Time after diagnosis in years)	# Alive at Beginning of Interval	# Dying During Interval	# Last Seen Alive During Interval "Withdrawals"	Effective # Exposed to Risk of Dying (B - 1/2 D)	Proportion Dying During Interval (C / E)	Proportion Surviving the Interval (1.0 - F)	Cumulative Survival Rates
0 - < 1							
1 - < 2							
2 - < 3							
3 - < 4							
4 - < 5							
5 or more							

The easiest way to explain a life table is to go through an example. We will use the same patient listing (table 27 on page 121) that we used in illustrating the direct method (page 123). We will calculate a 5-year survival rate using interval 1-year, 2-year, 3-year and 4-year survival rates in this example.

The column headings A-H will be used in explaining how to fill out the life table (table 28). The letters in parenthesis: (l), (d), (w), (l'), (q), (p) and (P or CP) are headings used by some computer programs. The description for filling out each column and row and an example are presented on pages 130-132. Notice that columns C and D should be filled out before column B.

If you are doing the example, you should check your tabulations in columns C and D against the filled-out life table (table 29) before you go on to column B.

Steps in Calculating Survival Rates: Actuarial Method (Tables 28 and 29)

Col	General Procedure	Example	Explanation
A	Fill in the intervals to be used for the survival rate calculations. Intervals should be mutually exclusive.	Since we are going to calculate annual rates, enter 0 - < 1 yr. in the first row. This notation means from the time of diagnosis up to, but not including, 1 year from diagnosis.	In each row we will calculate a 1-year survival rate. Patients who died or are "withdrawn alive" (those lost to followup or diagnosed less than a year ago) will be used for calculations in the interval during which they occur, but not in the following intervals.
	The last interval should be for those "left over" survivors for the whole study.	Follow with: 1 - <2 years, 2 - <3 years, 3 - <4 years and 4 - <5 years. Since we are finding a 5-year survival rate, the last interval is 5 years or more.	We could continue the rows up to 10 years if we wanted a 10-year survival rate, or make the intervals smaller (e.g., 0 to <3 months, 3 to <6 months, etc.).
B	Fill in the number of patients alive at the beginning of the interval for the first row.	All 41 of our patients are alive at the start of the study.	All patients in the study will be alive at the beginning (patients diagnosed at autopsy or reported based on death certificate information only are not included in a survival study).
		Complete cols C & D	
	For row 2, take the number from row 1, column B, subtract the number in row 1, column C, then subtract the number from column D.	The # in row 1 is 41. Subtract 5 (from column C), subtract 6 (from column D) to get 30. Put 30 into row 2, column B.	For the second row, we will be calculating a 1-year survival rate for those patients who survived the first year. Therefore, we subtract those who died and those withdrawn from the study alive.
	Repeat for the rest of the rows. Column B from the last row should equal the sum of columns C and D for that row.	Enter next rows: 30 - 2 - 2 = 26 for row 3 26 - 0 - 1 = 25 for row 4 25 - 2 - 7 = 16 for row 5 and finally, 16 - 2 - 1 = 13 which is equal to 2 + 11 (the total of column C+D for this row).	If less than 10 patients are left in column B, quit calculating; your rate will be too unreliable.

Steps in Calculating Survival Rates: Actuarial Method (Tables 28 and 29)

Col	General Procedure	Example	Explanation
C&D	Tabulate the patients who died during the interval and enter in the proper row in column C. Enter those still alive who withdrew during the interval in column D.	Patient 1 is alive after 10 years and would be counted in column D, row 6 (more than 5 years). Patient 2 would be tabulated in column C, row 1, patient 3 in column D, row 6, 4 in column C, row 2, etc.	It's easiest just to go down the list and put a tick mark where each patient should be tabulated, then add up the tick marks in each box.
	The sum of all entries in column C + column D should be the total of all the patients in the study.	The total of all entries in columns C and D should equal 41. Return to col. B in row 2.	Notice that you only have to look at years surviving and vital status to decide in what row and column the patient should be tabulated.
E	For each row, subtract 1/2 of column D from column B.	For row 1, column D is 6, 1/2 of 6 is 3, subtract 3 from 41 to get 38. For row 2, 1/2 of 2 is 1, so subtract 1 from column B to get 29. In row 3, 1/2 of 1 is 0.5. Subtract 0.5 from 26 to get 25.5. Next, 25 - (1/2 X 7) = 21.5; and 16 - (1/2 X 1) = 15.5.	This column is for the "effective number exposed to the risk of dying." Those patients still alive, but without a whole year of observation during that interval were, on the average, observed for 1/2 of the interval (they contributed only 1/2 a person-year at risk).
F	In each row, divide column C by column E. Do not round your number off to less than 3 places after the decimal.	In row 1, divide 5 (col C) by 38 (col E) to get 0.132. For row 2, 2/29 = 0.069, for row 3, 0/25.5 = 0.000. Row 4 = 0.093, row 5 = 0.129.	This is like calculating the proportion dying for each interval, but instead of dividing by the number starting the interval, we consider that some of the patients weren't observed for the whole interval, so we divide by an adjusted number (column E).

Steps in Calculating Survival Rates: Actuarial Method (Tables 28 and 29)

Col	General Procedure	Example	Explanation
G	Subtract column F from 1.000 to get the proportion surviving. Keep 3 places after the decimal in your answer	For row 1, 1.000 - 0.132 = 0.868 row 2, 1.000 - 0.069 = 0.931 row 3 = 1.000, row 4 = 0.907, row 5 = 0.871	Subtracting the proportion dying from 1.000 to get the proportion surviving is like subtracting the percent dying from 100 percent to get the percent surviving.
H	For the first row only, put the number from col. G into col. H. For subsequent rows, multiply column G in each row by column H in the row **above** to get column H.	Row 1 = 0.868 Row 2 = 0.868 X 0.931 = 0.808 Row 3, 0.808 X 1.000 = 0.808 Row 4, 0.808 X 0.907 = 0.733 Row 5, 0.733 X 0.871 = 0.638	Column G is the annual survival rate for patients who already survived the previous intervals. To take into account the risk of dying in these previous intervals, we multiply the annual survival rate by the survival rate for the previous intervals to get the *cumulative* survival rate from diagnosis to the end of the interval.
	Row 1 will contain the survival rate for the 1st interval, row 2 the survival rate for 2 years since diagnosis, and so on.	Therefore, the 1-year survival rate is 0.868, the 2-year rate is 0.808, the 5-year survival rate is 0.638.	We can multiply because of the rule in statistical probability theory that says if we want to get an overall estimate of the likelihood of two independent events both happening, we multiply the individual probabilities.
	If you prefer working with percentages, multiply the proportion by 100 and round to the nearest whole percent.	1 yr. = 0.868 X 100 = 87 percent so these proportions are the same as survival rates: 1 yr. = 87 percent 2 yr. = 81 percent 3 yr. = 81 percent 4 yr. = 73 percent 5 yr. = 64 percent	Thus, the probability of surviving for two years is the probability of surviving the first year times the probability of surviving the second year (for those who were still at risk entering the second year).

Table 29. Actuarial Life Table for Patients Diagnosed at My Hospital, 1978-87

A (i)	B (l)	C (d)	D (w)	E (l')	F (q)	G (p)	H (P or CP)
Interval of Observation (Time after diagnosis in years)	# Alive at Beginning of Interval	# Dying During Interval	# Last Seen Alive During Interval "Withdrawals"	Effective # Exposed to Risk of Dying (B - 1/2 D)	Proportion Dying During Interval (C / E)	Proportion Surviving the Interval (1.0 - F)	Cumulative Survival Rates
0 - < 1	41	5	6	38	.132	0.868	0.868
1 - < 2	30	2	2	29	.069	0.931	0.808
2 - < 3	26	0	1	25.5	.000	1.000	0.808
3 - < 4	25	2	7	21.5	.093	0.907	0.733
4 - < 5	16	2	1	15.5	.129	0.871	0.638
5 or more	13	2	11				

Notice that the 5-year survival rate of 64 percent (See column H in table 29 above) is somewhat higher than the 59 percent survival rate using the direct method (See pp. 123-124). There is a difference in the rates because we were able to use the experience of all 41 patients included in our study and not just the 22 patients diagnosed December 1983 or earlier and not lost to followup.

If you are doing the actuarial method of survival by hand, it is very useful to have a list of patients that have been sorted first by survival time, and then by vital status. Better still, if you do not have a computer to sort for you, prepare a data card for each observation (patient) and write down the variables that you are going to use: age, race, sex, date of diagnosis (treatment), date of last contact/death, vital status, stage, survival time.

3. Kaplan-Meier (Product Moment) Method for Calculating an Observed Survival Rate

The Kaplan-Meier method is recommended for those registries that have a computer that will do these calculations for you. It differs from the actuarial method in that a calculation is done every time someone dies. Because of that, it is a more exact description of the pattern of survival seen in your patients. It also differs in that patients who are withdrawn from the study are not used in ensuing calculations. They are dropped at the point at which they drop out of the study, and no estimation of contribution of person-years at risk is made. Unless you have a very small patient group, there are a large number of calculations required. For that reason, usually only those with a computer program that calculates these rates will use this method. It is presented here so that you will understand what the computer program is doing for you. If you don't have access to a computer program that does Kaplan-Meier and you have a large enough patient group (at least 20, but preferably 30), it is sufficient to use the actuarial method described above.

To begin the Kaplan-Meier procedure, first sort your patients by survival time in months and then by vital status as shown in table 30. The calculations are shown in table 31.

Table 30. CASES OF LOCALIZED COLON CANCER DIAGNOSED AT MY HOSPITAL, 1978-87

Obs	Sex	Age	Race/Ethnicity	DX Date	FUP Date	Status	Survival Time
11	Female	58	White	08/80	09/80	Alive	01 M
9	Male	72	White	03/79	04/79	Dead	01 M
28	Male	76	White	06/84	07/84	Dead	01 M
7	Male	81	White	11/79	12/79	Dead	01 M
26	Female	85	White	08/83	09/83	Dead	01 M
38	Female	91	White	07/87	10/87	Alive	03 M
39	Female	84	White	07/87	11/87	Alive	04 M
37	Female	72	White	04/87	12/87	Alive	08 M
2	Female	78	White	11/78	07/79	Dead	08 M
36	Female	72	White	03/87	12/87	Alive	09 M
40	Male	59	White	10/87	09/88	Alive	11 M
41	Female	66	White	12/87	12/88	Alive	12 M
4	Male	62	White	09/79	02/81	Dead	17 M
35	Female	97	White	02/86	10/87	Alive	20 M
10	Male	85	White	05/80	02/82	Dead	21 M
33	Male	67	Chinese	09/85	07/88	Alive	34 M
14	Male	84	White	03/82	03/85	Dead	36 M
12	Female	89	White	10/80	10/83	Dead	36 M
34	Female	71	Black	XX/85	08/88	Alive	37 M
32	Female	65	Hispanic	06/85	08/88	Alive	38 M
31	Female	78	Korean	06/85	10/88	Alive	40 M
29	Female	75	White	02/85	08/88	Alive	42 M
24	Female	69	Black	08/83	03/87	Alive	43 M
23	Female	64	Hispanic	06/83	01/87	Alive	43 M
30	Female	64	White	02/85	10/88	Alive	44 M
6	Male	81	White	09/79	02/84	Dead	53 M
27	Female	79	White	03/84	09/88	Dead	54 M
22	Female	66	Hispanic	04/83	03/88	Alive	59 M
17	Male	67	White	07/82	07/87	Alive	60 M
25	Male	68	White	08/83	08/88	Alive	60 M
20	Female	76	White	11/82	02/88	Alive	63 M
15	Female	64	White	04/82	01/88	Alive	69 M
21	Male	86	White	12/82	12/88	Alive	72 M
18	Female	60	White	08/82	09/88	Alive	73 M
19	Female	70	White	06/82	07/88	Alive	73 M
16	Male	72	White	05/82	02/89	Alive	78 M
8	Female	83	White	12/79	09/86	Dead	81 M
13	Female	75	White	12/80	12/88	Dead	96 M
5	Male	77	White	09/79	06/88	Alive	105 M
3	Female	69	White	01/79	08/88	Alive	115 M
1	Female	61	White	04/78	10/88	Alive	126 M

Steps in Calculating Survival Rates: Kaplan-Meier Method (Table 31)

Col.

A — Write the survival time in months in order from smallest value to largest value in column A, ending with 60 months if you want a 5-year survival rate (or 12 months for a 1-year rate, or 120 months for a 10-year survival rate). You do not have to list those months when no one died or withdrew. For example, in the current study no one died or withdrew during months 2, 5, 6, or 7, etc.

C & D — Tabulate all your patients into the appropriate row of either column C (died during that month) or column D (withdrew during that month).

B — Enter the number remaining in the study for each row.

For row 2, enter the number of patients in the study. For successive rows, subtract the number dying and withdrawing in the previous row from the number entering alive in the previous row.

E — For those months in which someone has died, calculate the proportion dying by dividing the number dying in that month by the number present through the whole interval (column C divided by (column B - column D).

F — Subtract the proportion dying from 1.000 for each row to get the proportion surviving that interval.

G — To compute the survival rate, for the first row in which someone died, copy the proportion surviving into the survival rate column. For the second row in which someone died, multiply the proportion surviving in that row by the survival rate from the **previous row in which someone died**. Enter the result into the survival rate for the row. Continue with the rest of the rows in which someone has died.

To find the 1-year survival rate, use the last computed survival rate just previous to 12 months. If no one has died before 12 months, use 100 percent. The same principle holds for the 2-year, 3-year, etc., survival rates.

The 1-year survival rate is 87 percent, the 2-year rate is 81 percent, the 3-year rate is also 81 percent, the 4-year rate is 74 percent, and the 5-year rate is 65 percent. This is almost exactly what we found by using the actuarial method. The Kaplan-Meier and the actuarial method will usually give very similar results.

Table 31. Kaplan-Meier Table for Localized Colon Cases Diagnosed at My Hospital, 1978-87

A	B	C	D	E	F	G
Time (months)	Entered Alive	Died	Withdrawn	Proportion Dying	Proportion Surviving	Cumulative Survival Rate
1	41	4	1	4/40=0.100	0.900	0.900
3	36		1			
4	35		1			
8	34	1	1	1/33=0.033	0.967	0.870
9	32		1			
11	31		1			
12	30		1			
17	29	1		1/29=0.034	0.966	0.840
20	28		1			
21	27	1		1/27=0.037	0.963	0.809
34	26		1			
36	25	2		2/25=0.080	0.920	0.744
37	23		1			
38	22		1			
40	21		1			
42	20		1			
43	19		2			
44	17		1			
53	16	1		1/16=0.062	0.938	0.698
54	15	1		1/15=0.067	0.933	0.651
59	14		1			
60	13					

If you are working with a small group of patients, it is recommended that you use the Kaplan-Meier method for calculating the observed survival rate. For larger groups of patients (30 or more), the actuarial method is an acceptable alternative which requires fewer calculations and the method that has been most commonly used in the past. The direct method described on pp. 123-124 is not recommended because it limits the number of patients that can be used in the study, and it does not use information from your more recent patients. If you are comparing your survival rate with someone else's rate, it is important that you choose the same method of calculating survival, as each method will give you a slightly different rate.

EXCLUDING NONCANCER DEATHS

If you look at the calculations for the observed survival rates from above, you notice that no consideration is taken of the fact that patients die from causes other than cancer. Observed survival rates underestimate survival from cancer because they group deaths from all causes in the calculations. There are two general ways to correct for this. For registries that are able to get good cause of death information, it is possible to calculate an *adjusted survival rate*. For registries where reliable and complete cause of death information is not available, it is possible to do an indirect adjustment for other causes of death by calculating a *relative survival rate*.

1. Adjusted Survival Rate

If you have good cause of death information (i.e., you know if patients died from the cancer under study), you may use any of the methods for calculating an observed survival rate with minor modifications.

a. Only count as deaths those patients who died from the cancer under study.

b. Consider those patients who died from the other causes to be withdrawn from the study at that point (i.e., tabulated with the withdrawn alive cases).

Table 32 gives an abbreviated patient listing to illustrate how to tabulate adjusted rates taking into account whether patients died with or without the cancer under study. Blank tables for calculating an adjusted actuarial rate and an adjusted Kaplan-Meier rate (for the first 21 months of survival) are given in tables 33 and 34.

Table 32. LOCALIZED COLON CANCER DIAGNOSED AT MY HOSPITAL 1978-87

Obs	Sex	Age	Cause of Death	DX Date	FUP Date	Status	Survival
1	Female	61		04/78	10/88	Alive	10 Y 06 M
2	Female	78	Heart Disease	11/78	07/79	Dead	0 Y 08 M
3	Female	69		01/79	08/88	Alive	9 Y 07 M
4	Male	62	Colon Cancer	09/79	02/81	Dead	1 Y 05 M
5	Male	77		09/79	06/88	Alive	8 Y 09 M
6	Male	81	Heart Disease	09/79	02/84	Dead	4 Y 05 M
7	Male	81	Unknown	11/79	12/79	Dead	0 Y 01 M
8	Female	83	Colon Cancer	12/79	09/86	Dead	6 Y 09 M
9	Male	72	Prostate Cancer	03/79	04/79	Dead	0 Y 01 M
10	Male	85	Rectal Cancer	05/80	02/82	Dead	1 Y 09 M
11	Female	58		08/80	09/80	Alive	0 Y 01 M

Table 33. Actuarial Life Table To Be Used for an Adjusted Survival Rate

A (i)	B (l)	C (d)	D (w)	E (l')	F (q)	G (p)	H (P or CP)
Interval of Observation (Time after diagnosis in years)	# Alive at Beginning of Interval	# Dying from This CA During Interval	# Last Seen Alive During Interval or Dying of Other Causes	Effective # Exposed to Risk of Dying from This CA (B - 1/2 D)	Proportion Dying from This CA During Interval (C / E)	Proportion Surviving This CA During the Interval (1.0 - F)	Cumulative Adjusted Survival Rates
0 - < 1	10		# 2, 9, 11				
1 - < 2		# 4, 10					
2 - < 3							
3 - < 4							
4 - < 5			# 6				
5 or more		# 8	# 1,3,5				

Table 34. Kaplan-Meier Life Table To Be Used for an Adjusted Survival Rate

A	B	C	D	E	F	G
Time (months)	Entered Interval Alive	Died From This CA	Withdrawn Alive or Died From Other Causes	Proportion Dying From This CA	Proportion Surviving This CA	Cumulative Adjusted Survival Rates
1	10		# 9, 11			
8			# 2			
17		# 4				
21		# 10				

Patients from table 32 would be tabulated as follows in both the actuarial and the Kaplan-Meier tables for adjusted survival rates:

Patient 7 would be dropped from the study (not tabulated) because it is not known if he died from colon cancer.

Patients 1, 3, 5, and 11 are alive and would be tabulated in column D in the appropriate row.

Patients 2, 6 and 9 would also be tabulated in column D because they died of other causes and would be treated as withdrawing from the study at that point.

Patients 4 and 8 would be tabulated in column C because they died from the cancer under study. Patients 1,3, 5 and 8 all survived beyond 5 years although patient 8 is known to have died in the seventh year.

Patient 10 died from what was recorded as rectal cancer. Colon and rectal cancers are sometimes misrecorded on death certificates. It would take further research to decide if this patient ever had rectal cancer, or really died from colon cancer. The study designer should make the decision on how to tabulate this patient. In tables 33 and 34 we have assumed that the patient died from "this cancer."

The remainder of tables 33 and 34 have not been completed since in real life we would not calculate survival rates on only 10 patients. Usually, at least data for 25 patients are necessary in order to calculate meaningful survival rates. Tables 33 and 34 are only shown to demonstrate how patients not dying from the cancer of interest would be handled.

2. Relative Survival Rate

Since in the real world most registries do not have good enough cause of death information, it is possible to *indirectly* adjust the observed survival rate to remove the effect of normal mortality. Remember that to combine survival experiences in the life table, we multiplied the survival rate for each interval by the survival rate from the previous interval. To account for the risk of dying from other causes, we divide the *observed survival rate* by the *expected (normal) rate.*

Expected survival rates can be obtained from standard life expectancy tables. For the United States, standard life tables for males and females for various race and ethnic groups are produced periodically. Some tables showing expected survival rates for 1970 and 1980 are shown in appendix 3. These tables are based on the mortality experience of the entire U. S. population including those who died from cancer. However, for calculating relative survival rates, using the U. S. life tables will yield reliable results for comparing the chance of patient groups escaping deaths due to cancer.

Some states may also produce life tables using their own mortality experience. If these are available for your state, they can be used for calculating the relative survival rate of your patients. In general, most computerized registries will have access to analytic packages in which relative survival rates will be produced using built-in life tables, and the registrar will not have to calculate rates manually. To understand the process of how relative survival rates are constructed, the following discussion is given.

First, calculate the observed survival rate by any of the methods presented above.

Then, calculate the 1-year relative survival rate:

- For each patient in the study group, look up the expected 1-year survival rate by age at diagnosis, race, sex, and year of diagnosis in a table of expected survival rates (see appendix 3).
- Average the expected survival rates for all of your cases.
- Divide the observed 1-year survival rate for the study group by the average expected 1-year survival rate to get the 1-year relative survival rate.

Next, to calculate 2-year 3-year, ... etc. relative survival rates:

- For each case, add 1 year to the age and 1 year to the date of diagnosis. Look up the new expected survival rate for the second year in the appropriate table.
- For each case, multiply the 1-year expected survival rate for the second year by the first year expected survival rate.
- Average these multiplied rates for all your cases.
- Repeat this process for your 3-year, 4-year, ... expected rates
- Divide the observed rates for each year by the expected rates for the corresponding year to get the relative rates.

Remember, add another year to the age for each case. Also, "age" the year of observation as well. For example, a patient diagnosed in 1975 and who survived 1 year will next be observed in 1976. Thus, the 1980 expected life table is now more appropriate to use for expected survival than the 1970 life table, since 1976 is closer to 1980 than to 1970 and therefore life expectancy in 1976 is more likely to be closer to life expectancy in 1980 than in 1970.

It is helpful to use a list of patients or a set of cards sorted by sex, race, age, and year of diagnosis to facilitate looking up the expected survival rate. Many tumor registry computer programs will provide you with a sorted list of patients.

Using our colon cancer listing from table 27, look up the expected 1-year normal survival for each of our 41 patients. To illustrate how this works, we first sort our patients by sex, race, age, and year of diagnosis. Table 35 on the next page shows the expected survival rates for the first 10 patients on our sorted list.

Table 35. Expected Survival Rates for First Ten Cases on Sorted List

LOCALIZED COLON CANCER DIAGNOSED AT MY HOSPITAL, 1978-87

Patient Data					Interval				
Obs	Sex	Age	Race	DX	1	2	3	4	5
24	Fem	69	Black	08/83	0.97190	0.96928	0.96646	0.96361	0.96101
34	Fem	71	Black	XX/85	0.96646	0.96361	0.96101	0.95868	0.95640
23	Fem	64	Hispanic	06/83	0.98772	0.98645	0.98507	0.98359	0.98198
32	Fem	65	Hispanic	06/85	0.98645	0.98507	0.98359	0.98198	0.98026
22	Fem	66	Hispanic	04/83	0.98507	0.98359	0.98198	0.98026	0.97843
31	Fem	78	Korean	06/85	0.95533	0.95005	0.94411	0.93761	0.93051
11	Fem	58	White	08/80	0.99258	0.99189	0.99111	0.99025	0.98933
18	Fem	60	White	08/82	0.99111	0.99025	0.98933	0.98838	0.98741
1	Fem	61	White	04/78	0.99025	0.98933	0.98838	0.98741	0.98641
15	Fem	64	White	04/82	0.98741	0.98641	0.98530	0.98405	0.98260

The first patient (#24) on our sorted list is a black female 69 years old diagnosed in 1983. Look in appendix 3 at the table for black females in the row that has 69-year-olds. Look down the column that covers time closest to the date of diagnosis (1980). Where the row and the column intersect is the expected 1-year survival rate for the first year after diagnosis for that patient, 0.97190.

We then need to fill in the columns for the expected 1-year survival for the rest of the intervals. To do this we add 1 year to the patient's age (she is a year older in the next interval), and add 1 year to year of diagnosis (we want to know what her expected survival is for the *following* year). We then look up her expected survival in the same table, for age 70, and again use the column for 1980. We repeat this for the rest of the intervals, adding a year to her age and to the diagnosis date each time. By the third year (year of diagnosis + 3 = 1986) we should move to the 1990 expected survival column if it is available.

The next patient (#34) is similar.

The next two patients (#23, 32) are similar except the table for female Hispanics should be used. Note that patient #31 is Korean, therefore, the table for "Other Race" females must be used.

After looking up the expected survival for each of the 41 patients (see table 36 on the next page), add up all the expected survivals in the first interval to get 38.726.

Divide by 41 to get the average expected 1-year survival rate for all our patients which is 0.945.

Divide the observed 1-year survival rate of 0.868 from our actuarial method example (table 29) by the expected 1-year survival rate of 0.945, 0.868/0.945 = 0.919, which is our 1-year relative survival rate.

To calculate the 2-year relative rate: Multiply the expected survival from the first interval by that from the second interval to get an expected survival rate for the two intervals combined. (This is equivalent to calculating the cumulative survival rate in the actuarial method.) Add up all the expected survival rates that result from that multiplication to get 36.511. Divide by 41, 36.511/41 =0.891. Divide the 2-year observed rate 0.808 (table 29) by the 2-year expected rate, 0.808/0.891 = 0.907.

For the third interval, multiply the result of the previous multiplication by the expected survival rate for the third interval, then proceed as for the 2-year relative rate. The 4-year and 5-year relative survival rates are computed in an equivalent fashion.

An alternative way of estimating the average expected normal survival rate if you are looking for the rate for just one time period (e.g., 5-year rate) is to look up the expected rate for 5 years for each of your patients and average their expected 5-year rates. However, this will give a less precise estimate of expected survival.

Thus, our 3-, 4-, and 5-year expected survival rates are 34.359/41 = 0.838, 32.276/41 = 0.787 and 30.266/41 = 0.738, respectively. Finally, our 3-, 4-, and 5-year relative rates are, respectively, 0.808/0.838 = 0.964, 0.733/0.787 = 0.931 and 0.638/0.738 = 0.864.

A comparison of these rates is as follows:

Interval	Observed Rate	Expected Rate	Relative Rate
1	0.868	0.945	0.919
2	0.808	0.891	0.907
3	0.808	0.838	0.964
4	0.733	0.787	0.931
5	0.638	0.738	0.864

Table 36. LOCALIZED COLON CANCER DIAGNOSED AT MY HOSPITAL, 1978-87

1-Year Normal Survival Rates						Expected Cumulative Survival Rate Since DX				
A	B	C	D	E	F	G	H	I	J	K
Obs	1-year	2-year	3-year	4-year	5-Year	1-year (B)	2-year (C x G)	3-year (D x H)	4-year (E x I)	5-year (F x J)
24	0.97190	0.96928	0.96646	0.96361	0.96101	0.97190	0.94204	0.91044	0.87731	0.84310
34	0.96646	0.96361	0.96101	0.95868	0.95640	0.96646	0.93129	0.89498	0.85800	0.82059
23	0.98772	0.98645	0.98507	0.98359	0.98198	0.98772	0.97433	0.95978	0.94403	0.92702
32	0.98645	0.98507	0.98359	0.98198	0.98026	0.98645	0.97172	0.95577	0.93855	0.92002
22	0.98507	0.98359	0.98198	0.98026	0.97843	0.98507	0.96890	0.95144	0.93066	0.91254
31	0.95533	0.95005	0.94411	0.93761	0.93051	0.95533	0.90761	0.85688	0.80342	0.74759
11	0.99258	0.99189	0.99111	0.99025	0.98933	0.99258	0.98453	0.97578	0.96627	0.95596
18	0.99111	0.99025	0.98933	0.98838	0.98741	0.99111	0.98145	0.97098	0.95970	0.94762
1	0.99025	0.98933	0.98838	0.98741	0.98641	0.99025	0.97968	0.96830	0.95611	0.94312
15	0.98741	0.98641	0.98530	0.98405	0.98260	0.98741	0.97399	0.95967	0.94436	0.92793
30	0.98741	0.98641	0.98530	0.98405	0.98260	0.98741	0.97399	0.95967	0.94430	0.92793
41	0.98530	0.98405	0.98260	0.98093	0.97908	0.98530	0.96958	0.95271	0.93454	0.91499
3	0.98093	0.97908	0.97706	0.97483	0.97240	0.98093	0.96041	0.93838	0.91476	0.88951
19	0.97908	0.97706	0.97483	0.97240	0.96973	0.97908	0.95662	0.93254	0.90680	0.87935
39	0.91461	0.90537	0.89509	0.88466	0.87441	0.91461	0.82806	0.74119	0.65570	0.57335
37	0.97483	0.97240	0.96973	0.96685	0.96363	0.97483	0.94792	0.91923	0.88876	0.85644
13	0.96685	0.96363	0.95985	0.95533	0.95005	0.96685	0.93169	0.89428	0.85433	0.81166
29	0.96685	0.96363	0.95985	0.95533	0.95005	0.96685	0.93169	0.89428	0.85433	0.81166
20	0.96363	0.95985	0.95533	0.95005	0.94411	0.96363	0.92494	0.88362	0.83948	0.79256
2	0.95533	0.95005	0.94411	0.93761	0.93051	0.95533	0.90761	0.85688	0.80342	0.74759
27	0.95005	0.94411	0.93761	0.93051	0.92287	0.95005	0.89695	0.84099	0.78255	0.72219
8	0.92287	0.91461	0.90537	0.89509	0.88466	0.92287	0.84407	0.76420	0.68403	0.60513
33	0.97863	0.97628	0.97375	0.97104	0.96816	0.97863	0.95542	0.93034	0.90340	0.87464
26	0.90537	0.89509	0.88466	0.87441	0.86383	0.90537	0.81039	0.71692	0.62689	0.54153
12	0.86383	0.85169	0.83769	0.82291	0.80802	0.86383	0.73572	0.61631	0.50717	0.40980
38	0.83769	0.82291	0.80802	0.79310	0.77772	0.83769	0.68934	0.55700	0.44176	0.34357
36	0.97483	0.97240	0.96973	0.96685	0.96363	0.97483	0.94792	0.91923	0.88876	0.85644
35	0.74827	0.73449	0.72141	0.70906	0.69745	0.74827	0.54960	0.39649	0.28114	0.19608
40	0.98395	0.98238	0.98067	0.97881	0.97684	0.98395	0.96661	0.94793	0.92784	0.90635
4	0.97881	0.97684	0.97477	0.97262	0.97032	0.97881	0.95614	0.93202	0.90650	0.87960
17	0.96782	0.96505	0.96195	0.95852	0.95484	0.96782	0.93399	0.89845	0.86118	0.82229
25	0.96505	0.96195	0.95852	0.95484	0.95099	0.96505	0.92833	0.88982	0.84964	0.80800
9	0.95099	0.94705	0.94297	0.93854	0.93358	0.95099	0.90064	0.84928	0.79708	0.74414
16	0.95099	0.94705	0.94297	0.93854	0.93358	0.95099	0.90064	0.84928	0.79708	0.74414
28	0.93358	0.92820	0.92238	0.91606	0.90901	0.93358	0.86655	0.79929	0.73220	0.66558
5	0.92820	0.92238	0.91606	0.90901	0.90114	0.92820	0.85615	0.78428	0.71242	0.64244
7	0.90114	0.89267	0.88387	0.87477	0.86493	0.90114	0.80442	0.71100	0.62196	0.53795
6	0.90114	0.89267	0.88387	0.87477	0.86493	0.90114	0.80442	0.71100	0.62196	0.53795
14	0.87477	0.86493	0.85408	0.84309	0.83226	0.87477	0.75661	0.64621	0.54481	0.45342
10	0.86493	0.85408	0.84309	0.83226	0.82125	0.86493	0.73872	0.62281	0.51834	0.42567
21	0.85408	0.84309	0.83226	0.82125	0.80942	0.85408	0.72007	0.59929	0.49217	0.39837
Total						38.72609	36.51075	34.35894	32.27627	30.26581

Notice that the relative survival rate is larger than the observed rate. Also, notice the relative rate can go up and down. This is because the relative rate is an attempt to estimate what the adjusted survival rate would be if we had good cause of death information and thus measure the decrease in survival due only to colon cancer. If there is no decrease, the relative rate would be 100 percent. Sometimes the relative survival is >100 percent because the patient group under study actually has a better survival experience than that of the general population. Survival varies according to other factors, such as socioeconomic status, rural vs. urban residence, etc. It is impossible to predict with total accuracy what the survival would be for our patient group if they didn't have cancer. Therefore, the adjusted survival rate is preferable if you have the information available and are not comparing adjusted rates to relative rates from another source.

