APPLIED STATISTICS FOR EDUCATIONAL RESEARCH

Dr. Dadan Rosana, M.Si.

FACULTY OF MATEMATHICS AND NATURAL SCIENCES
YOGYAKARTA STATE UNIVERSITY
2012

PREFACE

This is an introductory textbook for a first course in applied statistics and probability for un dergraduate students in the physics education. Statistical methods are an important tool in these activities because they provide the education researcher with both descriptive and analytical methods for dealing with the variability in observed data. Although many of the methods we present are fundamental to statistical analysis in other disciplines, such as education and management, the life sciences, and the social sciences, we have elected to focus on an physics education students-oriented audience. We believe that this approach will best serve students in physics education and will allow them to concentrate on the many applications of statistics in these disciplines. This book can be used for a single course, although we have provided enough material for two courses in the hope that more students will see the important applications of statistics in their everyday work and elect a second course. We believe that this book will also serve as a useful reference.

ORGANIZATION OF THE BOOK

Chapter 1 is an introduction to the field of statistics and how engineers use statistical methodology as part of the science education problem-solving process. This chapter also introduces the reader to some science education applications of statistics, including building empirical models,designing engineering experiments, and monitoring manufacturing processes. These topics are discussed in more depth in subsequent chapters.

Chapters 2, 3, 4, and 5 cover the basic concepts of probability, discrete and continuous random variables, probability distributions, expected values, joint probability distributions, and independence. We have given a reasonably complete treatment of these topics but have avoided many of the mathematical or more theoretical details.

Chapter 6 begins the treatment of statistical methods with random sampling; data summary and description techniques, including stem-and-leaf plots, histograms, box plots, and probability plotting; and several types of time series plots. Chapter 7 discusses point estimation of parameters. This chapter also introduces some of the important properties of estimators, the method of maximum likelihood, the method of moments, sampling distributions, and the central limit theorem.

Chapter 8 discusses interval estimation for a single sample. Topics included are confidence intervals for means, variances or standard deviations, and proportions and prediction and tolerance intervals. Chapter 9 discusses hypothesis tests for a single sample. Chapter 10 presents tests and confidence intervals for two samples. This material has been extensively rewriteten and reorganized. There is detailed information and examples of methods for determiningappropriate sample sizes. We want the student to become familiar with how these techniques are used to solve real-world engineering problems and to get some understanding of the con-cepts behind them. We give a logical, heuristic development of the procedures, rather than a formal mathematical one.

Chapters 11 present simple and multiple linear regression. We use matrix algebra throughout the multiple regression material because it is the only easy way to understand the concepts presented. Scalar arithmetic presentations of multiple regression are awkward at best, and we have found that undergraduate engineers are exposed to enoughmatrix algebra to understand the presentation of this material.

TABLE OF CONTENT

CHAPTER 1

THE ROLE OF STATISTICS IN EDUCATIONAL RESEARCH

The issue of the quality of education is increasingly becoming an area of interest and concern to many nations of the developing world, especially in Indonesian. This is so because many countries in this region have realized that education plays a crucial and pivotal role in development at national, regional and international levels. There is concern because quality of education seems to be either stagnating or deteriorating. It is also generally accepted that educational development in Indonesia has remained low in comparison with other regions of the world. There is interest in the issue of quality because it is an integral part of the development and monitoring of education systems the world over. There can be no argument over the fact that quality assessments have frequently been made on the basis of key indicators generated through the analysis of the statistics available. What has frequently not been appreciated, however, is that statistics is one of the essential, key instruments for the promotion of quality in education. This presentation attempts to highlight some points on how statistics has been, and can be used, to improve quality in education. So the paper is not meant to tell you anything new, but rather to rai se awareness on the importance and value of statistics in the development of quality in education

A. Learning Objectives

After careful study of this chapter you should be able to do the following:

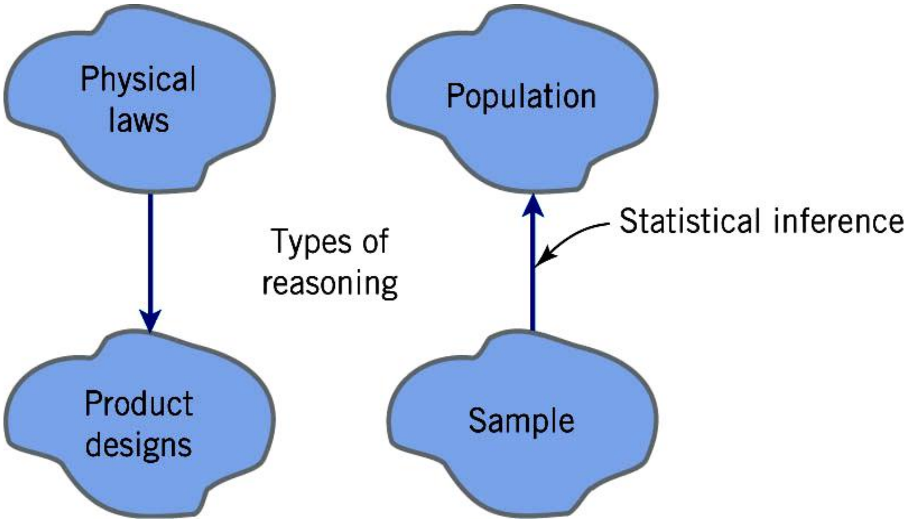
1. Identify the role that statistics can play in the science education problem-solving process
2. Discuss how variability affects the data collected and used for making educational research decisions
3. Explain the difference between enumerative and analytical studies

4. Discuss the different methods that scientist use to collect data
5. Identify the advantages that designed experiments have in comparison to other methods of collecting science education data
6. Explain the differences between mechanistic models and empirical models
7. Discuss how probability and probability models are used in science education

An Physicians is someone who solves problems of interest to society with the efficient application of scientific principles by:

- Refining existing products
- Designing new products or processes

Statistical techniques are useful for describing and understanding variability. By variability, we mean successive observations of a system or phenomenon do *not* produce exactly the same result. Statistics gives us a framework for describing this variability and for learning about potential sources of variability.



Basic Types of Studies

Three basic methods for collecting data:

1. A retrospective study using historical data
   - Data collected in the past for other purposes.
2. An observational study
   - Data, presently collected, by a passive observer.
3. A designed experiment
   - Data collected in response to process input changes.

B. The Nature of Data

Anything that can be counted or measured is called a variable. Knowledge of the different types of variables, and the way they are measured, play a crucial part in choice of coding and data collection. The measurement of variables can be categorized as categorical (nominal or ordinal scales) or continuous (interval or ratio scales).

Categorical measures can be used to identify change in a variable, however, should you wish to measure the magnitude of the change you should use a continuous measure.

A *nominal scale* allows for the classification of objects, individual and responses based on a common characteristic or shared property. A variable measured on the nominal scale may have one, two or more sub-categories depending on the degree of variation in the coding. Any number attached to a nominal classification is merely a label, and no ordering is implied: social worker, nurse, electrician, physicist, politician, teacher, plumber, etc.

An *ordinal scale* not only categorizes objects, individuals and responses into sub-categories on the basis of a common characteristic it also ranks them in descending order of magnitude. Any number attached to an ordinal classification is ordered, but the intervals between may not be constant: GCSE, A-level, diploma, degree, postgraduate diploma, higher degree, and doctorate.

The *interval scale* has the properties of the ordinal scale and, in addition, has a commencement and termination point, and uses a scale of equally spaced intervals in relation to the range of the variable. The number of intervals between the commencement and termination points is arbitrary and varies from one scale to another. In measuring an attitude using the Likert scale, the intervals may mean the same up and down the scale of 1 to 5 but multiplication is not meaningful: a rating of ‘4’ is not twice as ‘favourable’ as a rating of ‘2’.

In addition to having all the properties of the nominal, ordinal and interval scales, the *ratio scale* has a zero point. The ratio scale is an absolute measure allowing multiplication to be meaningful. The numerical values are ‘real numbers’ with which you can conduct mathematical procedures: a man aged 30 years is half the age of a woman of 60 years.

I.1.1.1.1.1.1.1.1 Categorical			I.1.1.1.1.1.1.1.2 ontinuous
Unitary	***Dichotomous***	***Polytomous***	***Interval or Ratio Scale***
Name Occupation Location Site	[1] . . . Yes [0] . . . No [1] . . . Good [0] . . . Bad [1] . . . Female [0] . . . Male	Attitudes (Likert Scale): [5] . . . strongly agree [4] . . . agree [3] . . . uncertain [2] . . . disagree [1] . . . strongly disagree Age: [4] . . . Old	Income (£000s per annum) Age (in years) Reaction Time (in seconds) Absence (in days)

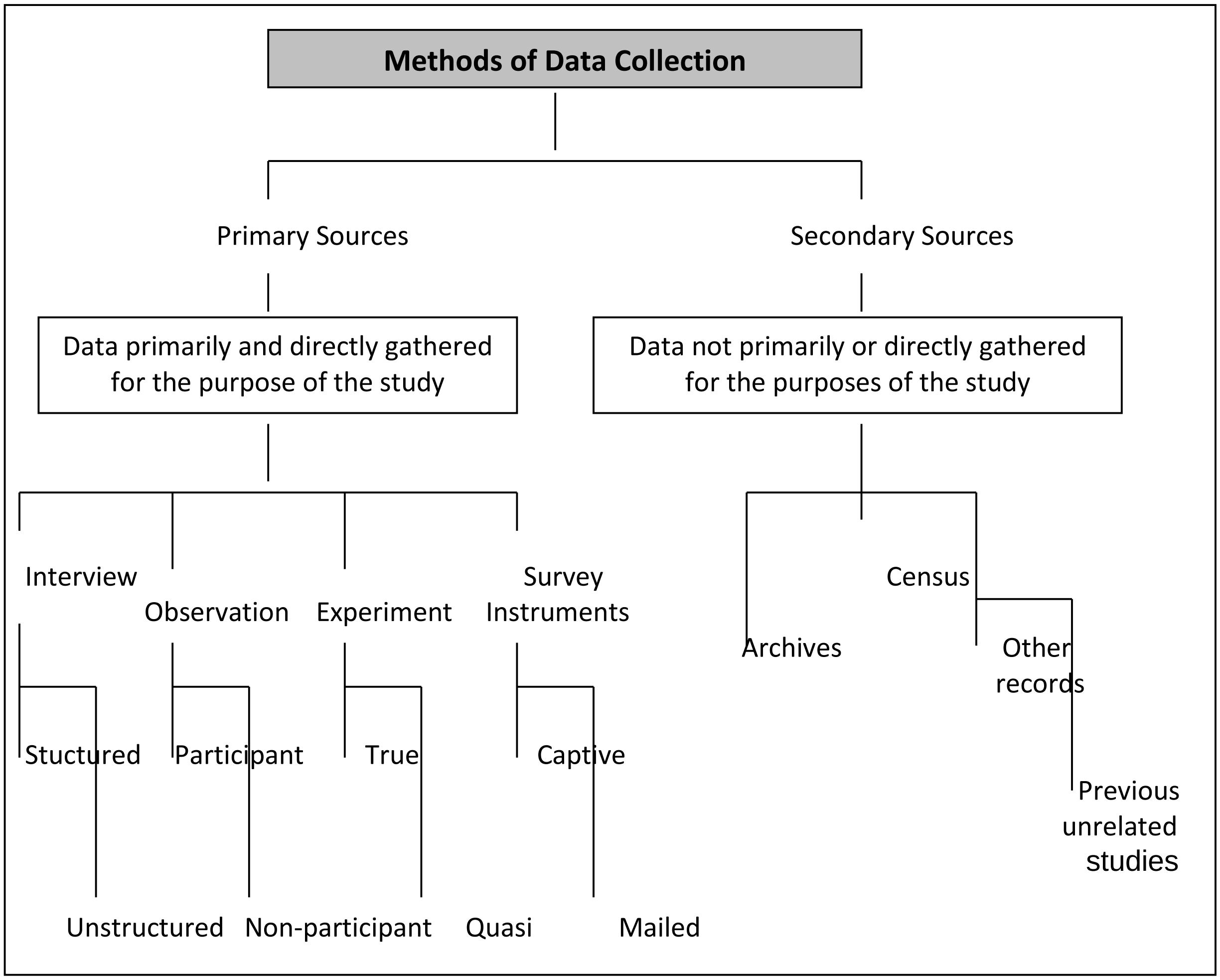
	[1] . . . Right [0] . . . Wrong [1] . . . Extrovert [0] . . . Introvert [1] . . . Psychotic [0] . . . Neurotic [1] . . . Assertive [0] . . . Passive [1] . . . Present [0] . . . Absent	[3] . . . Middle-aged [2] . . . Young [1] . . . Child Income: [3] . . . High [2] . . . Medium [1] . . . Low Socio-Economic Status: [5] . . . A [4] . . . B [3] . . . C1 [2] . . . C2 [1] . . . D [0] . . . E	Distance (in kilometres) Length (metres) Attitude (Thurstone & Cheve)

I.1.1.1.1.1.1.1.3 Qualitative	I.1.1.1.1.1.1.1.4 Quantitative
Sex (Male/Female) Age (Old/Young) Attitude (Favourable/Unfavourable) Attitude (Likert scale) Achieved Educational Level (High/Low) Style (Autocratic/Participative) Location (Urban/Rural) Performance (Good/Bad)	Age (in years) Attitude (Guttman scale) Attitude (Thurstone & Cheve scale) Performance (errors or faults per minute) Achieved Educational Level (number of years post-secondary school education)

Table I
A Two-Way Classification of Variables

1. Methods of Data Collection

The major approaches to gathering data about a phenomenon are from primary sources: directly from subjects by means of experiment or observation, from informants by means of interview, or from respondents by questionnaire and survey instruments. Data may also be obtained from secondary sources: information that is readily available but not necessarily directly related to the phenomenon under study. Examples of secondary sources include published academic articles, government statistics, an organization's archival records to collect data on activities, personnel records to obtain data on age, sex, qualification, length of service, and absence records of workers, etc. Data collected and analyzed from published articles, research papers and journals *may* be a primary source if the material is directly relevant to your study. For instance, primary sources for a study conducted using the Job Descriptive Index may be Hulin and Smith (1964-68) and Jackson (1986-90), whereas a study using an idiosyncratic study population, technique and assumptions, such as those published by Herzberg, et alia (1954-59), would be a secondary source.


Methods of Data Collection
Primary Sources
Secondary Sources
Data primarily and directly gathered for the purpose of the study
Data not primarily or directly gathered for the purposes of the study
Interview
Observation
Experiment
Survey Instruments
Census
Archives
Other records
Stuctured
Participant
True
Captive
Previous unrelated studies
Unstructured
Non-participant
Quasi
Mailed

1.1.1.1.1.1.1.2 Figure 2 Classification of Methods of Data Collection

1.1.1.1.1.1.1.3

2. *Procedures for Coding Data*

A coding frame is simply a set of instructions for transforming data into codes and for identifying the location of all the variable measured by the test or instrument. Primary data gathered from subjects and informants is amenable to control during the data collection phase. The implication is that highly structured data, usually derived from tests, questionnaires and interviews, is produced directly by means of a calibrated instrument or is readily produced from raw scores according to established rules and conventions. Generally, measures such as physical characteristics such as height and weight are measured on the ratio scale. Whereas psychological attributes such as measures of attitude and standard dimensions of personality are often based on questions to which there is no appropriate response. However, the sum of the responses is interpreted according to a set of rules and provides a numerical score on the interval scale but is often treated as though the measures relate the ratio scale. Norms are available for standard tests of physical and psychological attributes to establish the meaning of individual scores in terms of those derived from the general population. A questionnaire aimed at determining scores as a measure of a psychological attribute are said to be pre-coded; that is, the data reflects the coder's prior structuring of the population. The advantages of pre-coding are that it reduces time, cost and coding error in data handling. Ideally, the pre-coding should be sufficiently robust and discriminating as to allow data processing by computer.

A coding frame should include information for the variable to be measured:

- the source data (e.g. Question 7 or 'achieved educational level');
- a list of the codes (e.g. number of years post-secondary school education);
- column location of the variable on the coded matrix.

Example 1:
The following numbers represent students' scores on a physics test:
19,23,17,27,21,20,17,22,19,17,25,21,29,24

A frequency table shows the distribution or number of students who achieved a particular score on the physics test. In Example 1, three students achieved a score of 17

Physics Score	Frequency	Percent	Percentile
17	3	21.4	21.4

19	2	14.3	35.7
20	1	7.1	42.9
21	2	14.3	57.1
22	1	7.1	64.3
23	1	7.1	71.4
24	1	7.1	78.6
25	1	7.1	85.7
27	1	7.1	82.9
29	1	7.1	100.0
Totals	14	100.0	

The following are the most common statistics used to describe frequency distributions:

N – the number of scores in a population

n – the number of scores in a sample

Percent – the proportion of students in a frequency distribution who had a particular score. In Example 1, 21% of the students achieved a score of 17.

Percentile – The percent of students in a frequency distribution who scored at or below a particular score (also referred to as percentile rank). In Example 1, 79% of the students achieved a score of 24 or lower, so a score of 24 is at the 79th percentile.

Mean – The average score in a frequency distribution. In Example 1, the mean score is 21.5. (Abbreviations for the mean are M if the scores are from a sample of participants and μ if the scores are from a population of participants.)

Median – The score in the middle of frequency distribution, or the score at the 50th percentile. In Example 1, the median score is 21.

Mode – The score that occurs most frequently in the distribution. In Example 1, the mode is 17.

Range – The difference between the highest and lowest score in the distribution. In Example 1, the range is 12.

Standard Deviation – A measure of how much the scores vary from the mean. In the sample, the standard deviation is 3.76, indicating that the average difference between the scores and mean is around 4 points. The higher the standard deviation, the more different the scores are from one another and from the mean. (Abbreviations for the standard deviation are *SD* if the scores are from a sample and Σ if the scores are from a population.)

The mean, median and mode are called measures of central tendency because they identify a single score as typical or representative of all the scores in a frequency distribution.

The design of a coding frame is also determined by the approach we take in respect of the data: what the data signifies, and useful ways of understanding the data once collected. After Swift (1996), three approaches can be identified:

a. Representational Approach

The response of the informant is said to express the surface meaning of what is "out there" requiring the researcher to apply codes to reduce the data, whilst at the same time, reflecting this meaning as faithfully as possible. At this stage of the process, the data must be treated independently from any views the researcher may hold about underlying variables and meanings.

b. Anchored-in Approach

The researcher may view the responses as having additional and implicit meanings that come from the fact that the responses are dependent on the data-gathering context. For example, in investigating worker involvement, we might want to conduct this with a framework comprising of types of formal and informal worker/manager interactions. As a consequence, the words given by informants can be interpreted to produce codes on more than one dimension relating to the context: (1) nature of the contact: formal versus informal, intermittent versus continuous contact, etc. (2) initiator of contact: worker versus manager. The coding frame using this approach takes into account "facts" as being anchored to the situation, rather than treating the data as though they are context-free.

c. Hypothesis-Guided Approach

Although similar to the second approach, we may view the data as having multiple meanings according the paradigm or theoretical perspective from which they are approached (e.g. phenomenological or hermeneutic approach to investigating a human or social phenomenon). The hypothesis-guided approach recognizes that the data do not have just one meaning which refers to some reality approachable by analysis for the surface meaning of the words: words have multiple meanings, and "out there" is a multiverse rather than a universe. In the hypothesis-guided approach, the researcher might use the data, and other materials, to create or investigate variables that are defined in terms of the theoretical perspective and construct propositions. For example, a data set might contain data on illness and minor complaints that informants had experienced over a period of say, one year. Taking the hypothesis-guided approach, the illness data might be used as an indicator of occupational stress or of a reaction to transformational change. Hence, the coding frame is based on the researcher'' views and hypotheses rather than on the surface meanings of the responses.

C. Analysis of Individual Observations

In the analysis of individual observations, or ungrouped data, consideration will be given to all levels of measurement to determine which descriptive measures can be used, and under what conditions each is appropriate.

One of the most widely used descriptive measures is the ‘average’. One speaks of the ‘average age’, average response time’, or ‘average score’ often without being very specific as to precisely what this means. The use of the average is an attempt to find a single figure to describe or represent a set of data. Since there are several kinds of 'average', or measures of central tendency, used in statistics, the use of precise terminology is important: each ‘average’ must be clearly defined and labelled to avoid confusion and ambiguity. At least three kinds of common uses of the ‘average’ can be described:

1. An average provides a summary of the data. It represents an attempt to find one figure that tells more about the characteristics of the distribution of data than any other. For example, in a survey of several hundred undergraduates the average intelligence quotient was 105: this one figure summarizes the characteristic of intelligence.
2. The average provides a common denominator for comparing sets of data. For example, the average score on the Job Descriptive Index for British managers was found to be 144, this score provides a quick and easy comparison of levels of felt job satisfaction with other occupational groups.
3. The average can provide a measure of typical size. For example, the scores derived for a range of dimensions of personality can be compared to the norms for the group the sample was taken from; thus, one can determine the extent to which the score for each dimension is above, or below, that to be expected.

1. The Mode

The mode can be defined as the most frequently occurring value in a set of data; it may be viewed as a single value that is most representative of all the values or observation in the distribution of the variable under study. It is the only measure of central tendency that can be appropriately used to describe nominal data. However, a mode may not exist, and even if it does, it may not be unique:

1 2 3 4 5 6 7 8 9 10	. . . .	No mode
Y Y N Y N N N N Y	. . . .	Unimodal (N)
1 2 2 3 4 4 4 4 5 5	. . . .	Unimodal (4)
1 2 2 2 3 4 5 5 5 6	. . . .	Bimodal (2, 5)
1 2 2 3 4 4 5 6 6 7	. . . .	Multimodal (2, 4, 6)

With relatively few observations, the mode can be determined by assembling the set of data into an array. Large numbers of observations can be arrayed by means of Microsoft EXCEL, or other statistical software programs:

Subject	Reaction Time (in m/seconds)	Array
000123	625	460
000125	500	480
000126	480	500
000128	500	500
000129	460	500
000131	500	**500** Mode
000134	575	510
000137	530	525
000142	525	530
000144	500	575
000145	510	625

2. The Median

When a measurement of a set of observation is at least ordinal in nature, the observations can be ranked, or sorted, into an array whereby the values are arranged in order of magnitude with each value retaining its original identity. The median can be defined as the value of the middle item of a set of items that form an array in ascending or descending order of rank: the [N+1}/2 position. In simple terms, the median splits the data into two equal parts, allowing us to state that half of the subjects scored below the median value and half the subjects scored above the median value. If an observed value occurs more than once, it is listed separately each time it occurs:

Subject	**Reaction Time in m/secs**	**Reaction Time shown in array: shortest to longest**	**Reaction Time shown in array: longest to shortest**
000123	625	460	625
000125	500	480	575
000126	480	500	530
000128	500	500	525
000129	460	500	510
000131	500	500	**500** Median
000134	575	510	500
000137	530	525	500
000142	525	530	500
000144	500	575	480
000145	510	625	460

3. The Arithmetic Mean

Averages much more sophisticated than the mode or median can be used at the interval and ratio level. The arithmetic mean is widely used because it is the most commonly known, easily understood and, in statistics, the most useful measure of central tendency. The arithmetic mean is usually given the notion ∑ and can be computed:

$$\sum [x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 \ + \dots \ x_n] / N$$

where, x_1, x_2, x_3 . . . x_n are the values attached to the observations; and, N is the total number of observations:

Subject	000123 x1	000125 x2	000126 x3	000128 x3	000129 x4	000131 x5	000134 x6
Reaction Time (m/secs)	625	500	480	500	460	500	575

Using the above formula, the arithmetic mean can be computed:

$$\bar{x} = \sum (x)/N \ = \ 4695/8 \ = \ 586.875 \text{ m/secs.}$$

The fact that the arithmetic mean can be readily computed does not mean that it is meaningful or even useful. Furthermore, the arithmetic mean has the weakness of being unduly influenced by small, or unusually large, values in a data set. For example: five subjects are observed in an experiment and display the following reaction times: 120, 57, 155, 210 and 2750 m/secs. The arithmetic mean is 658.4 m/secs, a figure that is hardly typical of the distribution of reaction times.

CHAPTER 2
DESCRIPTIVE STATISTICS

A. Introduction

Descriptive statistics is the discipline of quantitatively describing the main features of a collection of data. Descriptive statistics are distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. This generally means that descriptive statistics, unlike inferential statistics, are not developed on the basis of probability theory. Even when a data analysis draws its main conclusions using inferential statistics, descriptive statistics are generally also presented. For example in a paper reporting on a study involving human subjects, there typically appears a table giving the overall sample size, sample sizes in important subgroups (e.g., for each treatment or exposure group), and demographic or clinical characteristics such as the average age, the proportion of subjects of each sex, and the proportion of subjects with related comorbidities.

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

Descriptive statistics are typically distinguished from inferential statistics. With descriptive statistics you are simply describing what is or what the data shows. With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

Descriptive Statistics are used to present quantitative descriptions in a manageable form. In a research study we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simply large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary. For instance, consider a simple number used to summarize how well a batter is performing in baseball, the batting average. This single number is simply the number of hits divided by the number of times at bat (reported to three significant digits). A batter who is hitting .333 is getting a hit one time in every three at bats. One batting .250 is hitting one time in four. The single number describes a large number of discrete events. Or, consider the scourge of many students, the Grade Point Average (GPA). This single number describes the general performance of a student across a potentially wide range of course experiences.

Every time you try to describe a large set of observations with a single indicator you run the risk of distorting the original data or losing important detail. The batting average doesn't tell you whether the batter is hitting home runs or singles. It doesn't tell whether she's been in a slump or on a streak. The GPA doesn't tell you whether the student was in difficult courses or easy ones, or whether they were courses in their major field or in other disciplines. Even given these limitations, descriptive statistics provide a powerful summary that may enable comparisons across people or other units.

Descriptive statistics (DS) characterize the shape, central tendency, and variability of a set of data. When referring to a population, these characteristics are known as parameters; with sample data, they are referred to as statistics.

The description of a data set includes, among, other things:

(a) Presentation of the data by tables and graphs.

(b) Examination of the overall shape of the graphed data for important features, including symmetry or departures from it.

(c) Scanning the graphed data for any unusual observation that seems to stick far out from the major mass of the data.

(d) Computation of numerical measures for a typical or representative value of the center of the data.

(e) Measuring the amount of spread or variation present in the data.

Data (plural) are the measurements or observations of a variable. A **variable** is a characteristic that can be observed or manipulated and can take on different values.

1. The Population and the Sample

1.1.1.1.2 *Population:* A population is a complete collection of all elements (scores, people measurements, and so on). The collection is complete in the sense that it includes all subjects to be studied.

Sample: A sample is a collection of observations representing only a portion of the population.

Simple Random Sample: A Simple Random Sample (SRS) of measurements from a population is the one selected in such a manner that every sample of size *n* from the population has equal chance (probability) of being selected, and every member of the population has equal chance of being included in the sample.

Example 2.1 To draw a SRS, consider the data below as our population. In a study of wrap breakage during the weaving of fabric (Technometrics, 1982, p63), one hundred pieces of yarn were tested. The number of cycles of strain to breakage was recorded for each yarn and the resulting data are given in the following table.

86	175	157	282	38	211	497	246	393	198
146	176	220	224	337	180	182	185	396	264
251	76	42	149	66	93	423	188	203	105
653	264	321	180	151	315	185	568	829	203
98	15	180	325	341	353	229	55	239	124
249	364	198	250	40	571	400	55	236	137
400	195	38	196	40	124	338	61	286	135
292	262	20	90	135	279	290	244	194	350
131	88	61	229	597	81	398	20	277	193
169	264	121	166	246	186	71	284	143	188

2. Graphical Description of Data

a. Stem-and-Leaf Plot

One useful way to summarize data is to arrange each observation in the data into two categories “stems and leaves”. First of all we represent all the observations by the same number of digits possibly by putting 0/s at the beginning or at the end of an observation as needed, or by rounding. If there are *r* digits in an observation, the first *x* ($1 \leq x \leq r$) of them constitute stems and last ($r - x$) digits called leaves are put against stems. If there are many observations in a stem (in a row), they may be represented by two rows by defining a rule for every stem.

Example 1.2 Weaver (1990) examined a galvanized coating process for large pipes. Standards call for an average coating weight of 200 lbs per pipe. These data are the coating weights for a random sample of 30 student.

216	202	208	208	212	202	193	208	206	206
206	213	204	204	204	218	204	198	207	218
204	212	212	205	203	196	216	200	215	202

b. Frequency Tables

When summarizing a large set of data it is often useful to classify the data into classes or categories and to determine the number of individuals belonging to each class, called the class frequency. A tabular arrangement of data by classes together with the corresponding frequencies is called a frequency distribution or simply a frequency table. Consider the following definitions:

Class Width: The difference between the upper and lower class limit of a given class.
Frequency: The number of observations in a class.
Relative Frequency: The ratio of the frequency of a class to the total number of observations in the data set.
Cumulative Frequency: The total frequency of all values less than the upper class limit.
Relative Cumulative Frequency: The cumulative frequency divided by the total frequency.

1.1.2 Example 1.3 **Consider the data in Example 1.2. The steps needed to prepare a frequency distribution for the data set are described below:**

Step 1: Range = Largest observation – Smallest observation

$= 218 - 193 = 25$.

Step 2: Divide the range between into classes of (preferably) equal width. A rule of thumb for the number of classes is $\sqrt{n}$.

$$\text{Class width} \approx \frac{\text{Range}}{\text{Number of classes}}$$

Since we have a sample of size 30, the number of classes in the histogram should be around $\sqrt{30} = 5.48$. In this case, the class width would be approximately $25/5.48 = 4.56 \approx 5$. The smallest observation is 193. The first class boundary may well start at 193 or little below it say at 190 (just to avoid the smallest observation, in general, falling on the class boundary). Thus the first class is given by (190, 195]. The second class is given by (195, 200]. Complete the class boundaries for all classes. In Statistica, the lower boundary of the first class is called the starting point, the class width or step size.

Step 3: For each class, count the number of observations that fall in that class. This number is called the class frequency.

Step 4: The relative frequency of a class is calculated by *f*/*n* where *f* is the frequency of the class and *n* is the number of observations in the data set.

Cumulative Relative Frequency of a class, denoted by *F,* is the total of the relative frequencies up to that class. To avoid rounding in every class, one may cumulate the frequencies up to a class and then divide by *n*. The resulting quantity Relative Cumulative Frequency (*F/n*) is just the same as Cumulative Relative Frequency. It is desirable in a frequency table. For the data in Example 2.2, we have the following frequency distribution:

Class	Count	*f*	*F*	Relative *f*	Relative *F*
(190, 195]	/	1	1	0.033...	0.033
(195, 200]	//	2	3	0.066...	0.100
(200, 205]	///// /////	10	13	0.333...	0.433
(205, 210]	///// ///	8	21	0.266...	0.700
(210, 215]	////	4	25	0.133...	0.833
(215, 220]	/////	5	30	0.166...	1.000
		30		1	

c. Graphs with Frequency Distributions

1) Frequency Histogram

A frequency histogram is a bar diagram where a bar against a class represents frequency of the class.

2) Frequency Plots

The data of Example 2.2 have been summarized by a frequency distribution in Figure 2.4. While we are in ***Basic Statistics and Tables,*** we may use Figure 2.4, frequency distribution to enter the midpoint of each interval in one column of the datasheet, another column to enter the count (frequency) of each interval (relative frequencies, cumulative relative frequencies can also be entered in two other columns).

Use frequency or relative frequency or cumulative relative frequency as vertical axis as needed by the graph.

(***a***) ***Frequency Plot***: If frequencies of classes are plotted against the mid values of respective classes, the resulting scatter graph is called a Frequency Plot.
(***b***) ***Frequency Curve***: If the dots of the frequency plot are joined by a smooth curve the resulting curve is called a frequency curve.
(c) Frequency Polygon: If the dots in a frequency plot are joined by lines, the resulting graph is called a Frequency Polygon. The polygon is sometimes extended to the midpoints of extreme adjacent classes (in both sides) with no frequencies.

d. Bar Chart and Pie Chart

Both bar and pie charts are used to represent discrete and qualitative data. A bar graph is a graphical representation of a qualitative data set. It gives the frequency (or relative frequency) corresponding to each category, with the height or length of the bar proportional to the category frequency (or relative frequency). The relative frequency of a category is calculated by f/n where f is the frequency of a category and n is the number of observations in the data set.

1) Bar Chart

To make a bar chart, the classes are marked along the horizontal axis and a vertical bar of height equal to the class frequency is erected over the respective classes.

2) Pie chart

A Pie chart is made by representing the relative frequency of a category by an angle of a circle determined by:

$$\text{Angle of a category} = \text{Relative frequency of the category} \times 360$$

e. Numerical Measures

Sometimes we are interested in a number which is representative or typical of the data set. Sample mean or median is such a number. Similarly, we define the range of the sample which gives some idea about the variation or dispersion of observations in the sample. The most important measure for dispersion is the sample standard deviation.

f. Box Plot

A box aligned with first and the third quartiles as edges, median at the appropriate place in the scale is called a box plot. It is extended to both directions up to the smallest and the largest values. These extensions may be called arms. This technique displays the structure of the data set by using the quartiles and the extreme values of a sample. The qurartiles Q_1, Q_2 and Q_3 are three values that divide the ordered sample observations in 4 quarters approximately.

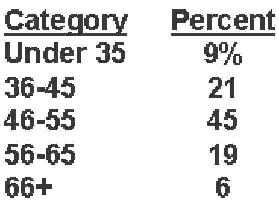
B. Univariate Analysis

Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

- the distribution
- the central tendency
- the dispersion

In most situations, we would describe all three of these characteristics for each of the variables in our study.

The Distribution. The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of persons who had each value. For instance, a typical way to describe the distribution of college students is by year in college, listing the number or percent of students at each of the four years. Or, we describe gender by listing the number or percent of males and females. In these cases, the variable has few enough values that we can list each one and summarize how many sample cases had the value. But what do we do for a variable like income or GPA? With these variables there can be a large number of possible values, with relatively few people having each one. In this case, we group the raw scores into categories according to ranges of values. For instance, we might look at GPA according to the letter grade ranges. Or, we might group income into four or five ranges of income values.



Category	Percent
Under 35	9%
36-45	21
46-55	45
56-65	19
66+	6

Table 1. Frequency distribution table.

One of the most common ways to describe a single variable is with a *frequency distribution*. Depending on the particular variable, all of the data values may be represented, or you may group the values into categories first (e.g., with age, price, or temperature variables, it would usually not be sensible to determine the frequencies for each value. Rather, the value are grouped into ranges and the frequencies determined.). Frequency distributions can be depicted in two ways, as a table or as a graph. Table 1 shows an age frequency distribution with five categories of age ranges defined. The same frequency distribution can be depicted in a graph as shown in Figure 2. This type of graph is often referred to as a histogram or bar chart.

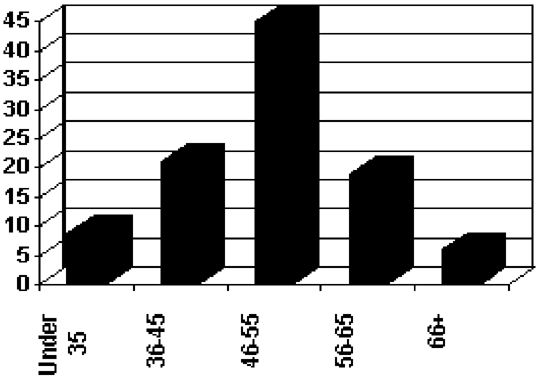


Figure 2. Frequency distribution bar chart.

Distributions may also be displayed using percentages. For example, you could use percentages to describe the:

- percentage of people in different income levels
- percentage of people in different age ranges
- percentage of people in different ranges of standardized test scores

Central Tendency. The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

The **Mean** or average is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values. For example, the mean or average quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

15, 20, 21, 20, 36, 15, 25, 15

The sum of these 8 values is 167, so the mean is 167/8 = 20.875.