In our example, the 3-year and 4-year relative survival rate actually increased over the 2-year relative survival rate since there were few deaths in that time period--in fact, no one died during year 3--which was a better experience than that enjoyed by the general population of those of similar age, race, and sex. Thus, we have the unusual situation in which survival seems to improve for cancer patients, and the 3-year relative survival rate is actually better than the 1-year relative rate. Such anomalies will occur from time to time since the relative rate is an attempt to correct for "normal mortality," and sometimes cancer patients do have a better experience than that of the general population, at least temporarily. These anomalies are more apt to occur with small numbers of patients.

If you want *expected population survival rates* for different time periods or for other races, you can contact the National Cancer Institute as follows:

The SEER Program
Cancer Statistics Branch
Surveillance Program
Division of Cancer Prevention and Control
National Cancer Institute
Executive Plaza North
Room 343J
Bethesda, MD 20892

Either relative or adjusted rates must be used when you compare the survival of your patients with another group of patients who may be different in factors that cause them to die for reasons other than cancer. As you can see it is important to use the same method for calculating survival that was used for the group with which you want to compare. Even so, if your survival is different from theirs, it is possible that this is due to factors other than differences in death from cancer. This is why clinical trials groups are set up to ascertain treatment effectiveness.

When presenting survival rates, it is important to consider their *standard error* which is discussed in section F.

Q8

When the calculation for an observed survival rate is done every time someone dies, that is called the _____ - _____ _____ or sometimes the _____ _____ _____.

Q9

Observed survival rates underestimate survival from cancer because they group deaths from all causes in the calculations. Two other calculations you might be able to use are:

1) the _____ _____ rate and

2) the _____ _____ rate.

Q10

You must have good cause of death information to use the _____ _____ rate because you count as deaths only those patients who died from the cancer under study. Patients who die from other causes are _____ from the study.

Q11

If the cause of death information is not good, it is still possible to adjust the observed survival rate by using _____ survival rates from standard life expectancy tables to account for the risk of dying from other causes. This is called a _____ _____ rate.

Answer: Q8

When the calculation for an observed survival rate is done every time someone dies, that is called the Kaplan - Meier method or sometimes the product moment method.

Answer: Q9

Observed survival rates underestimate survival from cancer because they group deaths from all causes in the calculations. Two other calculations you might be able to use are:

1) the adjusted survival rate and

2) the relative survival rate.

Answer: Q10

You must have good causes of death information to use the adjusted survival rate because you count as deaths only those patients who died from the cancer under study. Patients who die from other causes are withdrawn from the study.

Answer: Q11

If the cause of death information is not good, it is still possible to adjust the observed survival rate by using expected survival rates from standard life expectancy tables to account for the risk of dying from other causes. This is called a relative survival rate.

MEASURES OF RECURRENCE

Time to recurrence is obtained by subtracting the date of complete remission from the date of recurrence (or date of death or withdrawal without recurrence). Calculating summary measures for recurrence is analogous to calculating summary measures for survival.

1. Average or Median Time to Recurrence

The average or median time to recurrence may be calculated using the method for calculating average or median survival time (see page 122). The cautions about using average survival time also apply to average time to recurrence. Recurrence time is computed as date of *recurrence minus date of remission.*

2. Relapse Free Survival Rate

Either the actuarial method or the Kaplan-Meier method may be used to calculate a relapse free survival rate. Recurrences of the cancer are treated the same way as deaths in calculating the survival time, and patients with recurrences will be tabulated in the same column as those who died.

3. Recurrence Rate

Only patients who go into remission are used here. The starting point is the date of first complete remission. The end point is date of first recurrence. Deaths without recurrence are tabulated as withdrawn from the study. Notice, there is an additional column at the end (I - CP). This means to subtract the number in column H from 1.000. This will ensure that the recurrence rate will start at 0 percent and get larger, in contrast to a survival rate, which starts at 100 percent and gets smaller.

Table 37. Life Table for Calculating Recurrence Rates

A (i)	B (l)	C (d)	D (w)	E (l')	F (q)	G (p)	H (P or CP)	I (1 - CP)
Interval of Observation (Time after remission in years)	# in Remission at Beginning of Interval	# Recurring During Interval	# Last Seen Withdrawn Alive During Interval or Dying of Other Causes	Effective # Exposed to Risk of Recurrence (B - 1/2 D)	Proportion Recurring During Interval (C / E)	Proportion Not Recurring during the Interval (1.0 - F)	Cumulative Proportion	Recurrence Rates
0 - < 1								
1 - < 2								
2 - < 3								
3 - < 4								
4 - < 5								
5 or more								

PRESENTING SURVIVAL RESULTS

1. Graphically

As discussed in section B, the graph you select for presenting your data will depend on the message you wish to convey to your audience. For example, if you use the actuarial or Kaplan-Meier method for calculating survival, you will get interim survival rates. These can be graphed to show the pattern of survival or survival curve. The survival curve will allow your audience to see if the patients survived well for the first 3 years and then survival dropped off, or conversely, if survival dropped off rapidly in the first few years after diagnosis and then leveled off. Thus,

For survival times:

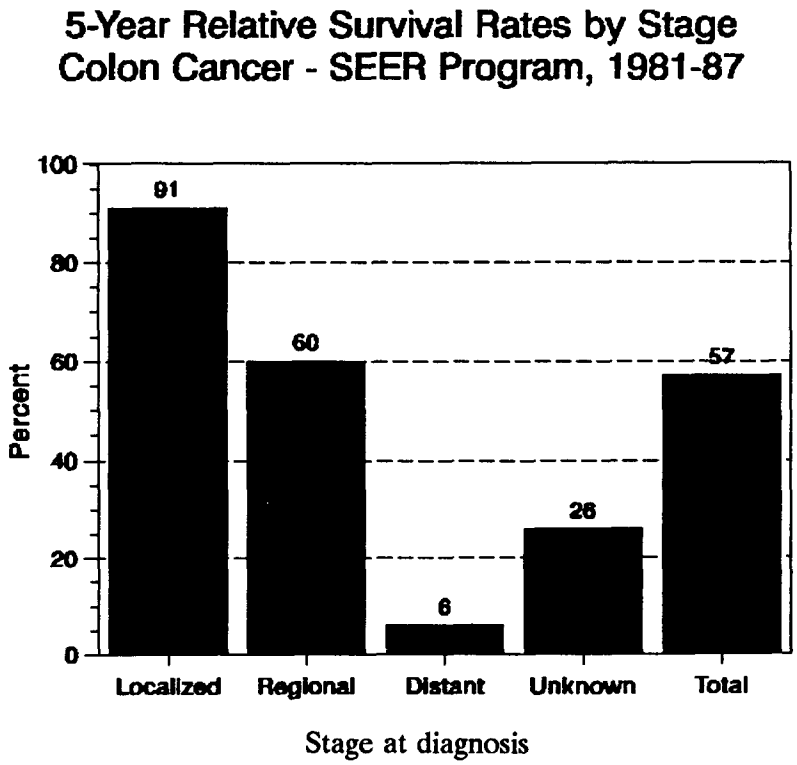
Mean or median survival time can be presented in a bar graph. Then you can look at the results for each group and easily compare them.

For survival rates:

Survival rates for a single period can be presented by bars, as in figure 19.

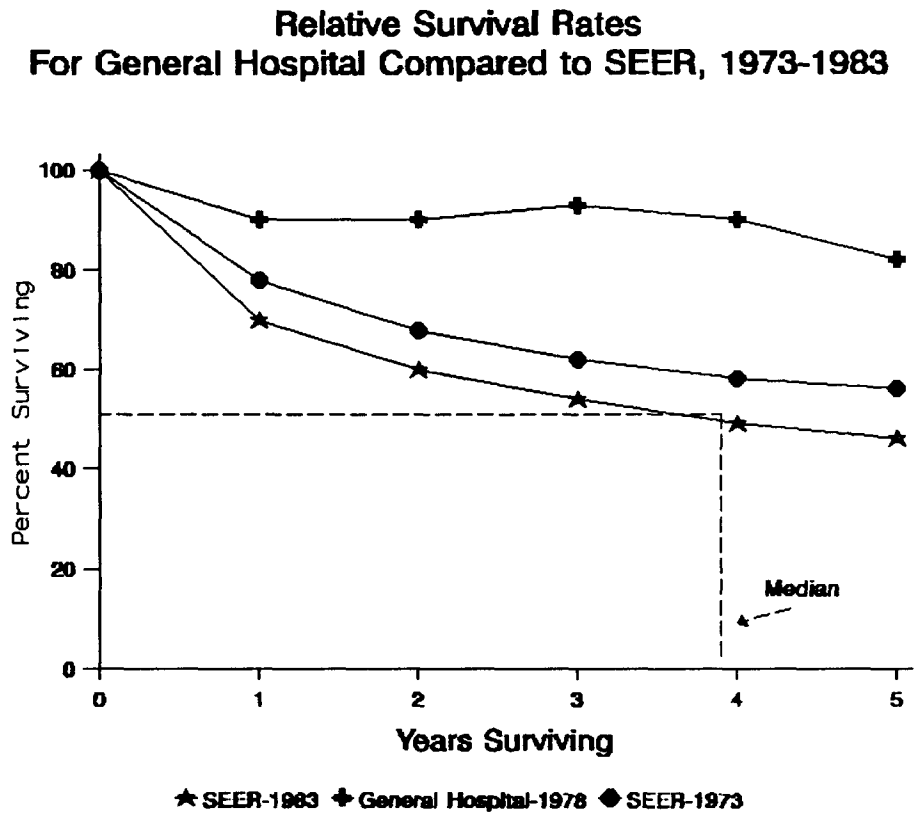
If you calculate interim rates using the actuarial or Kaplan-Meier method, it is better to use a line graph to emphasize the pattern of change over the time period. As you learned in section B, there are two types of scales used to present patterns of change over time, the arithmetic scale and the semilogarithmic scale.

Figure 19. Bar Graph for a Single Time Period



In either case, the graph starts with 100 percent surviving at the beginning of the study since we know that all our patients are alive. For actuarial survival, a slanted line is used to connect the points (see figure 20). This implies that survival intermediate to the points that we plotted changes gradually between those two points.

Figure 20. Line Graph for More Than One Time Period (Arithmetic Scale)



Note that if the survival rate falls below 50 percent, you can draw a line at 50 percent down to the time line, and read the median survival time off the graph.

For Kaplan-Meier survival the graph looks like a stair step. (See figure 21.) Since a calculation is made every time someone dies, the assumption is made that the survival is constant until the next death occurs.

Figure 21. Kaplan-Meier Survival Graph

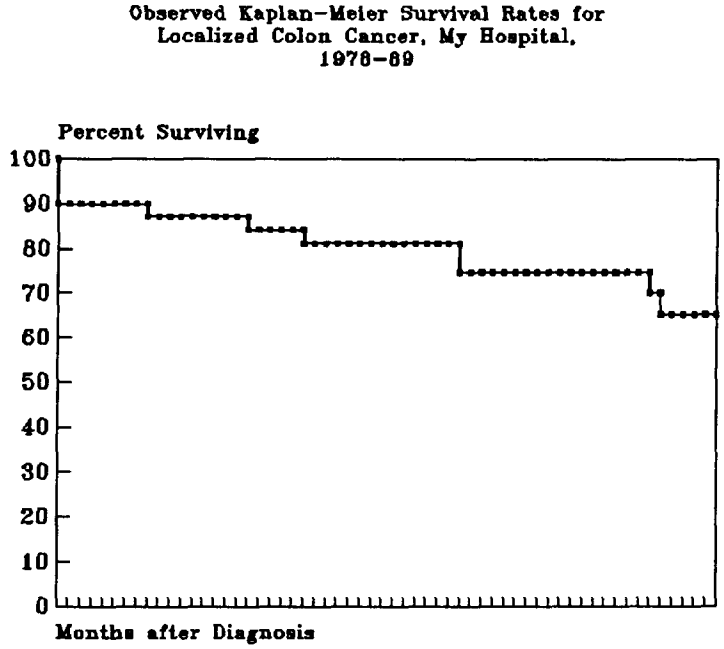
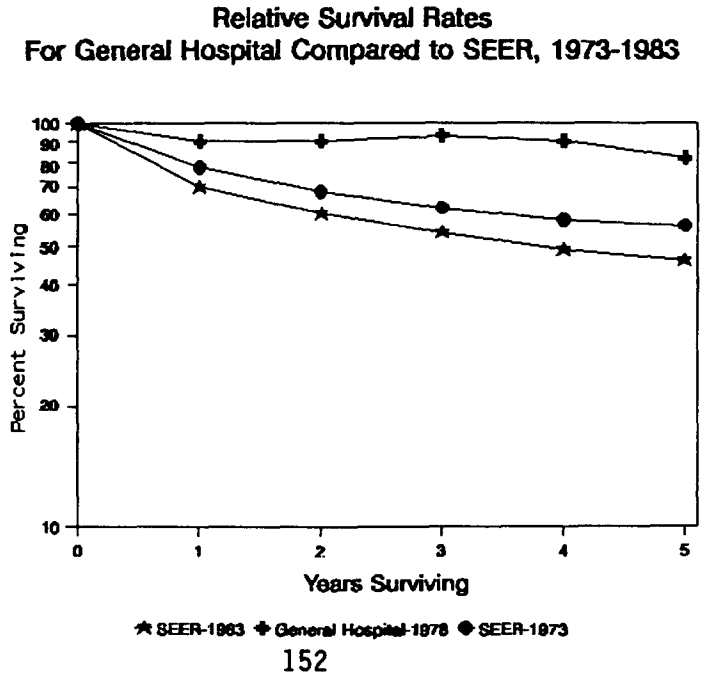


Figure 20 shows the survival rate graphed on an arithmetic scale. This emphasizes the numeric change in survival rate. Figure 22 shows the same information graphed on semi-logarithmic graph paper. You should use this if you want to emphasize the percent change in survival. See section B for more detail on using semilogarithmic graphs.

Figure 22. Line Graph for More Than One Time Period (Semilogarithmic Scale)



2. In a Report

A survival report must contain more that just the survival rate (or survival time). It must also contain a complete description of the patients, their disease, and their treatment. This will allow anyone reading the report (and especially anyone who wants to compare their survival results to yours) to be able to put the survival results into context. If you have excluded any patients (e.g., those not microscopically confirmed) make sure you make that clear in the report. If you have grouped patients (e.g., by stage) make sure you make clear what the criteria were for grouping the patients. The ACoS requires that you use comparison data. Make clear what that comparison data represent, and give the complete reference. Also, make sure that the comparison figures are calculated the same as your calculations and that the starting point is the same. ACoS survival may be calculated from treatment date, not diagnosis date, and will thus look artificially shorter if you don't make your starting time the same. Make sure you don't over interpret survival results and comparisons. Remember that differences are more likely to be due to differences in patient groups than differences in treatment efficacy. You must present your percent successful follow-up. If it is too low (<90 percent), your results are not reliable. Readers of your study must be able to judge the reliability of your results. You should report the number in each group that you used for calculating survival.

Refer to statistical information and hypothesis sections for methods of comparing survival results statistically.

For an additional discussion on the reporting of cancer survival see chapter 2 in the American Joint Committee on Cancer: Manual for Staging of Cancer--Fourth Edition.

Exercises

Use this listing of breast cancer cases to answer the questions on pages 155 and 156.

LISTING OF CASES OF FEMALE BREAST CANCER
DIAGNOSED WITH REGIONAL SPREAD AT MY HOSPITAL 1983-87

Ob	Age	Race	DX Date	FUP Date	Status	Survival*
1	77	Black	08/85	01/90	Alive	4 Y 05 M
2	54	Hispanic	04/87	09/87	Dead	0 Y 05 M
3	70	White	06/84	05/87	Dead	2 Y 11 M
4	57	White	08/85	06/91	Alive	4 Y 04 M
5	49	White	07/86	05/89	Alive	2 Y 10 M
6	79	White	12/84	08/86	Dead	1 Y 08 M
7	65	Black	05/84	05/88	Dead	4 Y 00 M
8	30	White	06/85	10/90	Alive	4 Y 06 M
9	32	Black	11/83	08/86	Dead	2 Y 09 M
10	54	White	04/83	01/90	Alive	6 Y 08 M
11	58	White	12/86	03/91	Alive	3 Y 00 M
12	79	White	03/83	05/88	Dead	5 Y 02 M
13	90	White	11/87	01/90	Alive	2 Y 01 M
14	62	White	03/83	03/91	Alive	6 Y 09 M
15	50	Chinese	10/84	02/86	Dead	1 Y 04 M
16	70	White	08/86	03/88	Dead	1 Y 07 M
17	76	White	01/86	02/89	Dead	3 Y 01 M
18	51	White	01/83	04/89	Dead	6 Y 03 M
19	61	White	03/84	02/91	Alive	5 Y 09 M
20	74	White	02/86	04/91	Alive	3 Y 10 M
21	30	White	01/87	03/90	Dead	2 Y 11 M
22	55	White	03/87	03/91	Alive	2 Y 09 M
23	51	White	05/85	08/90	Alive	4 Y 07 M
24	33	White	10/87	06/90	Alive	2 Y 02 M
25	52	White	01/85	02/90	Alive	4 Y 11 M

*Based on study cutoff date of 12/89

Q12

Perform an actuarial survival analysis on the breast cases on the previous page (e.g., fill out a life table) to compute the 5-year survival rate.

Actuarial Life Table for Breast Cancer Cases Diagnosed with Regional Spread at My Hospital 1983-87							
A	B	C	D	E	F	G	H
Interval of Observation (Time after diagnosis in years)	# Alive at Beginning of Interval	# Dying During Interval	# Last Seen Alive During Interval	Effective # Exposed to Risk of Dying (B - 1/2 D)	Proportion Dying During Interval (C / E)	Proportion Surviving the Interval (1.0 - F)	Cumulative Survival Rates
0 - < 1							
1 - < 2							
2 - < 3							
3 - < 4							
4 - < 5							
5 or more							

Q13

What is the expected 1-year normal survival rate for patient #1? What is the expected 2-year normal survival rate for patient #1? (Hint: Multiply the yearly rates for the first year and the second year after DX.)

Q14

The average expected normal 5-year survival rate for the group of regional breast cases is 0.978. Use this information and the 5-year observed survival rate from question 12 above to compute the 5-year relative survival rate.

Q15

Hospital B, our principal competitor, reports a relative survival rate higher than our observed rate for regional breast cancer. Which of the following reasons is the most likely explanation?

a. They calculated survival using a starting point of treatment date, and we used diagnosis date.

b. They are reporting relative rate while we are using an observed rate.

c. They used the Kaplan-Meier method to compute survival and we used the actuarial method.

d. They treat regional breast cancer better than we do.

Q16

When using the life table method to compute a 5-year observed survival rate we would have to exclude which of the following types of cases from our calculations.

a. Those who died before 5 years was up.

b. Those who were diagnosed less than 5 years ago.

c. Those lost to followup less than 5 years after diagnosis.

d. None of the above. We can include all the cases in a, b, and c.

Answers to Exercises

Answer: Q12 See the completed life table below.

Actuarial Life Table for Breast Cancer Cases Diagnosed with Regional Spread at My Hospital 1983-87							
A	B	C	D	E	F	G	H
Interval of Observation (Time after diagnosis in years)	# Alive at Beginning of Interval	# Dying During Interval	# Last Seen Alive During Interval	Effective # Exposed to Risk of Dying (B - 1/2 D)	Proportion Dying During Interval (C / E)	Proportion Surviving the Interval (1.0 - F)	Cumulative Survival Rates
0 - < 1	25	1	0	25.0	0.040	0.960	0.960
1 - < 2	24	3	0	24.0	0.125	0.875	0.840
2 - < 3	21	3	4	19.0	0.158	0.842	0.707
3 - < 4	14	1	2	13.0	0.077	0.923	0.653
4 - < 5	11	1	5	8.5	0.118	0.882	0.576
5 or more	5						

The 1-year survival rate is 96 percent, the 2-year rate is 84 percent, 3-year is 71 percent, 4-year is 65 percent, and the 5-year is 58 percent.

Answer: Q13 Look in appendix 3 at the table for the black females:

The expected normal survival rate for the first year is found at age 77 in the column for 1980 = 0.95091 (the expected normal 1-year survival rate).

To find the expected normal survival rate for the second year, add 1 year to age 77+1=78. Look in the column for 1980 for age 78 = 0.94718. To find the survival experience for the 2 years combined, multiply 0.95091 X 0.94718 = 0.9007 which is the 2-year expected normal survival rate.

Answer: Q14

To find the 5-year relative survival rate divide the 5-year observed rate by the 5-year average expected normal survival rate - 0.576/0.978 = 0.589 or 59 percent.

Answser: Q15

The correct answer is b. The relative survival rate will almost always be higher than the observed survival rate because the influence of normal mortality is removed.

Answer a is untrue because using a starting time of treatment date would artificially shorten survival time when compared to diagnosis date.

Answer c is untrue because Kaplan-Meier and actuarial methods will usually give very similar results (although you should still use the same method of calculation used for the group with which you are comparing).

Answer d. You hope this is untrue for obvious reasons. The goal of this exercise is to point out why you shouldn't jump to this conclusion.

Answer: Q16

d is correct. All cases described in a, b, and c may be included when using either the actuarial or Kaplan-Meier method for calculating observed survival.

SECTION E

ANALYTIC EPIDEMIOLOGY

SECTION E

ANALYTIC EPIDEMIOLOGY

In section C, you were introduced to the field of epidemiology and to some of the standard methods used to "describe" the distribution of disease in a study population--incidence, prevalence, and death rates--known as measures of risk.

The following section is concerned with analytic epidemiology which is the study of the methodology employed in investigating possible determinants (factors or causes) associated with the occurrence of diseases. The two general forms that analytic epidemiology may take are observational and experimental studies.

OBSERVATIONAL STUDIES

1. Cohort or Prospective Study

In a prospective study, a *group* of people (cohort) *without* disease are initially identified and characterized by a common experience or *exposure* (e.g., smoking). The group is then followed forward (prospectively) over a period of time to observe the development (incidence) of the disease under investigation. These studies are designed primarily to test a specific hypothesis. For example, populations such as those of Hiroshima and Nagasaki have been studied in order to evaluate the occurrence of leukemia and other cancers in persons exposed to atomic bomb radiation. In these studies, 125,000 and 111,150 people in the respective cities were identified. Thus, a major difficulty of cohort studies is the cost of the project because such studies involve recruitment of a population of large numbers of persons who must be followed during the course of the study.

If the factor under study is one to which only a small proportion of the population is exposed, it may be better to identify smaller groups for study. Hence, as an alternative, you might study persons exposed to large doses of x-ray given for a specific purpose such as ankylosis spondylitis or thymic enlargement to see if the risk of developing leukemia and other cancers is greater than in the general population.

2. Case-control or Retrospective Study

In retrospective studies, *two groups* are selected, one comprised of people *with* the disease of interest (cases) and the other of people with the same general characteristics but *without* the disease (controls). They are *compared* for possible differences in past exposure to factors hypothesized to be determinants of the disease in question.

This type of study can be done in the hospital setting or on a county, city, or state level where the population is limited and defined. All cases diagnosed with the disease between specified dates should be included. The control group of unaffected individuals believed to reflect the same characteristics as the population from which the affected group arose is selected for comparison. For example, young women with vaginal adenocarcinoma and nondiseased controls are compared in terms of exposure to DES (diethylstilbestrol) in utero. This methodology can be useful in the study of rare conditions.

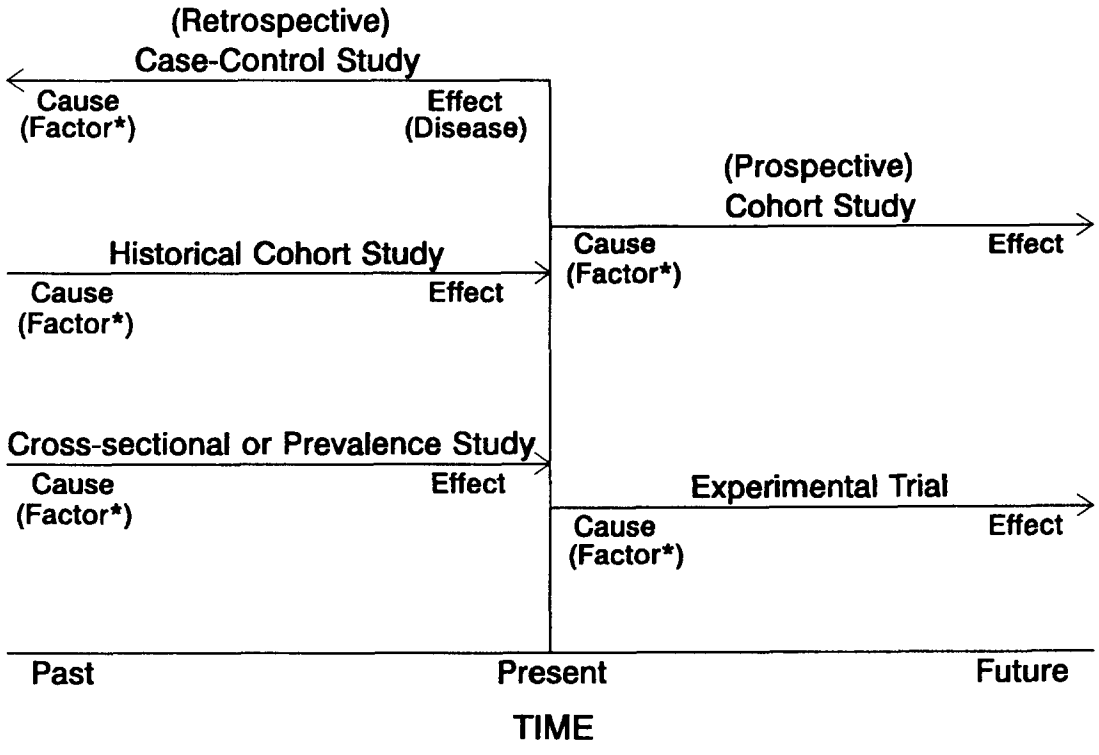
Sometimes it is possible to obtain a set of historical records in which people were previously classified into various groups (e.g., union records of persons retiring in 1960 or 1970 classified by job title) which can then be used to look at current disease status of the cohort. This is referred to as a retrospective cohort study or a historical cohort study or a retrospective prospective study.

EXPERIMENTAL STUDIES

In experimental epidemiology, the investigator studies the impact on the natural history of a disease by varying some factor which is under his/her control. Major applications include *intervention* trials to reduce risk factors in high-risk groups, screening for early stage of disease, and clinical trials of various treatment modalities. For example, a multiple risk factor intervention trial in which men at risk of myocardial infarction due to smoking, high cholesterol, or hypertension are counseled to modify their behavior; women at high risk to breast cancer because of family history are given Tamoxifen (chemoprevention).

The relationship of cohort (prospective), case-control (retrospective) studies, and experimental studies is shown in the figure below.

Figure 23. Schema for Analytic Epidemiologic Studies



*Hypothesized etiologic (causative) characteristic under study

Q1

The general forms that analytic epidemiologic studies may take are

1) _____ and 2) _____.

Q2

A study in which a group of people *without* cancer, but characterized by a common *exposure* are identified and followed over a period of time to observe the development of cancer might be called a:

1. _____ study

2. _____ study

Both of these are _____ studies.

Q3

One group of people *with* cancer and another group of people *without* cancer, but otherwise similar, are selected and then compared for possible differences of exposure to carcinogenic agents in the past, might be called a:

1) _____ study or

2) _____ study.

These, too, are _____ studies.

Q4

When an investigator studies the impact of some factor under his/her control on the natural history of disease, it is called an _____ study.

Answer: Q1

The general forms that analytic epidemiologic studies may take are 1) observational and 2) experimental.

Answer: Q2

A study in which a group of people *without* cancer, but characterized by a common *exposure* are identified and followed over a period of time to observe the development of cancer, might be called a:

1. cohort study or
2. prospective study.

Both of these are observational studies.

Answer: Q3

One group of people *with* cancer and another group of people *without* cancer, but otherwise similar, are selected and then compared for possible differences of exposure to carcinogenic agents in the past, might be called:

1. retrospective study or
2. case-control study.

These, too, are observational studies.

Answer: Q4

When an investigator studies the impact of some factor under his/her control on the natural history of disease, it is called an experimental study.