The **Median** is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score #250 would be the median. If we order the 8 scores shown above, we would get:

15,15,15,20,20,21,25,36

There are 8 scores and score #4 and #5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, you would have to interpolate to determine the median.

The **mode** is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores as shown above, and then count each one. The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the model. In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently.

Notice that for the same set of 8 scores we got three different values -- 20.875, 20, and 15 -- for the mean, median and mode respectively. If the distribution is truly normal (i.e., bell-shaped), the mean, median and mode are all equal to each other.

Dispersion. Dispersion refers to the spread of the values around the central tendency. There are two common measures of dispersion, the range and the standard deviation. The **range** is simply the highest value minus the lowest value. In our example distribution, the high value is 36 and the low is 15, so the range is 36 - 15 = 21.

The **Standard Deviation** is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values. The Standard Deviation shows the relation that set of scores has to the mean of the sample. Again lets take the set of scores:

15,20,21,20,36,15,25,15

to compute the standard deviation, we first find the distance between each value and the mean. We know from above that the mean is 20.875. So, the differences from the mean are:

15 - 20.875 = -5.875
20 - 20.875 = -0.875
21 - 20.875 = +0.125
20 - 20.875 = -0.875
36 - 20.875 = 15.125
15 - 20.875 = -5.875
25 - 20.875 = +4.125
15 - 20.875 = -5.875

Notice that values that are below the mean have negative discrepancies and values above it have positive ones. Next, we square each discrepancy:

-5.875 * -5.875 = 34.515625
-0.875 * -0.875 = 0.765625
+0.125 * +0.125 = 0.015625
-0.875 * -0.875 = 0.765625
15.125 * 15.125 = 228.765625
-5.875 * -5.875 = 34.515625
+4.125 * +4.125 = 17.015625
-5.875 * -5.875 = 34.515625

Now, we take these "squares" and sum them to get the Sum of Squares (SS) value. Here, the sum is 350.875. Next, we divide this sum by the number of scores minus 1. Here, the result is 350.875 / 7 = 50.125. This value is known as the **variance**. To get the standard deviation, we take the square root of the variance (remember that we squared the deviations earlier). This would be SQRT(50.125) = 7.079901129253.

Although this computation may seem convoluted, it's actually quite simple. To see this, consider the formula for the standard deviation:

$$\sqrt{\frac{\sum(X-\bar{X})^2}{(n-1)}} \qquad (2.1)$$

X= each score
$\bar{X}$ = the mean or average
n = the number of values
$\sum$ means we sum across the valuaes

In the top part of the ratio, the numerator, we see that each score has the the mean subtracted from it, the difference is squared, and the squares are summed. In the bottom part, we take the number of scores minus 1. The ratio is the variance and the square root is the standard deviation. In English, we can describe the standard deviation as:

the square root of the sum of the squared deviations from the mean divided by the number of scores minus one

Although we can calculate these univariate statistics by hand, it gets quite tedious when you have more than a few values and variables. Every statistics program is capable of calculating them easily for you. For instance, I put the eight scores into SPSS and got the following table as a result:

N	8
Mean	20.8750
Median	20.0000
Mode	15.00
Std. Deviation	7.0799
Variance	50.1250
Range	21.00

which confirms the calculations I did by hand above.

The standard deviation allows us to reach some conclusions about specific scores in our distribution. Assuming that the distribution of scores is normal or bell-shaped (or close to it!), the following conclusions can be reached:

- approximately 68% of the scores in the sample fall within one standard deviation of the mean
- approximately 95% of the scores in the sample fall within two standard deviations of the mean
- approximately 99% of the scores in the sample fall within three standard deviations of the mean

For instance, since the mean in our example is 20.875 and the standard deviation is 7.0799, we can from the above statement estimate that approximately 95% of the scores will fall in the range of 20.875-(2*7.0799) to 20.875+(2*7.0799) or between 6.7152 and 35.0348. This kind of information is a critical stepping stone to enabling us to compare the performance of an individual on one variable with their performance on another, even when the variables are measured on entirely different scales.

C. Descriptive statistics for measurements of a single variable

1. The basic idea

We now deal with descriptive statistics for measurements of a single variable. It is imagined that we have a large population of values from which we take samples. The population could consist of the diameters of automobile drive shafts produced in a given plant. To make sure the manufacturing equipment continues to operate satisfactorily, we measure the diameter of every tenth drive shaft.[1] The measurements over a given time period are called “samples” of the “population” of all drive shafts. The measurements will vary somewhat, both because of finite tolerances in the manufacturing equipment and because of uncertainties in the measurements themselves. From the samples, we wish to make judgments about the underlying population, i.e. the actual diameters of all drive shafts made. For example, the mean (average) of the samples is expected to be approximately the true unknown mean of the population. The accuracy of this sample estimate of the population mean would be expected to improve as the sample size is increased. For example, if we measured every other drive shaft, we would expect the mean of our measurements to become closer to the actual average diameter of all drive shafts than when we measured only 1/10 of them.

One of the primary objectives of statistics is to make quantitative statements. For example, rather than just saying that the average drive shaft diameter is approximately equal to the sample mean, we’d like to give a range of diameters within which the true mean lies with a probability of 95%.

2. The normal distribution

The most common assumption made in statistical treatments of data is that the probability of a particular value x deviating from the population mean μ is inversely proportional to the square of its deviation from the mean. This gives rise to the familiar “bell-shaped curve” normal probability density function:

[1] While we could measure every drive shaft, this is unnecessarily expensive.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{2.2}$$

where σ^2 the population variance, which is the mean of all values of $(x - \mu)^2$. The factor $1/\sigma\sqrt{2\pi}$ was chosen so that $\int_{-\infty}^{+\infty} f(x)dx = 1$. The probability that a given sample x lies between *a* and *b* is $\int_a^b f(x)dx$,[2] which gives the fundamental meaning of the probability density function f.

To illustrate the normal distribution, we present on the next page a MATLAB program to generate normally-distributed random numbers and compare the resulting histogram with equation 2.2. To save time, you can cut and paste this program into MATLAB's Editor, save in your working directory as ranhys.m, and then execute in MATLAB's Command window by typing >> **ranhys**. Try it for several values of the mean, variance and number of values, n. Notice how the histogram approaches the shape[3] of the normal distribution better and better as n is increased. A histogram for $\mu = 5$, $\sigma^2 = 2$ and n = 500 is given as Figure 2.1 on the next page.

```
% ranhys.m        W.R. Wilcox, Clarkson University, 1 June 2004.
% Comparison of a histogram of normally distributed
% random numbers with a normal distribution.
% n is the number of samples
% sigma is the sample standard deviation
% mu is the sample mean
% X is the vector of values
clear
n = input('Enter the number of values to be generated ');
mu = input('Enter the population mean ');
sigsq = input('Enter the population variance ');
sigma = sqrt(sigsq);
% Set the state for the random number generator
% (See >>help randn)
randn('state',sum(100*clock));
% Generate the random numbers desired
X = mu + sigma*randn(n,1);
% Plot the histogram with 10 bins (see >> help hist)
hist(X,10), xlabel('value'), ylabel('number in bin')
h = findobj(gca,'Type','patch');
set(h,'FaceColor','m','EdgeColor','w')
hold on
% Now create a curve for the normal distribution
% (with a maximum equal to 1/4 of the number of values n)
x = mu-4*sigma:sigma/100:mu+4*sigma;
```

[2] That is, the area under the f(x) curve between *a* and *b*.

[3] Compare **only** the shape, as here the maximum in the normal distribution is arbitrarily set to n/4.

```
f = 0.25*n*exp(-(x-mu).^2/2/sigma.^2);
plot(x,f), legend('random number', 'normal distribution')
title('Comparison of random number histogram with normal distribution
shape')
hold off
```

Do you get the same histogram if you use the same values again for μ , σ^2 and n? Examine the code until you understand why.

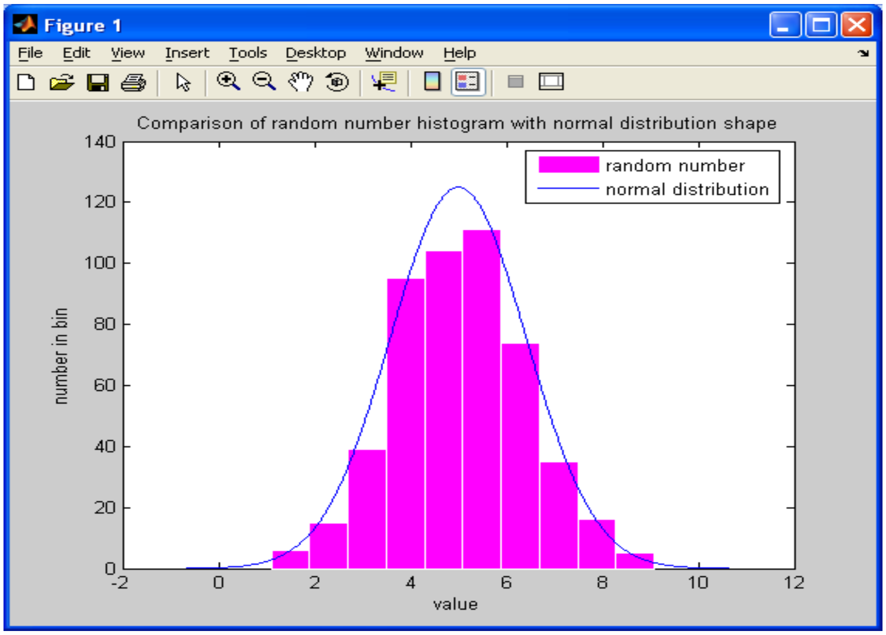


Figure 1
File Edit View Insert Tools Desktop Window Help
Comparison of random number histogram with normal distribution shape
random number
normal distribution
number in bin
value
0
20
40
60
80
100
120
140
-2
0
2
4
6
8
10
12

Figure 3. Sample histogram for μ = 5, σ^2 = 2 and n = 500.

See http://www.shodor.org/interactivate/activities/NormalDistribution/ for a graphical illustration of the influence of population standard deviation on the normal distribution and the influence of bin size on a histogram.

3. Tests to see if a population is normally distributed

Although normally distributed populations are common, many other distributions are known. If you have a set of data, how can you determine if the underlying population is normally distributed? The short answer is that you cannot be 100% sure, as is typical of questions in statistics. But there are several tests you can use to see if the answer is probably "yes."

Method 1: Prepare a histogram and see if it looks normal. This is only effective if the sample size is very large.

Method 2: A better method, particularly for smaller sample sizes, is to prepare a cumulative distribution plot. The cumulative distribution is the fraction F of the sample values that are less than or equal to each particular value x. A plot of F versus x can be compared to the cumulative distribution for a normal probability density function. Integrating equation 2.1 we obtain:

$$F = \int_{-\infty}^{x} f(t)dt = \frac{1}{2}\left(1 + \mathrm{erf}\,\frac{x-\mu}{\sqrt{2}\sigma}\right) \tag{2.3}$$

where t is a dummy variable for integration and $\mathrm{erf}(z) \equiv \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$ is called the error function, and is calculated by MATLAB using the command, for example, >> **erf(0.5)** .

Beginning below is a MATLAB program to generate normally distributed random numbers and plot the cumulative distribution versus that given by equation 2.2. Copy this into your MATLAB Editor, save it in your working directory as cumdist.m and execute >> **cumdist** in the MATLAB Command window. Test the program for different values of the mean, variance and number of values. Note how the resulting values become nearer and nearer the curve for a normal distribution (equation 2.2) as the number of values is increased.

```
% cumdist.m    W.R. Wilcox, Clarkson University, 2 June 2004.
% Comparison of cumulative distribution
% for normally distributed random numbers
% with integrated normal distribution
% mu = population mean
% sigma = square root of population variance
% n = number of samples from population
% X = vector of sample values
clear, clc
% Input the desired values of n, mu, sigma
n = input('Number of values to be generated:  ');
mu = input('Desired population mean:  ');
```

```
sigsq = input('Desired population variance:  ');
sigma = sqrt(sigsq);
% Set the state for the random number generator
% (See >>help randn)
randn('state',sum(100*clock));
% Generate the random numbers desired
X = mu + sigma*randn(n,1);
% Sort the numbers
X = sort(X);
j = 1:n;
% Generate the cumulative normal distribution curve:
x = mu-4*sigma:sigma/100:mu+4*sigma;
F = 1/2*(1+erf((x-mu)/sqrt(2)/sigma));
plot(X,(j-0.5)/n,x,F);
xlabel('x');
ylabel('fraction of values < x')
legend('samples','normal distribution','Location','SouthEast')
title('Cumulative distribution')
```

Method 3: An even better method is to plot the cumulative distribution on a scale that would give a straight line if the distribution were normal. This is done by making the vertical scale erfinv(2*F-1), where erfinv is the inverse error function (i.e., x = erfinv(y) satisfies y = erf(x)). Below is a MATLAB program that is the same as that above, except for the vertical scale. Copy it into your MATLAB Editor, save it in your working directory as cumdistp2.m and execute >> **cumdistp2** in the MATLAB Command window. Test the program for different values of the mean, variance and number of values. Note the resulting values become nearer and nearer the straight line for a normal distribution (equation 2.2) as the number of values is increased, except for the very high and the very low values.

```
% cumdistp2.m    W.R. Wilcox, Clarkson University, 2 November 2002.
% Plot of cumulative distribution
% for normally distributed random numbers
% on normal distribution probability scale
% mu = mean
% sigma = population standard deviation
% n = number of samples from population
% X = vector of sample x values
% F = fraction of values < x (cumulative distribution)
% z = erfinv(2F-1) (search erfinv in MATLAB help)
% For normal distribution get straight line for z versus x
% Note that F =(1 + erf z)/2
clear, clc
% Input the desired values of n, mu, sigma
n = input('Number of values to be generated:  ');
mu = input('Desired population mean:  ');
sigsq = input('Desired population variance:  ');
sigma = sqrt(sigsq);
% Set the state for the random number generator
% (See >>help randn)
randn('state',sum(100*clock));
% Generate the random numbers desired
X = mu + sigma*randn(n,1);
% Sort the numbers
X = sort(X);
```

```
% Generate z
j=(1:n)'; F=(j-1/2)/n; z=erfinv(2*F-1);
% Calculation of the normal distribution line:
Xn(1)=mu-2*sqrt(2)*sigma; Xn(2)=mu+2*sqrt(2)*sigma;
zn=[-2,2];
plot(X,z,'o',Xn,zn);
xlabel('x'); ylabel('z = erfinv(2F-1)');
title('Cumulative distribution using a normal probability scale')
legend('samples','normal distribution','Location','SouthEast')
```

An example for 100 values with a population mean of 0 and a variance of 1 is given in Figure 2.2.

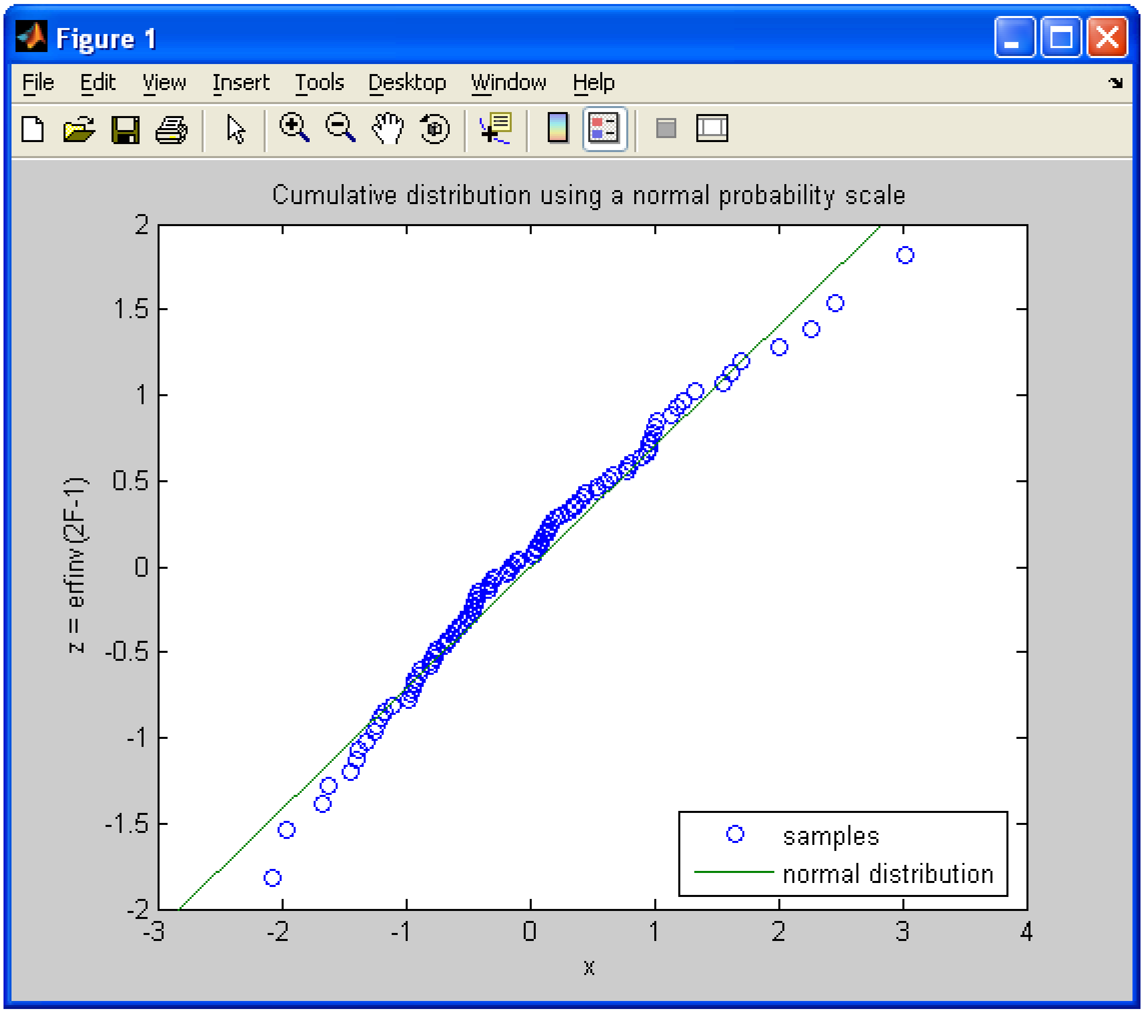


Figure 5. Normally distributed "data" from a population with a mean of 0 and a variance of 1.

Method 4: While Method 3 is useful, it is not quantitative. The "skewness" and "kurtosis" constitute quantitative measures of the normalcy of data. The qualitative definitions below were taken from the PROPHET StatGuide: Glossary at the Northwestern University Medical School. They refer to histograms of the data.

> "Skewness is a lack of symmetry in a distribution. Data from a positively skewed (skewed to the right) distribution have values that are bunched together below the mean, but have a long tail above the mean. (Distributions that are forced to be positive, such as annual income, tend to be skewed to the right.) Data from a negatively skewed (skewed to the left) distribution have values that are bunched together above the mean, but have a long tail below the mean."
>
> "Kurtosis is a measure of the heaviness of the tails in a distribution, relative to the normal distribution. A distribution with negative kurtosis (such as the uniform distribution) is light-tailed relative to the normal distribution, while a distribution with positive kurtosis (such as the Cauchy distribution) is heavy-tailed relative to the normal distribution."

Mathematically, skewness and kurtosis are measured via:[4]

skewness = $g_1 \equiv k_3/k_2^{3/2} = k_3/s^3$ and kurtosis = $g_2 \equiv k_4/k_2^2 = k_3/s^4$

(2.3)

where $s = k_2^{1/2}$ is the standard deviation and

$$k_2 \equiv \frac{nS_2 - S_1^2}{n(n-1)}$$

(2.4)

$$k_3 \equiv \frac{n^2 S_3 - 3nS_2S_1 + 2S_1^3}{n(n-1)(n-2)}$$

(2.5)

$$k_4 \equiv \frac{(n^3 + n^2)S_4 - 4(n^2 + n)S_3S_1 - 3(n^2 - n)S_2^2 + 12nS_2S_1^2 - 6S_1^4}{n(n-1)(n-2)(n-3)}$$

(2.2)

$$S_r \equiv \sum_{i=1}^{n} x_i^r$$

(2.7)

with x_i being the ith value of n samples from the population.

d. **Confidence limits on the mean**

From a sample, we can calculate the upper and lower limits for the unknown population mean μ with desired probability 1-α using the following equation:[4]

$$\mu = \bar{x} \pm \frac{ts}{\sqrt{n}}$$

(2.8)

[4] From `function t(nu,alpha)`
sections 2.43, 4.33 and 5.33 of "Statistical Analysis in Chemistry and the Chemical Industry," by C.A. Bennett and N.L. Franklin, Wiley, NY (1954).

where $\bar{x}$ is the sample mean, s is the sample standard deviation, and t is Student's t, which is a function of α and the degrees of freedom ν (nu). For a single variable, as considered here, ν=n-1. The relationship between α, ν and t is given by the Incomplete Beta Function, which in MATLAB is called by the name betainc (see >> **help betainc**) and is, specifically:[5]

$$\alpha = \text{betainc}\left(\frac{\nu}{\nu + t^2}, \frac{\nu}{2}, \frac{1}{2}\right) \tag{2.9}$$

To test your understanding, find α for t = 2.2281 for a sample of 11 values.

Unfortunately, MATLAB does not have an inverse incomplete beta function that would allow one to find t for a given α and ν. Consequently, the following MATLAB function was created that gives t to within 0.001.

```
% Calculation of Student's t from nu and alpha
% W.R. Wilcox, Clarkson University, April 2005
% nu is the degrees of freedom
% alpha is the fractional uncertainty
% A normal distribution of possible values is assumed
% Accurate to within 0.001
tp = 0.2:0.001:200;
res = abs(betainc(nu./(nu+tp.^2),nu/2,1/2)-alpha);
[minres,m]=min(res);
if tp(m) == 200
    fprintf('Student''s t is >= 200.  Choose a larger alpha.\n')
elseif tp(m) == 0.2
    fprintf('Student''s t is <= 0.2.  Choose a smaller alpha.\n')
else fprintf('Student''s t is %4.3f\n', tp(m));
end
end
```

This function was used to prepare Figure 2.3. The curve for ν = 1000 is indistinguishable from that for a normal distribution, which corresponds to $\nu = \infty$ and in MATLAB is given by:

$$t = \sqrt{2}\text{erfinv}(1 - \alpha) \tag{2.10}$$

It is interesting to compare the customary bell-shaped normal probability density function given by Equation 2.1 with that for Student's t, which is given by

$$f(x) = \frac{\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}}{sB\left(\frac{1}{2}, \frac{\nu}{2}\right)\sqrt{\nu}} \tag{2.11}$$

where $t = \frac{x - \bar{x}}{s}$, $\bar{x}$ is the sample mean, $\nu = n - 1$ is known as the degrees of freedom, and

[5] See equations 22.5.1 & 22.5.27 and section 22.7 in "Handbook of Mathematical Functions," edited by M. Abramowitz and I.A. Stegun, Dover, NY (1925).

$$B(z,w) = \int_0^1 y^{z-1}(1-y)^{w-1}dy$$

(2.12)

is the beta function, which MATLAB calculates using the function beta(z,w). The program on the next page permits one to input arbitrary values of the mean and standard deviation. The output is the probability density function for normal distribution and for selected values of ν (i.e., sample size minus 1). Figure 2.4 shows the result. Again note that the Student's t distribution has broader tails and approaches the normal distribution as the sample size increases.

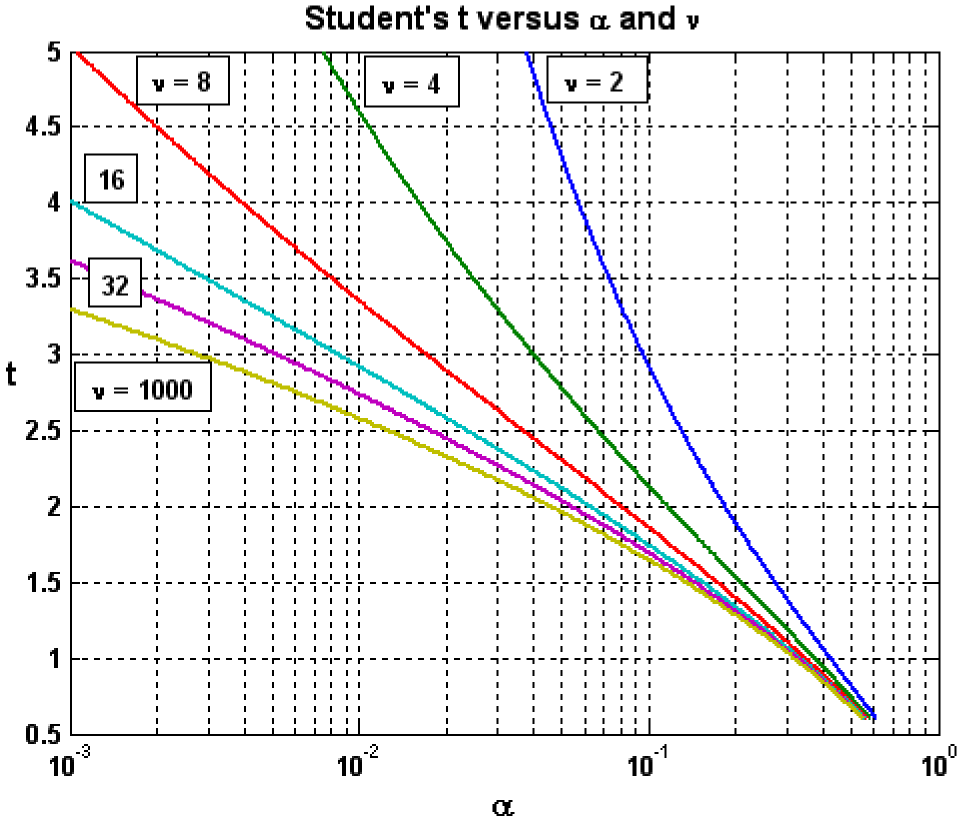


Figure 2.3. Plot of Student's t versus α and ν

```
% Norm_vs_stud_t.m
% William R. Wilcox, Clarkson University, November 2002
%Comparison of the probability density function for a normal distribution
%with that for a Student's t distribution, with a population mean mu
% and standard deviation sigma input.
mu = input('Enter the mean: ');
sigma = input('Enter the standard deviation: ');
x=mu-4*sigma:sigma/10:mu+4*sigma;
fx_norm = (1/sigma/sqrt(2*pi))*exp(-(x-mu).^2/2/sigma^2);
t=(x-mu)/sigma;
nu = 2;
fx_stud2 = 1/sigma*(1 + t.^2/nu).^(-(nu+1)/2)/sqrt(nu)/beta(1/2,nu/2);
nu = 4;
fx_stud4 = 1/sigma*(1 + t.^2/nu).^(-(nu+1)/2)/sqrt(nu)/beta(1/2,nu/2);
nu = 8;
```

```
fx_stud8 = 1/sigma*(1 + t.^2/nu).^(-(nu+1)/2)/sqrt(nu)/beta(1/2,nu/2);
nu = 12;
fx_stud12 = 1/sigma*(1 + t.^2/nu).^(-(nu+1)/2)/sqrt(nu)/beta(1/2,nu/2);
nu = 32;
fx_stud32 = 1/sigma*(1 + t.^2/nu).^(-(nu+1)/2)/sqrt(nu)/beta(1/2,nu/2);
plot(x,fx_norm,x,fx_stud2,x,fx_stud4,x,fx_stud8,x,fx_stud12,x,fx_stud32);
xlabel('x'); ylabel('f(x)');
titletext=['Probability density functions for mean of ',num2str(mu),' and SD
of ',num2str(sigma)];
title(titletext); legend('normal','stud2','stud4','stud8','stud12','stud32')
```

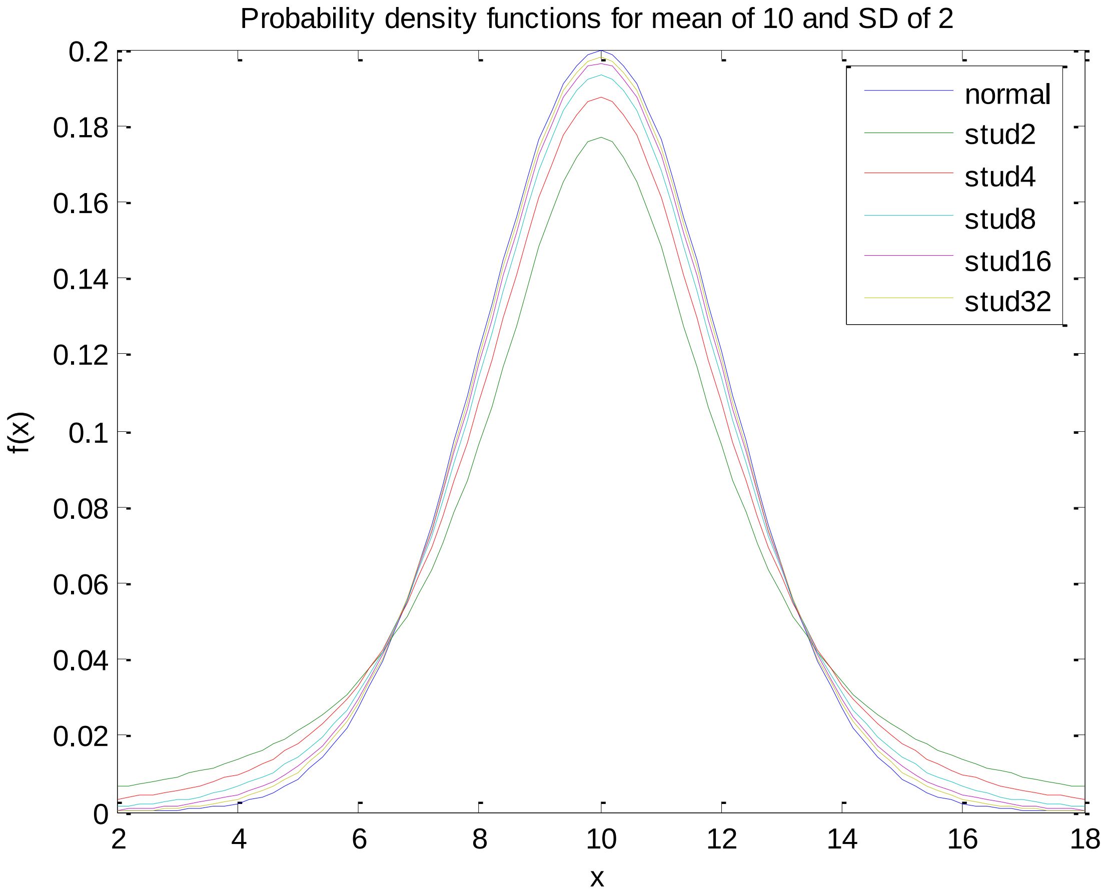


Figure 2.4. Comparison of probability density function f(x) for a normal probability (Eq 2.1) with that for Student's t (Eq 2.11) with sample sizes of 3, 5, 9, 17 and 33.

D. Summary: descriptive statistics function

The descriptive statistics function has been prepared to make it easy for you to do these calculations using MATLAB. Click on descriptive statistics function and go to File, Save as to save it in your working directory. To execute it, >> **descript(X)**, where X is the variable vector containing the data. To illustrate the use of this function, we consider the weight of adult field mice in St. Lawrence County. We trap mice using a **Have-a-heart trap**, weigh them, and then release them. The resulting data are contained in mouse weights for sample sizes of 2, 3, 4, 5, 10 and 100 (that's a lot of mice!). Save this file in your working directory.

Now test the function descript. Load mouse.mat into MATLAB by ***File, Import Data***, mouse, and then ***Finish*** in the Import Wizard window. Then the following:

```
>> descript(M2)
>> descript(M3)
>> descript(M4)
>> descript(M5)
>> descript(M10)
>> descript(M100)
```

What can you conclude from these results? [6]

A google search reveals that there are a variety of textbooks and websites dealing with statistics, ranging from theory to history to on-line computational engines. A couple of particularly useful web sites are:

Web Pages that Perform Statistical Calculations
NIST/SEMATECH e-Handbook of Statistical Methods

Statistics is a deep subject of great usefulness for engineers and scientists. Hopefully, you will want to learn more.

[6] From this we can conclude that as more mice were weighed the resulting distribution became more normal and that the difference between the confidence limits for the mean decreased. Note that these "data" were created using a population mean of 30 (g) and a variance of 25 ($\sigma = 5$). Knowing these parameters (which we usually don't) we can also see that as the sample size increases the sample mean becomes closer and closer to the population mean, and that the standard deviation of the sample approaches the square root of the population variance.

CHAPTER 3
PROBABILITY

A. Introduction

After careful study of this chapter you should be able to do the following:

a. Understand and describe sample spaces and events for random experiments with graphs, tables, lists, or tree diagrams

b. Interpret probabilities and use probabilities of outcomes to calculate probabilities of events in discrete sample spaces

c. Calculate the probabilities of joint events such as unions and intersections from the probabilities of individual events

d. Interpret and calculate conditional probabilities of events

e. Determine the independence of events and use independence to calculate probabilities

f. Use Bayes' theorem to calculate conditional probabilities

g. Understand random variables

B. Sample Spaces And Events

The **sample space** of a random experiment is a set *S* that includes all possible outcomes of the experiment; the sample space plays the role of the universal set when modeling the experiment. For simple experiments, the sample space may be precisely the set of possible outcomes. More often, for complex experiments, the sample space is a mathematically convenient set that includes the possible outcomes and perhaps other elements as well. For example, if the experiment is to throw a standard die and record the outcome, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$, the set of possible outcomes. On the other hand, if the experiment is to capture a cicada and measure its body weight (in milligrams), we might conveniently take the sample space to be $S = [0, \infty)$, even though most elements of this set are practically impossible.

Certain subsets of the sample space of an experiment are referred to as **events**. Thus, an event is a set of outcomes of the experiment. Each time the experiment is run, a given event *A* either *occurs*, if the outcome of the experiment is an element of *A*, or *does not occur*, if the outcome of the experiment is not an element of *A*. Intuitively, you should think of an event as a meaningful *statement* about the experiment.

The sample space *S* itself is an event; by definition it *always* occurs. At the other extreme, the empty set Ø is also an event; by definition it *never* occurs. More generally, if *A* and *B* are events in the experiment and *A* is a subset of *B*, then the occurrence of *A* *implies* the occurrence of *B*.

Set theory is the foundation of probability, as it is for almost every branch of mathematics. In probability, set theory is used to provide a language for modeling and describing random experiments.

Definition S : sample space, all possible outcomes

Example: tossing a coin, $S = \{H, T\}$

Example: reaction time to a certain stimulus, $S = (0, \infty)$

Sample space: may be countable or uncountable

Countable: put 1-1 correspondence with a subset of integers

Finite elements $\Rightarrow$ countable

Infinite elements $\Rightarrow$ countable or uncountable

Fact: There is only countable sample space since measurements cannot be made with infinite accuracy

Definition event: any measurable collection of possible outcomes, subset of S

If $A \subset S$, A occurs if outcome is in the set A.

$P(A)$: probability of an event (rather than a set)

Theorem A, B, C: events on S

(1). Commutativity $A \cup B = B \cup A$, $A \cap B = B \cap A$

(2). Associativity $A \cup (B \cup C) = (A \cup B) \cup C$, $A \cap (B \cap C) = (A \cap B) \cap C$

(3). Distributive Laws $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

(4). DeMorgan's Law $(A \cup B)^c = A^c \cap B^c$, $(A \cap B)^c = A^c \cup B^c$

Example: (a) If $S = (0,1]$, $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} [\frac{1}{i},1] = (0,1]$, $\bigcap_{i=1}^{\infty} A_i = \bigcap_{i=1}^{\infty} [\frac{1}{i},1] = \{1\}$.