COHORT OR PROSPECTIVE STUDIES

In cohort studies a group of people (cohort) without disease is identified, and, at the outset, demographic and physiologic characteristics and exposures are recorded for each member of the group. The cohort is then followed over time and the development of disease (incidence or mortality) is monitored carefully. Internal comparisons are made between disease rates among individuals exposed and those not exposed to factors of interest or between those with different baseline physiologic measures. Alternatively, disease rates among the study group may be compared to rates in the general population or another well-studied group.

1. Selection of Study Population(s)

    a. Entire state(s)
    b. Metropolitan area(s)
    c. Selected subgroup(s)

2. Comparison Groups

    a. Internal comparison groups
    b. General population
    c. Other well-studied cohorts

3. Strengths of the Cohort Group Approach

    a. Ideal time sequence (hypothesized cause *precedes* disease under study)
    b. Exposure can be accurately recorded at time of exposure (not based on recall of past events)

4. Problems Associated With the Cohort Study

    a. Duration (especially for rare diseases and those with long latency periods)
    b. Cost
    c. Initial nonresponse/subsequent attrition (losses to followup)
    d. Disease detection/diagnostic bias

Some examples of cohort studies are:

1. Framingham heart study
2. British prospective study of women using oral contraceptives
3. Follow-up study of fluoroscopy and subsequent breast cancer
4. Occupational cohort studies
5. Various prospective studies of cholesterol and cancer
6. Follow-up study of 50,000 college students to study the relation between exercise and coronary heart disease

Analysis of Results of Cohort Studies

In cohort studies, the group is divided into those exposed and those not exposed. The exposed group may be further divided into exposure levels (for example, heavy smokers vs. light smokers). The two groups are then compared with respect to their development of the disease of interest.

Table 38A. Format for Analysis of Cohort Studies

Exposure	Disease		
	Yes	No	Total
Yes	a	b	a + b
No	c	d	c + d

Table 38B. Occurrence of Lung Cancer Among Heavy Smokers vs. Nonsmokers

Exposure	Lung Cancer		
	Yes	No	Total
Heavy smokers	227	99,773	100,000
Non-smokers	7	99,973	100,000

Relative Risk (RR)

The measure of comparison of risk of the two groups is the relative risk. The relative risk of disease is the risk of disease in people exposed to a factor relative to the risk in people not exposed to a particular factor. In the above example, a study population of 100,000 heavy smokers and a like number of nonsmokers is used.

A relative risk greater than 1 implies a positive association of the disease with exposure to the factor; a relative risk of less than 1 implies a negative association of the disease with exposure to the factor.

$$RR = \frac{\textit{Disease rate in the exposed population}}{\textit{Disease rate in the nonexposed population}} = \frac{a/(a + b)}{c/(c + d)}$$

In the example of heavy smokers compared to nonsmokers shown in table 38B, we calculate

$$\frac{a/(a + b)}{c/(c + d)} = \frac{227/100{,}000}{7/100{,}000} = \frac{227}{7} = 32.4$$

The risk of lung cancer is 32 times as great for heavy smokers as it is for nonsmokers. This measure is known as the relative risk because it measures the risk (of lung cancer) of the exposed (heavy smokers) relative to that of the nonexposed (nonsmokers).

Attributable Risk (AR)

The difference between the disease rate in the exposed population and the rate in the non-exposed population is the absolute amount of disease which is "attributable to" the exposure. Thus, the attributable risk (AR) is obtained by subtracting the incidence of the disease among the nonexposed persons (7) from the total incidence among the exposed individuals (227). It is assumed that possible

other factors associated with this disease had an equal effect on the exposed and nonexposed groups. In our example, 227/100,000 - 7/100,000 = 220/100,000 that is, 220 of the 227 cancer cases (97 percent) that occurred in 1 year among 100,000 heavy smokers were attributable to heavy smoking. This calculation of attributable risk assumes (usually naively) a single factor etiology, and in our example that 7 of every 100,000 persons in the exposed group would have developed lung cancer even if they had not smoked based on the fact that 7 of every 100,000 nonsmokers developed lung cancer.

Population Attributable Risk (PAR)

The proportion of a disease in a population related to (attributable to) a given exposure is known as the *population* attributable risk (PAR) and is calculated according to the following formula:

$$PAR = \frac{PE\ (RR - 1)}{PE\ (RR - 1) + 1}$$

where PE = the proportion of the population exposed,
RR = the relative risk, and
PAR = the population attributable risk expressed as a percent.

The derivation of this formula involves higher mathematics and can be found in standard epidemiology text books. In our example, assuming 40 percent of the general population smokes (PE) and that the relative risk (RR) of lung cancer associated with the practice of smoking cigarettes is 9, then the *population* attributable risk (PAR) for smoking is:

$$PAR = \frac{0.40(9 - 1)}{0.40(9 - 1)+1} = \frac{0.40(8)}{0.40(8) + 1} = \frac{3.2}{3.2 + 1} = \frac{3.2}{4.2} = 76.2\%$$

that is, 76 percent of lung cancer in the general population is attributable to smoking assuming that 40 percent of the population smokes..

Comparison of Relative Risk and Attributable Risk

The relative risk is useful in determining the strength of an association between a factor and a disease. It is extremely important in etiologic research. However, it tells us little about the contribution of that factor to the total disease profile in the population (or how much the disease might be reduced in the community were the factor eliminated) because it does not reflect the extent of exposure to the factor in the general population. For instance, if smoking were associated with a dramatically increased relative risk of lung cancer but only a minute fraction of the United States population smoked, then the reduction in lung cancer deaths that might be expected to follow a successful antismoking campaign would be much less than it is in the current context of widespread smoking.

The relative frequency of different diseases will also influence the absolute impact of our campaign. We might launch an "exposure eradication" campaign on the basis of a large relative risk to a very small exposed segment of the population or, alternatively, on the basis of a small relative risk to a very large exposed segment of the population. This point is illustrated below. The data in table 39 show that elimination of smoking would prevent 114.4 lung cancer deaths per year and 500 coronary heart disease deaths per year among every 100,000 smokers. Because coronary heart disease is much more common (higher incidence) in the population, the actual number of lives saved (or deaths averted) would be greater for coronary heart disease than for lung cancer. Thus, although the relative risk associated with smoking is lower for coronary heart disease (2) than for lung cancer (9.9), the attributable risk for coronary heart disease is much higher, i.e., 500 vs. 114.4.

Table 39
Annual Death Rates for Lung Cancer and Coronary Heart Disease by Smoking Status, Males

Exposure Level	Annual Death Rate/100,000	
	Lung Cancer	Coronary Heart Disease
Cigarette Smokers	127.2	1,000
Nonsmokers	12.8	500

$$RR = \frac{127.2}{12.8} = 9.9 \qquad \frac{1,000}{500} = 2$$

$$AR = 127.2 - 12.8 = 114.4 \textit{ per } 100,000 \qquad 1,000 - 500 = 500 \textit{ per } 100,000$$

Remember, estimates of the reduction in disease rates to be expected from an attempt to reduce or eliminate a risk factor should not be limited to a single disease, since factors which contribute to one disease may contribute to other diseases as well--as in the case of smoking. Also remember that cohort studies are influenced by the duration of the study (time until diagnosis or death occurs) and attrition (loss to followup).

Q5

One of the strengths of a cohort group approach is (select one):

1. Low cost of such studies.
2. Exposure can be accurately recorded at time it happens.
3. Long latency periods irrelevant.
4. Diagnostic bias unlikely.

Q6

Groups that might be used for comparison purposes in a cohort study are:

1. _____.
2. _____.
3. _____.

Q7

Measures of the strength of an association between exposure to a particular factor and risk of a certain outcome are used in the analysis of _____.

Q8

The risk of lung cancer in people who smoke relative to the risk of lung cancer in people who do NOT smoke is called _____ _____.

Q9

A RR >1 implies a positive association of the disease with exposure to the factor. In the example of lung cancer the RR = 32 indicates that the risk of getting lung cancer is _____ for heavy smokers as it is for nonsmokers.

Q10

The absolute incidence of lung cancer among people who smoke is called the _____ _____. Instead of dividing the cancer rate in the exposed population by the cancer rate in the nonexposed population, you subtract the cancer rate in the _____ _____ from the cancer rate in the _____ _____ (the heavy smokers).

Q11

Previously we had determined that the attributable risk of developing lung cancer was 97 percent among heavy smokers. However, what if you wish to find out what proportion of cancer in a population is attributable to heavy smoking? What two counts would you need?

You would need:

1) the _____ and

2) the _____.

Answer: Q5

One of the strengths of a cohort group approach is that exposure can be accurately recorded at the time it happens. However, this approach is always costly, the latency period can be very long and, therefore, irrelevant, and diagnostic bias is likely.

Answer: Q6

Groups that might be used for comparison purposes are:

1. Internal comparison groups.
2. The general population.
3. Other well-studied cohorts.

Answer: Q7

Measures of the strength of an association between exposure to a particular factor and risk of a certain outcome, are used in the analysis of risk.

Answer: Q8

The risk of lung cancer in people who smoke relative to the risk of lung cancer in people who do NOT smoke is called relative risk.

Answer: Q9

A RR >1 implies a positive association of the disease with exposure to the factor. In the example of lung cancer the RR = 32 indicates that the risk of getting lung cancer is 32 times as great for heavy smokers as it is for nonsmokers.

Answer: Q10

The absolute incidence of lung cancer among people who smoke is called the attributable risk. Instead of dividing the cancer rate in the exposed population by the cancer rate in the nonexposed population, you subtract the cancer rate in the nonexposed population from the cancer rate of the exposed population (the heavy smokers).

Answer: Q11

You would need 1) the proportion of the population who were heavy smokers and 2) the relative risk of heavy smokers.

CASE-CONTROL OR RETROSPECTIVE STUDIES

In case-control studies patients with a disease (cases) are chosen, and suitable individuals without the disease (controls) are also selected. The two groups are compared for possible differences in past exposures or other characteristics thought to be related to the disease under study.

Data from the case-control study are conventionally arrayed as in table 40A so that cases and controls can be compared on exposure to a hypothesized etiologic factor:

Table 40A. Format for Analysis of Case-Control Studies

Exposure	Disease Status	
	Cases	Controls
Yes	a	b
No	c	d

Odds Ratio

The *incidence* of disease among the exposed and nonexposed *cannot be calculated* using case-control data because the cases and controls in the study rarely reflect the true proportions of diseased and nondiseased persons in the population. (Usually there are roughly equal numbers of cases and controls in the study, whereas there are many more nondiseased than diseased people in the population.) Therefore, relative risk of disease associated with exposure cannot be calculated directly in a case-control study as was shown for the cohort study. However, an estimate of the relative risk, known as the *odds ratio*, can be calculated if the proportion of diseased people in the general population is small compared to the proportion of nondiseased (almost always true). Recall the *true* relative risk using data from a cohort or incidence study is:

$$RR = \frac{a/(a + b)}{c/(c + d)}$$

Since in the general population a/(a+b) is approximately equal to a/b and c/(c+d) is appromimately equal to c/d, the formula for relative risk reduces to:

$$\frac{a/b}{c/d} = \frac{ad}{bc} = \textit{odds ratio (estimated risk)}$$

In this example, 100 men with lung cancer and 100 controls are interviewed regarding smoking history with the following results:

Table 40B. Smoking Status of Male Lung Cancer Cases and Controls

Exposure	Disease Status	
	Cases	Controls
Smokers	90	50
Non-smokers	10	50
Total	100	100

$$Odds\ ratio = \frac{ad}{bc} = \frac{90 \times 50}{50 \times 10} = \frac{4{,}500}{500} = 9$$

Since the odds ratio is an estimate of relative risk, one can conclude that these data show a nine-fold increased risk of lung cancer in smokers compared to nonsmokers.

"Matched" Case-Control Studies. Frequently controls are selected in a case-control study so as to be individually matched to the cases on characteristics such as age, sex, race, or socioeconomic status that are known to be related to the disease. Matching helps make the two groups similar with respect to factors other than the exposure of interest in the study and thereby is performed to reduce the likelihood of spurious associations. The investigator must be careful, however, not to overmatch, i.e., to match cases and controls on factors related to the exposure of interest; overmatching can artificially reduce, or may even eliminate, true exposure differences between diseased and nondiseased individuals in the population. It should be obvious that cases and controls cannot be compared in the analysis on any characteristics that have been matched.

The data in a matched pairs analysis are organized as shown below:

Table 41. Format for Analysis of Matched Case-Control Studies

Cases	Controls		
	Exposed	Not Exposed	Total
Exposed	r	s	a
Not Exposed	t	u	c
Total	b	d	

r = number of pairs in which both case and control are positive on exposure to the factor (concordant)

s = number of pairs in which the case but not the control is positive on exposure to the factor (discordant)

t = number of pairs in which the control but not the case is positive on exposure to the factor (discordant)

u = number of pairs in which both case and control are negative on exposure to the factor (concordant)

To compute the *odds ratio* (estimated relative risk) *for a matched series,* only the discordant pairs enter in the calculation.

$$Odds\ Ratio = \frac{s}{t}\ (provided\ t\ is\ not\ equal\ to\ 0\ ,\ i.e.\ ,\ t \neq 0)$$

Example[1]

One-hundred and seventy-five (175) women ages 15-44 admitted to a hospital in 1968 with thromboembolism were matched on age, sex, race, and date of admission with 175 controls. All women in the study were interviewed regarding use of oral contraceptives in the month preceding admission. The following results were obtained:

$$Odds\ Ratio = \frac{s}{t} = \frac{57}{13} = 4.4$$

One can conclude that these data show that women who have recently used oral contraceptives have a 4.4 times increased risk of admission for thromboembolism compared to nonusers.

Population attributable risk (PAR) (i.e., the proportion of all cases of the disease in the population that can be attributed to the exposure of interest) can be estimated from case-control studies as well as cohort studies, using the same formula:

$$PAR = \frac{PE\ (RR - 1)}{PE\ (RR - 1) + 1}$$

where PE = proportion of the population with a characteristic, and RR = relative risk (odds ratio estimate) associated with the characteristic.

[1]Sartwell, P.E. et al. American Journal of Epidemiology 90: 365-380, 1969.

Q12

The basic measure of risk of disease associated with an exposure calculated from a case-control study is the _____ _____.

Q13

In a case-control study of bladder cancer patients, truck drivers who smoked one to two packs of cigarettes per day were found to have an odds ratio of 6.8 compared to nontruck drivers who never smoked. Interpret these results.

Q14

In a case-control study of pancreatic cancer patients, controls were selected from other hospital patients admitted for gastrointestinal complaints. Is this a suitable control group, and if not, why not?

Answer: Q12

The basic measure of risk of disease associated with an exposure calculated from a case-control study is the odds ratio.

Answer: Q13

An odds ratio of 6.8 implies a strong association of bladder cancer with smoking and being employed as a truck driver suggesting a synergistic (multiplicative) effect between the two. This seems likely since truck drivers who spend long hours on the road tend to be heavy smokers. The same study showed elevated odds ratios for all smokers regardless of occupation and for truck drivers regardless of smoking habits.

Answer: Q14

The selection of patients with gastrointestinal complaints as controls for pancreatic cancer patients may not be appropriate since exposures resulting in some GI complaints such as cholecystitis may result in pancreatic cancer as well. Thus, patients and controls would end up reporting the same exposure and the resulting odds ratio would be close to 1.0, implying no association with the exposure.

SECTION F

STATISTICAL INFERENCE

SECTION F

STATISTICAL INFERENCE

In this section you will be introduced to the topic of statistical inference. Once you understand the way this inference process works, you will be introduced to concepts which are basic to inferential statistical analysis. You will then learn how sample statistics are used to predict what the true population parameters are and how reliable these estimates are.

Statistical inference is the process of drawing conclusions about populations based on data from limited samples. Medical knowledge is largely based on information from limited samples rather than entire populations. Health care workers should, therefore, be aware of the reliability of such information, and of conclusions based on the inferential process. As we study inferential statistics we begin to see its importance in our daily lives.

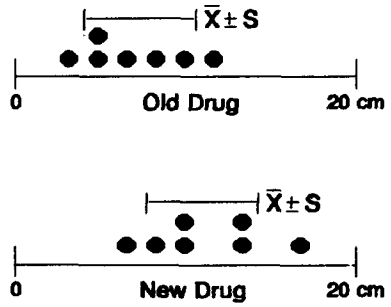
POPULATIONS VERSUS SAMPLES

To explain the use of samples to estimate the population, we will use two hypothetical examples.

Example 01:

Suppose you heard, at a meeting of your cancer committee, that some researchers believe giving a specific chemotherapeutic agent shrinks tumor size. The researchers gave the new drug to seven different patients and the standard drug normally used to seven other patients. To show their study results, they calculated the average change in tumor size for each of the two groups. The results are shown in the figure below.

Figure 24. Decrease in Tumor Size: New Drug vs. Old Drug

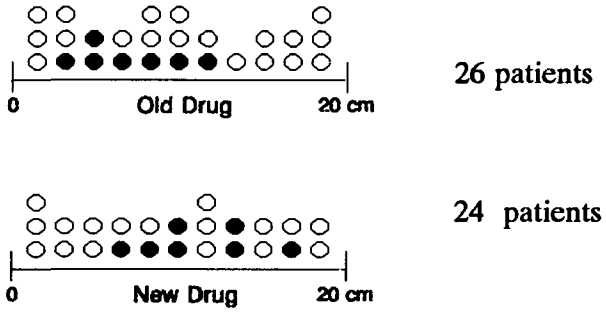


Source: Adapted with permission from SA Glantz, Primer of Biostatistics, 2nd Edition (1987), New York, McGraw-Hill, Inc.

The results show that in the seven people treated with the new drug, tumor size shrank more than in the group receiving the old drug, thus leading the researchers to conclude that the new drug was an effective chemotherapeutic agent *for those seven patients*! However, the researchers wanted to reduce tumor size for all patients with the disease, not just the few in the study group. *Statistical inference* is the process by which these researchers can answer the question "How likely is the new drug to shrink tumors in *all* people who received it?" based on the limited experience of their study.

Suppose that these same researchers could give the new drug to half of the entire population with disease (in this case, 50 patients), the old drug to the other half, and then measure any resulting changes in tumor size with the results, plotted below.

Figure 25. Decrease in Tumor Size for All Patients with Disease: New Drug vs. Old Drug.



Source: Adapted with permission from SA Glantz, Primer of Biostatistics, 2nd Edition (1987), New York, McGraw-Hill, Inc.

In figure 25, the individuals treated in the original study group are depicted by shaded circles; the patients added to the study are shown by unshaded circles. You can see that once the number of patients given the two drugs increases, there is no longer a difference in tumor shrinkage between the two groups. Now the researchers would have to conclude that the new drug was *not* a more effective chemotherapeutic agent than the old drug! What has happened? This *sample* of seven original patients from the population of all people with disease turned out not to be representative of how the whole population responded to the drug. The researchers would certainly like to know why this happened. By using a set of inferential statistical procedures known as *tests of hypotheses*, they could estimate how likely they were to select such an unrepresentative sample. Put another way, tests of hypotheses would allow the researchers to estimate how likely they were to erroneously conclude that the new drug was more effective in shrinking tumors when the relationship was actually due to selecting study subjects who were not representative of the population as a whole, and not to the effect of the drug itself. Clearly their study needed to address how applicable their study results, based on fourteen individuals, would be for a larger, target population.

As tumor registrars, you will want to be able to understand and evaluate the results of clinical trials and epidemiologic studies in order to keep abreast of new developments in cancer prevention and treatment. The rest of this section covers the building blocks of statistical inference. In the following section, section G, we will take up the topic of *statistical hypothesis testing.*

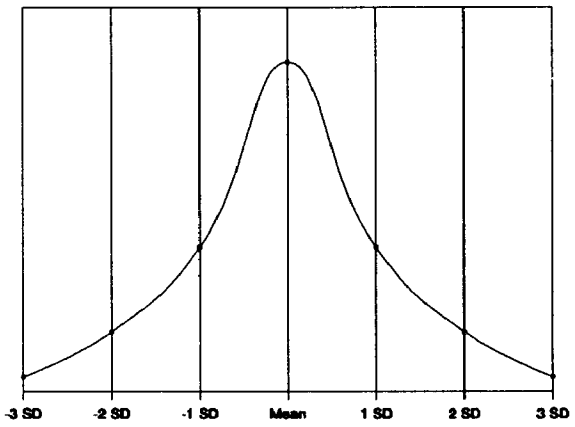
THE NORMAL DISTRIBUTION

In order to apply the technique of statistical inference, we must first understand the concept of a normal distribution, one of the most important frequency distributions in statistics. In appearance it is a symmetrical bell-shaped curve. Measurable characteristics occurring in nature--man, animals, and plants--tend to follow certain patterns. For instance, the frequency distribution of variables such as blood pressure, pulse rate, height, and serum cholesterol tend to take the shape of a *normal distribution* with little deviation from the average. *Normally distributed* means that if you were to measure the variable on every person in the population, you would find the frequency distribution would display a "normal" pattern with most of the measurements near the center of the frequency. You would also be able to completely describe the population, with respect to that variable, by calculating the mean and standard deviation of the values.

The Normal Curve

The frequency distribution of the normal population when plotted on arithmetic graph paper forms a curve with most of the observations near the center of the frequency distribution, and fewer and fewer observations occurring as you look further out in the tails. There are certain characteristics of a normal curve. First, each normal curve is bell-shaped and symmetrical about the mean. Second, the mean, median, and mode are identical. Third, *the width of the curves* depends on the standard deviation[1] (SD) or spread of values outward from the mean in both directions. It is possible to have multiple normal distributions with the same mean, median, and mode, but different standard deviations.

Figure 26. The Normal Curve



In a normal distribution, the following percentages of observed values will always lie between the mean minus a number of standard deviations (SD) and the mean plus a number of standard deviations.

---

[1] standard deviation--The different observations around the mean such that 95 percent of the observations lie between the mean and ± 1.96 standard deviations from a normal distribution.

Plus or Minus Standard Deviation	Percent Observations	
1 SD	68.27 percent	One-half (50 percent) of the observations will be
1.5 SD	86.64 percent	within ±0.6745 standard deviations of the mean.
2 SD	95.45 percent	
2.5 SD	98.76 percent	Ninety-five percent of the observations will be within
3 SD	99.73 percent	±1.96 standard deviations of the mean.

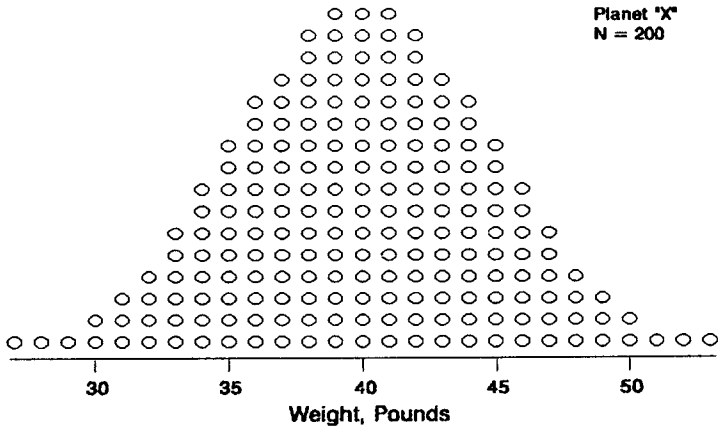
For calculation of the standard deviation, see section B, p. 74.

Medical decisions about categorizing individuals as having a disease or not and needing treatment or not require that some index of what is "normal" be available. A so-called "normal range" for a medical variable encompasses the values for a healthy population group. The ranges adopted will usually enclose about 95 percent of the values of randomly selected healthy people. Therefore, when a variable follows the normal distribution, *a medical "normal range" for that variable is simply the mean value plus or minus roughly 2 standard deviations*. Normal ranges often differ among age groups, sexes, and even geographic areas. For example, a "normal" serum cholesterol level varies between men and women and differs among age groups. When you see a normal range given for a variable, look for the population to which this range refers.

Example 02:

Suppose we want to summarize data about two hypothetical populations: women from planet "X" and women from planet "Y." There are only 200 women on "X" and 150 women on "Y," so we were able to record the weights of both *entire populations*. The resulting data for women from planet "X" are plotted in a frequency distribution in figure 27. You can easily see that most women on "X" weigh between 35 and 45 pounds. The remaining few weigh about 5 pounds more or five pounds less.

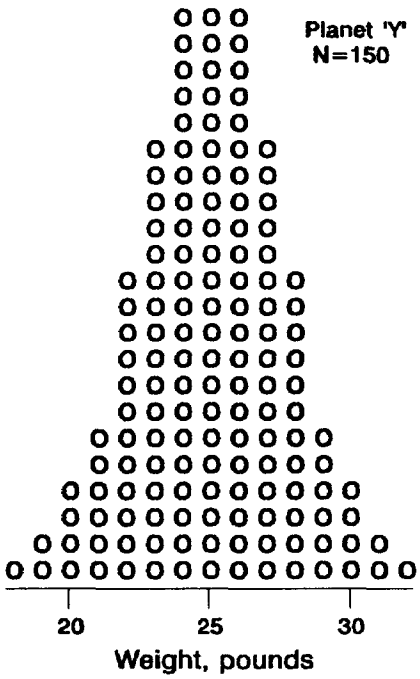
Figure 27. Frequency Distribution of Weights for All Women from Planet "X"



Source: Adapted with permission from SA Glantz, Primer of Biostatistics, 2nd Edition (1987), New York, McGraw-Hill, Inc.

The frequency distribution of weights of all 150 women from planet "Y" is shown below.

Figure 28. Frequency Distribution of Weights for All Women from Planet "Y"



Source: Adapted with permission from SA Glantz, Primer of Biostatistics, 2nd Edition (1987), New York, McGraw-Hill, Inc.

You can see that most women from "Y" weigh about 25 pounds, and that very few weigh less than 20 pounds or more than 30 pounds.

If you compare the two frequency distributions you notice that women from "Y" weigh less than those from "X" and that they also have less *variability* in their weights than do women from "X". Recall how to calculate a range. By doing so you can see that while most women from "X" range between 30 and 50 pounds, the range for women from "Y" is between 20 and 30 pounds. Also notice that despite differences in population size, average weight, and amount of variability in weight, the *pattern of the distributions* are virtually the same. It might not occur to you at first, but if you look more carefully, you can see that in both distributions, an individual is *more likely* to be near the middle of the distribution than to be far away from it. Also, each individual is *just as likely* to be either lighter or heavier than average. There is no tendency towards being only heavier or only lighter than average.

We now have carefully examined our raw data; therefore we can *reduce* this information about weight to a few summary statistics, namely the *mean* and *standard deviation*.

Table 42. Summary Statistics for Weight of Women from Two Planets

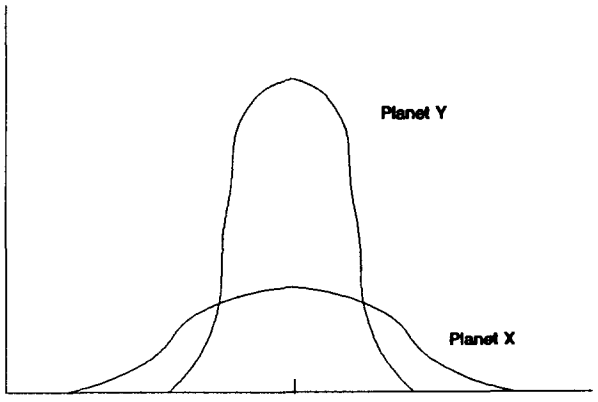
Population	Population Size	Population Mean	Population Std. Deviation
Women from "X"	200	40 lbs.	5.0 lbs.
Women from "Y"	150	25 lbs.	2.5 lbs.

We can express these results in narrative form by saying that the mean weight for women from "X" is 40 plus or minus 5 pounds, and the mean weight for women from "Y" is 25 plus or minus 2.5 pounds. Now we have summarized our earlier impressions, based on looking at the raw data, that women from "X" are heavier than women from "Y." Looking back at figures 27 and 28, if we were to count how many individual women from "X" fell within *one standard deviation of the mean*, we would find that approximately 68 percent of them weighed between 35 and 45 pounds. Similarly, 68 percent of the women from "Y" would weight between 22.5 and 27.5 pounds. If, for each population, we counted the number of women who fell between *two standard deviations of the mean*, we would find that about 95 percent of the women from "X" weighed between 30 and 50 pounds, and that 95 percent of the women from "Y" weighed between 20 and 30 pounds.

While it is true that the two populations have different mean weights and different amounts of variability in weight, the *patterns* of the two frequency distributions are actually similar to each other.

The population means and standard deviations completely define the shapes of curves. The curve for planet "X" (fig. 27) is wider and flatter than the curve for planet "Y" (fig. 28) because its standard deviation is twice as large. The positions of the curves on the x-axis are determined by the *mean* weight for each population.

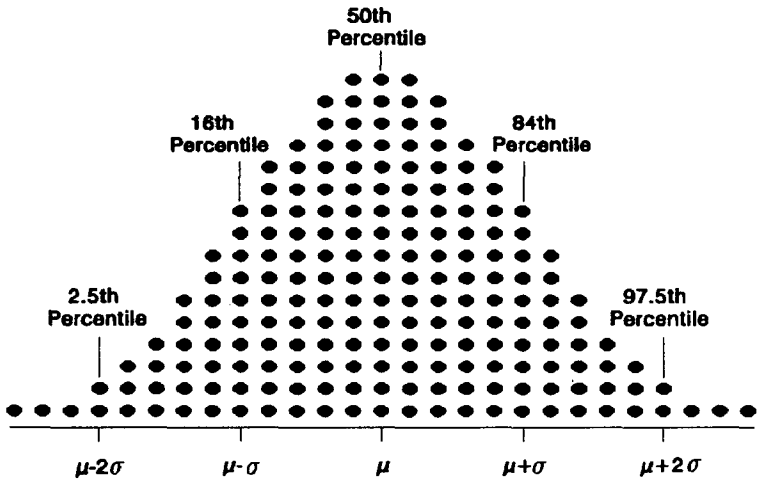
Example of Two Curves With the
Same Mean and Different
Standard Deviations



Percentiles of a Normal Distribution

Now that we have established how important normal ranges are for decision-making, e.g., medical decision-making, how do we tell if a variable being studied is normally distributed in the first place? An easy method for indicating the dispersion of values is to compute several *percentile points* of a population to see how close they are to those of a normal distribution. Figure 29, shows the values of percentile points for a normal distribution.

Figure 29. Percentile Points of the Normal Distribution



Source: Adapted with permission from SA Glantz, Primer of Biostatistics, 2nd Edition (1987), New York, McGraw-Hill, Inc.

Using the frequency distribution of weights of women from planet "X" in figure 27 above, we wish to find the values associated with the 2.5th, 16th, 50th, 85th and 97.5th percentiles.