(b) If $S = (0,1)$, $\bigcap_{i=1}^{\infty} A_i = \bigcap_{i=1}^{\infty} [\frac{1}{i},1] = \phi$.

Definition (a) A, B are disjoint if $A \cap B = \phi$.

(b) $A_1, A_2 \cdots$ are pairwise disjoint if $A_i \cap A_j = \phi$, $\forall i \neq j$.

Definition $A_1, A_2 \cdots$ are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = S$, then $A_1, A_2 \cdots$ form a partition of S.

Number of Outcomes of an Event

As an example, we may have an event *E*defined as
E = "day of the week"

We write the "number of outcomes of event*E*" as *n*(*E*).
So in the example, *n*(*E*) = 7, since there are 7 days in the week.

Addition Rule

Let E_1 and E_2 be **mutually exclusive** events (i.e. there are no common outcomes).
Let event *E* describe the situation where either event E_1 **or** event E_2 will occur.
The number of times event *E* will occur can be given by the expression:

$n(E) = n(E_1) + n(E_2)$

Tip
In counting and probability,**"OR**" usually requires us to**ADD**.
where
$n(E)$ = Number of outcomes of event E
$n(E_1)$ = Number of outcomes of event E_1
$n(E_2)$ = Number of outcomes of event E_2
[We see more on mutually exclusive events later in this chapter.]

Example 1
Consider a set of numbers S = {-4, -2, 1, 3, 5, 6, 7, 8, 9, 10} Let the events E_1, E_2 and E_3 be defined as:

E = choosing a negative or an odd number from S;
E_1= choosing a negative number from S;
E_2 = choosing an odd number from S.
Find $n(E)$.

Example 2
In how many ways can a number be chosen from 1 to 22 such that
(a) it is a multiple of 3 or 8? (b) it is a multiple of 2 or 3?

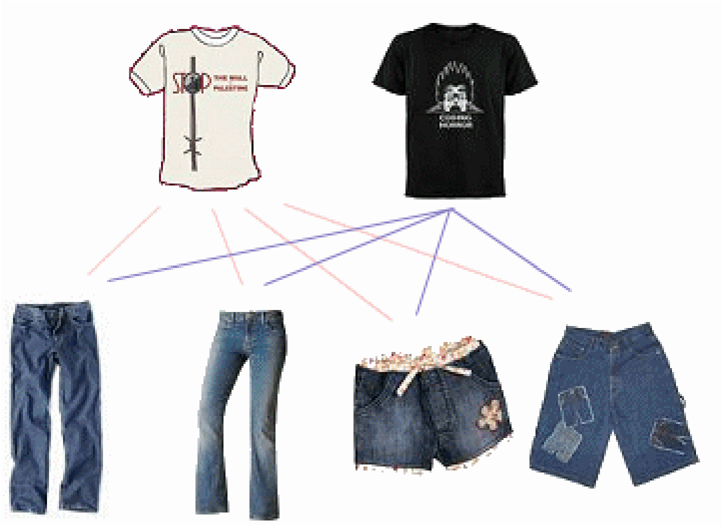
Multiplication Rule

Now consider the case when two events E_1 and E_2 are to be performed and the events E_1 and E_2 are **independent** events i.e. one does not affect the other's outcome.

Example

Say the only clean clothes you've got are 2 t-shirts and 4 pairs of jeans. How many different combinations can you choose?

Answer
We can think of it as follows:



We have 2 t-shirts and with each t-shirt we could pick 4 pairs of jeans. Altogether there are 2 × 4 = 8 possible combinations.

We could write

E_1 = "choose t-shirt" and

E_2 = "choose jeans"

Multiplication Rule in General

Suppose that event E_1 can result in any one of $n(E_1)$ possible outcomes; and for each outcome of the event E_1, there are $n(E_2)$ possible outcomes of event E_2.

Together there will be $n(E_1) \times n(E_2)$ possible outcomes of the two events.

Tip

In counting and probability,**"AND"** usually requires us to **MULTIPLY**.

That is, if event E is the event that both E_1 and E_2 **must**occur, then

$n(E) = n(E_1) \times n(E_2)$

In our example above,

$n(E_1) = 2$ (since we had 2 t-shirts)

$n(E_2) = 4$ (since there were 4 pairs of jeans)

So total number of possible outcomes is given by:

$n(E) = n(E_1) \times n(E_2) = 2 \times 4 = 8$

Example 3

What is the total number of possible outcomes when a pair of coins is tossed?

Example 4

The life insurance policies of an insurance company are classified by:

age of the insured:

under 25 years,

between 25 years and 50 years,
over 50 years old;
sex;
marital status:
single or
married.
What is the total number of classifications?

C. Basic Probability Theory

A: event in S, $P: A \to [0,1]$, $0 \le P(A) \le 1$: probability of A
Domain of P: all measurable subsets of S

Definition $\Im$ (sigma algebra, σ-algebra, Borel field): collection of subsets of S satisfies (a) $\phi \in \Im$ (b) if $A \in \Im \Rightarrow A^c \in \Im$ (closed under complementation)
(c) if $A_1, A_2, \cdots \in \Im \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \Im$ (closed under countable unions)

Properties (a) $S \in \Im$ (b) $A_1, A_2, \cdots \in \Im \Rightarrow \bigcap_{i=1}^{\infty} A_i \in \Im$

Example: (a) $\{\phi, S\}$: trivial σ-algebra
(b) smallest σ-algebra that contains all of the open sets in $S = \sigma$ {all open sets in S }= $\bigcap_{\alpha} \Im_\alpha$ (intersection on all possible σ-algebra)

Definition Kolmogorov Axioms (Axioms of Probability)
Given $(S, \Im)$, probability function is a function P with domain $\Im$ satisfies
(a) $P(A) \ge 0$, $\forall A \in \Im$ (b) $P(S) = 1$ (c) If $A_1, A_2, \cdots \in \Im$, pairwise disjoint
$P(\cup A_i) = \sum P(A_i)$ (Axiom of countable additivity)

Exercise: axiom of finite additivity + continuity of P (if $A_n \downarrow \phi \Rightarrow P(A_n) \to 0$) $\Rightarrow$ axiom of countable additivity

Theorem If $A \in \Im$, P: probability
(a) $P(\phi) = 0$ (b) $P(A) \le 1$ (c) $P(A^c) = 1 - P(A)$

Theorem (a) $P(B \cap A^c) = P(B) - P(A \cap B)$
(b) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
(c) $A \subset B$, then $P(A) \le P(B)$

Bonferroni's Inequality $P(A \cap B) \ge P(A) + P(B) - 1$

Example: (a) $P(A)=0.95=P(B)$, $P(A\cap B)\geq P(A)+P(B)-1=0.9$
(b) $P(A)=0.3$, $P(B)=0.5$, $P(A\cap B)\geq 0.3+0.5-1=-0.2$, useless but correct

Theorem (a) $P(A)=\sum_{i=1}^{\infty}P(A\cap C_i)$, for any partition $C_1,C_2,\cdots$

(b) $P(\bigcup_{i=1}^{\infty}A_i)\leq\sum_{i=1}^{\infty}P(A_i)$ for any $A_1,A_2,\cdots$ (Boole's inequality)

General version of Bonferroni inequality: $P(\cap A_i)\geq\sum P(A_i)-(n-1)$

Counting

	without replacement	with replacement
Ordered	P_r^n	n^r
Unordered	C_r^n	$\binom{n+r-1}{r}$

Let A_n be a sequence of sets. The set of all points $\omega\in\Omega$ that belong to A_n for infinitely many values of n is known as the *limit superior* of the sequence and is denoted by $\limsup_{n\to\infty}A_n$ or $\overline{\lim}_{n\to\infty}A_n$.

The set of all points that belong to A_n for all but a finite number of values of n is known as the *limit inferior* of the sequence $\{A_n\}$ and is denoted by $\liminf_{n\to\infty}A_n$ or $\underline{\lim}_{n\to\infty}A_n$. If $\underline{\lim}_{n\to\infty}A_n=\overline{\lim}_{n\to\infty}A_n$, we say that the limit exists and write $\lim_{n\to\infty}A_n$ for the common set and call it the *limit set*.

We have $\underline{\lim}_{n\to\infty}A_n=\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty}A_k\subseteq\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty}A_k=\overline{\lim}_{n\to\infty}A_n$.

If the sequence $\{A_n\}$ is such that $A_n\subseteq A_{n+1}$, for $n=1,2,\cdots$, it is called *nondecreasing*; if $A_n\supseteq A_{n+1}$, $n=1,2,\cdots$, it is called *nonincreasing*. If the sequence A_n is nondecreasing, or nonincreasing, the limit exists and we have

$\lim_n A_n=\bigcup_{n=1}^{\infty}A_n$ if A_n is nondecreasing and

$\lim_n A_n=\bigcap_{n=1}^{\infty}A_n$ if A_n is nonincreasing.

Theorem Let $\{A_n\}$ be a nondecreaing sequence of events in S; that is $A_n \in S$, $n = 1,2,\cdots$, and $A_n \supseteq A_{n-1}$, $n = 2,3,\cdots$. Then

$$\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n) = P(\bigcup_{n=1}^{\infty} A_n) .$$

Proof. Let $A = \bigcup_{j=1}^{\infty} A_j$. Then

$$A = A_n + \sum_{j=n}^{\infty} (A_{j+1} - A_j) .$$

By countable additivity we have

$$P(A) = P(A_n) + \sum_{j=n}^{\infty} P(A_{j+1} - A_j) ,$$

and letting $n \to \infty$, we see that

$$P(A) = \lim_{n\to\infty} P(A_n) + \lim_{n\to\infty} \sum_{j=n}^{\infty} P(A_{j+1} - A_j) .$$

The second term on the right tends to zero as $n \to \infty$ since the sum $\sum_{j=1}^{\infty} P(A_{j+1} - A_j) \le 1$ and each summand is nonnegative. The result follows.

Corollary Let $\{A_n\}$ be a nonincreasing sequence of events in S. Then

$$\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n) = P(\bigcap_{n=1}^{\infty} A_n) .$$

Proof. Consider the nondecreasing sequence of events $\{A_n^c\}$. Then

$$\lim_{n\to\infty} A_n^c = \bigcup_{j=1}^{\infty} A_j^c = A^c .$$

It follows from the above Theorem that

$$\lim_{n\to\infty} P(A_n^c) = P(\lim_{n\to\infty} A_n^c) = P(\bigcup_{j=1}^{\infty} A_n^c) = P(A^c) .$$

Hence, $\lim_{n\to\infty}(1 - P(A_n)) = 1 - P(A)$.

Example (Bertrand's Paradox) A chord is drawn at random in the unit circle. What is the probability that the chord is longer than the side of the equilateral triangle inscribed in the circle?

<u>Solution 1</u>. Since the length of a chord is uniquely determined by the position of its midpoint, choose a point C at random in the circle and draw a line through C and O, the center of the circle. Draw the chord through C perpendicular to the line OC . If l_1 is the length of the

chord with C as midpoint, $l_1 > \sqrt{3}$ if and only if C lines inside the circle with center O and radius $1/2$. Thus $P(A) = \pi(1/2)^2/\pi = 1/4$.

Solution 2. Because of symmetry, we may fix one endpoint of the chord at some point P and then choose the other endpoint P_1 at random. Let the probability that P_1 lies on an arbitrary arc of the circle be proportional to the length of this arc. Now the inscribed equilateral triangle having P as one of its vertices divides the circumference into three equal parts. A chord drawn through P will be longer than the side of the triangle if and only if the other endpoint P_1 of the chord lies on that one-third of the circumference that is opposite P. It follows that the required probability is $1/3$.

Solution 3. Note that the length of a chord is determined uniquely by the distance of its midpoint from the center of the circle. Due to the symmetry of the circle, we assume that the midpoint of the chord lies on a fixed radius, OM, of the circle. The probability that the midpoint M lies in a given segment of the radius through M is then proportional to the length of this segment. Clearly, the length of the chord will be longer than the side of the inscribed equilateral triangle if the length of OM is less than $radius/2$. It follows that the required probability is $1/2$.

Question: What's happen? Which answer(s) is (are) right?

Example: Consider sampling $r = 2$ items from $n = 3$ items, with replacement. The outcomes in the ordered and unordered sample spaces are these.

Unordered	{1,1}	{2,2}	{3,3}	{1,2}	{1,3}	{2,3}
Probability	1/6	1/6	1/6	1/6	1/6	1/6
Ordered	(1,1)	(2,2)	(3,3)	(1,2), (2,1)	(1,3), (3,1)	(2,3), (3,2)
Probability	1/9	1/9	1/9	2/9	2/9	2/9

Which one is correct?

Hint: The confusion arises because the phrase "with replacement" will typically be interpreted with the sequential kind of sampling, leading to assigning a probability 2/9 to the event {1, 3}.

D. Conditional Probability and Independence

Definition Conditional probability of A given B is $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$**,** provided $P(B) > 0$.

Remark: (a) In the above definition, B becomes the sample space and $P(B \mid B) = 1$. All events are calibrated with respect to B.
(b) If $A \cap B = \phi$ then $P(A \cap B) = 0$ and $P(A \mid B) = P(B \mid A) = 0$. Disjoint is not the same as independent.

Definition A and B are independent if $P(A \mid B) = P(A)$.
(or $P(A \cap B) = P(B)P(A)$)

Example: Three prisoners, A, B, and C, are on death row. The governor decides to pardon one of the three and chooses at random the prisoner to pardon. He informs the warden of his choice but requests that the name be kept secret for a few days.

The next day, A tries to get the warden to tell him who had been pardoned. The warden refuses. A then asks which of B or C will be executed. The warden thinks for a while, then tells A that B is to be executed.

Warden's reasoning: Each prisoner has a 1/3 chance of being pardoned. Clearly, either B or C must be executed, so I have given A no information about whether A will be pardoned.

A**'s reasoning:** Given that B will be executed, then either A or C will be pardoned. My chance of being pardoned has risen to 1/2.

Which one is correct?

Bayes' Rule $A_1, A_2, \cdots$: partition of sample space, B: any set,

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(A_i)P(B \mid A_i)}{\sum P(B \mid A_j)P(A_j)}.$$

Example: When coded messages are sent, there are sometimes errors in transmission. In particular, Morse code uses "dots" and "dashes", which are known to occur in the proportion of 3:4. This means that for any given symbol,

$$P(dot \quad sent) = \frac{3}{7} \text{ and } P(dash \quad sent) = \frac{4}{7}.$$

Suppose there is interference on the transmission line, and with probability 1/8 a dot is mistakenly received as a dash, and vice versa. If we receive a dot, can we be sure that a dot was sent?

Theorem If $A \perp B$ then (a) $A \perp B^c$, (b) $A^c \perp B$, (c) $A^c \perp B^c$.

Definition $A_1, \cdots, A_n$: mutually independent if any subcollection $A_{i1}, \cdots, A_{ik}$ then

$$P(\bigcap_{j=1}^{k} A_{ij}) = \prod_{j=1}^{k} P(A_{ij}) .$$

E. Random Variable

Definition Define $X : S \rightarrow$ new sample space $\chi \subset \Re$.

X : random variable, $X : S \rightarrow \Re$, $(S, P) \rightarrow (\chi, P_X)$, where P_X : induced probability function on χ in terms of original P by

$$P_X (X = x_i) = P(\{s_j \in S : X(s_j) = x_i\}) ,$$

and P_X satisfies the Kolmogorov Axioms.

Example: Tossing three coins, X : # of head

S =	{HHH ,	HHT,	HTH,	THH,	TTH,	THT,	HTT,	TTT}
X :	3	2	2	2	1	1	1	0

Therefore, $\chi = \{0,1,2,3\}$, and

$$P_X (X = 1) = P(\{s_j \in S : X(s_j) = 1\}) = P\{TTH, THT, HTT\} = \frac{3}{8} .$$

F. Distribution Functions

With every random variable X , we associate a function called the cumulative distribution function of X .

Definition The cumulative distribution function or cdf of a random variable X , denoted by $F_X(x)$, is defined by $F_X(x) = P_X(X \le x)$, for all x .

Example: Tossing three coins, X : # of head $(X = 0,1,2,3)$, the corresponding c.d.f. is

$$F_X(x) = \begin{cases} 0 & if \quad -\infty < x < 0 \\ 1/8 & if \quad 0 \le x < 1 \\ 1/2 & if \quad 1 \le x < 2 \\ 7/8 & if \quad 2 \le x < 3 \\ 1 & if \quad 3 \le x < \infty \end{cases} ,$$

where F_X : (a) is a step function

(b) is defined for all x , not just in $\chi = \{0,1,2,3\}$

(c) jumps at $x_i \in \chi$, size of jump $= P(X = x_i)$

(d) $F_X(x) = 0$ for $x < 0$; $F_X(x) = 1$ for $x \ge 3$

(e) is right-continuous (is left-continuous if $F_X(x) = P_X(X < x)$)

Theorem $F(x)$ is a c.d.f. $\Leftrightarrow$ (a) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$.

(b) $F(x)$: non-decreasing
(c) $F(x)$: right-continuous

Example: Tossing a coin until a head appears. Define a random variable X : # of tosses required to get a head. Then

$$P(X = x) = (1-p)^{x-1} p, \qquad x = 1,2,\cdots, \qquad 0 < p < 1.$$

The c.d.f. of the random variable X is

$$F_X(x) = P(X \le x) = 1-(1-p)^x, \qquad x = 1,2,\cdots.$$

It is easy to check that $F_X(x)$ satisfies the three conditions of c.d.f.

Example: A continuous c.d.f. (of logistic distribution) is $F_X(x) = \frac{1}{1+e^{-x}}$, which satisfies the three conditions of c.d.f.

Definition (a) X is continuous if $F_X(x)$ is continuous.
(b) X is discrete if $F_X(x)$ is a step function.

Definition X and Y are identical distributed if $\forall\ A \in \Im$, $P(X \in A) = P(Y \in A)$.

Example: Tossing a fair coin three times. Let X : # of head and Y : # of tail. Then

$$P(X = k) = P(Y = k), \quad \forall\ k = 0,1,2,3.$$

But for each sample point $s \in \Omega$, $X(s) \ne Y(s)$.

Theorem X and Y are identical distributed $\Leftrightarrow F_X(x) = F_Y(x)$, $\forall\ x$.

F. Density and Mass Function

Definition The probability mass function (p.m.f.) of a discrete random variable X is

$$f_X(x) = P(X = x) \quad \text{for all } x.$$

Example: For the geometric distribution, we have the p.m.f.

$$f_X(x) = P(X = x) = \begin{cases} (1-p)^{x-1} p & x = 1,2,\cdots \\ 0 & otherwise \end{cases}.$$

And $P(X = x) =$ size of jump in c.d.f. at x,

$$P(a \le X \le b) = \sum_{k=a}^{b} f_X(k) = \sum_{k=a}^{b} (1-p)^{k-1} p,$$

$$P(X \le b) = \sum_{k=1}^{b} f_X(k) = F_X(b) \ .$$

Fact: For continuous random variable (a) $P(X = x) = 0$, $\forall\ x$.

(b) $P(X \le x) = F_X(x) = \int_{-\infty}^{x} f_X(t)dt$. Using the Fundamental Theorem of Calculus, if $f_X(x)$ is continuous, we have

$$\frac{d}{dx}F_X(x) = f_X(x).$$

Definition The probability density function or pdf, $f_X(x)$, of a continuous random variable X is the function that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \quad \text{for all } x.$$

Notation: (a) $X \sim F_X(x)$, X is distributed as $F_X(x)$.
(b) $X \sim Y$, X and Y have the same distribution.

Fact: For continuous, $P(a < X < b) = P(a < X \le b) = P(a \le X < b) = P(a \le X \le b)$.

Example: For the logistic distribution

$$F_X(x) = \frac{1}{1+e^{-x}},$$

we have $f_X(x) = F_X'(x) = \dfrac{e^{-x}}{(1+e^{-x})^2}$, and

$$P(a < X < b) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(x)dx - \int_{-\infty}^{a} f_X(x)dx = \int_{a}^{b} f_X(x)dx.$$

Theorem $f_X(x)$: pdf (or pmf) of a random variable if and only if
(a) $f_X(x) \ge 0$, $\forall\ x$.
(b) $\sum f_X(x) = 1$ (pmf) or $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (pdf).

Fact: For any nonnegative function with finite positive integral (or sum) can be turned into a pdf (or pmf)

$$\because \int f(x)dx = k < \infty \text{ then } g(x) = \frac{f(x)}{k}.$$

CHAPTER 4
PROBABILITY DISTRIBUTIONS 1
(DISCRETE RANDOM VARIABLES)

A. Discrete Random Variables

1. Basic concept of random variable

- In the world, lots of *numerical values* come from outcomes controlled by *random* probability.
- Examples:
  - wins or losses of money based on *random* outcomes of tossing coins;
  - differences between bus arriving times and scheduled ones due to *random* traffic conditions;
  - measures of body temperatures due to *random* day-to-day body conditions, etc.
- The numerical values are *variable* and *random* in nature due to the *occurrence probabilities* of them. They so may be called *random variables*.

2. Formal definition of random variable

- A brief review of the concept of *function* ---

  A function *f* with domain *A* and range *B* is a mapping from *A* to *B* such that each value $a \in A$ is mapped to a unique value $b \in B$, denoted as $f(\boldsymbol{a})$, i.e., we write $f(a) = b$.

- ***Definition 4.1*** ---

  A random variable is a *real-valued function of outcomes*, as illustrated in the following:

  $$\text{outcomes} \xrightarrow[\text{random variable}]{} \text{real values.}$$

- The *domain* of a random variable is the sample space, and the *range* is the set of real numbers.
- Examples of random variables:
  - the *sum* of the outcomes of tossing two dices;
  - the *money* you win in a fair card-drawing game;
  - the *number* of students who fail in a course;
  - the *number* of accidents in a city in a year, etc.
- We may assign probabilities to random variables according to the probabilities of related outcomes.

Consider a dice with the following information:
X = Output 1 with the probability of 1/2
X = Output 2 with the probability of 1/3
X = Output 3 with the probability of 1/6

Hence, $E(X) = \sum_{x \in X} x.P(x)$

$$E(X) = 1.(1/2) + 2.(1/3) + 3(1/6) = 1 + 2/3 = 5/3$$

Thus, any variable that has probabilities of equaling different values is a discrete random variable. We calculate the average or expected value of that discrete random variable using the formula above. As an exercise, let Y be the discrete random variable equal to the sum of the faces after rolling two standard six-sided dice. Show that E(Y) = 7.

In addition to expectation, we define a term called variance for discrete random variables. (This calculation is rarely made in computer science for algorithm analysis, but I am including it for completeness sake with respect to the topic of discrete random variables.) Variance is simply a definition which roughly gauges, "how spread out" the distribution of the discrete random variable is. Here is the formula:

$$Var(X) = \sum_{x \in X} (x - E(X)^2 p_x$$

For our example above, we have

$$Var(X) = (1 - 5/3)^2 (1/2) + (2 - 5/3)^2 (1/3) + (3 - 5/3)^2 (1/6) = 5/9$$

Also, the standard deviation of a discrete random variable is simply defined as the square root of its variance. An alternate way to calculate variance is as follows:

$$Var(X) = E(X^2) - [E(X)]^2$$

We define E(X^2) as follows: $E(X^2) = \sum_{x \in X} x^2 p_x$

Example 4.1

In tossing two dices, what is the probability of obtaining a sum X smaller than 5 points?

Solution:

- The event A for $X < 5$ is $A = \{(1, 1), (1, 2), (2, 1), (1, 3), (3, 1), (2, 2)\}$ which has 6 elements.
- The sample space S has 36 elements, as is well known.
- So, according to Fact 2.3 of Chapter 2, the desired probability is:

$$P(A) = \frac{\#\ points\ in\ A}{\#\ points\ in\ S} = 6/36.$$

- By Definition 4.1, X here is just a random variable, with its domain being S and A being a subset of it.

A note about the notation for events:

Subsequently, we use the notation $X < 5$ to denote the event A itself and write $P(A)$ as $P\{X < 5\}$. Notations for more general cases, like $P\{X \leq n\}$, are similarly defined.

Example 4.2

Three balls are drawn from an urn containing 3 white, 3 red, and 5 black balls. Suppose we win one dollar for each white ball drawn and lose one for each red ball drawn. What is the probability we win money in a draw?

Solution:

- *Key concept*: define first a random variable X as the amount of money won in a draw.
- And the probability of winning money is $P\{X > 0\}$ where $X > 0$ is the event.

- Experiment = drawing 3 balls.
- #outcomes in the sample space $S = C(11, 3)$.
- Possible values of $X = 0, \pm 1, \pm 2, \pm 3$ (i.e., winning or losing at most 3 dollars).
- So the desired probability is $P\{X > 0\} = P\{X = 1, 2, 3\}$.
- Now,
  $P\{X = 0\} = P\{$draw 3 black, or draw 1 white, 1 red, & 1 black$\}$
  $= [C(5, 3) + C(3, 1)C(3, 1)C(5, 1)] \div C(11, 3)$ (by Fact 2.3)
  $= 55/165$.
- $P\{X = 1\} = P\{$draw 1 white & 2 black, or draw 2 white & 1 red$\}$
  $= P\{X = □1\} = P\{$draw 1 red & 2 black, or draw 2 red & 1 white$\}$
  $= [C(3, 1)C(5, 2) + C(3, 2)C(3,1)] \div C(11, 3) = 39 /165$. (by Fact 2.3)
- Similarly with the reasoning details omitted (check by yourself!), we have
  $P\{X = 2\} = P\{X = □2\}$
  $= C(3, 2)C(5, 1) \div C(11, 3) = 15/165$. (by Fact 2.3)
- And
  $P\{X = 3\} = P\{X = □3\}$
  $= C(3, 3) \div C(11, 3) = 1/165$. (by Fact 2.3)
- Finally, the desired probability of winning money is:
  $P\{X = 1, 2, \text{or } 3\}$
  $= P\{X = 1\} + P\{X = 2\} + P\{X = 3\}$ (by mutual exclusiveness)
  $= 39/165 + 15/165 + 1/165$
  $= 55/165$
  $= 1/3$.

Notation:
P – this means "probability"
A, B, C, … – these stand for events
A^c, $\overline{A}$ – this means the complement of A
P(A) – this means the probability that A occurs
S – this is used to denote the sample space

Example:
Q: Bob rolls a 6-sided die. What is the sample space of this procedure?
A: S = {1, 2, 3, 4, 5, 6}

Q: Sue measures how many coin flips it takes to get 3 heads. What is the sample space of this procedure?
A: S = {3, 4, 5, …} = {all integers > 2}

Q: Fred sees what proportion of cars on his block are SUVs. What is the sample space of this procedure?
A: S = {any real number between 0 and 1} = [0,1]

Since events come in a lot of different ways, there are 3 general approaches to finding the probabilities for events. The method that is most useful depends on the situation.

Approach #1: Relative Frequency Approximation
For procedures that can be repeated over and over again, we can estimate the probability of an event A by using the following:

$$p = \frac{\text{Number of Times A Occurred}}{\text{Total Number of Trials}}$$

From theoretical arguments (see “Law of Large Numbers”, p.141), it turns out that this value *p* will get closer to *P*(*A*) as the number of trials gets larger.

Approach #2: Classical Approach
For procedures with equally likely outcomes (e.g. rolling a die, flipping a coin, etc.), we can find *P*(*A*) directly, by computing:

$$P(A) = \frac{\text{Number of Ways } A \text{ can Occur}}{\text{Total Number of Simple Event Outcomes}}$$

Approach #3: Subjective Probability
For procedures that cannot be repeated, and do not have equally likely outcomes, the true probability of an event is usually not able to be determined. In situations like this, we can *estimate* the probability using our knowledge and experience of the subject. For instance, we could ask “What is the probability that the Columbus Blue Jackets will win the Stanley Cup this year?” No one knows the true probability, but people who know a lot about hockey could give a ballpark figure.

Examples:
A situation where Approach #1 is used is in baseball. If we want to know the probability that a player will get a hit when they go up to bat, we cannot use Approach #2 because the outcomes are not equally likely. We could use Approach #3, but that would be subjective. However, by dividing the number of hits by the number of at-bats gives the *batting average,* which is an estimate of the true probability of getting a hit.

A situation where Approach #2 is used is something like rolling a die. Each face is equally likely to turn up, so we can find probabilities using this approach. Let's say A is the event of rolling an even number. What is P(A)?

$$P(A) = \frac{\text{Number of Ways } A \text{ can Occur}}{\text{Total Number of Simple Event Outcomes}} = \frac{3}{6}\frac{\text{(getting 2, 4, or 6)}}{\text{(6 possible outcomes)}} = \frac{1}{2} = 0.5$$

3. Another Example

Q: Joe flips one coin 3 times and records the 3 outcomes. What is the sample space?
A: S = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Q: What is the probability of getting 1 or 2 heads?
A: Since all outcomes are equally likely (we assume the coin is fair), we can use Approach #2.

$$P(A) = \frac{\text{Number of Ways } A \text{ can Occur}}{\text{Total Number of Simple Event Outcomes}} = \frac{6}{8}\frac{\text{(HHT, HTH, HTT, THT, TTH, THH)}}{\text{(Total of 8 outcomes)}} = \frac{3}{4} = 0.75$$

Q: What is the probability of the complement of the previous event? (i.e. P(not 1 or 2 heads))

A: There are 2 ways this can happen (HHH or TTT), so it will be $\frac{2}{8} = \frac{1}{4} = 0.25$

4. Odds

Another popular way to describe probabilities is with *odds*. Odds are ratios of success to failure, or vice versa.

Odds against A – this is the ratio $\frac{P(A^c)}{P(A)}$, often written in the form *a* : *b*, where *a* and *b* are integers with no common factor.

Odds in favor of A – this is the ratio $\frac{P(A)}{P(A^c)}$, which would be written as *b* : *a*.

Payoff Odds – this is the ratio $\frac{\text{Net Profit}}{\text{Amount Bet}}$, written as (Net Profit) : (Amount Bet).

Example:

Q: You are playing the slots. It costs $5 to pull the lever. The prize if you win is $500. The probability of winning is 0.20, and the

probability of losing is 0.80. Find all three odds listed above.

A: Let A be the event of winning the game.

$$\text{Odds Against Winning} = \frac{P(A^c)}{P(A)} = \frac{0.80}{0.20} = \frac{4}{1} \Rightarrow \text{Odds against winning is 4:1}$$

$$\text{Odds in Favor of Winning} = \frac{P(A)}{P(A^c)} = \frac{0.20}{0.80} = \frac{1}{4} \Rightarrow \text{Odds in favor of winning is 1:4}$$

$$\text{Payoff Odds} = \frac{\text{Net Profit}}{\text{Bet Amount}} = \frac{\text{Prize - Amount of Bet}}{\text{Amount of Bet}} = \frac{495}{5} = \frac{99}{1} \Rightarrow \text{Payoff Odds is 99:1}$$

So the Odds Against and Odds in Favor are telling you that when you play the game 5 times, you will win once and lose 4 times,

on average. The Payoff Odds is telling you that if you win, you'll get back 99 times what you bet. This game would be a good one

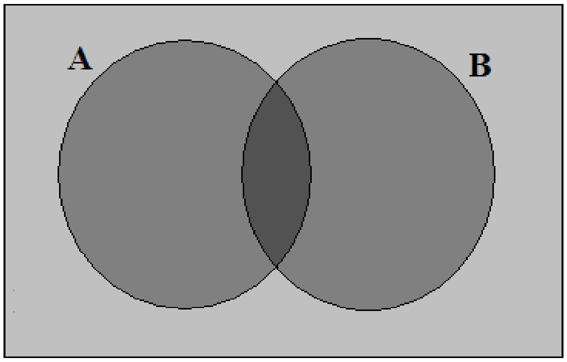
to play, because it takes you about 5 games (on average) to win, but you win 99 times what you bet.

5. Addition Rule

Compound Event – An event that is comprised of two or more *simple events*. Generally, compound events are written in terms of their simple events.
For example, the event "It will rain or snow today" could be written as "**A or B**", where A is the event that it rains today and B is the event that it snows today. So this event would happen if it rained today, snowed today, or both.
Another type of event is of the form "**A and B**", in which the event only occurs if *both* A and B occur.
For example, the event "It is at least 70 degrees and sunny outside" could be written as "A and B", where A is the event that it is at least 70 degrees outside, and B is the event that it is sunny.

a. The Formal Addition Rule

To see how this rule is derived, let's examine a *Venn Diagram*. The area within each circle corresponds to the probability of that event occurring. Where the two circles overlap (dark grey), both A and B occur. However the area, say, in circle A that does not overlap B (grey) would be when A occurs but B does not. The area outside of both circles (light grey) corresponds to neither A nor B occurring.



How would we find P(A or B) then? We want the area within the two circles (the grey and dark grey areas) because that's where A happens, B happens, or they both happen. What we could do is add together the area in circle A, and the area of circle B. The problem is that we could the overlapping area (dark grey) twice. That means we need to subtract it. Using the fact that the area in circle A is P(A), the area in circle B is P(B), and the overlap is P(A and B), we get the formal addition rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

This rule works for any events A and B. Anytime you know three of the four quantities in the equation, you can solve for the fourth.

b. Disjoint Events

Disjoint Events are events that cannot both happen at the same time. For example, let A be the event that a traffic light is green, and B be the event that the traffic light is red. The event "A and B" cannot happen, because traffic lights are never green and red at the same time. If two events are disjoint, then **P(A and B) = 0**. Disjoint events are also often called *mutually exclusive events*.

c. Addition Rule for Disjoint Events

Using the fact that P(A and B) = 0 for disjoint events, we can rewrite the formal addition rule as:

$$P(A \text{ or } B) = P(A) + P(B)$$

d. Complementary Events

Recall from the previous section that for an event A, its complement A^c is the event that A does *not* occur. Since A^c only happens when A does not, and vice versa, P(A and A^c) = 0. In other words, A and A^c are disjoint. Therefore, P(A or A^c) = P(A) + P(A^c), by the addition rule for disjoint events. But what is P(A or A^c)? This means the probability that either A happens, or A does not happen. This probability is 1, since *something* has to happen, whether it is A or not. Therefore, we have a trio of equivalent formulas:

$$P(A) + P(A^c) = 1$$
$$P(A^c) = 1 - P(A)$$
$$P(A) = 1 - P(A^c)$$

<u>Examples</u>: For the following questions, imagine we are drawing one card from a deck of 52 cards.
Q: What is the probability of drawing a queen?
A: Using Approach #2 from the previous section, and letting A be the event in question,

$$P(A)=\frac{\text{Number of Ways } A \text{ can Occur}}{\text{Total Number of Simple Event Outcomes}}=\frac{4}{52}=\frac{1}{13}$$

Q: What is the probability of drawing a diamond?
A: Using Approach #2 from the previous section, and letting B be the event in question,

$$P(B)=\frac{\text{Number of Ways } B \text{ can Occur}}{\text{Total Number of Simple Event Outcomes}}=\frac{13}{52}=\frac{1}{4}$$

Q: What is the probability of drawing a queen of diamonds?
A: This is the event "A and B", and we can use Approach #2 again:

$$P(A \text{ and } B)=\frac{\text{Number of Ways } A \text{ and } B \text{ can Occur}}{\text{Total Number of Simple Event Outcomes}}=\frac{1}{52}$$

Q: What is the probability of drawing a queen or a diamond?
A: We could could up the total number of cards that fit this bill (13 diamonds – one of which is a queen – and the other 3 queens = 16
possible cards out of 52), or we can just use the formal addition rule:

$$P(A \text{ or } B)=P(A)+P(B)-P(A \text{ and } B)=\frac{4}{52}+\frac{13}{52}-\frac{1}{52}=\frac{16}{52}=\frac{4}{13}$$

Q: What is the probability of drawing a number card? (Aces included)
A: Let's call this event C. There are 4 of each number 1 to 10, for a total of 40 out of 52 cards.

$$P(C)=\frac{\text{Number of Ways } C \text{ can Occur}}{\text{Total Number of Simple Event Outcomes}}=\frac{40}{52}=\frac{10}{13}$$

Q: What is the probability of drawing a number card or a queen?
A: Now we can use the addition rule for disjoint events, since A and C can't happen at the same time.

$$P(A \text{ or } C)=P(A)+P(C)=\frac{4}{52}+\frac{40}{52}=\frac{44}{52}=\frac{11}{13}$$

6. Multiplication Rule: The Basics

Now that we can find "A or B" probabilities, we focus on how to find "A and B" probabilities. Intuitively, for A and B to happen, we need two things to take place:

a. A needs to happen

b. Given that A happened, B needs to happen

This leads us to…

a. The Formal Multiplication Rule

$$P(A \text{ and } B) = P(A)\cdot P(B|A)$$

Here, P(B|A) is what is called a *conditional probability*. It stands for the probability that B happens *given* A already happened. (The vertical bar means "given") Since "A and B" is the same as "B and A", we can also write the formula as:

$$P(A \text{ and } B) = P(B)\cdot P(A|B)$$

Which form you use depends on what information is available.