In the following calculations, these Greek symbols are used:

μ - population mean (lower case mu)
σ - population standard deviation (lower case sigma)

Beginning with the 2.5th percentile, let us find the associated values. First, we determine which individual observation corresponds to the 2.5th percentile. Since there are a total of 200 women from "X," we convert the *percentile* to a *proportion*: 2.5 divided by 100 = 0.025, and multiply by 200: 0.025 x 200 = 5. Second, we work our way from the left-hand side of figure 27 to the right, counting off observations until we reach the 5th. The value corresponding to the 5th observation is 30 pounds.

Now we can compare our *observed* value for the 2.5th percentile with the value we would *expect* if the population were normally distributed. The expected value is found by subtracting 2 σ from the population mean, according to the formula given in figure 29. We know that the mean weight of women from "X" is 40, and the standard deviation is 5, therefore: $\mu - 2\sigma = 40 - (2 \times 5) = 30$, *exactly the same* as the observed value for the 2.5th percentile! (See table 43 below.)

Now let's find the observed value associated with the 16th percentile and compare it to the expected value. As before, we convert the percentile to a proportion: 16 divided by 100 = 0.16, and multiply this by 200, the total number of observations: 0.16 x 200 = 32. Working our way from the left-hand side of the frequency distribution in figure 27, we count off 32 observations and find that the value associated with the 32nd observation is 35 pounds. Looking again at the formulas given in figure 29, we see that the value expected for the 16th percentile of a normal distribution is (population mean - population standard deviation): 40 - 5 = 35. Again, the observed and expected values are exactly the same!

If we continued on to find the observed values for the 50th, 84th, and 97.5th percentiles of the frequency distribution of weights of women from "X," we would find that each value corresponds *exactly* to the value expected for a normally distributed population. This can be seen in the completed table below.

Table 43. Observed and Expected Values for Percentiles

Percentile	Observation #	Value Observed	Value Expected
2.5th	5	30	$\mu - 2\sigma = 40 - 10 = 30$
16.0th	32	35	$\mu - \sigma = 40 - 5 = 35$
50.0th	100	40	$\mu = 40$
84.0th	168	45	$\mu + \sigma = 40 + 5 = 45$
97.5th	195	50	$\mu + 2\sigma = 40 + 10 = 50$

In this example, the values associated with the percentiles are exactly the same as those expected on the basis of the mean and standard deviation of the population. This result occurred because we had carefully devised data for this example. In more realistic situations, when the observed values are not too different from the expected values, you may conclude that the data you have closely approximate the normal distribution and that the population mean and standard deviation do a good job of describing the population.

Why do you want to know if your data are from a population that is normally distributed? The answer is that many tests of hypotheses used in statistical inference are valid only if the population which is being studied approximately follows the normal distribution. However, not all distributions are normally distributed. For example, variations may be skewed resulting in an asymmetrical distribution which cannot be well described by its mean and standard deviation. In such a case, tests of significance which rely on an assumption of a normal distribution with equal mean, median, and mode do not apply.

Q1

When drawing conclusions about populations based on data from limited samples, the process is known as _____ _____. It is a concept we all use in our daily lives.

Q2

We are able to apply the above concept because of one of the important frequency distributions in statistics, the _____ _____.

Q3

If we were to observe variables such as blood pressure, pulse rate, height, and serum cholesterol for the entire population, the frequency distribution of these variables would take the shape of a _____ _____ and, if plotted, would form a _____ _____.

Q4

The normal curve is _____ about the mean, also, the mean, median, and mode are _____. Only the width of the curve may vary depending on the spread of values outward from the mean in both directions.

Q5

The spread of values outward from the mean in both directions is called the _____ _____ such that 95 percent of the observations lie between the mean and ±1.96 _____ _____ from a normal distribution.

Q6

How do you tell if a variable is normally distributed in the first place?

_____

_____

Q7

Why do you want to know if your data are from a population that is normally distributed?

_____

Answer: Q1

When drawing conclusions about populations based on data from limited samples, the process is known as statistical inference. It is a concept we all use in our daily lives.

Answer: Q2

We are able to apply the above concept because of one of the important frequency distributions in statistics, the normal distribution.

Answer: Q3

If we were to observe variables such as blood pressure, pulse rate, height, and serum cholesterol for the entire population, the frequency distribution of these variables would take the shape of a normal distribution and, if plotted would form a bell-shaped curve.

Answer: Q4

The normal curve is symmetrical about the mean, also the mean, median, and mode are identical. Only the width of the curve may vary depending on the spread of values outward from the mean in both directions.

Answer: Q5

The spread of values outward from the mean in both directions is called the standard deviation such the 95 percent of the observations lie between the mean and ± 1.96 standard deviations from a normal distribution.

Answer: Q6

To tell if a variable is normally distributed in the first place, determine whether the shape of the distribution is symmetrical (bell-shaped), and the mean, median are equal, and then compute several percentile points of the population to see how close they are to those of the normal distribution.

Answer: Q7

You want to know if your data are from a population that is normally distributed because many tests of hypotheses used in statistical inference are valid only if the population which is being studied approximately follows the normal distribution.

SAMPLE DISTRIBUTIONS

Up until now, everything we have done has been exact because we were able to examine every member of the populations we have studied. The real world does not contain only 200 women! Instead, we are limited to examining *samples* of individuals drawn from the population in which we are actually interested. In doing so, we hope that our sample is *representative* of the entire population, so that our conclusions about this sample can be extended to the larger group.

In example 01 of this section, we saw a sample of seven patients who received the new drug turn out not to be representative of how the population of individuals with the disease responded to the therapy. This may have occurred because the researchers drew these individuals from a very sick group of patients in their hospital, only to discover that they responded better to the drug than did patients with less advanced disease. Another explanation could be that these patients were all in a certain age group, which conferred an advantage as far as drug efficacy was concerned. These and other kinds of explanations are known as *confounders*. Instead of attributing the relationship between drug use and tumor shrinkage to the drug itself, the researchers would have to consider whether *prognostic factors*, such as age and progression of disease are more likely to explain the relationship. Confounding effects are very common in clinical and epidemiological research, and it is not possible to eliminate all of them. However, one very important preventive measure for avoiding them is through *random sampling*.

Random Sampling

In a *random sample*, every individual in the population has an equal and independent chance of being selected for the sample. Consider example 02. Suppose we were not able to weigh every woman on planet "X"? Instead, the funds available from our interplanetary research grant permit us to collect data for only 10 of the 200 women in the population. How do we select these 10 women? To ensure that we get a *random sample* of 10 women, we could write each woman's name on a card, put all of the cards in a hat, mix them thoroughly, and draw out one card. After writing down the name on the card on a list, we return the card to the hat, shuffle again, and draw out a second card. We would do this again and again until we had a list of 10 different names. This method ensures that every woman on the planet had an *equal and independent chance* of being selected. Since the cards printed with each woman's name were shuffled thoroughly before each draw, each woman had an *equal chance* of having the card with her name on it picked from the hat. *Independent chance* means that the probability of selecting one woman of a particular weight is not affected by the woman selected before her.

Another method of random sampling is to use a table of random numbers. A table of random numbers (see appendix 2A) contains several rows and columns of the digits zero through nine. The order of the digits follows no defined pattern, hence, it is "random." This means that entering the table at any point gives you an equal chance that any one of the digits zero through nine will be located there. Similarly, there is again an equal chance that any of 10 digits will occupy the immediately adjacent position on the page. To use a table of random numbers, consider our cancer researcher's study of new and old drugs.

The researcher has 14 patients whom she would like to randomly assign to receive either the new drug or the old drug. She turns to her table of random numbers (appendix 2A), and closing her eyes,

lets her finger fall to a starting point on the page. The new drug can be assigned to those study subjects for whom the digit was even, and the old drug to those for whom the digit was odd. The sequence of numbers our researcher selected and the drug assigned (N=new, O = old) are as follows:

Patient number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Random # drawn for each patient	7	8	5	4	2	4	2	7	8	5	1	3	6	6
Treatment group assigned	O	N	O	N	N	N	N	O	N	O	O	O	N	N

This process ensures that the sequence of drug assignments is random in order and that each study subject has an equal chance of being assigned either drug. Random number generators are available for use with computers and generally used by statisticians. In this example using random assignment, eight patients were randomly assigned the new drug and six the old drug.

CALCULATING SAMPLE STATISTICS

Population, Mean, and Standard Deviation

In the real world, since we can no longer measure every individual in the population of interest, we cannot calculate a *population mean* or a *population standard deviation* as we did for the women from planets "X" and "Y." Instead, we must *estimate* these population values from limited samples. These estimates of population values are called the *sample mean* and the *sample standard deviation.* These values are calculated in virtually the same way as the population mean and standard deviation described in section B, "Descriptive Statistics."

In the calculation of sample statistics, the following symbols are used:

$\bar{X}$ - *sample mean*

S_x - *sample standard deviation*

$S_{\bar{x}}$ - *standard error of the sample mean*

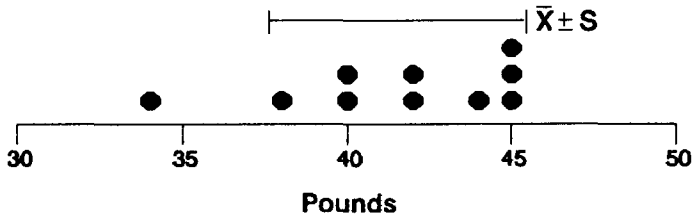
$$\textit{Sample mean } (\bar{X}) = \frac{\textit{sum of values of observations in sample}}{\textit{number of observations in sample}}$$

$$\textit{Sample standard deviation } (S_x) = \sqrt{\frac{\textit{sum of (value of observation in the sample - mean)}^2}{\textit{number observations in the sample - 1}}}$$

The sample mean and standard deviation calculated from a random sample are *estimates* of the mean and standard deviation of the entire population from which the sample was selected.

Returning to women's weights from planet "X," let us randomly sample 10 women from the entire population of 200.

Figure 30. Distribution of Weights for Sample of 10 Women from Planet "X"

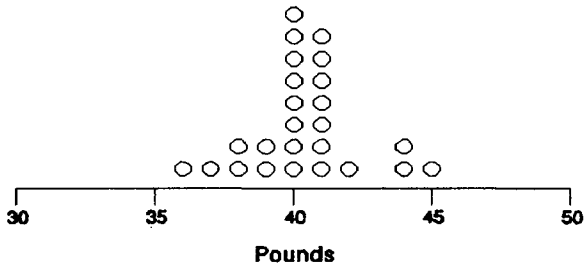


Source: Adapted with permission from SA Glantz, Primer of Biostatistics, 2nd Edition (1987), New York, McGraw-Hill, Inc.

The sample mean for these data is 41.5 pounds; the standard deviation is 3.6 pounds. These values are similar to the population mean and standard deviation, which were 40 pounds and 5 pounds, respectively.

If we continued to draw random samples of ten women from the population and calculated their means and standard deviations, we would find that each sample mean and standard deviation is *similar* to but not the same as the population parameters. We would also see that the sample statistics differ from one another. If we plotted the means of 25 such random samples, for example, we would get a distribution like the one below, in figure 31.

Figure 31. Distribution of Means of 25 Random Samples of Weights of 10 Women from Planet "X"



Source: Adapted with permission from SA Glantz, Primer of Biostatistics, 2nd Edition (1987), New York, McGraw-Hill, Inc.

Do you notice anything familiar about the shape of this distribution? You can see that the 25 sample means are distributed in a "bell shaped," normal fashion. We can, therefore, summarize this distribution of sample means by computing its mean and standard deviation. The mean of the 25 sample means is found by summing the 25 sample means and dividing by 25, the number of samples. The standard deviation of the sample means depicted above = 1.6 pounds. This "standard deviation of the means of random samples" is known as the *standard error of the mean*. It is a very important statistic, which measures how precisely a sample mean estimates the true population mean. Because the sample means are approximately normally distributed, 95 times out of 100, the true population mean will lie somewhere between the sample mean ± 1.96 standard errors of the mean. This situation is exactly the same as in example 02, when we saw that 95 percent of the weights of women from planet "X" fell between 30 and 50 pounds, or within *2 standard deviations of the mean weight* of 40 pounds.

The formula for calculating the standard error of the sample mean is:

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}}$$

where S_x equals the standard deviation of the sample and n is the sample size. An example should make the formula clear. Suppose that the mean age at diagnosis for a sample of 100 breast cancer patients is 59 years, and the standard deviation is 7 years. The standard error of the sample mean would be:

$$S_{\bar{x}} = \frac{7}{\sqrt{100}} = 0.7$$

Therefore, the sample mean, $\bar{X}$, ± 1.96 $S_{\bar{x}}$, i.e., ± 1.4 years, will capture the true population mean age of all breast cancer patients 95 percent of the time. In this example, we are 95 percent confident that the true (population) mean age of all breast cancer patients is 59 ±1.4 years or between 57.6 and 60.4 years. This is called a confidence interval and will be discussed more fully in the next section.

Proportions and Rates

Not all data are *continuous*, that is, a continuum of values from lowest to highest, e.g., tumor size and weight, but may be expressed in terms of *discrete* values such as rates and proportions.

A proportion is simply the number in a category divided by the total number in the entire series, for example:

$$\frac{\textit{Number of males}}{\textit{Number of males and females}} \quad \textit{or} \quad \frac{\textit{Number alive}}{\textit{Number alive and dead}}$$

Standard errors can also be calculated for sample estimates of proportions. Suppose that 14 percent of a sample of 20 patients receiving a new chemotherapy drug survived 5 years. We would like to know how well our sample estimate of the proportion surviving 5 years approximates the true rate we would observe if, instead of just 20 patients, we could examine all patients treated with the new drug. The formula for calculating the standard error of a sample proportion is:

$$S_p = \sqrt{\frac{p\ (1\ -\ p)}{n}}$$

where p is the sample estimate of the proportion of patients who survived 5 years and n is the sample size, which is 20 in this case. Thus,

$$S_p = \sqrt{\frac{0.14\ (1\ -\ 0.14)}{20}} = \sqrt{\frac{0.1204}{20}} = \sqrt{0.00602} = 0.078$$

Thus, we are 95 percent confident that the population survival rate is $0.14 \pm 1.96\ (0.078)$ or between 0 and 29.3 percent. Notice, with a small sample size (n) our standard error is large.

Rates describe the rapidity with which a given event occurs, such as a mortality rate and survival rate. Calculation of mortality and survival rates are discussed in sections C and D of this manual.

SETTING CONFIDENCE INTERVALS

Population Mean

We have seen that the distribution of sample means approximately follows the normal distribution, and, therefore, that the true population mean lies within about two standard errors of the mean 95 percent of the time. We will now use the standard error of the mean to set *confidence intervals* around an estimate of the population mean. *Confidence intervals* estimate the range of values that include the actual population mean (μ).

The following expression is used for setting a 95 percent confidence interval around a sample mean:

$$Pr\ [\ \bar{X}\ -\ 1.96\ S_{\bar{x}} < \mu < \bar{X}\ +\ 1.96\ S_{\bar{x}}\ ] = 0.95$$

The left-hand side of the expression is used to compute the *lower bound* of the confidence interval; the right-hand side is used to compute the *upper bound*. You can see that the standard error of the mean is multiplied by the value 1.96 in order to obtain these bounds. The interval between the lower and upper bounds is called the *confidence interval*. The "Pr" in the expression stands for probability, and simply means that if you were to take 100 random samples of women's weights and constructed these upper and lower bounds for each sample, you could expect 95 of the 100 confidence intervals

to contain the true population mean of 40 pounds, and 5 of the 100 confidence intervals to miss it.

To make this clearer, we will look at two examples using our interplanetary friends, the women from planet "X." To calculate the *95 percent confidence interval* for a random sample of 10 women from "X" (you can refer to the actual distribution of these 10 weights in figure 30), four simple steps are followed.

Step

1. Calculate the sample mean. This has already been done for the random sample of 10 weights shown in figure 30, and was found to be 41.5 pounds.

2. Calculate the sample standard deviation. This also has already been done, and was found to be 3.6 pounds.

3. Calculate the standard error of the sample mean, using the formula:

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}} = \frac{3.6}{\sqrt{10}} = 1.14 \text{ pounds}$$

4. Plug the sample mean and standard error of the mean into the expression for obtaining a 95 percent confidence interval:

$$\textit{The lower limit is } L_1 = \bar{X} - (1.96)(S_{\bar{x}}) = 41.5 - (1.96)(1.14) = 39.3 \text{ pounds}$$

$$\textit{The upper limit is } L_2 = \bar{X} + (1.96)(S_{\bar{x}}) = 41.5 + (1.96)(1.14) = 43.7 \text{ pounds}$$

We express our results by saying that 95 percent of the time, our confidence interval contains the true population mean, or that the mean weight of women in the population lies somewhere between 39.3 and 43.7 pounds 95 percent of the time.

Consider a second sample of 10 randomly selected women from planet "X." Suppose the sample mean for this group was 36 pounds and the standard deviation was 5 pounds. Beginning with step number three, calculate the standard error of the sample mean (1.58 pounds).

Next, plug the sample mean and standard error of the mean into the expression for obtaining the 95 percent confidence interval:

$$L_1 = 36 - (1.96)(1.58) = 32.9 \text{ pounds} \qquad L_2 = 36 + (1.96)(1.58) = 39.1 \text{ pounds}$$

This result tells us that the mean weight of women in the population lies somewhere between 32.9 and 39.1 pounds. The standard error of the first sample was 1.14, while that for the second was 1.58. The second sample has a *wider* confidence interval than does the first. This means that there is a greater range of values within which the population mean lies. Therefore, the second sample does not provide as precise an estimate of the population mean as does the first sample.

Another important point about confidence intervals is that the higher the level of confidence, the wider the interval. If being right 95 times out of 100 is not enough, and you wanted to be even more sure that your confidence interval covered the true population mean, you could set a *99 percent confidence interval* around the sample mean. By doing this you could be sure that 99 times out of 100, the true population mean would lie within the confidence interval. However, a 99 percent confidence interval is 1.3 times as wide as a 95 percent confidence interval. Thus, you get greater confidence that your interval covers the true mean but could be much less certain what the true value of the mean actually is because of the wider interval!

SETTING CONFIDENCE INTERVALS

Proportions and Rates

A confidence interval on a sample mean concerns only the *mean* of the population from which the sample was selected. It does not enclose a proportion of the population. For example, in a case where a 95 percent confidence interval of 31.5 to 44.8 months was found for mean survival time in patients receiving a new cancer therapy, *you could not say* that 95 percent of the survival times are enclosed within those bounds. Instead, you *could* say that there is 95 percent *certainty* that the confidence interval of 31.5 to 44.8 months *contains the mean survival in the underlying population* from which the sample of patients was selected.

Confidence intervals can also be used to see how reliable a sample proportion (p) is at estimating a population proportion. Remember that the distribution of sample means follows the normal distribution. Because the sample means are normally distributed, we were able to calculate confidence intervals for sample estimates of the population mean. There is also a distribution for proportions, which follows what is called *the binomial distribution*. When the sample size is large, the binomial distribution approximates the normal curve. This allows us to use confidence intervals to estimate a population proportion based on a sample proportion. The binomial distribution is applicable to data for proportions where there are only two possible outcomes, for example, success or failure, survival or death, treated or not treated, early diagnosis or late diagnosis, etc. The proportion of the population having the characteristic under study is represented by p, while all others are represented by $1\text{-}p$ since you either have the characteristic or you don't.

Suppose a cancer registrar working in a population-based registry was involved in a study of endocrine surgery (bilateral orchiectomy) for treatment of prostate cancer. She reviewed a random sample of 125 records and found that 32 (26 percent) of patients with advanced prostate cancer were treated with bilateral orchiectomy. She now wishes to use this sample proportion to estimate, with

95 percent confidence, the proportion of the prostate cancer patient population who received this therapy. The procedure for calculating the 95 percent confidence interval is analogous to that for the confidence interval for the mean, and the formula is:

$$95\%\ \textit{confidence interval} = p \pm 1.96 \sqrt{\frac{p\ (1\ -\ p)}{n}}$$

where p is the sample proportion.

Therefore, the 95 percent confidence interval (CI) for the proportion of patients treated with bilateral orchiectomy is $0.26 \pm 1.96 \sqrt{\frac{0.26\ (1-0.26)}{125}} = 0.26 \pm 0.077$.

The lower bound of the confidence interval is 0.26 - 0.077 = 0.183; the upper bound of the confidence interval is 0.26 + 0.077 = 0.337. Thus, the registrar would have 95 percent certainty that in the underlying population of prostate cancer patients, the proportion receiving endocrine surgery is somewhere in the range of 18.3 and 33.7 percent, based on the assumption that treatment patterns among hospitals follow a normal distribution.

Q8

Confounding effects are very common in clinical and epidemiological research and the best way to avoid them is through _____ _____.

Q9

Every individual in the study population has an equal and independent chance of being selected in a _____ _____.

Q10

In the real world we cannot measure every individual in the population of interest, so we estimate the total population values from limited samples by calculating the _____ _____ and the _____ _____ _____.

Q11

The standard deviation of the means of random samples is known as the _____ _____ _____ _____ _____.

Q12

There is also a distribution for sample proportions called the _____ _____.

Q13

When data cannot be expressed in terms of discrete values, sample statistics can be calculated for _____ and _____.

Q14

We can set up _____ _____ to estimate the range of values that include the actual population mean.

Answer: Q8

Confounding effects are very common in clinical and epidemiological research and the best way to avoid them is through random sampling.

Answer: Q9

Every individual in the study population has an equal and independent chance of being selected in a random sample.

Answer: Q10

In the real world we cannot measure every individual in the population of interest, so we *estimate* the total population values from limited samples by calculating the sample mean and the sample standard deviation.

Answer: Q11

The standard deviation of the means of random samples is known as the standard error of the mean.

Answer: Q12

There is also a distribution for sample proportions called the binomial distribution.

Answer: Q13

When data cannot be expressed in terms of discrete values, sample statistics can be calculated for proportions and rates.

Answer: Q14

We can set up confidence intervals to estimate the range of values that include the actual population mean.

SECTION G

STATISTICAL HYPOTHESIS TESTING

SECTION G

STATISTICAL HYPOTHESIS TESTING

INTRODUCTION

So far, we have used a variety of descriptive statistics such as the mean, median, and standard deviation to summarize data, and the standard error of the mean to estimate how reliably a sample mean estimates a population mean. We have used the standard error of the mean to set confidence intervals around sample means so that we can say that 95 percent of the time the true population mean lies within the range of values enclosed by the confidence interval. Similarly, we have used sample proportions to estimate population proportions and confidence intervals to see how reliable these estimates are.

We are now ready to learn how statistical methods are used to test *scientific hypotheses*. These statistical techniques are called *tests of significance*. In cancer research, the scientific hypothesis being tested is often whether different treatments (surgery, chemotherapy or radiation protocols, etc.) have an effect on some variable (tumor shrinkage, survival time). Different statistical tests are employed from those used with continuous data, such as tumor size, when the variable of interest is a proportion, such as the proportion of breast cancer patients surviving 10 years.

In the course of preparing annual reports or patient care evaluation studies, tumor registrars are likely to encounter published studies reporting "significantly different survival rates" as a result of some new cancer-directed therapy. In this section, you will become acquainted with the statistical hypothesis testing process for both proportions and continuous data. This is intended as an introduction to the hypothesis testing process so that you can familiarize yourself with basic statistical methods used in clinical and epidemiological research publications. You should not expect to go ahead and carry out studies of your own yet. Therefore, the exercises at the end of this section emphasize study evaluation skills rather than actually carrying out statistical tests.

You are probably already aware of many questions or hypotheses currently being investigated in cancer research and other areas. Is alcohol consumption related to breast cancer risk? Does eating oat bran lower serum cholesterol? Does AZT slow progression of AIDs in patients diagnosed early with HIV infection?

WHAT IS A HYPOTHESIS?

In clinical and epidemiological research studies, a *hypothesis* is a statement which claims a relationship exists between a study variable and an outcome variable. An epidemiologist hypothesizes that pesticide exposure poses a risk for developing leukemia. Her hypothesis is that the study variable, pesticide exposure, is related to the outcome variable, leukemia. A clinician wishes to demonstrate that a new chemotherapy treatment protocol for treating ovarian cancer is more effective than the standard therapy. His hypothesis is that the new protocol results in longer survival than the standard one. The study variable is the chemotherapy protocol and the outcome variable is survival.

In testing hypotheses using statistical techniques, the hypothesis is actually posed in the opposite way to what is really being investigated. For example, if the clinician hypothesizes that new treatment A is superior to standard treatment B, he would actually state his study hypothesis as follows: *Treatment A is the same as treatment B*. In statistics, this is called the *null hypothesis* or hypothesis of no difference (abbreviated H_0). In this situation, the objective of statistical hypothesis testing is to *reject* the null hypothesis in favor of the *alternative hypothesis*. Here the alternative hypothesis is that treatment A is NOT the same as treatment B. Forming the null and alternative hypotheses is a critical step in carrying out clinical trials and epidemiological research.

HYPOTHESIS TESTING

Testing for Differences Between Two Populations Means ($\mu_e - \mu_s$)

A common hypothesis in clinical trials research is that some new therapeutic agent confers better survival than does another therapeutic agent.

Thus, we often are concerned with comparing *two population means* in assessing, for example, the relative effectiveness of two treatments. We may have a standard drug (treatment "s") for a disease, and we may wish to compare it with a new or experimental drug (treatment "e") that has yet to be tested. Our objective will be to estimate the value of $\mu_e - \mu_s$ where μ_e is the average response to the new product and μ_s is the average response to the standard treatment. The μ's are population mean values indicating what the average response would be if the treatments were administered to all potential recipients of these treatments.

In order to compare the two treatments, it is necessary to collect two sets of data, one for each treatment. We shall use our sample means, ($\bar{X}_e$ and $\bar{X}_s$) and their difference ($\bar{X}_e - \bar{X}_s$) to estimate the difference in the population means ($\mu_e - \mu_s$) in which we are primarily interested.

In designing or planning our study, there are two approaches to be considered:

1. Analysis of Paired Observations

Sometimes we can use both treatments on the same subject. For example, in testing a product for pain relief we can first use one treatment and later the second treatment on the same patient and compare the results. Employing this procedure results in the collection of pairs of observations on each of a number of subjects, and we study the difference in treatment results ($X_e - X_s$) for each subject.

If we try the two agents on the same person and take the difference in response to each, we may anticipate that the difference in results will primarily reflect the difference in effectiveness of the two treatments. In contrast, if we compare the response to the *new treatment* on one person and the response to the *standard treatment* on a second person, we may not be sure how much of the difference in results will be due to the difference in the effectiveness of the two treatments and how much may be due to the difference between the two subjects in their sensitivity to drugs of the kind being tested. Since our interest is in the difference in effect of treatment, this suggests that there may be advantages to collecting paired data on the same subjects. In fact, if the same individual can receive both treatments, or if two "similar" individuals can be paired, we can obtain the same amount of information for estimation of ($\mu_e - \mu_s$) with a smaller size study than would be necessary in a study without pairing.

2. Analysis of Independent Samples

There are times when pairing is either not practical or not possible. Perhaps a treatment will have a long-term effect that will prevent use of the second drug on the same individual. Also, if the same person cannot be used in both treatments, it may not be easy to find a partner with the necessary "similar" characteristics (age, sex, race, stage of disease, etc.) for treatment with the second drug. Furthermore, it is not always known which are the important characteristics on which to match pairs of individuals. In these situations, we would employ independent samples to estimate the difference in effectiveness of the two treatment procedures.

Thus, we may use treatment "e" (experimental) on one group of persons and treatment "s" (standard) on a second independent group. This can be accomplished by randomly assigning half of those available for the study to one treatment and the other half to the second treatment.

Q1

In research studies, a statement which claims a relationship between a study variable and an outcome variable is called a _____. You actually state that there is no difference, the _____ _____.

Q2

A common hypothesis in clinical trials research is that some new therapeutic agent confers better survival than does another therapeutic agent. To compare the relative effectiveness of the two treatments you would compare the two _____ _____ to determine if there was a difference between the two _____.

Q3

What is the value of pairing observations of two agents on the same person?

_____

_____

Q4

Why is it not always possible to do a paired study?

_____

_____

Answer: Q1

In research studies, a statement which claims a relationship between a study variable and an outcome variable is called a hypothesis. You actually state that there is no difference, the null hypothesis.

Answer: Q2

A common hypothesis in clinical trials research is that some new therapeutic agent confers better survival than does another therapeutic agent. To compare the relative effectiveness of the two treatments you would compare the two population means to determine if there was a difference between the two means.

Answer: Q3

The value of pairing observations of two agents on the same person is that the same amount of information can be obtained for estimation of the difference in population means ($\mu_e - \mu_s$), but with a smaller size study than would be necessary if pairing were not possible.

Answer: Q4

It is not always possible to pair observations because one drug may have a long term effect which makes giving the second drug to the same individual impossible. Further, in paired studies it is not always clear which patient characteristics should be "matched."

CALCULATING HYPOTHESIS TESTS

Confidence Intervals for Differences Between Two Population Means--t Test

We shall now proceed to consider how to obtain confidence intervals for the difference between two population means and also how to test hypotheses or claims about the magnitude of the difference, such as that possibly advanced by the manufacturer of the new product. Analysis of paired data will be presented first, followed by the analysis of data from two independent samples.

1. Paired t Test

Data shown in table 44 are provided by Colton[1] on the effect of placebo and hydrochlorothiazide on the systolic blood pressure of 11 hypertensive patients. We wish to find the average difference in blood pressure employing hydrochlorothiazide compared with placebo. For simplicity let us call this average difference μ_d. The average difference (μ_d) is numerically the same as the difference between the two population means ($\mu_p - \mu_h$) where X_p and X_h stand for individual observations using placebo and hydrochlorothiazide, respectively.

Our attention will be directed toward the column of 11 differences in blood pressure readings for the 11 subjects, i.e., the blood pressure following placebo (X_p) minus the blood pressure following the use of hydrochlorothiazide (X_h). Each of the 11 differences will be designated d.

The average difference, $\bar{d}$, is 24.0 millimeters of mercury. Our calculated $\bar{d}$, or average difference, is our best estimate of μ_d which equals $\mu_p - \mu_h$. (If you take the time, you can confirm that this average difference, $\bar{d} = 24.0$, is equal to the difference between the two treatment averages, i.e., $\bar{X}_p$ minus $\bar{X}_h$ equals 24.0).

The paired t-test statistic is:

$$t = \frac{\textit{average difference of paired means}}{\textit{standard error of the difference}}$$

The formula is:

$$t = \frac{\bar{d}}{s_{\bar{d}}}$$

---

[1]Theodore Colton, Statistics in Medicine, Little, Brown and Co., Boston, 1974.