<u>Example</u>: Students can take a standardized test at three test centers A, B, and C. Suppose that after the most recent test, 500 students went to A, 200 went to B, and 300 went to C. Furthermore, the proportion of students who passed the exam were 50%, 80%, and 75%, respectively.

Q: What is the probability that a randomly selected student took the test at center B?
A: There are a total of 1000 students, and 200 went to B. Thus P(B) = 200/1000 = 0.20.

Q: What is the probability that a student who took the test at B passed the exam?
A: Now, we want to find the probability of passing *given* that the student took the test at B. We are told in the problem that 80% of
the students at center B passed. Thus we have: P(Pass | B) = 0.80.

Q: What is the probability that a student both took the test at B and passed?
A: Using the Multiplication Rule, P(Pass and B) = P(B)P(Pass | B) = (0.20)(0.80) = 0.16.

b. Independent Events

Two events are called **independent** if the occurrence of one does not affect the chances of the other one occurring. Statistically, what this means is that A and B independent → **P(A | B) = P(A)**. In other words, the probability of A happening *given* that B happened is just the same as if we didn't know whether B happened (because the occurrence of B has no effect on the occurrence of A).

<u>Note</u>: if A and B are disjoint, then we know that only one can occur. Thus, knowing that B happened tells you that A definitely did not happen, and we have P(A | B) ≠ P(A). Thus disjoint events are *never* independent events.

c. Multiplication Rule for Independent Events

Using the fact that P(A | B) = P(A) for independent events, we see that the formal multiplication rule turns into:

$$P(A \text{ and } B) = P(A)\cdot P(B)$$

d. The Law of Total Probability

This rule is very intuitive, and is useful for finding probabilities of events. To explain it, we refer to the test center example above.

Q: What is the probability of a randomly selected student passing the exam?

A: From the information above, we can find out (similar to the previous example) that:

P(A) = 0.50	P(B) = 0.20	P(C) = 0.30
P(Pass | A) = 0.50	P(Pass | B) = 0.80	P(Pass | C) = 0.75

We want to find P(Pass). What are the possible scenarios where a student passes the exam? They could take the test at A and pass, they could take it at B and pass, or they could take it at C and pass.

So P(Pass) = P(A and Pass OR B and Pass OR C and Pass). But each of those 3 scenarios are disjoint, because a student can't take the test at more than one center. Therefore, by the addition rule we can add these probabilities as follows:

P(Pass) = P(A and Pass) + P(B and Pass) + P(C and Pass)

Then, by the multiplication rule, we can find all of these probabilities:

P(A and Pass) = P(A)·P(Pass | A) = (0.50)(0.50) = 0.25
P(B and Pass) = 0.16
P(C and Pass) = P(C)·P(Pass | C) = (0.30)(0.75) = 0.225

Thus P(Pass) = 0.25 + 0.16 + 0.225 = 0.635

In general, if you have disjoint events B_1, B_2, … , B_n that represent every possible outcome of a procedure, then you can write:

$$P(A) = P(A \text{ and } B_1) + P(A \text{ and } B_2) + \ldots + P(A \text{ and } B_n) = \sum_{i=1}^{n} P(A \text{ and } B_i)$$

The most common way to use this rule is if you have two events A and B, then:
$P(A) = P(A \text{ and } B) + P(A \text{ and } B^c)$

Examples:

A telemarketing company makes phone calls to potential customers all across the U.S. For each call, the probability of the customer answering the phone is 0.20. For the next couple of questions, assume calls are independent of each other.

Q: Let's say the company makes 10 phone calls. What is the probability that all of them are answered?

A: P(10 calls answered) = P(1st call answered AND 2nd call answered AND … AND 10th call answered)

= P(1st call answered)·P(2nd call answered)· … ·P(10th call answered) ← (Mult. Rule for Independent Events)

= (0.20)·(0.20)· … ·(0.20) = 0.20^{10} = 0.0000001024

Not very likely, is it?

Q: Let's say the company makes 2 phone calls. What is the probability that exactly one of them is answered?

A: First, note that from the Complement Rule, the probability that a call is *not* answered is 1 – 0.20 = 0.80. Thus:

P(1 call answered) = P(1st call answered and 2nd call not answered OR 1st call not answered and 2nd call answered)

= P(1st call answered and 2nd call not answered) + P(1st call not answered and 2nd call answered) ← (Add. Rule)

= P(1st call answered)·P(2nd call not answered) + P(1st call not answered)·P(2nd call answered) ← (Mult. Rule)

= (0.20)·(0.80) + (0.80)·(0.20) = 0.16 + 0.16 = 0.32

Q: Now suppose that if a customer answers the phone, their chance of buying the product is 0.10. (Note that if they do not answer the phone, their chance of buying it is 0). What is the overall chance of a telemarketer selling the product when they call a home?

A: In the question, we are told P(Buying | Call Answered) = 0.10 and P(Buying | Not Answered) = 0. We want to find P(Buying).

From the Law of Total Probability,

P(Buying) = P(Buying and Call Answered) + P(Buying and Not Answered)

= P(Call Answered)·P(Buying | Call Answered) + P(Not Answered)·P(Buying | Not Answered) ← (Mult. Rule)

= (0.20)·(0.10) + (0.80)·(0) = 0.02.

Note: If this seemed complicated, try just replaced "Buying" with A, "Call Answered" with B, and "Not Answered" with B^c. Then the calculations above follow directly from the Law of Total Probability written before.

B. Distribution Function

1. Definition of cumulative distribution function

- ***Definition 4.2***

  The *cumulative distribution function* (*cdf*), or simply *distribution function, F* of a random variable *X* is defined as

$$F(b) = P\{X \leq b\} \qquad \forall\ □\infty < b < \infty.$$

- Notes: here "$\forall$" means "for all"; *b* is a real number; and the notation "$\{X \leq b\}$" is considered as an *event* as mentioned before.

Example 4.3

In a game of tossing a fair coin, let *X* be a random variable with its value defined as +5 (winning 5 dollars) if a head (H) appears and as −3 (losing 3 dollars) if a tail (T) appears. Derive the cdf $F(b)$ of *X* and draw a diagram for it.

Solution:

- The random variable *X* takes only two *discrete* values of −3 and +5.
- Concept used to derive $F(b)$: enumerate all cases for *b*.
  - For *b* < □3, $F(b) = P\{X \leq b\} = P\{$□$\} = 0$ where □ means the empty set, because neither outcome T nor H will "yield" any value of *X* smaller than □3.
  - For □3 ≤ *b* < +5, $F(b) = P\{X \leq b\} = P\{\text{T}\} = 1/2$ because only the value of $X = -3$ corresponding to the outcome T lies in the range of □3 ≤ *b* < +5.

    (Note: here, the notation T in $P\{\text{T}\}$ above is considered as *an event* including just an element, namely, the set {T}. Similar interpretations will be applied to subsequent discussions.)
  - For $+5 \leq b$, $F(b) = P\{X \leq b\} = P\{\text{T, H}\} = 1$ because $X = -3$ and +5 when T and H appear, respectively, and both values of □3 and +5 are ≤ *b*~~)~~.
- A cdf diagram for the random variable *X* is shown in Fig. 4.1. Note the *continuity* condition at the discrete point of $b = -3$ or +5.

2. Notes about *limit points* ---

- In Fig. 4.1, the hollow circle ○ at the right end of the middle line segment for $F(b) = 1/2$, for example, means the "limit point" 5□, which is *the largest real value smaller than 5*.
- Formally, the limit point b^- is defined as $\lim_{n\to\infty}(b - \frac{1}{n})$ and may be regarded to be located *right to the left* of the point at *b*.

- Similarly, b^+ is defined as $\lim_{n\to\infty}(b + \frac{1}{n})$. Such points do *not* appear in Fig. 4.1.

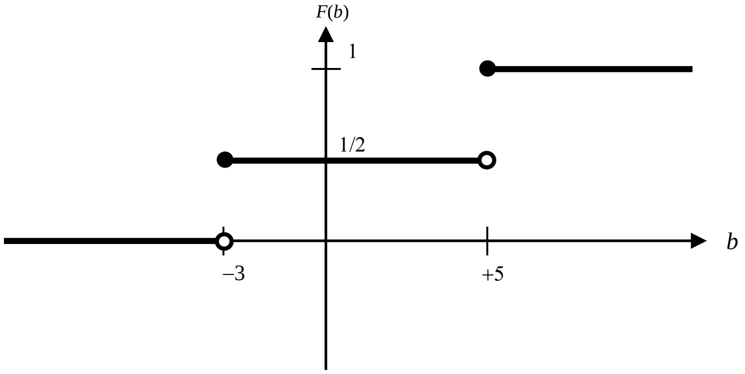


Fig. 4.1 Cumulative distribution function $F(b)$ for Example 4.3.

3. Some properties of the cdf

The following properties are intuitive in general (can be seen to be true from Fig. 4.1 above); for proof, see the reference book, or prove them by yourself.

- ***Property 4.1 ---***

  $F(a) \le F(b)$ if $a < b$, i.e., F is a *nondecreasing* function.

- ***Property 4.2 ---***

  $$\lim_{b\to+\infty} F(b) = 1.$$

- ***Property 4.3 ---***

  $$\lim_{b\to-\infty} F(b) = 0.$$

- ***Property 4.4 ---***

  For any b and any *decreasing* sequence $b_1, b_2, b_3, \ldots$ which converges to b, it is true that $\lim_{n\to\infty} F(b_n) = F(b)$ (i.e., F is *right continuous*).

- A note: $F(b)$ with $b = 5$, for example for Fig. 4.1 above, is just denoted by the solid circle ● at the left end of the right line segment for $F(b) = 1$.

4. Some facts

- All probability questions can be answered in terms of the cdf. Some examples are the following facts.
- ***Fact 4.1 ---***

  $P\{a < X \le b\} = F(b)$ □ $F(a)$.

Proof: easy to prove from the definition of the cdf, and the fact $\{X \le b\} = \{a < X \le b\} \text{U} \{X \le a\}$ where $\{a < X \le b\}$ and $\{X \le a\}$ are mutually exclusive.

- ***Fact 4.2 ---***

  $P\{X < b\} = F(b^{□})$

  (note: there is no sign of "=" in $X < b$).

  Proof: left as an exercise.

- ***Fact 4.3 ---***

  $P\{X = b\} = F(b) - F(b^{-})$

  (note: this value is the "jump from b^{-} to b ").

  Proof: left as an exercise.

- ***Fact 4.4 ---***

  $P\{X > b\} = 1 - F(b)$.

  Proof: left as an exercise.

- Note: $F(b^{-}) = P\{X < b\} \ne P\{X \le b\} = F(b)$. An example can be seen from Fig. 4.1 where $1/2 = F(5^{-}) \ne F(5) = 1$.

Example 4.4

Given a cdf as follows (as illustrated by Fig. 4.2):

$$\begin{array}{lll}
F(x) = 0 & \text{for } x < 0; & \text{(A)} \\
\quad = x/2 & \text{for } 0 \le x < 1; & \text{(B)} \\
\quad = 2/3 & \text{for } 1 \le x < 2; & \text{(C)} \\
\quad = 11/12 & \text{for } 2 \le x < 3; & \text{(D)} \\
\quad = 1 & \text{for } 3 \le x, & \text{(E)}
\end{array}$$

compute the values $P\{2 < X \le 4\}$, $P\{X < 3\}$, $P\{X = 1\}$, and $P\{X > 1/2\}$.

Solution:

- $P\{2 < X \le 4\} = F(4)$ □ $F(2)$ (by Fact 4.1)
  $= 1$ □ $11/12 = 1/12$.
- $P\{X < 3\} = F(3^{□}) = 11/12$. (by Fact 4.2 and (D) above)
- $P\{X = 1\} = P\{X \le 1\}$ □ $P\{X < 1\}$ (by Fact 4.3)
  $= F(1)$ □ $F(1^{-})$ (by definition and Fact 4.2)
  $= 2/3$ □ $1/2$ (by (C) and (*B*) above)
  $= 1/6$.
- $P\{X > 1/2\} = 1$ □ $P\{X \le 1/2\} = 1$ □ $F(1/2) = 3/4$. (by Fact 4.4)

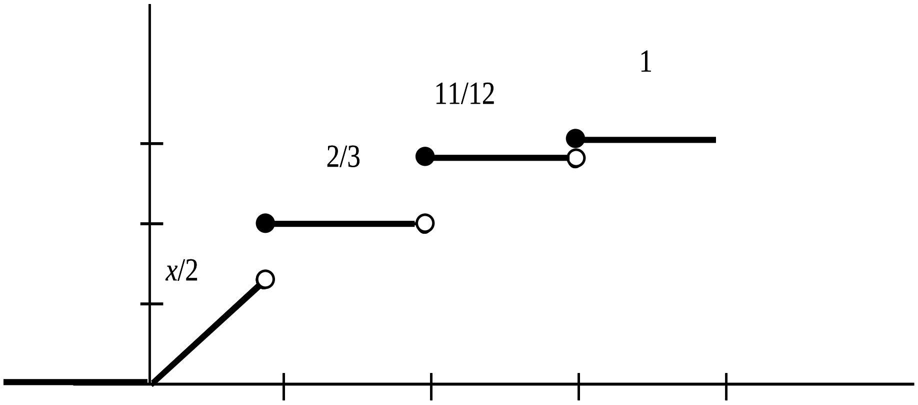


Fig. 4.2 Cumulative distribution function $F(x)$ for Example 4.4

C. Discrete Random Variable

1. *Definitions of discrete random variable and probability mass function* ---

- A random variable which can take only a countable number of possible values is said to be *discrete*.
- ***Definition 4.3 ---***
  The *probability mass function* (*pmf*) $p(a)$ of a discrete random variable X is defined as

$$p(a) = P\{X = a\}.$$

Example 4.5 ---

In a game of tossing two coins played by a child with a parent, if two heads appear, then the child wins two dollars from the parent; if one head and one tail appear, he/she wins one dollar, and if two tails appear, he/she wins nothing. By taking the money the child wins from the parent as a random variable X, what is the pmf $p(a)$ of X?

Solution:

- Events --- $A = \{(\mathrm{H}, \mathrm{H})\}$, $B = \{(\mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{H})\}$, and $C =\{(\mathrm{T}, \mathrm{T})\}$.
- Corresponding random variable values $X = 2, 1, 0$.
- Sample space $S = \{(\mathrm{H}, \mathrm{H}), (\mathrm{H}, \mathrm{T}), (\mathrm{T}, \mathrm{H}), (\mathrm{T}, \mathrm{T})\}$.
- Corresponding probabilities $P(A) = 1/4$, $P(B) = 1/2$, $P(C) = 1/4$ according to Fact 2.3:

$$P(E) = \frac{\#\ points\ in\ an\ event}{\#\ points\ in\ S}.$$

- Therefore, $P\{X = 2\} = 1/4$, $P\{X = 1\} = 1/2$, and $P\{X = 0\} = 1/4$.
- That is, the pmf $p(a)$ of X are $p(2) = 1/4$, $p(1) = 1/2$, $p(0) = 1/4$.
- Note: $p(2) + p(1) + p(0) = 1$ as ensured by Axiom 2.

- A graphic diagram for the pmf is shown in Figure 4.3.

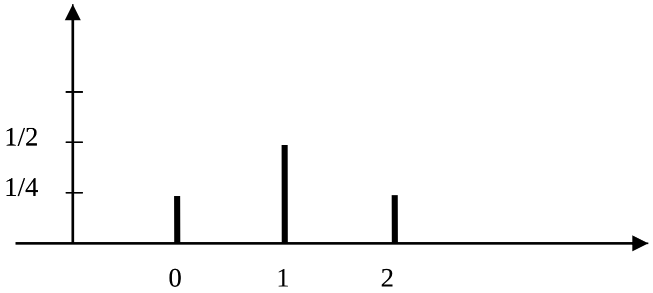


Fig. 4.3 An example of probability mass functions.

2. **Properties of discrete random variables**

If a discrete random variable X assumes one of the values of x_1, x_2, ..., then we have the following obvious properties:

- $p(x_i) \geq 0$ for all $i = 1, 2, ...;$
- $p(x) = 0$ for all other values;
- $\sum_{i=1}^{\infty} p(x_i) = 1.$

Why? According to the definitions of random variable and pmf as well as the three axioms of probability.

3. **Relation between the pmf and the cdf**

- ***Fact 4.5 ---***

The cdf $F(x)$ of a discrete random variable X can be expressed in terms of the pmf $p(x)$ by

$$F(a) = \sum_{all\ x_i \leq a} p(x_i)$$

Why? Because $F(a) = P\{X \leq a\} = P\{\text{all } x_i \leq a\} = \sum_{all\ x_i \leq a} p(x_i)$.

- The cdf as described above is a *step function* (like that of Fig. 4.1 but not that of Fig. 4.2; the latter is not discrete!)

Example 4.6 (Example 4.5 revisited)

Find the cdf $F(a)$ of the random variable X described in Example 4.5.