Our approach to obtaining a 95 percent confidence interval for μ_d is identical to that used in section F for finding a confidence interval for a population mean, μ_x. The only difference is that we shall process the d values and their mean, $\bar{d}$, rather than X values and their mean, $\bar{X}$.

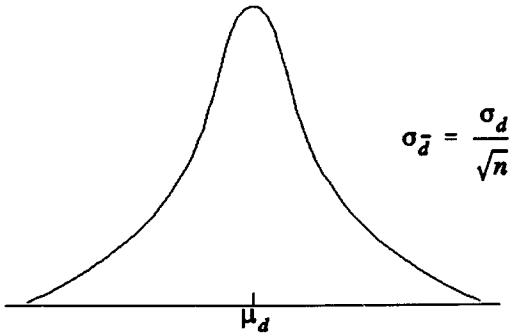
Table 44. Comparison of Paired Means: Effect of Placebo and Hydrochlorothiazide on Systolic Blood Pressure of 11 Hypertensive Patients (Systolic Blood Pressure in mm Hg)

Patient	Placebo X_p	Hydrochlorothiazide X_h	Difference d
FB	211	181	30
IF	210	172	38
PG	210	196	14
HF	203	191	12
RR	196	167	29
LP	190	161	29
BK	191	178	13
IF	177	160	17
MK	173	149	24
MT	170	119	51
JM	163	156	07
			Σd = 264 mm Hg
			$\bar{d}$ = 24.0 mm Hg

$\sum_d (d - \bar{d})^2 = 1{,}714$

$$S_d^2 = \frac{\sum (d - \bar{d})^2}{(n - 1)} = \frac{1714}{10} = 171.4 \text{ mm}^2 \text{ Hg}$$

From Section F, we know that sample means (here $\bar{d}$'s) follow a normal distribution about μ_d with standard error of $\bar{d}$ equal to $\sigma_{\bar{d}} = \frac{\sigma_d}{\sqrt{n}}$ where σ_d is the standard deviation of the distribution of d values.



Distribution of $\bar{d}$

To calculate the 95 percent confidence interval (CI) for μ_d using our sample means, we first estimate

$\sigma_{\bar{d}}$ by using $S_{\bar{d}} = \frac{S_d}{\sqrt{n}}$.

Here, $S_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n - 1}} = \sqrt{\frac{1{,}714}{10}} = \sqrt{171.4} = 13.09$ mm Hg.

Therefore, $S_{\bar{d}} = \frac{13.09}{\sqrt{11}} = \frac{13.09}{3.32} = 3.94$ mm Hg.

Then we find our 95 percent CI by calculating $t_{0.05} \cdot S_{\bar{d}}$ where t is based on the statistic

$$t = \frac{\textit{difference in sample means}}{\textit{standard error of difference in sample means}}.$$

In this example we have not calculated t but we will use a table of t values (appendix 2B) to determine what value of t would be significant at the 95 percent level designated as $t_{0.05}$. To use the t table we need to know the *degrees of freedom* from our sample. Since we have 11 d values, the degrees of freedom are n - 1 = 10. This is based on the fact that if we know the total of the differences and any ten values of d, we automatically know the eleventh value. From the t table (appendix 2B) we find $t_{0.05,10df}$ by reading down the column headed "Degrees of Freedom" to find the number "10" and across that row to the column headed "0.050" to find the value of 2.228 (which rounds to 2.23).

Therefore, our 95% H_0CI $= t_{0.05,10df} \cdot S_{\bar{d}} = (2.23)(3.94) = 8.78$.

The lower limit of our 95 percent confidence interval for μ_d is found by taking $\bar{d}$ - 95% CI = 24.0 - 8.78 = 15.22 mm Hg and the upper limit is $\bar{d}$ + 95 percent CI = 24.0 + 8.78 = 32.78 mm Hg. Often we use the notation 95% H_0CI to indicate 95 percent confidence intervals around the null hypothesis of no treatment difference.

Our conclusion is that we are 95 percent confident that the average blood pressure of the population of patients on hydrochlorothiazide would be between 15.22 and 32.78 mm Hg lower than the blood pressure of the population of patients on placebo.

Test of Hypotheses on Values of μ_d

We would reject at the 5 percent level of significance any claim (or hypothesis) that μ_d is either below 15.22 or above 32.78 mm Hg. We would therefore reject the null hypothesis that there is no difference in blood pressure following use of hydrochlorothiazide compared with the use of a placebo since 0.0 (the null value) falls below the lower limit of our confidence interval. We could not reject at the 5 percent level of significance any hypothesized value of μ_d between 15.22 and 32.78 mm Hg.

2. Unpaired t Test

Consider the data below on uterine weights of two groups of rats, one group treated by estrogens and the other group untreated. Animals were sacrificed in order to permit excision and weighing of each uterus.

Table 45. Uterine Weights (mg) of Rats Treated With an Estrogen Compared With Untreated Controls--Two Independent Samples

Estrogen Treated	Untreated
X_T - mg	X_U - mg
33	18
35	23
21	20
23	17
31	22
24	16
29	12
30	28
26	24
$\Sigma X_T = 252$	$\Sigma X_U = 180$
$n_T = 9$	$n_U = 9$
$\bar{X}_T = 28$ mg	$\bar{X}_U = 20$ mg
$\Sigma(X_T - \bar{X}_T)^2 = 182$	$\Sigma(X_U - \bar{X}_U)^2 = 186$

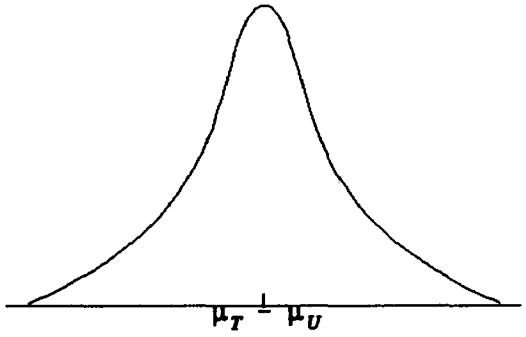
We wish to estimate, as before, a confidence interval for the difference between μ_T and μ_U, the population mean weights of treated and untreated rats. Our estimate will again be based on a t-test statistic. The formula for the unpaired t test is:

$$t = \frac{\bar{X}_T - \bar{X}_U}{S(\bar{X}_T - \bar{X}_U)}$$

For the method to be described below, we make the assumption that the variability in uterine weights of treated and untreated rats is the same, i.e., $\sigma_{X_T} = \sigma_{X_U} = \sigma_X$.

We will estimate $\mu_T - \mu_U$ by using $\bar{X}_T - \bar{X}_U$, the difference between the sample means. As with individual X's the difference between means of independent samples ($\bar{X}_T - \bar{X}_U$) follows a normal distribution about $\mu_T - \mu_U$ with $\sigma_{\bar{x}_T - \bar{x}_U}$, the standard deviation (or standard error) of $\bar{X}_T - \bar{X}_U$ is

equal to $\sigma_x\sqrt{\frac{1}{n_T} + \frac{1}{n_U}}$ where n_T = number in sample of treated rats and n_U = number in untreated sample.



Distribution of $\bar{x}_T - \bar{x}_U$

Calculation of 95 percent Confidence Interval for $\mu_T - \mu_U$

From table 45 we calculate $\bar{X}_T = \sum X_T/n_T = \frac{252}{9} = 28$ mg,

$\bar{X}_U = \sum X_U/n_U = \frac{180}{9} = 20$ mg

and $\bar{X}_T - \bar{X}_U = 8$ mg.

Since we do not know σ_x, we estimate $\sigma_{\bar{x}_T - \bar{x}_U}$ by using the formula $S_p\sqrt{\frac{1}{n_T} + \frac{1}{n_U}}$ where S_p (our sample estimate of σ_x) brings together or *pools* the information on variability from both samples of n_T and n_U observations, respectively. We could have estimated σ_x from each sample separately, if we wished.

The formula for S_p which combines or pools data from the two samples is

$$S_p = \sqrt{\frac{\sum (X_T - \bar{X}_T)^2 + \sum (X_U - \bar{X}_U)^2}{n_T + n_U - 2}} = \sqrt{\frac{182 + 186}{9 + 9 - 2}} = \sqrt{\frac{368}{16}} = \sqrt{23} = 4.80$$

As before when we had to estimate a σ from sample data, we use t instead of 1.96 in calculating our 95% H_0CI. Since we have $(n_T - 1)$ df for estimating σ_x in the treated sample and $(n_U - 1)$ df for estimating σ_x from observations in the untreated sample, our estimate S_p based on the combined information has a total of $((n_T - 1) + (n_U - 1) = (n_T + n_U - 2))$ df. Our 95% H_0CI will then be equal to:

$$t_{0.05} \left(S_p \sqrt{\frac{1}{n_T} + \frac{1}{n_U}} \right)$$

t has $(n_T + n_U - 2)$ df = (9 + 9 - 2) df = 16 df.

From the t table, $t_{0.05,16df} = 2.120$.

Our 95% H_0CI is $t_{0.05} \left(S_p \sqrt{\frac{1}{n_T} + \frac{1}{n_U}}\right) = (2.12)(4.80) \sqrt{\frac{1}{9} + \frac{1}{9}} = (2.12)(4.80)(0.471) = 4.79.$

Thus, our 95 percent confidence interval for $\mu_{X_T} - \mu_{X_U}$ is:

Lower Limit: $(\bar{X}_T - \bar{X}_U)$ - 95 percent H_0CI = 8 - 4.79 = 3.21 mg

Upper Limit: $(\bar{X}_T - \bar{X}_U)$ + 95 percent H_0CI = 8 + 4.79 = 12.79 mg.

Therefore, we are 95 percent confident that the average weight of uteri of estrogen treated rats is between 3.21 and 12.79 mg heavier than uteri of untreated rats.

Tests of Hypotheses on Values of $\mu_T - \mu_U$

We would reject at the 5 percent level of significance any claim (or hypothesis) that $\mu_T - \mu_U$ is less than 3.21 (including the null hypothesis of no difference) or greater than 12.79 mg.

Sample Size and the t Test

If the confidence intervals for μ_d (in the case of paired samples), or for $\mu_T - \mu_U$ (using independent samples) are too broad for the needs of the investigator, these may be reduced in length by increasing the sample size. If estimates of the population standard deviations are available, it is possible to use tables available in statistics texts to determine the required size of samples to meet specifications on how close the sample estimates should be to the universal values with a defined level of confidence.

The degree to which we can reduce the size of a paired study compared to a study with independent samples will depend on the value of σ_x^2 and on how well we are able to do our pairing. This is discussed in standard statistical texts.

Difference in Rates and Proprotions--**z Test**

Hypothesis testing can also be carried out using sample proportions rather than sample means. Suppose, in a clinical trial of new drug A versus standard drug B, researchers found that 10 of a total of 37 patients (27.0 percent) randomized to receive drug A and 8 of the 42 patients (19.0 percent) randomized to receive drug B survived for 5 years after treatment. The researchers would like to *reject the null hypothesis* that there is no difference in the two drugs with respect to 5-year survival, in favor of *accepting the alternative hypothesis* that the two drugs are different in effect.

There is a standard statistical test which these researchers can employ to see if there is a *statistically significant difference between the two proportions*, known as the **z-test**. The test statistic is:

$$z = \frac{\textit{difference of 2 sample proportions}}{\textit{standard error of difference of 2 sample proportions}}$$

Therefore, the form of this statement used to compute the z statistic is

$$z = \frac{p_1 - p_2}{\sqrt{p\,(1 - p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where p equals the proportion of *all patients* who survived for 5 years regardless of which drug they received. p_1 and p_2 are the *observed proportions* of patients surviving 5 years for drug A and drug B, respectively. n_1 and n_2 are the sample sizes of the drug A group and the drug B group.

It is conventional to require that np is greater than or equal to ($\geq$) 5 where

n = the number in a sample

p = proportion in a specific category.

In order to test the null hypothesis that drug A and drug B are not different, the researchers simply plug their numbers into the formula. Their data are presented in the table below.

Table 46. Survival Time for Drug A and Drug B Recipients

Survival Times	Drug A	Drug B	Total
5 years or more	10	8	18
< 5 years	27	34	61
Total	37	42	79

The values required for the z test are computed as follows:

$$p_1 = \frac{10}{37} = 0.270 \qquad n_1 = 37$$

$$p_2 = \frac{8}{42} = 0.190 \qquad n_2 = 42$$

$$p = \frac{10 + 8}{37 + 42} = \frac{18}{79} = 0.228$$

Therefore,

$$z = \frac{p_1 - p_2}{\sqrt{p\,(1 - p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.270 - 0.190}{\sqrt{0.228\ (1 - 0.228)\left(\frac{1}{37} + \frac{1}{42}\right)}}$$

$$= \frac{0.080}{\sqrt{0.228\ (0.772)\ (0.027 + 0.024)}}$$

$$= \frac{0.080}{\sqrt{0.0090}} = \frac{0.080}{0.095} = 0.842$$

How can we interpret the calculated z-value of 0.842?

Using our knowledge of the normal distribution, we realize that in only one time in 20 (5 percent of the time) will the z-value from the standard normal curve exceed the value ±1.96 (see page 182). The z-value calculated above in effect indicates how far out on a standard normal curve the difference in observed proportion (0.270 - 0.190 = 0.080) is. Conventionally, we reject the null hypothesis when the z-value is at least 1.96. From the z-table in appendix 2C we find that a value of 0.842 has a probability of 0.30 by reading down the first column to 0.8 and across that row to the column headed 0.04 to find the value for 0.84 which is 0.2995. Therefore, a z-value of 0.842 or larger could easily have occurred due to chance if null is true and cannot be regarded as inconsistent with the hypothesis.

In statistical jargon, if the z-value is 1.96 or greater reflecting the difference in the proportion of patients surviving, we say the difference is significant at the 5 percent level. You will often see this written as $P < 0.05$.

Just to be certain that you understand how to use the z test, consider another example. The following data are from a study of survival rates for breast cancer in black and white women. The study was conducted using data from a *population-based* registry, and is therefore representative of the experience in the population.

Table 47. Observed Frequencies of Black and White Women with Breast Cancer Surviving for 5 Years

Survival Time	White	Black	Total
5 years or more	285	178	463
< 5 years	114	118	232
Total	399	296	695

What is the null hypothesis for this study? There is no difference in 5-year survival rates between black and white women with breast cancer. What is the alternative hypothesis? Black women have *different* 5-year survival rates for breast cancer than do white women. The *study variable* is race and the *outcome variable* is 5-year survival.

In order to apply the z test, we first check to see if the data meet the criterion that *np* for each sample ≥ 5. Since p_1 is the proportion of whites surviving 5 years out of the total number of whites studied, and p_2 is the proportion of blacks surviving 5 years out of the total number of blacks studied, it follows:

$p_1 = 285/399 = 0.714$
$p_2 = 178/296 = 0.601$

n_1p_1 = (sample size of whites) (proportion of whites surviving 5 yrs) = 399 (0.714) = 285
n_2p_2 = (sample size of blacks) (proportion of blacks surviving 5 yrs) = 296 (0.601) = 178

Both values clearly exceed five, so we can go ahead and use the z test. First we obtain our numbers to plug in

$p_1 = 0.714 \quad n = 399$
$p_2 = 0.601 \quad n = 296$

$p = \frac{285 + 178}{399 + 296} = 0.666$ Now we can go ahead and solve for z:

$$z = \frac{p_1 - p_2}{\sqrt{p\,(1 - p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$= \frac{0.714 - 0.601}{\sqrt{0.666\ (1-0.666)\left(\frac{1}{399} + \frac{1}{296}\right)}}$$

$$= \frac{0.113}{\sqrt{(0.666)\ (0.334)\ (0.0025 + 0.0034)}}$$

$$= \frac{0.113}{\sqrt{(0.666)\ (0.334)\ (0.0059)}}$$

$$= \frac{0.113}{\sqrt{0.0013}} = \frac{0.113}{0.036} = 3.14$$

This value exceeds 1.96, the critical value of z. We can therefore conclude that the sample of white breast cancer patients have a *statistically significantly* higher 5-year survival rate than the black breast cancer patients, and we therefore reject the null hypothesis that survival rates of black women with breast cancer are the same as that in white women.

Establishing a statistically significant difference alone is not the end of statistical hypothesis testing. As readers of the medical literature, you must also believe that the design of the study was carefully constructed so that doubts about bias do not creep into interpreting the test results. What might be confounding the results presented for white and black breast cancer survival rates? On average, are white and black women diagnosed at the same stage of disease? Is the white population younger than the black population? Also, are the two groups receiving similar treatment for the disease? Think about the effect of any of these prognostic factors such as stage, age at diagnosis, and treatment on 5-year survival rates. Would you expect patients with metastatic breast cancer to survive as long as patients with localized disease? If black women tend to be diagnosed with more progressive disease, couldn't that at least partially explain their lower survival rate? If the white women in the study tended to be younger than blacks, perhaps they have fewer additional health problems which could contribute to their higher survival rate. Finally, if the two groups are receiving different treatment for the same stage of disease, could that be affecting survival rates? Perhaps you can think of additional explanations for the discrepancy in survival rates. The point is that demonstrating a statistically significant difference alone does not constitute an adequate analysis of data. When you read the medical literature, you should look for possible explanations of the findings

in addition to use of the appropriate statistical test and presentation of a *P value*. P is the probability of rejecting the null hypothesis when it is actually true.

A value of z as large as that calculated (3.1) actually would result in an even greater significance level, meaning the the probability of the observed difference in proportions surviving being due to chance is only 1 percent or (0.01) rather than 5 percent. The distribution of possible values of the z statistic is presented in table form in many basic statistics textbooks and here in appendix 2C. One looks up the value of z (obtained in a test) in the table in order to find out if it is statistically significant.

Difference Between More Than Two Means--**Chi-Square Test**

The z test applies to situations when there are only two groups of interest, for example, black and white women, or drug A and drug B. You can probably imagine that there are many situations when there are more than two groups or outcomes of interest. For example, if you wanted to compare survival rates by stage of disease, you would have to look at the proportion surviving 5 years with, e.g., localized, regional and distant disease. In other words, you would have three groups at which to look. A statistical hypothesis test which can handle more than two samples is called the *chi-square test*. This test can also be used instead of the *z* test for the case where only two groups or outcomes are being compared. Instead of using proportions, the actual *counts* are employed. These are counts of *observed* numbers of individuals for a particular cell of a table (you will see an example shortly) and the *expected* numbers based on the null hypothesis. All of this will become clearer after looking at some examples. First though, you must be introduced to the test itself.

The chi-square test statistic (χ^2) is defined as:

$$\chi^2 = \textit{sum of } \frac{(\textit{observed - expected number of individuals in a cell})^2}{\textit{expected number of individuals in a cell}}$$

The actual computational formula is $\chi^2 = \sum \frac{(O - E)^2}{E}$

where O is the observed number (frequency) in a given cell and E is the expected number for that cell. The larger the differences in observed and expected frequencies, the larger will be the value of the calculated chi-square.

1. Application of Chi-Square Test for Two Groups

We will now apply the *chi-square test* to the problem of survival rates for black and white breast cancer patients. The data for this study are repeated in the table below (2 x 2 tables).

Table 47. Observed Frequencies of Black and White Womem with Breast Cancer Surviving for 5 Years

Survival Time	White	Black	Total
5 years or more	285	178	463
< 5 years	114	118	232
Total	399	296	695

The numbers 285, 114, 178 and 118 correspond to the four *cells* in the table. These numbers are the *observed* frequencies or counts, that is, the numbers found in the study data. But where do the *expected* frequencies come from? Recall that the *null hypothesis* for this study stated that there was no difference in the five-year survival rates between black and white women and that observed study differences are the result of chance.

Table 48. Expected Frequencies of 5-Year Survival in White and Black Women with Breast Cancer

Survival Time	White	Black	Total
5 years	265.76	197.24	463.0
< 5 years	133.24	98.76	232.0
Total	399	296	695

Thus, we can calculate the expected value of any cell in the table by multiplying appropriate row and column totals and dividing by the grand total. For example, if there is no difference between white and black women, we would expect the proportion of women surviving 5 years (463/695) to be the same in both white and blacks. Thus we can calculate the expected size of the white 5-year survivors to be (463/695 X 399 = 265.76) and of blacks to be (463/695 X 296 = 197.24). This is equivalent to $\frac{row\ total\ X\ column\ total}{grand\ total}$.

Note that, as in previous chapters, our "expected" counts do not have to be whole numbers.

Just by examining the observed and expected frequencies, you should be able to see that fewer black women survived for 5 years (178) than would be expected (197.24) if the survival rates were the same for the two groups of women. The reverse is seen for white women. You can formalize this observation by carrying out a *chi-square test.*

Chi-square test:

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(285 - 265.76)^2}{265.76} + \frac{(178 - 197.24)^2}{197.24} + \frac{(114 - 133.24)^2}{133.24} + \frac{(118 - 98.76)^2}{98.76}$$

$$= \frac{(19.24)^2}{265.76} + \frac{(-19.24)^2}{197.24} + \frac{(-19.24)^2}{133.24} + \frac{(19.24)^2}{98.76}$$

$$= \frac{370.18}{265.76} + \frac{370.18}{197.24} + \frac{370.18}{133.24} + \frac{370.18}{98.76}$$

$$= 1.39 + 1.88 + 2.78 + 3.75$$

$$= 9.80$$

To determine whether this value of the *chi-square* test statistic is large or small, we must look at how often *chi-square* values of a given size are exceeded when the null hypothesis is true. This is called the *chi-square distribution with 1 degree of freedom*, and is appropriate to use in the case of 2x2 data such as we have here (two races-black and white; two outcomes-survived 5 years, did not survive 5 years).

Just as we learned that there are tables of values of the t- and z-test statistics for looking up how "large" or "small" the value of the t- or z-statistic we obtained is, there is an analogous table of values for looking up how "large" or "small" our value of the *chi-square* statistic is. The **z** test statistic was only applicable when comparing two proportions. Since the *chi-square* test is applicable for more than two proportions, the distribution of the test statistic depends on the number of groups being compared (number of races = two) and the number of outcomes (survived 5 yrs., did not survive five years = two).

The number of comparisons made is reflected by a number which is called *degrees of freedom*. Since the table has r rows or outcomes and c columns, the degrees of freedom (v) is calculated by $v = (r-1)(c-1)$. For instance, for a 2x2 table, the degrees of freedom $v = (2-1)(2-1) = 1$.

When we look at the values of the chi-square statistic in appendix 2D, we find by reading across the row for one degree of freedom that our value of 9.80 exceeds that of the highest value in the row. This means that the probablility of observing such a large value is even less than 0.005 percent since that is the probability of observing a value of 7.88 or greater. In fact the critical value of *chi-square with one degree of freedom* only needs to be 3.84 for the null hypothesis of no difference in survival rates to be rejected at the 0.05 level since the probability of observing a value greater than 3.84 is five percent or less. Therefore, we conclude the data we observed in table 47 are unlikely to occur when the null hypothesis of no difference in the survival rate is true, and that there is a statistically significant difference in the survival rates for the two samples. We therefore conclude the difference in sample survival rates between blacks and whites with breast cancer is greater than zero.

The Yates Correction for Continuity

Statisticians have found that for 2x2 tables with one degree of freedom, the value of the χ^2 statistic leads to P-values that are smaller than they should be, resulting in a tendency to conclude that a difference exists when the data do not actually support this. This has to do with theoretical considerations which we need not concern ourselves with here. Instead, you should remember that when analyzing 2x2 tables ONLY, the following computational formula should be used to obtain the value of χ^2.

$$\chi^2 = \sum \frac{(|O - E| - .5)^2}{E}$$

Applying this new formula to the observed and expected counts of black and white 5-year survivors[1]

$$\chi^2 = \frac{(|285 - 265.76| - 0.5)^2}{265.76} + \frac{(|178 - 197.24| - 0.5)^2}{197.24} + \frac{(|114 - 133.24| - 0.5)^2}{133.24}$$

$$+ \frac{(|118 - 98.76| - 0.5)^2}{98.76}$$

$$= \frac{(18.74)^2}{265.76} + \frac{(18.74)^2}{197.24} + \frac{(18.74)^2}{133.24} + \frac{(18.74)^2}{98.76}$$

$$= 1.32 + 1.78 + 2.64 + 3.56 = 9.30$$

This value is smaller than 9.80, the value we found previously using the earlier formula for computing the χ^2 value. Of course, 9.30 is still much greater than the critical value of 3.84 required for us to reject the null hypothesis that there is no statistically significant difference. Therefore, again we conclude the difference in 5-year survival rates between the black and white women with breast cancer in the study is significant at $P < 0.05$.

---

[1]The absolute value sign | | means perform the indicated operation and make the result a positive number.

2. Application of Chi-Square Tests to Larger Tables

Now let's see how a chi-square test can be applied to analyze more than two treatments or outcomes. Suppose a researcher looked at survival outcomes for patients randomized to receive three different treatments. The data are presented in the table below (3 x 2 tables).

Table 49. Observed Frequencies of Patients Surviving 5 Years after Receiving Treatments A, B, and C

	Survival Outcome		
Treatment	≥ 5 years	< 5 years	Total
A	16	21	37
B	13	23	36
C	17	27	44
Total	46	71	117

The expected frequencies are calculated as follows:

Table 49 shows that out of a total of 117 patients, a total of 46 survived for 5 years. Therefore, the *total proportion of 5-year survivors* = 46/117 = 0.393. If there is no difference in survival rates between the three groups, the *same proportion of 5-year survivors should occur in each treatment group*! This total proportion of 5-year survivors is used to calculate the expected proportion for each treatment group. Since the total proportion *not surviving 5 years is 1-0.393 = 0.607*, this proportion is used to calculate the expected number *not surviving* 5 years for each treatment group. The expected frequencies are found in table 50.

Table 50. Expected Frequencies of Patients Surviving 5 Years After Receiving Treatments A, B, and C

	Survival Outcome		
Treatment	5 Years	< 5 Years	Total
A	37 x 0.393 = 14.54	37 x 0.607 = 22.46	37
B	36 x 0.393 = 14.15	36 x 0.607 = 21.85	36
C	44 x 0.393 = 17.29	44 x 0.607 = 26.71	44
Total	45.98	71.02	117

Note that the row and column totals are the same as they were for the observed table (except for rounding error). Since we are working with a 3 x 2 table (3 rows = treatment; 2 columns = survival

outcome) and not a 2x2 table, the **Yates continuity correction** does not apply here. The χ^2 test statistic is therefore computed as:

$$\chi^2 = \sum \frac{(O - E)^2}{E} = \frac{(16 - 14.54)^2}{14.54} + \frac{(21 - 22.46)^2}{22.46} + \frac{(13 - 14.15)^2}{14.15}$$

$$+ \frac{(23 - 21.85)^2}{21.85} + \frac{(17 - 17.29)^2}{17.29} + \frac{(27 - 26.71)^2}{26.71}$$

$$= 0.147 + 0.095 + 0.093 + 0.060 + 0.005 + 0.003 = 0.403$$

Since the table has three rows and two columns, the degrees of freedom for looking up our value of the χ^2 test statistic are $v = (r-1)(c-1) = (3-1)(2-1) = 2$. If you were to look up the χ^2 value 0.403 with two degrees of freedom in the table of values for the χ^2 distribution, you would find that our value of 0.403 is much smaller than the critical value of 5.991 required to reject the null hypothesis of no difference in survival between the different treatment groups. We therefore cannot reject the null hypothesis, and we conclude that there is no association between treatment and 5-year survival.

TYPE I AND TYPE II ERRORS - WHAT DO P VALUES REALLY MEAN?

What do P values really mean? By now you should be accustomed to thinking of a P value as the chance of obtaining a critical value of a test statistic, such as χ^2 or z, when the treatments have actually the same effect (no difference). The P value quantifies the probability or chance of mistakenly concluding that the treatment had an effect when only random variation (chance) is operating. Thus, when we obtain a P value of $P < 0.05$ and conclude that this means there is a statistically significant difference between treatments, we are in effect accepting that *1 in 20 times* our conclusions will be wrong when in reality the null hypothesis is actually true. This type of mistake is called a *type I error*. Concluding that a treatment *did not have an effect* when it actually did constitutes another kind of mistake, called a *type II error*. Type II errors commonly occur when studies involve just a few patients and, therefore, do not have the power to detect a difference because of small sample sizes, even when a treatment did have an effect.

Q5

Hypothesis testing requires establishing _____ _____ through a number of tests.

Q6

For the purpose of testing the null hypothesis that there is no difference between two sample means you apply the ___ _____.

Q7

For the purpose of testing the null hypothesis that there is no difference between two sample proportions you apply the ___ _____.

Q8

For the purpose of testing the null hypothesis that there is no difference between more than two means you apply the _____ _____ _____.

Answer: Q5

Hypothesis testing requires establishing confidence intervals through a number of tests.

Answer: Q6

For the purpose of testing the null hypothesis that there is no difference between two sample means you apply the t test.

Answer: Q7

For the purpose of testing the null hypothesis that there is no difference between two sample proportions you apply the z test.

Answer: Q8

For the purpose of testing the null hypothesis that there is no difference between more than two means you apply the chi square test.

APPENDIX 1

NOTATION, FORMULAE, AND MATHEMATICAL OPERATIONS USED IN STATISTICS

APPENDIX 1

NOTATION, FORMULAE, AND MATHEMATICAL OPERATIONS USED IN STATISTICS

I. Notation and Formulae

A. General symbols

X	A variable or the value of a variable. Other English letters may also be used such as Y.
$\sum$	Capital Greek letter sigma; carry out the process of addition or summation, e.g.: $\sum X = X_1 + X_2 + \bullet\bullet\bullet + X_n$
$\sqrt{X}$	Take the square root of X.
X^2	Square X (multiply X by X).
$\lvert X - Y \rvert$	Absolute value of the difference between two values X and Y = the difference between X and Y without regard to the sign. (The value of this expression is always positive.)
∞	Infinity
$<$	Less than, e.g., $X < Y$ means the value of X is less than the value of Y.
$\leq$	Less than or equal to
$>$	Greater than
$\geq$	Greater than or equal to
CI	Confidence interval
df	Degrees of freedom (n-1)
H_0	The null hypothesis = a particular hypothesis to be tested.
t	Probability of difference between two sample means
μ	Population mean
σ	Standard deviation of population mean
z	Value representing the number of standard deviations from the mean

B. General formulae

1. Ratio

a. A ratio of two numbers is the quotient obtained by dividing the first by the second, e.g., the ratio of 8 to 4 is 8/4 or 2; the ratio of 4 to 8 is 4/8 or 1/2; the ratio of a to b is a/b where a is any real number and b is any real number not equal to zero.

b. Formula:

$$ratio = \frac{number\ in\ a\ category}{number\ in\ another\ category}$$

2. Proportion

a. A proportion is a statement of the equality of two ratios, e.g., 2/4 = 1/2, 4/2 = 8/4, a/b = c/d.

b. Formula:

$$proportion = \frac{number\ in\ a\ category}{total\ number}$$

e.g., number of males/number in the population (male + female);
number of deaths/total population

3. Percent

A proportion multiplied by 100 e.g.,

$$\frac{100\ deaths}{100{,}000\ population} = \frac{0.001}{1.0}\ shows\ the\ ratio\ equality$$

$$\frac{100\ deaths}{100{,}000\ population} = 0.001\ usual\ reporting$$

$$0.001 \times 100 = 0.1\%\ expressed\ as\ a\ percent$$

C. Notation and formulae for central tendency and variation

$\bar{X}$ sample mean = $\sum X/n$, i.e.,

$$\frac{X_1 + X_2 + X_3 + X_n}{n} \quad \textit{estimates } \mu$$

range $X_{largest} - X_{smallest}$

S^2_x $\quad$ *sample variance* $= \sum\frac{(X-\bar{X})^2}{n-1} = \frac{\sum X^2 - (\sum X)^2/n}{n-1}$

estimates σ^2_x

S_x = s.d.(x) = *sample standard deviation* $= \sqrt{\textit{sample variance}} = \sqrt{S_x^2} = \sqrt{\frac{\sum (X-\bar{X})^2}{n-1}}$

estimates σ_x

n the number of values or observations in a sample.

(n-1) (n-1) is used in the denominator instead of the actual number of observations when we wish to measure the **variability** of observations from the mean. If we have three observations, ($\sum X = 15$), we only have two observations that actually have freedom to demonstrate the desired underlying variability. We do not have three because once we have the value of two observations and how far they are from the mean of 5 based on a total of 15, we can figure out what the third observation **must be** and how far it is from the mean. So, in general, if we know the values of (n-1) observations, the nth is predetermined and can add nothing to the information on variability. We are, therefore, left with what we call (n-1) degrees of freedom in our assessment of underlying variability.

χ^2 Chi square (capital Greek letter chi)

D. Notation and formulae for test for proportions

O Observed frequency in a cell of a contingency table

E Expected frequency in a cell of a contingency table =

$$= \frac{\textit{row marginal total x column marginal total}}{\textit{total in table}}$$

$\frac{\sum (O - E)^2}{E}$ used to test the equality of two or more proportions

For 2 x 2 contingency tables use: $X^2 = \frac{\Sigma(|O - E| - 1/2)^2}{E}$

or the computational form: $X^2 = \frac{\left(|ad - bc| - \frac{n}{2}\right)^2 n}{(a + c)(b + d)(a + b)(c + d)}$

where a, b, c, d, n are the
entries in the 2 X 2 table as shown below.

a	b	a + b
c	d	c + d
a + c	b + d	n

E. Notation for survival analysis

P_k k^{th} time interval cumulative survival rate (e.g., P_5 could be a 5-year cumulative survival rate)

$= p_1 \times p_2 \times \bullet \bullet \bullet p_k$ where p_i = proportion surviving the i^{th} time interval

II. Mathematical Operations[1]

A. Basic operations on numbers

1. Addition: 2 + 2 = 4

2. Subtraction: 4 - 2 = 2

   If a larger number is subtracted from a smaller number, the result is a negative number: 2 - 4 = -2

3. Multiplication: 2 x 2 = 4

   a. Other signs that also mean multiply: 2 • 2, 2(2), (2)(2)

[1]Adapted from materials presented by Marilyn C. Hurst, MS, CTR, at the National Tumor Registrars Annual Meeting, 1983.

b. Some special rules for multiplication:

1) Anything multiplied by 1 remains unchanged: $2 \times 1 = 2$

2) The results of multiplying a number by 0 is 0: $2 \times 0 = 0$

3) Multiplying a positive number by a positive number results in a positive number: $(2)(2) = 4$

4) Multiplying a negative number by a negative number results in a positive number: $(-2)(-2) = 4$

5) Multiplying a positive number by a negative number results in a negative number: $(2)(-2) = -4$

4. Division: $4/2 = 2$

   a. Other signs that also mean divide: $4/2$, $4 \div 2$

   b. Some special rules for division:

      1) Anything divided by 1 remains unchanged: $2/1 = 2$

      2) 0 divided by anything is 0: $0/2 = 0$

      3) Do not divide by 0, the result is infinity. $1/0 = \infty$

      4) A positive number divided by a positive number results in a positive number: $(4)/(2) = 2$

      5) A negative number divided by a negative number results in a positive number: $(-4)/(-2) = 2$

      6) Dividing a number by a number of opposite sign results in a negative number:
      $(-4)/2 = -2$, $(4)/(-2) = -2$

5. Exponentiation (raising to a power): $2^2 = 4$

   a. Squaring a number means raising to the power of 2. It means multiplying the number by itself, e.g., $2^2 = 2 \times 2 = 4$

   b. Special rules about squaring:

      1) $1^2 = 1$

      2) $0^2 = 0$

      3) Any number squared is a positive number: $(-2)^2 = 4$

6. Exponentiation (taking a root): $\sqrt{4} = 2$

   a. Taking the square root means to find the number which when multiplied by itself gives you the number inside the square root sign. For example, we know from above that 2 squared = 4; therefore the square root of 4 is 2.

      To obtain square roots, one can look them up in square root tables. Also, many inexpensive calculators will calculate square roots.

   b. Another sign that means take the square root: $(4)^{1/2}$

   c. Some special rules about square roots:

      1) Square roots may be either positive or negative.

      2) The square root of 1 = 1 or -1.

      3) The square root of 0 = 0.

7. Absolute value: $|2 - 4| = 2$, $|4 - 2| = 2$

   The absolute value sign, | |, means perform the indicated operation and make the result a positive number.

8. Order of operations

   Example: $75 - (2 \times 5^2) + 8^2/16 = ?$ How does one decide the order in which the indicated mathematical operations should be performed?

   a. First perform operations that are in parentheses.

   b. Next exponentiate (powers and roots) in any order.

   c. Next multiply and divide in any order.

   d. Finally, add and subtract in any order.

      For our example:

      1) First, working inside the parentheses (2×5^2), we know from our rules to solve 5^2 first, resulting in

         $75 - (2 \times 25) + 8^2/16$.

2) Continuing to work inside the parentheses, we solve 2 x 25, resulting in

   $75 - (50) + 8^2/16$.

3) Next we exponentiate resulting in: $75 - 50 + 64/16$.

4) Then we divide: $75 - 50 + 4$

5) Finally, we add and subtract: $25 + 4 = 29$.

A helpful phrase for remembering order of operations:

Please Excuse My Dear Aunt Sally.

P = parentheses
E = exponentiation
M = multiply
D = divide
A = add
S = subtract

B. Substance of Algebra

1. Constants, variables, and coefficients

   a. In algebra we use letters to stand for numbers. When we use letters to stand for numbers, we follow the same order of operations rules as when using integers or real numbers (zero an exception).

   b. A constant always has the same value, e.g., 2, 1/2, -3, 4^3

   c. A variable can have many values because it can change depending on the situation.

      e.g., what is the value of $5X^2 - 4XY + Y^2$? The value will depend on the values we assign to X and Y.

      If X=3 and Y=2, what is the value of the expression?

first substitute	$5(3^2) - 4(3)(2) + 2^2$
then raise to the power	$5(9) - 4(3)(2) + 4$
then multiply	$45 - 24 + 4$
then add and subtract	answer 25

      e.g., what is the value of $(a + b)(a - b)$ if a=0.2 and b=0.03?

first substitute	$(0.2 + 0.03)(0.2 - 0.03)$
then work inside parentheses	$(0.23)\quad(0.17)$
multiply	$(0.23)(0.17) = 0.0391$ answer

d. A coefficient is a constant written as a prefix to a variable. e.g., in the expression $4X + 2Y$, 4 is the coefficient of X and 2 is the coefficient of Y; 5 is the coefficient of X^2 in the preceding example. When you know the value assigned to the variables, you multiply that value by the coefficient.

e.g., if $X=2$, $4X = 4(2) = 8$

Note: If there is no coefficient prefixing a variable, it is understood to equal one (1), thus X means 1X.

2. Rules to be followed in simplifying algebraic expressions:

a. Only similar terms may be added, subtracted, multiplied or divided.

e.g., 3 chairs + 4 chairs = 7 chairs $3X + 4X = 7X$

83 chairs + 2 books cannot be simplified any further nor can $(3X + 2Y)$

simplify $3X + 4X + Y - 2X - 3Y$
answer: $(3X + 4X - 2X) + (Y - 3Y) = 5X - 2Y$

add $3X^2 + 2X - 2 + 6X - 4 + X^2$
answer: $4X^2 + 8X - 6$

b. Exponents in multiplication
Recall our earlier example 2^4 which is shorthand for $2 \cdot 2 \cdot 2 \cdot 2$ or $(2)(2)(2)(2)$ which is equal to 16.

For any value of X, $X^4 = X \cdot X \cdot X \cdot X$

$X^2 \cdot X^3$ is the same as $(X \cdot X)(X \cdot X \cdot X)$ is the same as $X \cdot X \cdot X \cdot X \cdot X = X^5$

The rule is: $X^m \cdot X^n = X^{m+n}$ where m and n are any exponents.

e.g., $a^2 \cdot a^3 \cdot a^8 = a^{13}$ $X \cdot X^3 = X^4$ (an exponent of one is understood if no exponent is included)

Note: Any number or algebraic expression (except zero), which has a zero exponent has a numerical value of one (1).

$2^0 = 1$ $X^0 = 1$ $3X^0 = 3$ $2X^0Y = 2Y$ $X^0Y^0 = 1$

c. Exponents in division

Evaluate $2^5/2^2$ This example can be rewritten as:

$$\frac{2^5}{2^2} = \frac{2\cdot2\cdot2\cdot2\cdot2}{2\cdot2} = \frac{32}{4} = 8 \quad \text{and } 8 = 2^3, \text{ therefore}$$

The rule is: $x^m / X^n = X^{m-n}$ where m is the exponent of the dividend and n is the exponent of the divisor.

e.g., $X^4/X^2 = X^{4-2} = X^2$ $X^3Y^7/X^0Y^2 = X^{3-0}Y^{7-2} = X^3Y^5$

3. Equations

a. An equation is a statement that two algebraic expressions are equal. We all agree that 1 + 3 = 4. This is an equation. But, 2 + 2 = 4; 5 - 1 = 4; (2)(2) = 4; 8/2 = 4; $2^6/2^4 = 4$ are also equations. Let us consider the following:

X + 3 = 4 How do we solve this equation algebraically, i.e., for X?

Rule: If we have two expressions which are equal, they will still be equal if we treat both sides the same way.

e.g., If we add 3 to both sides of X + 3 = 4, we still have two equal expressions. Consider:

1 + 3 = 4. Add 3 to both sides. 1 + 3 + 3 = 4 + 3

Do we still have an equation, i.e., both sides equal? YES, 7 = 7.

We want to know what X equals in X + 3 = 4. So let us subtract 3 from both sides.

	X + 3 - 3	= 4 - 3 now simplify
	X + 0	= 1
	X	= 1
To solve	Y + 2	= 4 for Y, subtract 2 from each side and simplify.
	Y + 2 - 2	= 4 - 2
	Y + 0	= 2
	Y	= 2
To solve	Z - 1	= 4 for Z, add 1 to each side and simplify.
	Z - 1 + 1	= 4 + 1
	Z + 0	= 5
	Z	= 5

b. Transposition: In an equation, we can move a term on the right of the equal sign to the left of the equal sign as long as we change the sign of the term and vice versa.

e.g., X + 3 = 4
3 is a positive number, change its sign to a minus sign and move it to the right side.

X = 4 - 3 What does X equal? X = 1

Y + 2 = 4
2 is a positive number, change its sign to a minus sign and move it to the right side.

Y = 4 - 2 What does Y equal? Y = 2

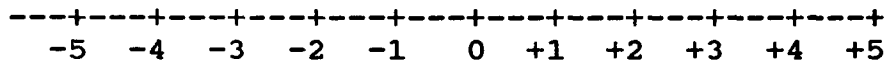
Z - 1 = 4
- 1 is a negative number, change its sign to a plus sign and move it to the right side.

Z = 4 + 1 What does Z equal? Z = 5
Transposing is only an apparent process. It is a shortcut approach to actually adding or subtracting the same value from both sides of an equation.

c. Solutions to equations can be checked for accuracy by substituting in the original equation.

4. Operations on Signed Numbers

a. The Number Line



b. The signs indicate the direction, right or left of zero, the units are to be counted. The value of a number without regard to its sign is called its absolute value. e.g., $10 earned (+$10) and $10 spent (-$10) both have the same absolute value.

If the class average on an exam were 82, someone whose grade is 87 is 5 points above (+5) the average, while someone whose grade is 80 is 2 points below (-2) the average.

c. Rules for performing operations on signed numbers:

1) Addition

To add two numbers having like signs, find the sum of their absolute values and prefix the common sign.

(+2) + (+5) = ?
the absolute values are 2 and 5, therefore 2 + 5 = 7
the common sign is +, therefore prefix a + before 7.
(+2) + (+5) = +7 answer

To add two numbers having unlike signs, find the difference of their absolute values and prefix the sign of the greater.

(+2) + (-5) = ?
the absolute values are 2 and 5 and their difference is 3
the sign of the larger absolute value (in this case the 5) is negative, therefore prefix a - before the 3.
(+2) + (-5) = -3 answer

2) Subtraction

Mentally change the sign of the subtrahend and proceed as in algebraic addition (above).

subtract: 2 from 5 No signs are indicated, therefore, positive signs are understood. 2 is the subtrahend. Change +2 to -2 and proceed as for addition.
5 - 2 = 3 or (+5) + (-2) = (+3) answers 3 or (+3)

subtract: (-3) from (+5) The subtrahend is (-3). Change the -3 to +3 and proceed as for addition.
(+5) + (+3) = 8 answer

3) Multiplication

The product of two numbers having like signs is positive.

(+2)(+3) = (+6) (-4)(-5) = (+20) or
(2)(3) = 6 (-4)(-5) = 20

The product of two numbers having unlike signs is negative.

(-2)(+3) = (-6) (+2)(-3) = (-6) (-2)(3) = (-6)

4) Division

The quotient of two numbers having like signs is positive.

$(+10)/(+5) = (+2)$ $(-10)/(-5) = (+2)$ $10/5 = 2$ $(-10)/(-5) = 2$

The quotient of two numbers having unlike signs is negative.

$(+10)/(-5) = (-2)$ $(-10)/(+5) = (-2)$

Example

$5X + 2 - X = 2X + 6$ first, transpose
$5X - X - 2X = 6 - 2$
$2X = 4$
$2X = 4$ divide both sides by (+2)

$$\frac{2X}{2} = \frac{4}{2}$$

$X = 2$ answer

What if we decided to transpose the variable to the right side?

$2 - 6 = 2X - 5X + X$
$- 4 = -2X$ divide both sides by (-2)

$$\frac{-4}{(-2)} = \frac{-2X}{(-2)}$$

Recall our rule for the division of like signed numbers.

$2 = X$ answer

As long as we follow our rules carefully, the solution may be arrived at in more than one way.

Example

$3X + 4 - X = 5X + 10$ first, transpose
$3X - X - 5X = 10 - 4$
$- 3X = 6$ divide both sides by (-3)

$$\frac{-3X}{(-3)} = \frac{6}{(-3)}$$

Recall our rule for the division of unlike signed numbers.

$X = -2$ answer

Checking these solutions in the original equations

$5X + 2 - X = 2X + 6$ answer $X = (+2)$
$5(+2) + 2 - (+2) = 2(+2) + 6$
$10 + 2 - (+2) = 4 + 6$
$10 + 0 = 10$
$10 = 10$ and our solution is correct

What if we made an error and thought our answer was $X = (+1)$?

$5X + 2 - X = 2X + 6$
$5(+1) + 2 - (+1) = 2(+1) + 6$
$5 + 2 - (+1) = 2 + 6$
$7 - 1 \neq 8$
$6 \neq 8$ and the solution $X = (+1)$ is wrong

IMPORTANT ALWAYS check in the original equation!

APPENDIX 2

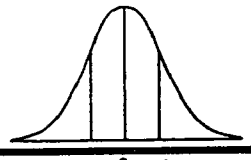
STATISTICAL TABLES

Appendix 2A
RANDOMLY ASSORTED DIGITS

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067
20	75884	12952	84318	95108	72305	64620	91318	89872	45375	85436
21	16777	37116	58550	42958	21460	43910	01175	87894	81378	10620
22	46230	43877	80207	88877	89380	32992	91380	03164	98656	59337
23	42902	66892	46134	01432	94710	23474	20423	60137	60609	13119
24	81007	00333	39693	28039	10154	95425	39220	19774	31782	49037
25	68089	01122	51111	72373	06902	74373	96199	97017	41273	21546
26	20411	67081	89950	16944	93054	87687	96693	87236	77054	33848
27	58212	13160	06468	15718	82627	76999	05999	58680	96739	63700
28	70577	42866	24969	61210	76046	67699	42054	12696	93758	03283
29	94522	74358	71659	62038	79643	79169	44741	05437	39038	13163
30	42626	86819	85651	88678	17401	03252	99547	32404	17918	62880
31	16051	33763	57194	16752	54450	19031	58580	47629	54132	60631
32	08244	27647	33851	44705	94211	46716	11738	55784	95374	72655
33	59497	04392	09419	89964	51211	04894	72882	17805	21896	83864
34	97155	13428	40293	09985	58434	01412	69124	82171	59058	82859
35	98409	66162	95763	47420	20792	61527	20441	39435	11859	41567
36	45476	84882	65109	96597	25930	66790	65706	61203	53634	22557
37	89300	69700	50741	30329	11658	23166	05400	66669	48708	03887
38	50051	95137	91631	66315	91428	12275	24816	68091	71710	33258
39	31753	85178	31310	89642	98364	02306	24617	09609	83942	22716
40	79152	53829	77250	20190	56535	18760	69942	77448	33278	48805
41	44560	38750	83635	56540	64900	42912	13953	79149	18710	68618
42	68328	83378	63369	71381	39564	05615	42451	64559	97501	65747
43	46939	38689	58625	08342	30459	85863	20781	09284	26333	91777
44	83544	86141	15707	96256	23068	13782	08467	89469	93842	55349
45	91621	00881	04900	54224	46177	55309	17852	27491	89415	23466
46	91896	67126	04151	03795	59077	11848	12630	98375	52068	60142
47	55751	62515	21108	80830	02263	29303	37204	96926	30506	09808
48	85156	87689	95493	88842	00664	55017	55539	17771	69448	87530
49	07521	56898	12236	60277	39102	62315	12239	07105	11844	01117

Enter the table in any row or column and continue either vertically or horizontally.

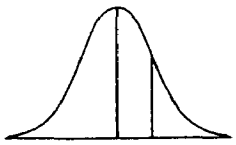
Appendix 2B
DISTRIBUTION OF t (TWO-TAILED TESTS)



Degrees of Freedom	Probability of a Larger Value, Sign Ignored								
	0.500	0.400	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	1.000	1.376	3.078	6.314	12.706	25.452	63.657		
2	0.816	1.061	1.886	2.920	4.303	6.205	9.925	14.089	31.598
3	.765	0.978	1.638	2.353	3.182	4.176	5.841	7.453	12.941
4	.741	.941	1.533	2.132	2.776	3.495	4.604	5.598	8.610
5	.727	.920	1.476	2.015	2.571	3.163	4.032	4.773	6.859
6	.718	.906	1.440	1.943	2.447	2.969	3.707	4.317	5.959
7	.711	.896	1.415	1.895	2.365	2.841	3.499	4.029	5.405
8	.706	.889	1.397	1.860	2.306	2.752	3.355	3.832	5.041
9	.703	.883	1.383	1.833	2.262	2.685	3.250	3.690	4.781
10	.700	.879	1.372	1.812	2.228	2.634	3.169	3.581	4.587
11	.697	.876	1.363	1.796	2.201	2.593	3.106	3.497	4.437
12	.695	.873	1.356	1.782	2.179	2.560	3.055	3.428	4.318
13	.694	.870	1.350	1.771	2.160	2.533	3.012	3.372	4.221
14	.692	.868	1.345	1.761	2.145	2.510	2.977	3.326	4.140
15	.691	.866	1.341	1.753	2.131	2.490	2.947	3.286	4.073
16	.690	.865	1.337	1.746	2.120	2.473	2.921	3.252	4.015
17	.689	.863	1.333	1.740	2.110	2.458	2.898	3.222	3.965
18	.688	.862	1.330	1.734	2.101	2.445	2.878	3.197	3.922
19	.688	.861	1.328	1.729	2.093	2.433	2.861	3.174	3.833
20	.687	.860	1.325	1.725	2.086	2.423	2.845	3.153	3.850
21	.686	.859	1.323	1.721	2.080	2.414	2.831	3.135	3.819
22	.686	.858	1.321	1.717	2.074	2.406	2.819	3.119	3.792
23	.685	.858	1.319	1.714	2.069	2.398	2.807	3.104	3.767
24	.685	.857	1.318	1.711	2.064	2.391	2.797	3.090	3.745
25	.684	.856	1.316	1.708	2.060	2.385	2.787	3.078	3.725
26	.684	.856	1.315	1.706	2.056	2.379	2.779	3.067	3.707
27	.684	.855	1.314	1.703	2.052	2.373	2.771	3.056	3.690
28	.683	.855	1.313	1.701	2.048	2.368	2.763	3.047	3.674
29	.683	.854	1.311	1.699	2.045	2.364	2.756	3.038	3.659
30	.683	.854	1.310	1.697	2.042	2.360	2.750	3.030	3.646
35	.682	.852	1.306	1.690	2.030	2.342	2.724	2.996	3.591
40	.681	.851	1.303	1.684	2.021	2.329	2.704	2.971	3.551
45	.680	.850	1.301	1.680	2.014	2.319	2.690	2.952	3.520
50	.680	.849	1.299	1.676	2.008	2.310	2.678	2.937	3.496
55	.679	.849	1.297	1.673	2.004	2.304	2.669	2.925	3.476
60	.679	.848	1.296	1.671	2.000	2.299	2.660	2.915	3.460
70	.678	.847	1.294	1.667	1.994	2.290	2.648	2.899	3.435
80	.678	.847	1.293	1.665	1.989	2.284	2.638	2.887	3.416
90	.678	.846	1.291	1.662	1.986	2.279	2.631	2.878	3.402
100	.677	.846	1.290	1.661	1.982	2.276	2.625	2.871	3.390
120	.677	.845	1.289	1.658	1.980	2.270	2.617	2.860	3.373.
∞	.6745	.8416	1.2816	1.6448	1.9600	2.2414	2.5758	2.807	3.2905

Reprinted by permission from Maxine Merrington's "Table of Percentage Points of the t-Distribution," *Biometriko* 32 (1942):300

Appendix 2C
CUMULATIVE NORMAL FREQUENCY DISTRIBUTION
(area under standard normal curve from 0 to z)



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3428	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000									

Appendix 2D

CUMULATIVE DISTRIBUTION OF CHI-SQUARE

Degrees of Freedom	Probability of a Larger Value												
	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
1	....	....	....	....	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.57	40.00

CUMULATIVE DISTRIBUTION OF CHI-SQUARE

Degrees of Freedom	Probability of a Larger Value												
	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	33.66	39.34	45.62	51.80	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	42.94	49.33	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	52.29	59.33	66.98	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	61.70	69.33	77.58	85.53	90.53	95.02	100.42	104.22
80	51.17	53.54	57.15	60.39	64.28	71.14	79.33	88.13	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	80.62	89.33	98.64	107.56	113.14	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	90.13	99.33	109.14	118.50	124.34	129.56	135.81	140.17

Condensed from table with 6 significant figures by Catherine M. Thompson, by permission of the editor of Biometrika.
Read the significance P from the normal table A3. Use only one tail.

For numbers of degrees of freedom greater than 100, calculate the approximate normal deviate $z = \sqrt{2x^2} - \sqrt{2(df) - 1}$.

APPENDIX 3

EXPECTED SURVIVAL RATE TABLES

Sources: National Cancer Institute, DCPC/SP/CST, EPN, Room 343J, Bethesda, MD 20892, (301) 496-8510.
National Center for Health Statistics, 6525 Belcrest Road, Hyattsville, MD 20782

HOW TO USE EXPECTED SURVIVAL RATE TABLES

The following tables contain the expected 1-year normal survival rates for whites, blacks, American Indians, Japanese, Chinese, Hawaiians, Filipinos, Hispanics, residents of Puerto Rico, and other races for 1970 and 1980. Separate tables are supplied for males and females, and for ages 0 years old to 118 years old. The expected life tables give the probability that a person of a certain age will live 1 more year. These tables are used to calculate the relative survival rate (see section D on survival for more details). The table closest to the calendar year of interest should be used, for example, for someone diagnosed in 1968, the 1970 life table should be used. If the patient survives until 1976, then the 1980 table should be used to calculate expected survival between 1975 and 1976.

To calculate the 1-year relative survival rate:

Look up the expected 1-year survival rate by age at diagnosis and year of diagnosis in the appropriate table for each patient in the study group.

Average the expected survival rates for your cases.

Divide the observed 1-year survival rate by the expected 1-year survival rate to get the 1-year relative survival rate.

To calculate 2-year, 3-year, ... etc. survival rates

For all cases, add 1 year to the age and 1 year to the date of diagnosis. Look up the new expected survival rate in the appropriate table.

Multiply the 1-year expected survival rate from the second year by the 1-year survival rate from the first year.

Average these multiplied rates for all your cases for each year.

Divide the observed survival rate by the average expected survival rate.

Repeat this process for the rest of the intervals.

In each case add another year to the age and year of diagnosis.

Multiply the expected normal survival from the previous years.

Average the results from these cases.

Divide the observed survival rate by the average expected survival rate.

EXPECTED 1-YEAR SURVIVAL RATES
WHITE MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.98407	.98769	30	.99837	.99834	60	.97970	.98238	90	.79944	.80942
1	.99901	.99908	31	.99837	.99835	61	.97770	.98067	91	.78424	.79611
2	.99926	.99934	32	.99834	.99834	62	.97569	.97881	92	.76886	.78136
3	.99941	.99947	33	.99826	.99831	63	.97372	.97684	93	.75350	.76547
4	.99949	.99957	34	.99817	.99825	64	.97174	.97477	94	.73826	.74939
5	.99954	.99961	35	.99804	.99816	65	.96968	.97262	95	.72334	.73383
6	.99957	.99963	36	.99790	.99804	66	.96742	.97032	96	.70891	.71999
7	.99959	.99966	37	.99773	.99791	67	.96486	.96782	97	.69512	.70689
8	.99963	.99970	38	.99754	.99776	68	.96188	.96505	98	.68186	.69455
9	.99968	.99976	39	.99731	.99760	69	.95848	.96195	99	.66940	.68297
10	.99973	.99981	40	.99706	.99739	70	.95482	.95852	100	.65776	.67216
11	.99973	.99981	41	.99677	.99713	71	.95093	.95484	101	.64691	.66209
12	.99965	.99972	42	.99643	.99684	72	.94672	.95099	102	.63683	.65276
13	.99948	.99954	43	.99604	.99652	73	.94216	.94705	103	.62750	.64412
14	.99923	.99929	44	.99559	.99618	74	.93724	.94297	104	.61889	.63616
15	.99896	.99904	45	.99509	.99580	75	.93192	.93854	105	.61096	.62883
16	.99870	.99882	46	.99457	.99537	76	.92622	.93358	106	.60368	.62210
17	.99848	.99863	47	.99399	.99486	77	.92015	.92820	107	.59700	.61593
18	.99835	.99849	48	.99337	.99427	78	.91371	.92238	108	.59089	.61029
19	.99825	.99837	49	.99273	.99361	79	.90691	.91606	109	.58531	.60514
20	.99818	.99825	50	.99201	.99294	80	.89988	.90901	110	.58021	.60045
21	.99809	.99814	51	.99122	.99225	81	.89270	.90114	111	.57557	.59617
22	.99805	.99807	52	.99037	.99150	82	.88569	.89267	112	.57135	.59228
23	.99807	.99807	53	.98945	.99066	83	.87930	.88387	113	.56751	.58874
24	.99813	.99811	54	.98844	.98973	84	.87434	.87477	114	.56403	.58553
25	.99821	.99817	55	.98739	.98875	85	.86637	.86493	115	.56086	.58262
26	.99828	.99823	56	.98625	.98773	86	.85463	.85408	116	.55800	.57998
27	.99834	.99828	57	.98491	.98662	87	.84201	.84309	117	.55540	.57760
28	.99837	.99832	58	.98337	.98536	88	.82847	.83226	118	.55305	.57543
29	.99839	.99833	59	.98161	.98395	89	.81422	.82125			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
WHITE FEMALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.98770	.99035	30	.99926	.99935	60	.99023	.99111	90	.83514	.85169
1	.99925	.99923	31	.99922	.99932	61	.98934	.99025	91	.81991	.83769
2	.99939	.99949	32	.99916	.99928	62	.98849	.98933	92	.80402	.82291
3	.99950	.99963	33	.99910	.99923	63	.98770	.98838	93	.78778	.80802
4	.99959	.99970	34	.99904	.99917	64	.98693	.98741	94	.77162	.79310
5	.99965	.99972	35	.99898	.99910	65	.98611	.98641	95	.75598	.77772
6	.99970	.99974	36	.99890	.99901	66	.98511	.98530	96	.74116	.76271
7	.99973	.99977	37	.99879	.99891	67	.98380	.98405	97	.72729	.74827
8	.99976	.99979	38	.99866	.99881	68	.98212	.98260	98	.71434	.73449
9	.99979	.99982	39	.99851	.99870	69	.98006	.98093	99	.70218	.72141
10	.99980	.99983	40	.99833	.99857	70	.97783	.97908	100	.69076	.70906
11	.99981	.99984	41	.99815	.99842	71	.97541	.97706	101	.68010	.69745
12	.99978	.99981	42	.99795	.99826	72	.97264	.97483	102	.67017	.68658
13	.99974	.99975	43	.99775	.99808	73	.96941	.97240	103	.66095	.67645
14	.99966	.99968	44	.99753	.99789	74	.96576	.96973	104	.65243	.66703
15	.99958	.99960	45	.99730	.99769	75	.96174	.96685	105	.64456	.65832
16	.99950	.99953	46	.99704	.99746	76	.95744	.96363	106	.63732	.65027
17	.99945	.99948	47	.99675	.99720	77	.95285	.95985	107	.63068	.64285
18	.99942	.99946	48	.99647	.99690	78	.94800	.95533	108	.62458	.63603
19	.99942	.99945	49	.99615	.99657	79	.94281	.95005	109	.61901	.62978
20	.99941	.99944	50	.99581	.99624	80	.93723	.94411	110	.61392	.62406
21	.99941	.99943	51	.99545	.99590	81	.93113	.93761	111	.60928	.61883
22	.99940	.99943	52	.99505	.99553	82	.92441	.93051	112	.60505	.61406
23	.99940	.99942	53	.99461	.99512	83	.91689	.92287	113	.60120	.60971
24	.99939	.99942	54	.99413	.99468	84	.90835	.91461	114	.59771	.60577
25	.99939	.99942	55	.99362	.99421	85	.89852	.90537	115	.59453	.60217
26	.99939	.99942	56	.99308	.99372	86	.88739	.89509	116	.59165	.59889
27	.99936	.99941	57	.99246	.99319	87	.87558	.88466	117	.58904	.59593
28	.99934	.99940	58	.99179	.99258	88	.86299	.87441	118	.58668	.5932
29	.99931	.99937	59	.99106	.99189	89	.84952	.86383			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
BLACK MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.97371	.97703	30	.99548	.99592	60	.97199	.97123	90	.83088	.83239
1	.99865	.99852	31	.99544	.99576	61	.97018	.96942	91	.82325	.82383
2	.99888	.99890	32	.99534	.99559	62	.96826	.96748	92	.81612	.81352
3	.99905	.99914	33	.99510	.99540	63	.96622	.96548	93	.80945	.80112
4	.99919	.99930	34	.99479	.99517	64	.96405	.96349	94	.80323	.78764
5	.99930	.99937	35	.99444	.99491	65	.96174	.96154	95	.79744	.77446
6	.99938	.99945	36	.99408	.99461	66	.95930	.95956	96	.79209	.76726
7	.99945	.99951	37	.99370	.99428	67	.95670	.95740	97	.78713	.76056
8	.99951	.99957	38	.99334	.99391	68	.95395	.95489	98	.78257	.75437
9	.99956	.99963	39	.99297	.99352	69	.95103	.95196	99	.77838	.74865
10	.99958	.99967	40	.99256	.99309	70	.94794	.94859	100	.77453	.74338
11	.99957	.99966	41	.99211	.99261	71	.94466	.94499	101	.77101	.73854
12	.99948	.99959	42	.99162	.99206	72	.94119	.94134	102	.76779	.73410
13	.99935	.99943	43	.99110	.99143	73	.93752	.93798	103	.76486	.73004
14	.99914	.99922	44	.99052	.99071	74	.93363	.93492	104	.76218	.72633
15	.99892	.99901	45	.98992	.98993	75	.92952	.93186	105	.75974	.72294
16	.99868	.99880	46	.98926	.98910	76	.92518	.92846	106	.75752	.71986
17	.99839	.99858	47	.98849	.98819	77	.92059	.92463	107	.75551	.71705
18	.99804	.99835	48	.98760	.98720	78	.91574	.92001	108	.75368	.71450
19	.99765	.99809	49	.98662	.98616	79	.91063	.91434	109	.75203	.71218
20	.99723	.99779	50	.98511	.98512	80	.90523	.90732	110	.75053	.71007
21	.99683	.99749	51	.98413	.98406	81	.89955	.89913	111	.74917	.70817
22	.99650	.99721	52	.98309	.98291	82	.89357	.89047	112	.74794	.70645
23	.99624	.99700	53	.98199	.98165	83	.88728	.88286	113	.74683	.70489
24	.99607	.99685	54	.98081	.98028	84	.88067	.87698	114	.74582	.70347
25	.99589	.99670	55	.97956	.97884	85	.87303	.87128	115	.74491	.70219
26	.99572	.99654	56	.97822	.97738	86	.86476	.86441	116	.74409	.70104
27	.99559	.99638	57	.97681	.97592	87	.85616	.85718	117	.74335	.69999
28	.99553	.99623	58	.97530	.97444	88	.84747	.84929	118	.74268	.69904
29	.99550	.99608	59	.97370	.97289	89	.83898	.84072			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
BLACK FEMALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.97778	.98073	30	.99840	.99852	60	.98361	.98423	90	.85903	.87344
1	.99880	.99873	31	.99832	.99843	61	.98236	.98305	91	.85060	.86381
2	.99902	.99913	32	.99822	.99832	62	.98102	.98183	92	.84294	.85328
3	.99920	.99934	33	.99804	.99820	63	.97957	.98064	93	.83637	.84184
4	.99936	.99952	34	.99784	.99806	64	.97802	.97950	94	.83020	.82973
5	.99949	.99956	35	.99760	.99789	65	.97635	.97842	95	.82445	.81721
6	.99958	.99963	36	.99736	.99769	66	.97456	.97728	96	.81909	.80830
7	.99965	.99969	37	.99711	.99748	67	.97264	.97592	97	.81413	.79978
8	.99970	.99973	38	.99687	.99725	68	.97058	.97413	98	.80954	.79175
9	.99973	.99975	39	.99661	.99702	69	.96837	.97190	99	.80530	.78423
10	.99974	.99976	40	.99635	.99676	70	.96599	.96928	100	.80140	.77721
11	.99973	.99976	41	.99605	.99648	71	.96345	.96646	101	.79783	.77070
12	.99970	.99973	42	.99570	.99615	72	.96072	.96361	102	.79455	.76466
13	.99965	.99969	43	.99532	.99579	73	.95780	.96101	103	.79155	.75909
14	.99959	.99963	44	.99487	.99538	74	.95467	.95868	104	.78882	.75395
15	.99952	.99957	45	.99438	.99495	75	.95131	.95640	105	.78632	.74923
16	.99943	.99951	46	.99388	.99448	76	.94773	.95385	106	.78405	.74490
17	.99934	.99944	47	.99338	.99398	77	.94390	.95091	107	.78198	.74093
18	.99926	.99938	48	.99295	.99345	78	.93980	.94718	108	.78010	.73731
19	.99916	.99932	49	.99253	.99290	79	.93543	.94246	109	.77840	.73400
20	.99906	.99926	50	.99218	.99235	80	.93076	.93650	110	.77685	.73099
21	.99896	.99919	51	.99157	.99179	81	.92578	.92959	111	.77545	.72824
22	.99887	.99912	52	.99092	.99118	82	.92047	.92249	112	.77418	.72574
23	.99880	.99905	53	.99023	.99050	83	.91482	.91665	113	.77303	.72347
24	.99875	.99898	54	.98948	.98974	84	.90880	.91256	114	.77199	.72142
25	.99870	.99891	55	.98867	.98893	85	.90204	.90894	115	.77105	.71955
26	.99864	.99882	56	.98780	.98808	86	.89391	.90409	116	.77020	.71786
27	.99857	.99874	57	.98686	.98720	87	.88546	.89832	117	.76944	.71634
28	.99852	.99867	58	.98586	.98628	88	.87676	.89114	118	.76874	.71496
29	.99846	.99860	59	.98478	.98530	89	.86787	.88262			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
AMERICAN INDIAN, ALEUTIAN AND ESKIMO MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.96786	.99592	30	.99404	.99619	60	.97916	.98369	90	.84227	.89892
1	.99727	.99736	31	.99385	.99615	61	.97816	.98261	91	.82978	.89458
2	.99793	.99843	32	.99362	.99606	62	.97708	.98147	92	.81775	.89025
3	.99842	.99906	33	.99335	.99592	63	.97590	.98026	93	.80656	.88596
4	.99879	.99936	34	.99305	.99576	64	.97461	.97897	94	.79653	.88194
5	.99905	.99947	35	.99274	.99561	65	.97317	.97759	95	.78790	.87820
6	.99923	.99952	36	.99242	.99546	66	.97134	.97613	96	.78024	.87472
7	.99936	.99957	37	.99209	.99529	67	.96928	.97457	97	.77305	.87151
8	.99943	.99962	38	.99177	.99508	68	.96709	.97290	98	.76636	.86854
9	.99946	.99964	39	.99146	.99486	69	.96477	.97113	99	.76016	.86580
10	.99943	.99961	40	.99114	.99460	70	.96234	.96923	100	.75444	.86329
11	.99934	.99957	41	.99082	.99431	71	.95985	.96722	101	.74916	.86099
12	.99916	.99953	42	.99049	.99400	72	.95733	.96509	102	.74431	.85888
13	.99886	.99942	43	.99013	.99368	73	.95482	.96281	103	.73986	.85696
14	.99839	.99919	44	.98974	.99337	74	.95232	.96040	104	.73579	.85521
15	.99775	.99880	45	.98931	.99306	75	.94979	.95784	105	.73206	.85361
16	.99697	.99827	46	.98882	.99275	76	.94718	.95513	106	.72866	.85215
17	.99613	.99769	47	.98827	.99244	77	.94439	.95227	107	.72557	.85083
18	.99533	.99713	48	.98766	.99211	78	.94132	.94929	108	.72275	.84963
19	.99468	.99667	49	.98700	.99174	79	.93786	.94619	109	.72018	.84854
20	.99421	.99634	50	.98631	.99131	80	.93393	.94220	110	.71785	.84755
21	.99395	.99613	51	.98561	.99082	81	.92946	.93812	111	.71574	.84666
22	.99385	.99601	52	.98493	.99027	82	.92436	.93393	112	.71382	.84585
23	.99387	.99596	53	.98427	.98965	83	.91847	.92963	113	.71209	.84512
24	.99396	.99597	54	.98364	.98896	84	.91161	.92528	114	.71052	.84445
25	.99408	.99600	55	.98301	.98821	85	.89922	.92088	115	.70909	.84386
26	.99419	.99603	56	.98237	.98740	86	.88951	.91647	116	.70781	.84331
27	.99425	.99607	57	.98168	.98655	87	.87875	.91206	117	.70664	.84283
28	.99424	.99612	58	.98092	.98565	88	.86710	.90766	118	.70559	.84238
29	.99417	.99617	59	.98009	.98470	89	.85483	.90328			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
AMERICAN INDIAN, ALEUTIAN AND ESKIMO FEMALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.97459	.99627	30	.99676	.99832	60	.98869	.99152	90	.86839	.92004
1	.99770	.99761	31	.99662	.99826	61	.98817	.99082	91	.85645	.91493
2	.99823	.99860	32	.99644	.99820	62	.98761	.99008	92	.84505	.90977
3	.99863	.99919	33	.99623	.99814	63	.98702	.98930	93	.83455	.90458
4	.99893	.99948	34	.99599	.99806	64	.98637	.98848	94	.82526	.89938
5	.99916	.99962	35	.99572	.99796	65	.98565	.98762	95	.81730	.89417
6	.99931	.99969	36	.99545	.99785	66	.98508	.98672	96	.81067	.88895
7	.99942	.99974	37	.99517	.99771	67	.98393	.98577	97	.80485	.88372
8	.99949	.99974	38	.99491	.99757	68	.98269	.98475	98	.79941	.87848
9	.99953	.99976	39	.99466	.99741	69	.98132	.98364	99	.79438	.87325
10	.99954	.99980	40	.99443	.99722	70	.97982	.98241	100	.78973	.86801
11	.99952	.99977	41	.99421	.99703	71	.97815	.98104	101	.78546	.86277
12	.99948	.99975	42	.99400	.99686	72	.97631	.97951	102	.78154	.85787
13	.99940	.99974	43	.99380	.99670	73	.97428	.97784	103	.77794	.85331
14	.99930	.99964	44	.99360	.99651	74	.97206	.97600	104	.77465	.84906
15	.99916	.99949	45	.99341	.99631	75	.96964	.97400	105	.77164	.84514
16	.99899	.99933	46	.99321	.99605	76	.96701	.97184	106	.76890	.84151
17	.99880	.99920	47	.99302	.99578	77	.96417	.96950	107	.76641	.83817
18	.99859	.99907	48	.99283	.99550	78	.96110	.96701	108	.76413	.83510
19	.99837	.99894	49	.99264	.99525	79	.95778	.96437	109	.76207	.83228
20	.99815	.99886	50	.99243	.99502	80	.95415	.96158	110	.76020	.82970
21	.99794	.99882	51	.99219	.99482	81	.95010	.95880	111	.75850	.82735
22	.99776	.99880	52	.99193	.99464	82	.94549	.95533	112	.75696	.82520
23	.99759	.99875	53	.99163	.99445	83	.94014	.95164	113	.75556	.82324
24	.99745	.99868	54	.99129	.99424	84	.93388	.94775	114	.75430	.82146
25	.99732	.99862	55	.99092	.99398	85	.92244	.94362	115	.75316	.81984
26	.99721	.99859	56	.99052	.99366	86	.91337	.93927	116	.75212	.81836
27	.99711	.99855	57	.99010	.99325	87	.90322	.93472	117	.75119	.81703
28	.99700	.99846	58	.98966	.99275	88	.89215	.92997	118	.75035	.81582
29	.99689	.99838	59	.98919	.99218	89	.88042	.92507			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
JAPANESE MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.98744	.99889	30	.99939	.99926	60	.99086	.99263	90	.82033	.86344
1	.99957	.99915	31	.99939	.99926	61	.98975	.99178	91	.79795	.85389
2	.99963	.99939	32	.99937	.99927	62	.98851	.99080	92	.77500	.84415
3	.99968	.99957	33	.99934	.99929	63	.98714	.98965	93	.75222	.83428
4	.99972	.99971	34	.99928	.99930	64	.98569	.98835	94	.73039	.82431
5	.99975	.99980	35	.99921	.99929	65	.98422	.98690	95	.71018	.81427
6	.99976	.99985	36	.99912	.99926	66	.98292	.98527	96	.69195	.80417
7	.99977	.99988	37	.99902	.99921	67	.98180	.98349	97	.67563	.79404
8	.99977	.99989	38	.99891	.99915	68	.98061	.98155	98	.66093	.78389
9	.99976	.99988	39	.99879	.99905	69	.97931	.97945	99	.64733	.77373
10	.99974	.99984	40	.99864	.99893	70	.97787	.97718	100	.63457	.76357
11	.99972	.99981	41	.99847	.99878	71	.97625	.97475	101	.62267	.75341
12	.99969	.99978	42	.99827	.99862	72	.97437	.97216	102	.61161	.74322
13	.99965	.99975	43	.99802	.99844	73	.97219	.96942	103	.60137	.73300
14	.99960	.99970	44	.99774	.99823	74	.96965	.96654	104	.59190	.72344
15	.99954	.99965	45	.99744	.99801	75	.96673	.96307	105	.58318	.71453
16	.99946	.99957	46	.99714	.99778	76	.96345	.95935	106	.57517	.70625
17	.99938	.99948	47	.99685	.99754	77	.95986	.95541	107	.56782	.69859
18	.99929	.99938	48	.99660	.99728	78	.95602	.95126	108	.56108	.69151
19	.99919	.99931	49	.99637	.99701	79	.95197	.94689	109	.55493	.68498
20	.99911	.99925	50	.99617	.99673	80	.94766	.94190	110	.54932	.67899
21	.99904	.99921	51	.99595	.99646	81	.94287	.93587	111	.54420	.67349
22	.99902	.99918	52	.99571	.99619	82	.93733	.92938	112	.53954	.66846
23	.99903	.99916	53	.99542	.99591	83	.93069	.92244	113	.53531	.66386
24	.99907	.99915	54	.99506	.99562	84	.92259	.91507	114	.53146	.65966
25	.99914	.99915	55	.99461	.99530	85	.90693	.90730	115	.52797	.65584
26	.99922	.99917	56	.99407	.99493	86	.89384	.89915	116	.52480	.65236
27	.99929	.99920	57	.99343	.99448	87	.87852	.89066	117	.52194	.64920
28	.99934	.99923	58	.99269	.99396	88	.86102	.88186	118	.51934	.64632
29	.99938	.99925	59	.99183	.99335	89	.84151	.87278			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
JAPANESE FEMALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.99297	.99912	30	.99963	.99967	60	.99548	.99597	90	.84909	.89853
1	.99953	.99934	31	.99959	.99963	61	.99510	.99562	91	.82849	.89024
2	.99961	.99954	32	.99952	.99957	62	.99464	.99520	92	.80720	.88179
3	.99968	.99969	33	.99944	.99951	63	.99410	.99470	93	.78589	.87321
4	.99974	.99979	34	.99933	.99945	64	.99349	.99409	94	.76529	.86454
5	.99978	.99985	35	.99922	.99940	65	.99279	.99337	95	.74602	.85581
6	.99981	.99988	36	.99911	.99936	66	.99197	.99252	96	.72846	.84703
7	.99983	.99990	37	.99900	.99934	67	.99102	.99152	97	.71262	.83823
8	.99984	.99992	38	.99892	.99933	68	.98996	.99036	98	.69823	.82941
9	.99985	.99992	39	.99885	.99931	69	.98878	.98904	99	.68489	.82059
10	.99986	.99991	40	.99879	.99927	70	.98748	.98756	100	.67202	.81177
11	.99986	.99989	41	.99874	.99921	71	.98603	.98590	101	.65967	.80295
12	.99985	.99987	42	.99868	.99913	72	.98445	.98407	102	.64817	.79413
13	.99984	.99985	43	.99860	.99901	73	.98272	.98207	103	.63746	.78532
14	.99983	.99982	44	.99850	.99888	74	.98084	.97989	104	.62754	.77651
15	.99981	.99981	45	.99838	.99874	75	.97879	.97754	105	.61836	.76771
16	.99978	.99978	46	.99823	.99858	76	.97656	.97503	106	.60990	.75948
17	.99975	.99977	47	.99806	.99842	77	.97409	.97235	107	.60212	.75181
18	.99972	.99974	48	.99786	.99825	78	.97132	.96952	108	.59497	.74469
19	.99969	.99972	49	.99764	.99808	79	.96817	.96656	109	.58842	.73809
20	.99965	.99969	50	.99742	.99792	80	.96450	.96346	110	.58244	.73201
21	.99963	.99967	51	.99720	.99776	81	.96022	.95928	111	.57697	.72640
22	.99961	.99965	52	.99699	.99760	82	.95514	.95422	112	.57198	.72125
23	.99961	.99965	53	.99682	.99743	83	.94905	.94869	113	.56744	.71652
24	.99962	.99966	54	.99667	.99725	84	.94166	.94269	114	.56332	.71220
25	.99964	.99966	55	.99653	.99708	85	.92750	.93623	115	.55956	.70825
26	.99965	.99966	56	.99639	.99690	86	.91571	.92934	116	.55616	.70464
27	.99966	.99966	57	.99624	.99671	87	.90192	.92208	117	.55307	.70136
28	.99966	.99967	58	.99604	.99650	88	.88613	.91449	118	.55027	.69837
29	.99965	.99968	59	.99579	.99626	89	.86843	.90663			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
CHINESE MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.99019	.99882	30	.99945	.99933	60	.98342	.99000	90	.78591	.85178
1	.99977	.99915	31	.99939	.99931	61	.98132	.98886	91	.76939	.84359
2	.99978	.99944	32	.99932	.99927	62	.97905	.98758	92	.75275	.83533
3	.99978	.99965	33	.99922	.99924	63	.97666	.98614	93	.73629	.82700
4	.99978	.99980	34	.99911	.99920	64	.97418	.98454	94	.72030	.81863
5	.99977	.99989	35	.99899	.99915	65	.97161	.98276	95	.70496	.81021
6	.99976	.99993	36	.99886	.99909	66	.96875	.98079	96	.69036	.80177
7	.99975	.99991	37	.99872	.99901	67	.96621	.97863	97	.67649	.79332
8	.99974	.99988	38	.99858	.99891	68	.96346	.97628	98	.66319	.78486
9	.99973	.99987	39	.99843	.99882	69	.96051	.97375	99	.65057	.77641
10	.99972	.99990	40	.99828	.99871	70	.95734	.97104	100	.63880	.76849
11	.99970	.99993	41	.99812	.99859	71	.95396	.96816	101	.62786	.76112
12	.99968	.99995	42	.99793	.99847	72	.95033	.96523	102	.61771	.75428
13	.99965	.99995	43	.99773	.99834	73	.94645	.96159	103	.60832	.74794
14	.99960	.99989	44	.99750	.99818	74	.94229	.95774	104	.59967	.74209
15	.99955	.99979	45	.99723	.99798	75	.93781	.95366	105	.59171	.73670
16	.99949	.99968	46	.99691	.99775	76	.93298	.94938	106	.58440	.73175
17	.99943	.99956	47	.99654	.99749	77	.92775	.94435	107	.57771	.72721
18	.99936	.99948	48	.99610	.99719	78	.92208	.93875	108	.57159	.72305
19	.99931	.99943	49	.99557	.99685	79	.91590	.93279	109	.56600	.71925
20	.99927	.99940	50	.99496	.99649	80	.90915	.92647	110	.56090	.71579
21	.99926	.99939	51	.99427	.99609	81	.90176	.91985	111	.55626	.71264
22	.99928	.99940	52	.99350	.99565	82	.89364	.91297	112	.55204	.70976
23	.99931	.99942	53	.99267	.99517	83	.88472	.90587	113	.54821	.70715
24	.99937	.99943	54	.99178	.99464	84	.87486	.89858	114	.54473	.70478
25	.99943	.99943	55	.99080	.99405	85	.85814	.89111	115	.54157	.70263
26	.99947	.99943	56	.98971	.99340	86	.84563	.88350	116	.53871	.70068
27	.99950	.99942	57	.98845	.99269	87	.83205	.87574	117	.53612	.69892
28	.99951	.99941	58	.98700	.99190	88	.81747	.86787	118	.53377	.69732
29	.99949	.99937	59	.98532	.99101	89	.80203	.85988			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
CHINESE FEMALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.99339	.99916	30	.99947	.99967	60	.99319	.99504	90	.85294	.90578
1	.99974	.99941	31	.99941	.99964	61	.99247	.99453	91	.83264	.89981
2	.99979	.99962	32	.99934	.99958	62	.99168	.99393	92	.81128	.89378
3	.99982	.99975	33	.99927	.99950	63	.99084	.99324	93	.78955	.88771
4	.99985	.99983	34	.99920	.99943	64	.98994	.99245	94	.76826	.88161
5	.99988	.99987	35	.99915	.99938	65	.98896	.99154	95	.74821	.87548
6	.99990	.99989	36	.99910	.99935	66	.98771	.99051	96	.72988	.86934
7	.99991	.99988	37	.99905	.99933	67	.98645	.98935	97	.71337	.86318
8	.99992	.99988	38	.99901	.99930	68	.98508	.98806	98	.69845	.85701
9	.99993	.99988	39	.99898	.99924	69	.98358	.98662	99	.68470	.85084
10	.99993	.99989	40	.99893	.99913	70	.98198	.98503	100	.67157	.84467
11	.99993	.99989	41	.99888	.99901	71	.98030	.98329	101	.65841	.83852
12	.99992	.99986	42	.99881	.99890	72	.97856	.98141	102	.64556	.83277
13	.99990	.99983	43	.99873	.99879	73	.97676	.97940	103	.63363	.82740
14	.99988	.99982	44	.99861	.99870	74	.97493	.97725	104	.62252	.82242
15	.99984	.99983	45	.99847	.99863	75	.97301	.97492	105	.61223	.81781
16	.99980	.99984	46	.99829	.99856	76	.97098	.97216	106	.60271	.81355
17	.99975	.99985	47	.99807	.99848	77	.96876	.96924	107	.59393	.80963
18	.99971	.99983	48	.99782	.99837	78	.96626	.96615	108	.58585	.80602
19	.99968	.99980	49	.99754	.99823	79	.96339	.96281	109	.57844	.80271
20	.99966	.99977	50	.99724	.99806	80	.96007	.95878	110	.57165	.79969
21	.99966	.99976	51	.99692	.99785	81	.95620	.95442	111	.56544	.79692
22	.99965	.99974	52	.99659	.99763	82	.95163	.94977	112	.55977	.79440
23	.99966	.99972	53	.99627	.99740	83	.94616	.94487	113	.55461	.79210
24	.99966	.99969	54	.99596	.99716	84	.93953	.93973	114	.54990	.79001
25	.99965	.99967	55	.99563	.99689	85	.92673	.93440	115	.54562	.78811
26	.99964	.99967	56	.99527	.99659	86	.91597	.92890	116	.54173	.78638
27	.99962	.99966	57	.99487	.99626	87	.90325	.92326	117	.53820	.78481
28	.99958	.99966	58	.99439	.99590	88	.88847	.91752	118	.53500	.78339
29	.99953	.99968	59	.99383	.99550	89	.87165	.91169			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
HAWAIIAN MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.99430	.99797	30	.99749	.99802	60	.96972	.97340	90	.72114	.89481
1	.99688	.99832	31	.99751	.99799	61	.96649	.97121	91	.71347	.88885
2	.99822	.99866	32	.99748	.99795	62	.96311	.96897	92	.70637	.88236
3	.99892	.99896	33	.99742	.99789	63	.95971	.96668	93	.69982	.87553
4	.99930	.99922	34	.99731	.99781	64	.95636	.96435	94	.69380	.86845
5	.99952	.99944	35	.99713	.99769	65	.95306	.96196	95	.68827	.86122
6	.99965	.99960	36	.99687	.99754	66	.94969	.95952	96	.68320	.85387
7	.99973	.99971	37	.99650	.99733	67	.94609	.95701	97	.67857	.84646
8	.99978	.99978	38	.99601	.99708	68	.94209	.95445	98	.67434	.83901
9	.99981	.99981	39	.99548	.99679	69	.93759	.95135	99	.67048	.83156
10	.99981	.99983	40	.99480	.99645	70	.93251	.94789	100	.66696	.82414
11	.99979	.99984	41	.99412	.99607	71	.92679	.94455	101	.66377	.81677
12	.99973	.99983	42	.99342	.99563	72	.92039	.94132	102	.66086	.80950
13	.99961	.99980	43	.99271	.99514	73	.91329	.93831	103	.65823	.80270
14	.99941	.99973	44	.99197	.99458	74	.90547	.93563	104	.65583	.79636
15	.99911	.99960	45	.99118	.99395	75	.89689	.93325	105	.65367	.79047
16	.99869	.99941	46	.99032	.99323	76	.88756	.93111	106	.65170	.78502
17	.99819	.99917	47	.98939	.99242	77	.87746	.92921	107	.64993	.77998
18	.99766	.99893	48	.98839	.99152	78	.86660	.92762	108	.64832	.77535
19	.99718	.99871	49	.98734	.99053	79	.85501	.92621	109	.64687	.77108
20	.99682	.99852	50	.98627	.98945	80	.84276	.92380	110	.64555	.76717
21	.99660	.99837	51	.98520	.98828	81	.82994	.92133	111	.64437	.76360
22	.99652	.99826	52	.98415	.98702	82	.81668	.91886	112	.64330	.76033
23	.99657	.99819	53	.98310	.98568	83	.80315	.91637	113	.64233	.75734
24	.99671	.99815	54	.98200	.98425	84	.78954	.91388	114	.64146	.75462
25	.99689	.99813	55	.98076	.98275	85	.77604	.91137	115	.64067	.75215
26	.99707	.99812	56	.97929	.98118	86	.76289	.90884	116	.63996	.74990
27	.99723	.99810	57	.97749	.97954	87	.75012	.90630	117	.63932	.74786
28	.99736	.99808	58	.97529	.97759	88	.73764	.90373	118	.63874	.74600
29	.99745	.99805	59	.97269	.97552	89	.72531	.90021			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
HAWAIIAN FEMALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.99768	.99846	30	.99832	.99915	60	.98243	.98710	90	.79137	.91938
1	.99868	.99874	31	.99818	.99914	61	.98123	.98612	91	.78058	.91499
2	.99919	.99899	32	.99806	.99913	62	.97968	.98511	92	.76978	.91065
3	.99945	.99922	33	.99797	.99911	63	.97769	.98408	93	.75895	.90633
4	.99959	.99940	34	.99792	.99907	64	.97525	.98302	94	.74801	.90189
5	.99966	.99955	35	.99788	.99902	65	.97239	.98194	95	.73683	.89730
6	.99970	.99965	36	.99785	.99895	66	.96919	.98083	96	.72535	.89251
7	.99973	.99971	37	.99781	.99883	67	.96579	.97971	97	.71364	.88759
8	.99975	.99974	38	.99773	.99865	68	.96226	.97856	98	.70177	.88262
9	.99977	.99976	39	.99760	.99843	69	.95864	.97739	99	.68970	.87763
10	.99979	.99977	40	.99741	.99818	70	.95491	.97621	100	.67794	.87263
11	.99980	.99978	41	.99716	.99791	71	.95102	.97500	101	.66701	.86764
12	.99979	.99979	42	.99683	.99761	72	.94689	.97377	102	.65684	.86297
13	.99978	.99979	43	.99641	.99729	73	.94246	.97251	103	.64741	.85862
14	.99974	.99977	44	.99591	.99697	74	.93764	.97124	104	.63869	.85459
15	.99968	.99971	45	.99534	.99662	75	.93235	.96991	105	.63065	.85085
16	.99961	.99963	46	.99470	.99626	76	.92653	.96830	106	.62325	.84740
17	.99951	.99953	47	.99400	.99587	77	.92012	.96658	107	.61646	.84422
18	.99940	.99945	48	.99325	.99545	78	.91310	.96476	108	.61023	.84130
19	.99927	.99938	49	.99244	.99500	79	.90542	.96282	109	.60454	.83863
20	.99914	.99933	50	.99141	.99452	80	.89709	.96077	110	.59934	.83618
21	.99902	.99929	51	.99038	.99399	81	.88812	.95859	111	.59460	.83394
22	.99892	.99927	52	.98936	.99343	82	.87853	.95628	112	.59028	.83190
23	.99884	.99926	53	.98836	.99281	83	.86838	.95338	113	.58636	.83004
24	.99879	.99924	54	.98740	.99214	84	.85775	.94876	114	.58279	.82834
25	.99875	.99923	55	.98651	.99142	85	.84678	.94384	115	.57955	.82681
26	.99871	.99922	56	.98568	.99064	86	.83564	.93880	116	.57661	.82541
27	.99865	.99920	57	.98491	.98982	87	.82444	.93373	117	.57395	.82414
28	.99856	.99918	58	.98417	.98895	88	.81324	.92873	118	.57154	.82299
29	.99845	.99916	59	.98337	.98804	89	.80204	.92393			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
FILIPINO MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.98994	.99913	30	.99941	.99934	60	.99177	.99380	90	.79367	.91254
1	.99962	.99941	31	.99941	.99934	61	.99090	.99314	91	.77453	.90844
2	.99964	.99963	32	.99942	.99934	62	.98984	.99236	92	.75566	.90433
3	.99966	.99977	33	.99945	.99934	63	.98857	.99146	93	.73745	.90021
4	.99967	.99984	34	.99948	.99934	64	.98705	.99042	94	.72023	.89608
5	.99967	.99987	35	.99951	.99933	65	.98530	.98925	95	.70414	.89195
6	.99968	.99988	36	.99953	.99934	66	.98313	.98794	96	.68918	.88808
7	.99968	.99988	37	.99952	.99933	67	.98141	.98648	97	.67524	.88447
8	.99968	.99990	38	.99947	.99931	68	.97953	.98487	98	.66206	.88112
9	.99968	.99992	39	.99938	.99929	69	.97747	.98312	99	.64963	.87802
10	.99968	.99993	40	.99924	.99925	70	.97522	.98122	100	.63804	.87515
11	.99967	.99992	41	.99906	.99916	71	.97276	.97919	101	.62727	.87252
12	.99964	.99988	42	.99885	.99904	72	.97005	.97701	102	.61728	.87009
13	.99960	.99982	43	.99862	.99892	73	.96702	.97476	103	.60805	.86787
14	.99954	.99978	44	.99837	.99881	74	.96364	.97199	104	.59954	.86583
15	.99946	.99973	45	.99810	.99868	75	.95985	.96902	105	.59172	.86397
16	.99937	.99964	46	.99783	.99853	76	.95561	.96587	106	.58454	.86227
17	.99927	.99955	47	.99755	.99835	77	.95087	.96258	107	.57797	.86072
18	.99918	.99945	48	.99724	.99814	78	.94562	.95915	108	.57196	.85931
19	.99911	.99937	49	.99691	.99790	79	.93981	.95561	109	.56648	.85803
20	.99907	.99931	50	.99654	.99764	80	.93336	.95198	110	.56148	.85687
21	.99907	.99929	51	.99613	.99735	81	.92613	.94825	111	.55692	.85581
22	.99910	.99930	52	.99570	.99705	82	.91796	.94446	112	.55278	.85486
23	.99916	.99934	53	.99527	.99673	83	.90864	.94060	113	.54902	.85399
24	.99923	.99939	54	.99486	.99640	84	.89796	.93669	114	.54561	.85321
25	.99930	.99943	55	.99445	.99605	85	.87904	.93274	115	.54252	.85250
26	.99936	.99944	56	.99404	.99569	86	.86444	.92875	116	.53971	.85185
27	.99939	.99941	57	.99360	.99530	87	.84836	.92473	117	.53717	.85127
28	.99941	.99938	58	.99309	.99487	88	.83099	.92069	118	.53488	.85075
29	.99941	.99935	59	.99249	.99437	89	.81263	.91662			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
FILIPINO FEMALES

AG		1970		AGE	1970		AGE	1970		AGE	1970
0	.99122	.99917	30	.99961	.99971	60	.99638	.99721	90	.87318	.96192
1	.99963	.99944	31	.99958	.99967	61	.99600	.99698	91	.84981	.96030
2	.99967	.99965	32	.99955	.99965	62	.99551	.99671	92	.82483	.95868
3	.99971	.99978	33	.99951	.99963	63	.99488	.99637	93	.79924	.95706
4	.99974	.99985	34	.99946	.99958	64	.99409	.99596	94	.77422	.95554
5	.99976	.99989	35	.99941	.99953	65	.99314	.99547	95	.75097	.95413
6	.99977	.99993	36	.99935	.99949	66	.99151	.99490	96	.73023	.95282
7	.99978	.99996	37	.99929	.99946	67	.99049	.99424	97	.71214	.95161
8	.99979	.99996	38	.99923	.99943	68	.98938	.99351	98	.69636	.95049
9	.99979	.99996	39	.99918	.99942	69	.98821	.99271	99	.68226	.94945
10	.99979	.99996	40	.99912	.99939	70	.98701	.99185	100	.66909	.94851
11	.99978	.99994	41	.99906	.99935	71	.98586	.99095	101	.65605	.94764
12	.99978	.99990	42	.99899	.99930	72	.98482	.98982	102	.64335	.94684
13	.99977	.99989	43	.99890	.99924	73	.98390	.98860	103	.63155	.94611
14	.99976	.99989	44	.99880	.99915	74	.98310	.98729	104	.62059	.94545
15	.99975	.99987	45	.99869	.99904	75	.98239	.98590	105	.61043	.94485
16	.99973	.99984	46	.99857	.99888	76	.98167	.98443	106	.60103	.94430
17	.99971	.99982	47	.99846	.99872	77	.98085	.98292	107	.59238	.94380
18	.99969	.99981	48	.99835	.99856	78	.97981	.98137	108	.58442	.94334
19	.99967	.99980	49	.99826	.99841	79	.97843	.97979	109	.57712	.94293
20	.99964	.99978	50	.99818	.99829	80	.97662	.97819	110	.57043	.94256
21	.99963	.99976	51	.99809	.99820	81	.97429	.97657	111	.56431	.94222
22	.99962	.99974	52	.99798	.99813	82	.97130	.97495	112	.55873	.94191
23	.99961	.99973	53	.99786	.99806	83	.96738	.97332	113	.55364	.94164
24	.99962	.99972	54	.99771	.99799	84	.96218	.97169	114	.54901	.94139
25	.99962	.99972	55	.99754	.99791	85	.95105	.97006	115	.54480	.94118
26	.99963	.99973	56	.99735	.99781	86	.94086	.96843	116	.54097	.94100
27	.99963	.99973	57	.99715	.99769	87	.92811	.96680	117	.53750	.94083
28	.99963	.99973	58	.99693	.99756	88	.91259	.96517	118	.53435	.94068
29	.99962	.99974	59	.99668	.99739	89	.89423	.96354			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
HISPANIC MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.99371	.99800	30	.99559	.99669	60	.98488	.98388	90	.82078	.85771
1	.99646	.99847	31	.99554	.99671	61	.98354	.98245	91	.81101	.85059
2	.99790	.99890	32	.99549	.99672	62	.98203	.98090	92	.80232	.84343
3	.99866	.99924	33	.99544	.99671	63	.98038	.97924	93	.79482	.83623
4	.99907	.99949	34	.99539	.99666	64	.97859	.97746	94	.78814	.82900
5	.99929	.99965	35	.99533	.99660	65	.97666	.97555	95	.78189	.82175
6	.99942	.99973	36	.99524	.99653	66	.97462	.97353	96	.77611	.81446
7	.99950	.99976	37	.99510	.99644	67	.97255	.97138	97	.77076	.80717
8	.99954	.99975	38	.99489	.99635	68	.97032	.96910	98	.76583	.79987
9	.99954	.99974	39	.99463	.99625	69	.96790	.96671	99	.76130	.79257
10	.99952	.99973	40	.99432	.99613	70	.96527	.96421	100	.75714	.78809
11	.99946	.99971	41	.99398	.99598	71	.96243	.96160	101	.75334	.77638
12	.99935	.99966	42	.99363	.99580	72	.95933	.95875	102	.74986	.76526
13	.99918	.99954	43	.99330	.99558	73	.95596	.95550	103	.74668	.75475
14	.99892	.99931	44	.99299	.99532	74	.95231	.95207	104	.74379	.74487
15	.99858	.99899	45	.99270	.99501	75	.94834	.94849	105	.74115	.73562
16	.99815	.99862	46	.99241	.99467	76	.94403	.94475	106	.73876	.72700
17	.99764	.99823	47	.99212	.99429	77	.93934	.94072	107	.73658	.71898
18	.99709	.99784	48	.99181	.99386	78	.93423	.93568	108	.73461	.71155
19	.99653	.99749	49	.99148	.99339	79	.92865	.93027	109	.73282	.70469
20	.99601	.99719	50	.99113	.99286	80	.92254	.92454	110	.73120	.69837
21	.99558	.99695	51	.99076	.99229	81	.91583	.91852	111	.72973	.69256
22	.99529	.99677	52	.99039	.99166	82	.90842	.91228	112	.72840	.68722
23	.99516	.99665	53	.99000	.99097	83	.90025	.90584	113	.72720	.68234
24	.99517	.99656	54	.98958	.99022	84	.89125	.89925	114	.72611	.67788
25	.99527	.99653	55	.98911	.98939	85	.87622	.89253	115	.72513	.67381
26	.99541	.99653	56	.98855	.98849	86	.86534	.88572	116	.72424	.67010
27	.99553	.99656	57	.98788	.98749	87	.85405	.87882	117	.72344	.66672
28	.99560	.99660	58	.98705	.98639	88	.84263	.87184	118	.72272	.66365
29	.99562	.99665	59	.98605	.98519	89	.83142	.86480			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
HISPANIC FEMALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.99523	.99903	30	.99911	.99935	60	.99014	.99183	90	.84342	.88817
1	.99736	.99926	31	.99905	.99938	61	.98929	.99094	91	.83506	.88183
2	.99846	.99944	32	.99897	.99938	62	.98829	.98996	92	.82783	.87546
3	.99904	.99957	33	.99887	.99932	63	.98712	.98889	93	.82141	.86908
4	.99935	.99967	34	.99875	.99923	64	.98576	.98772	94	.81536	.86268
5	.99953	.99974	35	.99862	.99914	65	.98418	.98645	95	.80974	.85626
6	.99963	.99979	36	.99846	.99906	66	.98216	.98507	96	.80452	.84983
7	.99969	.99982	37	.99828	.99899	67	.98044	.98359	97	.79970	.84339
8	.99973	.99985	38	.99808	.99894	68	.97856	.98198	98	.79525	.83695
9	.99975	.99988	39	.99785	.99888	69	.97653	.98026	99	.79116	.83051
10	.99976	.99988	40	.99761	.99881	70	.97433	.97843	100	.78740	.82448
11	.99975	.99985	41	.99737	.99872	71	.97192	.97647	101	.78396	.81886
12	.99972	.99980	42	.99714	.99863	72	.96928	.97439	102	.78081	.81364
13	.99967	.99976	43	.99695	.99853	73	.96638	.97219	103	.77794	.80881
14	.99960	.99972	44	.99679	.99843	74	.96318	.96992	104	.77532	.80444
15	.99950	.99969	45	.99665	.99831	75	.95965	.96706	105	.77293	.80023
16	.99937	.99966	46	.99651	.99817	76	.95576	.96402	106	.77076	.79646
17	.99924	.99960	47	.99637	.99801	77	.95147	.96080	107	.76879	.79299
18	.99909	.99953	48	.99619	.99779	78	.94674	.95741	108	.76700	.78982
19	.99896	.99943	49	.99597	.99753	79	.94152	.95278	109	.76537	.78693
20	.99886	.99935	50	.99570	.99722	80	.93576	.94778	110	.76390	.78428
21	.99880	.99930	51	.99537	.99688	81	.92941	.94249	111	.76257	.78187
22	.99880	.99926	52	.99498	.99650	82	.92242	.93696	112	.76136	.77968
23	.99884	.99924	53	.99452	.99609	83	.91475	.93124	113	.76027	.77769
24	.99891	.99922	54	.99400	.99564	84	.90639	.92536	114	.75929	.77588
25	.99899	.99921	55	.99343	.99515	85	.89264	.91934	115	.75839	.77424
26	.99907	.99922	56	.99284	.99462	86	.88282	.91322	116	.75759	.77275
27	.99912	.99923	57	.99222	.99403	87	.87270	.90703	117	.75686	.77140
28	.99915	.99927	58	.99157	.99337	88	.86255	.90078	118	.75621	.77018
29	.99914	.99931	59	.99089	.99264	89	.85268	.89449			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
PUERTO RICAN RESIDENT MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.99355	.99592	30	.99756	.99744	60	.98336	.98356	90	.82403	.84771
1	.99660	.99799	31	.99750	.99741	61	.98200	.98234	91	.81397	.84040
2	.99809	.99915	32	.99741	.99735	62	.98052	.98105	92	.80484	.83338
3	.99882	.99957	33	.99729	.99728	63	.97893	.97967	93	.79682	.82672
4	.99919	.99966	34	.99713	.99722	64	.97721	.97820	94	.78960	.82048
5	.99939	.99967	35	.99695	.99713	65	.97536	.97660	95	.78283	.81466
6	.99949	.99967	36	.99674	.99700	66	.97325	.97484	96	.77654	.80926
7	.99955	.99969	37	.99650	.99680	67	.97111	.97290	97	.77072	.80425
8	.99957	.99972	38	.99624	.99657	68	.96879	.97077	98	.76534	.79963
9	.99956	.99973	39	.99596	.99637	69	.96631	.96842	99	.76039	.79537
10	.99954	.99969	40	.99565	.99617	70	.96364	.96584	100	.75583	.79145
11	.99950	.99968	41	.99531	.99594	71	.96081	.96300	101	.75166	.78786
12	.99943	.99966	42	.99494	.99568	72	.95779	.95988	102	.74784	.78457
13	.99933	.99955	43	.99454	.99539	73	.95459	.95645	103	.74435	.78157
14	.99921	.99945	44	.99411	.99505	74	.95118	.95273	104	.74116	.77883
15	.99906	.99926	45	.99366	.99466	75	.94754	.94869	105	.73826	.77633
16	.99889	.99899	46	.99320	.99421	76	.94361	.94433	106	.73561	.77405
17	.99869	.99872	47	.99272	.99369	77	.93933	.93966	107	.73321	.77199
18	.99849	.99844	48	.99225	.99310	78	.93462	.93469	108	.73103	.77011
19	.99830	.99825	49	.99176	.99248	79	.92941	.92945	109	.72905	.76841
20	.99811	.99817	50	.99127	.99186	80	.92362	.92333	110	.72726	.76686
21	.99795	.99811	51	.99076	.99124	81	.91716	.91635	111	.72563	.76546
22	.99782	.99802	52	.99022	.99061	82	.90998	.90904	112	.72416	.76419
23	.99773	.99795	53	.98964	.98995	83	.90204	.90150	113	.72283	.76305
24	.99767	.99786	54	.98901	.98924	84	.89331	.89382	114	.72162	.76201
25	.99764	.99777	55	.98832	.98848	85	.87877	.88608	115	.72053	.76107
26	.99763	.99767	56	.98755	.98766	86	.86823	.87834	116	.71955	.76022
27	.99762	.99757	57	.98668	.98676	87	.85721	.87063	117	.71866	.75946
28	.99762	.99749	58	.98570	.98577	88	.84596	.86293	118	.71786	.75877
29	.99760	.99746	59	.98459	.98471	89	.83479	.85524			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
PUERTO RICAN RESIDENT FEMALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.99360	.99773	30	.99894	.99935	60	.98979	.99167	90	.84762	.87241
1	.99672	.99868	31	.99889	.99927	61	.98887	.99089	91	.83877	.86628
2	.99822	.99931	32	.99882	.99917	62	.98785	.99001	92	.83088	.86040
3	.99895	.99962	33	.99873	.99907	63	.98671	.98904	93	.82409	.85481
4	.99932	.99973	34	.99864	.99899	64	.98545	.98795	94	.81803	.84959
5	.99951	.99976	35	.99852	.99892	65	.98402	.98671	95	.81235	.84471
6	.99962	.99976	36	.99840	.99886	66	.98213	.98579	96	.80709	.84019
7	.99968	.99978	37	.99828	.99880	67	.98029	.98451	97	.80221	.83599
8	.99971	.99980	38	.99814	.99872	68	.97827	.98289	98	.79771	.83212
9	.99973	.99982	39	.99800	.99866	69	.97608	.98097	99	.79357	.82855
10	.99972	.99985	40	.99786	.99857	70	.97374	.97869	100	.78977	.82527
11	.99971	.99981	41	.99771	.99846	71	.97127	.97618	101	.78629	.82226
12	.99969	.99977	42	.99756	.99833	72	.96869	.97339	102	.78310	.81951
13	.99965	.99974	43	.99740	.99818	73	.96602	.97044	103	.78019	.81699
14	.99961	.99971	44	.99722	.99802	74	.96323	.96727	104	.77753	.81469
15	.99957	.99965	45	.99703	.99785	75	.96028	.96379	105	.77511	.81260
16	.99952	.99963	46	.99681	.99765	76	.95708	.95991	106	.77291	.81069
17	.99947	.99964	47	.99656	.99742	77	.95353	.95559	107	.77091	.80896
18	.99941	.99964	48	.99627	.99715	78	.94950	.95073	108	.76909	.80739
19	.99936	.99960	49	.99593	.99686	79	.94487	.94534	109	.76744	.80596
20	.99930	.99956	50	.99555	.99654	80	.93955	.93944	110	.76595	.80467
21	.99925	.99954	51	.99514	.99620	81	.93347	.93317	111	.76460	.80349
22	.99921	.99954	52	.99470	.99582	82	.92661	.92657	112	.76337	.80243
23	.99917	.99955	53	.99423	.99543	83	.91899	.91976	113	.76227	.80147
24	.99914	.99951	54	.99375	.99502	84	.91068	.91284	114	.76126	.80060
25	.99911	.99947	55	.99323	.99459	85	.89711	.90589	115	.76036	.79982
26	.99909	.99942	56	.99267	.99411	86	.88743	.89901	116	.75954	.79911
27	.99906	.99941	57	.99206	.99359	87	.87740	.89222	117	.75880	.79847
28	.99903	.99942	58	.99138	.99301	88	.86723	.88550	118	.75814	.79789
29	.99899	.99941	59	.99063	.99237	89	.85720	.87883			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
OTHER RACE MALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.98407	.98769	30	.99837	.99834	60	.97970	.98238	90	.79944	.80942
1	.99901	.99908	31	.99837	.99835	61	.97770	.98067	91	.78424	.79611
2	.99926	.99934	32	.99834	.99834	62	.97569	.97881	92	.76886	.78136
3	.99941	.99947	33	.99826	.99831	63	.97372	.97684	93	.75350	.76547
4	.99949	.99957	34	.99817	.99825	64	.97174	.97477	94	.73826	.74939
5	.99954	.99961	35	.99804	.99816	65	.96968	.97262	95	.72334	.73383
6	.99957	.99963	36	.99790	.99804	66	.96742	.97032	96	.70891	.71999
7	.99959	.99966	37	.99773	.99791	67	.96486	.96782	97	.69512	.70689
8	.99963	.99970	38	.99754	.99776	68	.96188	.96505	98	.68186	.69455
9	.99968	.99976	39	.99731	.99760	69	.95848	.96195	99	.66940	.68297
10	.99973	.99981	40	.99706	.99739	70	.95482	.95852	100	.65776	.67216
11	.99973	.99981	41	.99677	.99713	71	.95093	.95484	101	.64691	.66209
12	.99965	.99972	42	.99643	.99684	72	.94672	.95099	102	.63683	.65276
13	.99948	.99954	43	.99604	.99652	73	.94216	.94705	103	.62750	.64412
14	.99923	.99929	44	.99559	.99618	74	.93724	.94297	104	.61889	.63616
15	.99896	.99904	45	.99509	.99580	75	.93192	.93854	105	.61096	.62883
16	.99870	.99882	46	.99457	.99537	76	.92622	.93358	106	.60368	.62210
17	.99848	.99863	47	.99399	.99486	77	.92015	.92820	107	.59700	.61593
18	.99835	.99849	48	.99337	.99427	78	.91371	.92238	108	.59089	.61029
19	.99825	.99837	49	.99273	.99361	79	.90691	.91606	109	.58531	.60514
20	.99818	.99825	50	.99201	.99294	80	.89988	.90901	110	.58021	.60045
21	.99809	.99814	51	.99122	.99225	81	.89270	.90114	111	.57557	.59617
22	.99805	.99807	52	.99037	.99150	82	.88569	.89267	112	.57135	.59228
23	.99807	.99807	53	.98945	.99066	83	.87930	.88387	113	.56751	.58874
24	.99813	.99811	54	.98844	.98973	84	.87434	.87477	114	.56403	.58553
25	.99821	.99817	55	.98739	.98875	85	.86637	.86493	115	.56086	.58262
26	.99828	.99823	56	.98625	.98773	86	.85463	.85408	116	.55800	.57998
27	.99834	.99828	57	.98491	.98662	87	.84201	.84309	117	.55540	.57760
28	.99837	.99832	58	.98337	.98536	88	.82847	.83226	118	.55305	.57543
29	.99839	.99833	59	.98161	.98395	89	.81422	.82125			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

EXPECTED 1-YEAR SURVIVAL RATES
OTHER RACE FEMALES

AGE	1970	1980	AGE	1970	1980	AGE	1970	1980	AGE	1970	1980
0	.98770	.99035	30	.99926	.99935	60	.99023	.99111	90	.83514	.85169
1	.99925	.99923	31	.99922	.99932	61	.98934	.99025	91	.81991	.83769
2	.99939	.99949	32	.99916	.99928	62	.98849	.98933	92	.80402	.82291
3	.99950	.99963	33	.99910	.99923	63	.98770	.98838	93	.78778	.80802
4	.99959	.99970	34	.99904	.99917	64	.98693	.98741	94	.77162	.79310
5	.99965	.99972	35	.99898	.99910	65	.98611	.98641	95	.75598	.77772
6	.99970	.99974	36	.99890	.99901	66	.98511	.98530	96	.74116	.76271
7	.99973	.99977	37	.99879	.99891	67	.98380	.98405	97	.72729	.74827
8	.99976	.99979	38	.99866	.99881	68	.98212	.98260	98	.71434	.73449
9	.99979	.99982	39	.99851	.99870	69	.98006	.98093	99	.70218	.72141
10	.99980	.99983	40	.99833	.99857	70	.97783	.97908	100	.69076	.70906
11	.99981	.99984	41	.99815	.99842	71	.97541	.97706	101	.68010	.69745
12	.99978	.99981	42	.99795	.99826	72	.97264	.97483	102	.67017	.68658
13	.99974	.99975	43	.99775	.99808	73	.96941	.97240	103	.66095	.67645
14	.99966	.99968	44	.99753	.99789	74	.96576	.96973	104	.65243	.66703
15	.99958	.99960	45	.99730	.99769	75	.96174	.96685	105	.64456	.65832
16	.99950	.99953	46	.99704	.99746	76	.95744	.96363	106	.63732	.65027
17	.99945	.99948	47	.99675	.99720	77	.95285	.95985	107	.63068	.64285
18	.99942	.99946	48	.99647	.99690	78	.94800	.95533	108	.62458	.63603
19	.99942	.99945	49	.99615	.99657	79	.94281	.95005	109	.61901	.62978
20	.99941	.99944	50	.99581	.99624	80	.93723	.94411	110	.61392	.62406
21	.99941	.99943	51	.99545	.99590	81	.93113	.93761	111	.60928	.61883
22	.99940	.99943	52	.99505	.99553	82	.92441	.93051	112	.60505	.61406
23	.99940	.99942	53	.99461	.99512	83	.91689	.92287	113	.60120	.60971
24	.99939	.99942	54	.99413	.99468	84	.90835	.91461	114	.59771	.60577
25	.99939	.99942	55	.99362	.99421	85	.89852	.90537	115	.59453	.60217
26	.99939	.99942	56	.99308	.99372	86	.88739	.89509	116	.59165	.59889
27	.99936	.99941	57	.99246	.99319	87	.87558	.88466	117	.58904	.59593
28	.99934	.99940	58	.99179	.99258	88	.86299	.87441	118	.58668	.59324
29	.99931	.99937	59	.99106	.99189	89	.84952	.86383			

Source: National Cancer Institute, DCPC/SP/CST
National Center for Health Statistics

BIBLIOGRAPHY

BIBLIOGRAPHY

American Statistics Index: *A Comprehensive Guide and Index to the Statistical Publications of the U.S. Government*. (Issued annually). Washington, DC: Congressional Information Service.

Beahrs, O.H; Henson, D.E; Hutter, R.V.P; Kennedy, B.J (eds.), *American Joint Committee on Cancer: Manual for Staging of Cancer*, 4th ed. Philadelphia: J. B. Lippincott, 1992.

Bross, I.D.J. *Scientific Strategies to Save Your Life: A Statistical Approach to Primary Prevention*. Statistics: Textbooks and Monographs, Vol. 35. New York: Marcel Dekker, Inc., 1981.

Bureau of the Census *Statistical Abstract of the United States*. U.S. Department of Commerce. (Issued annually). Washington, DC: U.S. Government Printing Office.

Castle, W.M. *Statistics in Operation*. New York: Churchill Livingstone, 1979. Congressional Information Service (CIS) Index to Publications of the United States Congress. Washington, DC.

Colton, T. *Statistics in Medicine*. Boston: Little, Brown and Co, 1974

Congressional Information Service. (Issued annually). Cutler, S.J. and Young, J.L., Jr. *Third National Cancer Survey: Incidence Data*. National Cancer Institute Monograph 41. DHEW, Pub. No. (NIH) 75-787. Washington, D.C.: U.S. Government Printing Office, 1975. OUT OF PRINT

Dorn H.F. and Cutler, S.J. *Morbidity from Cancer in the United States. Part I: Variation in Incidence by Age, Sex, Race, Marital Status and Geographic Region. Part II: Trends in Morbidity Association with Income and Stage of Diagnosis*. Public Health Service Pub. No. 590. Washington, D.C.: U.S. Government Printing Office, 1959. OUT OF PRINT

Feinleib, M.; Fabsitz, R.; and Sharrett, A.R. *Mortality from Cardiovascular and Non-cardiovascular Diseases for US Cities 1949-1950, 1959-1961, 1969-1971*. DHEW (NIH). Washington D.C: U.S. Government Printing Office, 1979.

Gehlbach, S.H. *Interpreting the Medical Literature: A Clinician's Guide*. Lexington, MA: The Collamore Press, 1982.

Glantz, S.A. *Primer of Biostatistics*, 2nd edition. New York: McGraw-Hill, Inc., 1987.

Goldstone, L.A. *Understanding Medical Statistics*. London: William Heinemann Medical Books Ltd., 1983.

Hopkins, K.D. and Glass, Gene V. *Basic Statistics for the Behavioral Sciences*. University of Colorado. New Jersey: Prentice-Hall Inc., 1978.

Jaegar, R. and Law, A. *Statistics as a Spectator Sport*. Inverness, California: Edgepress, 1982.

Leaverton, P.E. *A Review of Biostatistics: A Program for Self-Instruction*, 2nd ed. Boston: Little, Brown, 1978.

Mattson, D.E. *Statistics: Difficult Concepts, Understandable Explanations*. St. Louis: Mosby, 1981.

McKay, F.W.; Hanson, M.R.; and Miller, R.W. *Cancer Mortality in the United States: 1950-1977*. National Cancer Institute Public Health Monograph 59. DHHS (NIH) Pub. No. 82-2435. Washington, DC: U.S. Government Printing Office, 1982.

Merrington, M. "Table of Percentage Points of the t-Distribution." *Biometrika* **32**:300, 1942.

Miller, B.A.; Gloeckler Ries, L.A.; Hankey, B.F.; Kosary, C.L.; and Edwards, B.K, eds. *Cancer Statistics Review, 1973-1989*. DHHS (NIH) Pub. No. 92-2789. Bethesda, MD: 1992.

Moore, D.S. *Statistics: Concepts and Controversies*. San Francisco: W.H. Freeman, 1979.

National Cancer Institute. *Directory of Cancer Research Information Resources*. DHEW (NIH) Bethesda, MD: 1979.

National Center for Health Statistics *Catalog of Public Use Data Tapes from the National Center for Health Statistics*. Hyattsville, MD: 1980.

National Center for Health Statistics *Data Systems of the National Center for Health Statistics. Programs and Collection Procedures*. Series 1, No. 16 DHHS. Washington DC: U.S. Government Printing Office, 1981.

National Center for Health Statistics: *Vital Statistics of the United States, 1980, Vol. II, Mortality, Part B*. DHHS Pub. No. (PHS) 85-1102. Public Health Service. Washington, DC: U. S. Government Printing Office, 1985. (Issued annually).

Phillips, D.S. *Basic Statistics for Health Science Students*. San Francisco: W.H. Freeman, 1978.

Ries, L.G.; Pollack, E.S.; and Young, J.L., Jr. "Cancer Patient Survival: Surveillance, Epidemiology and End Results Program, 1973-79." *J. Natl Cancer Inst.* **70**(4): 693-707, 1983.

Riggan, W.B.; Van Bruggen, J.; Acquavella, J.F.; Beaubier, J.; and Mason, T.J. *U.S. Cancer Mortality Rates and Trends, 1950-1979*. 3 vols. National Cancer Institute/U. S. Environmental Protection Agency, Interagency Agreement on Environmental Carcinogenesis. Washington, DC: U.S. Government Printing Office, 1983.

Sartwell, P.E., et. al. *American Journal of Epidemiology*. **90**:365-380, 1969.

SEER Program. *Cancer Incidence and Mortality in the United States, 1972-1981.* Prepared by Biometry Branch, Division of Cancer Prevention and Control, National Cancer Institute. Edited by J.W. Horm, A.J. Asire, J.L. Young, Jr., and E.S. Pollack. DHHS (NIH) Pub. No. 85-1837. Revised November 1984. Bethesda: U.S. Government Printing Office, November 1984.

Sondik, E.J.; Young, J.L., Jr.; Horm, J.W.; and Gloeckler Ries, L.A., eds. *1985 Annual Cancer Statistics Review.* DHHS (NIH) Pub. No. 86-2789. Bethesda, MD: 1986.

Statistical Reference Index . . . Annual. *A Selective Guide to American Statistical Publications from Private Organizations and State Government Sources.* Washington DC: Congressional Information Service.

Tanur, J.M. et al. *Statistics: A Guide to Biological and Health Sciences.* San Francisco: Holden-Day, 1977.

Waterhouse, J.; Muir, C.; Shanmugaratnam, K.; and Powell, J. *Cancer Incidence in Five Continents*, Vol. IV. International Agency for Research of Cancer, Lyon. IARC Sci Pub No. 42, 1982.

Young, J.L., Jr.; Percy, C.L.; Asire, A.J. eds. *Surveillance, Epidemiology, and End Results Program: Incidence and Mortality Data, 1973-77.* National Cancer Institute Monograph 57, DHHS, Pub. No. (NIH) 81-2330. Washington, DC: U.S. Government Printing Office, 1981.

Young, J.L., Jr.; Ries, L.G.; and Pollack, E.S. "Cancer Patient Survival among Ethnic Groups in the United States." *J. Natl Cancer Inst.* **73**(2), 341-352, 1984.

Zeisset, P.T. *1970 Census of the Population. Index to Selected 1970 Census Reports.* Bureau of Census, U.S. Department of Commerce. Washington, DC: U.S. Government Printing Office.

INDEX

Book 7 Index