Solution:

- The pmf $p(a)$ of X are $p(2) = 1/4$, $p(1) = 1/2$, $p(0) = 1/4$.
- By the above-mentioned pmf-cdf relation, the cdf values $F(a)$ are:

for $a < 0$, $F(a) = 0$;
for $0 \le a < 1$, $F(a) = p(0) = 1/4$;
for $1 \le a < 2$, $F(a) = p(0) + p(1) = 1/4 + 1/2 = 3/4$;
for $2 \le a$, $F(a) = p(0) + p(1) + p(2) = 1/4 + 1/2 + 1/4 = 1$.

- A graphic diagram of the above cdf F is given in Fig. 4.4.
- Note the “continuity jump” at $a = 0, 1$, or 2.

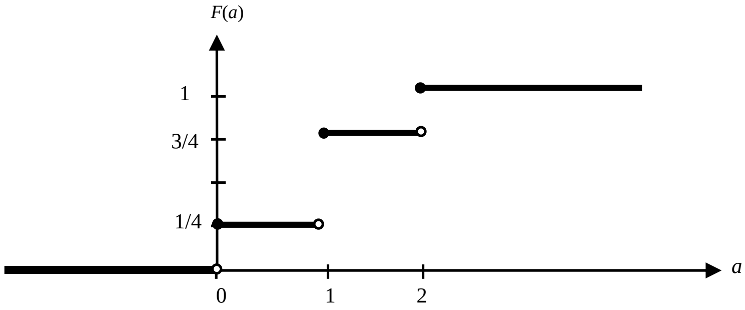


Fig. 4.4 Graphic diagram of the cdf of Example 4.6.

D. Expectation (Mean)

Example 4.7 ---

Find the mean □ of the random variable X whose values are taken to be the outcomes of tossing a fair die.

Solution:

- The outcomes are 1 through 6 which are also the values of the random variable X.
- The mean may be computed as □ = (1 + 2 + … + 6)/6 = 21/6 = 7/2 = 3.5.
- But this means □ actually is computed under the assumption that the six faces of the die will appear with equal probabilities 1/6.
- That is, □ is computed as a *weighted sum* of the outcome values with the probability values as the weights in the following way:

□ = 1×(1/6) + 2×(1/6) + … + 6×(1/6) = 7/2 = 3.5.

Example 4.8 (Example 4.5 revisited again) ---

The game of tossing two fair coins by a child as described in Example 4.5 is obviously *unfair* because the child will never lose in each guess (at least, win nothing). If now two heads or two tails appear, the child wins two dollars from the parent, and if one head and one tail appear, he/she loses one to the parent. How will you judge the game now? Fair or not? If not, how will you correct the game further to make it fair?

Solution:

- Again, let the money the child wins in each tossing be a random variable X.

- As done in Example 4.5, we can get the following probabilities:
  $P\{X = +2\} = P\{(\text{H, H}), (\text{T, T})\} = 1/4 + 1/4 = 1/2$;
  $P\{X = -1\} = P\{(\text{H, T}), (\text{T, H})\} = 1/4 + 1/4 = 1/2$.
  (Note: here we use $P\{(\text{H, H}), (\text{T, T})\}$ to represent $P(A)$ with $A = \{(\text{H, H}), (\text{T, T})\}$, and so on, just like using $P\{\text{H}\}$ to represent $P(B)$ with $B = \{\text{H}\}$ as mentioned before.)
- That is, the pmf is: $p(+2) = 1/2$ and $p(-1) = 1/2$.
- To judge whether a game is fair, one way is to compute the *average* (mean) □ of money the child wins in each tossing, which, as done in Example 4.7, may be computed as

$$
\begin{aligned}
\square &= (+2)\times P\{X = +2\} + (-1)\times P\{X = -1\} \\
&= (+2)\times p(+2) + (-1)\times p(-1) \\
&= (+2)\times(1/2) + (-1)\times(1/2) \\
&= +1/2.
\end{aligned}
$$

- That is, *in the long run*, the child will win 1/2 dollar from each tossing *on the average*. So, after 100 tossings, for example, he/she wins *about* $(1/2)\times 100 = 50$ dollars. The game is so still *unfair*!
- To make it fair, an obvious way is to lose *two* dollars, instead of one, to the parent when the case of a head and a tail appears; or to win just one dollar instead of two when two heads or two tails appear. Then, the average □ may be computed to be 0 (check this by yourself) which means that the game is fair now.

Definition of the expectation (mean) of a random variable

- From the above discussions, we may have the following reasonable definition for the mean of a random variable.
- ***Definition 4.4***

  Given a discrete random variable X with pmf $p(x)$, the *mean* (or *expectation, expected value*) of X, denoted by $E[X]$, is defined as

$$E[X] = \sum_{x:p(x)>0} xp(x)\,.$$

- Comment: in words, the mean of X is a *weighted average* of the possible values that X can take with the weights being taken to be the probabilities of the values ("average" is used as an undefined term in this course).

E. Expectation of a Function of a Random Variable

1. Functions of random variables

- Sometimes we need ways to compute more complicated data from those of existing random variables. Then functions of random variables may be used.
- For example, let X and Y be the numbers of red and yellow tokens, respectively, which you win in a casino, and they are used as substitutes for 10 and 1 dollars, respectively.

Then, the dollars you win totally is also a random variable Z which may be expressed to be a function g of the two random variables X and Y as $Z = g(X, Y) = 10X + Y$.

2. How to compute the values of a function of a random variable?

Given a discrete random variable X and a function $g(X)$ of X, how do we compute $E[g(X)]$?

(1) First way: by use of the *definition* of expectation.

(2) Second way: by use of a *proposition* derived later.

Example 4.9 (computing function values of random variables by definition) ---

Let random variable X takes one of the values □1, 0, 1 with respective probabilities $P\{X = □1\} = 0.2$, $P\{X = 0\} = 0.5$, $P\{X = 1\} = 0.3$, and let $g(X) = X^2$. Compute the expectation value $E[g(X)]$.

Solution:

- Let $Y = \mathrm{g}(X) = X^2$.
- The pmf $p(y)$ of Y is:

$$P\{Y = 1\} = p(1) = P\{X = □1 \text{ or } +1\} \qquad (\because y = x^2 = 1 \Rightarrow x = \pm 1)$$

$$= P\{X = -1\} + P\{X = +1\} \qquad \text{(by mutual exclusiveness)}$$

$$= 0.2 + 0.3 = 0.5;$$

$$P\{Y = 0\} = p(0) = P\{X = 0\} = 0.5. \qquad (\because y = x^2 = 0 \Rightarrow x = 0)$$

- Therefore, $E[X^2] = E[Y] = \sum_{y:p(y)>0} yp(y_i) = 1\times(0.5) + 0\times(0.5) = 0.5$.

Proposition

If random variable X takes the values x_i, $i \geq 1$, with respective probability $p(x_i)$, then for any real-valued function g,

$$E[g(X)] = \sum_i g(x_i)p(x_i).$$

Proof:

- First, divide all the values of $g(x_i)$ into groups, each group being with identical values of $g(x_i)$, denoted as y_j.
- Therefore,

$$\sum_i g(x_i)p(x_i) = \sum_{i:g(x_i)=y_1} y_1 p(x_i) + \sum_{i:g(x_i)=y_2} y_2 p(x_i) + \ldots$$

$$= y_1 \times \sum_{i:g(x_i)=y_1} p(x_i) + y_2 \times \sum_{i:g(x_i)=y_2} p(x_i) + \ldots$$

$$= y_1 \times P\{g(X) = y_1\} + y_2 \times P\{g(X) = y_2\} + \ldots$$

(∵ $\sum_{i:g(x_i)=y_j} p(x_i)$ is the sum of probabilities for the event $g(X) = y_j$ to occur)

$$= \sum_j y_j P\{g(X) = y_j\}$$

$= E[g(X)].$ (by the definition of $E[g(X)]$)

Example 4.10

Let random variable X takes one of the values □1, 0, 1 with probabilities $P\{X =$ □$1\} = 0.2$, $P\{X = 0\} = 0.5$, $P\{X = 1\} = 0.3$, and let $g(X) = X^2$, computer $E[g(X)]$.

Solution:

- By Proposition 4.1, we have

$$E[X^2] = (\square 1)^2 \times p(-1) + 0^2 \times p(0) + 1^2 \times p(1)$$
$$= (\square 1)^2 \times (0.2) + 0^2 \times (0.5) + 1^2 \times (0.3)$$
$$= 0.5$$

which is the same as that computed in Example 4.9!

3. Linearity property of the expectation function

- ***Corollary 4.1***

  If a and b are two constants, then

$$E[aX + b] = aE[X] + b.$$

Proof:

$$E[aX + b] = \sum_{x:p(x)>0} (ax+b)p(x) \qquad \text{(by Proposition 4.1)}$$

$$= a \sum_{x:p(x)>0} xp(x) + b \sum_{i:p(x)>0} p(x)$$

$= aE[X] + b.$

(by the definition of expectation and Axiom 2: $\sum_{i:p(x)>0} p(x) = 1$)

- Note: the notation "i: $p(x)>0$" under the summation sign □ means that only those discrete values x with non-zero $p(x)$ are dealt with.
- Comments*:*
  - The expectation function $E[\cdot]$ may be regarded as a *linear* operator according to the above corollary.
  - $E[X]$ is also called the *first moment* of X.

4. Definition of the moment function

- ***Definition 4.5***

  The *n*th moment of X is defined as

$$E[X^n] = \sum_{x:p(x)>0} x^n p(x) .$$

- The moment function is useful in many engineering application areas.

5. **Other interesting averages of numbers ---**
   - In daily life, the *mean* may be used roughly a simple *representative value* of a group of numerical data, showing the "overall magnitude" or the "trend" of the data values. Here, in this course it is formally defined as the weighted average of the possible values of a random variable.
   - However, the mean sometimes is *not* a good representation of a data group in certain applications. There are "*averages of other senses*" for various uses.

6. **An example of improper use of the mean ---**

   Two groups of students took a test and their scores are shown in Table 4.1. How should we evaluate their achievements? Which group is better? A common answer is to use the means in the following way.
   - The two groups' mean scores may be computed easily to be 87 and 103, respectively.
   - And so we may say that group *B* has a better achievement accordingly.
   - However, an inspection of the table data reveals that the larger mean score value of Group *B* is contributed mainly by the large value of 3**5**7 of a member in the group; the other members as a whole actually are not so good as those of Group *A*.
   - Then, is there another way of evaluation using a single representative value related to the data of each group?
   - An answer is to use the *median* instead of the mean, as described next.

7. **An informal definition of the median**

   Simply speaking, the *median m* of a group of numerical data is the value such that the number of data values larger than *m* is *equal to* the number of those smaller than *m*. A formal definition of median for random variables will be given later.

8. **An example of using the median in replacement of the mean**

   For the last example immediate above, we try to use the median in the following way.
   - After sorting the data in Table 4.1 to become Table 4.2, the medians of the two groups can be found easily to be 86 and 75, respectively.
   - Therefore, judging from the two median values 86 and 75, we get a conclusion, *contrary* to that mentioned previously, that Group *A*, instead of *B*, has a better achievement.

- Translation of the two terms mean, median:

Table 4.1 Test scores of two groups of students.

Group *A*	Group *B*
86	75
72	38
112	357
113	77
91	79
48	42
87	53
sum=609	sum=721
mean=87	mean=103

Table 4.2 Test scores of two groups of students.

Group *A*	Group *B*
72	38
48	42
86	53
median=87	median=75
113	77
91	79
112	357
sum=609	sum=721
mean=87	mean=103

9. Formal definition of the median of a random variable ---

- In the last example, a group of data values may be regarded as the outcomes of a random variable X and the informal definition of its median m --- "the number of data values larger than m is *equal to* the number of those smaller than m" --- means that the value of X is just as likely to be larger than m as it is to be smaller, or equivalently, that the probability for $X > m$ and that for $X < m$ are equal, leading to the following formal definition for the median.

- ***Definition 4.6*** ---

  Given a discrete random variable X with cdf F, the *median* of X is defined as the value m such that $F(m) = 1/2$.

- Comments:
  - In words, a random variable is just as likely to be larger than its median as it is to be smaller.
  - Sometimes, due to the *discreteness of* the random variable, the exact value of m

for $F(m) = 1/2$ to be true is not available, but can only be estimated in such a way that $F(m)$ is as close to 1/2 as possible.

Example 4.11 (computing the median for a discrete random variable)

Find the median m of the random variable X whose values are taken to be the outcomes of tossing a fair die. (Note that the mean of this X has been computed to be 3.5 in Example 4.7.)

Solution:

- Obviously, the pmf for X is: $p(1) = 1/6$, $p(2) = 1/6$, …, $p(6) = 1/6$.
- By the definition of cdf, it is easy to see that the cdf F for X is: $F(1) = 1/6$, $F(2) = 1/6 + 1/6 = 1/3$, $F(3) = 1/6 +1/6 + 1/6 = 1/2$, and so on.
- Therefore, the median of X is 3, which is different from the mean of X already known to be 3.5.

10. Two other types of means: geometric and harmonic means

- The above-mentioned mean of numerical data actually is the so-called *arithmetic mean*, because there are two other types of means, namely, *geometric mean* and *harmonic mean* which have respective significant applications.

F. Variance

1. Concept of variance

- Another property of a random variable other than the mean and median is its *variance* which describes the *degree of scatter* of the random variable values. The larger the variance, the larger the scatter.
- Conceptually, if the values of the random variable are all the same in the extreme case, then the variance of the random variable should be zero.

2. Definition of the variance of a random variable

- ***Definition 4.7 ---***

  If X is a random variable with mean □, then the *variance* of X, denoted by Var(X), is defined by

  $$\mathrm{Var}(X) = E[(X □ □)^2].$$

- The variance is computed after a *normalization* of the random variable values with respect to the mean.

3. An alternative formula for computing the variance

- ***Proposition 4.2***

  The value of Var(X) may be computed alternatively by

$$\mathrm{Var}(X) = E[X^2] \;\square\; (E[X])^2.$$

Proof:

$$
\begin{aligned}
\mathrm{Var}(X) &= E[(X - \square)^2] \\
&= \sum_x (x - \square)^2 p(x) && \text{(by the definition of mean)} \\
&= \sum_x (x^2 - 2\square x + \square^2) p(x) \\
&= \sum_x x^2 p(x) - 2\square \sum_x x p(x) + \square^2 \sum_x p(x) && \text{(by Corollary 4.1)} \\
&= E[X^2] - 2\square E[X] + \square^2 \\
&\qquad \text{(by the definition of mean and } \textstyle\sum_x p(x) = 1 \text{ coming from Axiom 2)} \\
&= E[X^2] - 2\square^2 + \square^2 && \text{(by the definition of mean)} \\
&= E[X^2] - \square^2.
\end{aligned}
$$

- Comments:
  - In words, the above proposition says that the variance of a random variable is equal to the expected value of X^2 minus the square of its expected value.
  - Use of this proposition is often the easiest way to compute Var(X).

Example 4.12

Compute Var(X) if X represents the outcome of rolling a fair die.

Solution:

- By Proposition 4.1, $E[X^2] = 1^2\times(1/6) + 2^2\times(1/6) + \ldots + 6^2\times(1/6) = 91/6$.
- Also, we know from the result of Example 4.7 that $E[X] = 3.5 = 7/2$.
- By Proposition 4.2, $\mathrm{Var}(X) = E[X^2] \;\square\; (E[X])^2 = 91/6 \;\square\; (7/2)^2 = 35/12$.

Corollary 4.2

If a and b are constants, then

$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X).$$

Proof:

By Corollary 4.1, we have $E[aX + b] = aE[X] + b = a\square + b$. Accordingly, we have

$$
\begin{aligned}
\mathrm{Var}(aX + b) &= E[(aX + b - E[aX + b])^2] && \text{(by the definition of variance)} \\
&= E[(aX + b - a\square - b)^2] \\
&= E[a^2(X - \square)^2]
\end{aligned}
$$

$= a^2E[(X - \square)^2]$ (by Corollary 4.1)

$= a^2\mathrm{Var}(X)$. (by the definition of variance)

4. Definition of standard deviation

- ***Definition 4.8***

  The square root of Var(X), $\sqrt{\mathrm{Var}(X)}$, is called the *standard deviation* of X, and is denoted as SD(X), i.e.,

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}\ .$$

G. The Bernoulli and Binomial Random Variables

1. Assumptions for the following discussions

- Given a trial with an outcome of success or failure, define a random variable X to be

  $X = 1$ if the outcome = a success; and
  $X = 0$ if the outcome = a failure.

- And assume the following pmf for random variable X:

  $p(0) = P\{X = 0\} = 1 \square p$; and
  $p(1) = P\{X = 1\} = p$ (4.1)

  where p is the probability of success in a trial.

2. Definitions of Bernoulli and binomial random variables

- ***Definition 4.9 ---***

  A random variable X is said to be a Bernoulli random variable if its pmf is described by (4.1) above for some p such that $0 < p <1$.

- ***Definition 4.10 ---***

  If X represents the number of successes in n independent trials with p as the probability of success and $1 \square p$ as that of failure in a trial, then X is called a binomial random variable with parameters (n, p).

- A comment: a Bernoulli random variable is just a binomial random variable with parameters $(1, p)$

3. The pmf of a binomial random variable

- ***Fact 4.6***

  The pmf $p(i)$ for a binomial random variable X with parameters (n, p) is:

$$\begin{aligned} p(i) &= P\{X = i\} \\ &= P\{\#\text{successes in } n \text{ trials} = i\} \\ &= C(n, i)p^i(1 \square p)^{n-i}, \;\; \forall\, i = 1, 2, \ldots \end{aligned} \tag{4.2}$$

Why? Think about it by yourself using a similar reasoning used in Example 3.11.

Example 4.13 ("wheel of fortune")

A game called "wheel of fortune" often played in casinos goes like: bet a number N within 1 through 6, and then roll 3 dies; if N appears i times, $i = 1, 2, 3$, then the player win i units; otherwise, the player loses one unit. Is this game fair?

Solution:

- A trial = a roll of a die here.
- Success in a trial = N appears in the rolling result.
- $P\{N$ appears in a trial$\} = 1/6$.
- Let X = units won by the player ("–" means "lose", and "+" means "win").
- Let Y = #times that N appears in the 3 rollings.
- Then, Y is a binomial random variable with parameters (3, 1/6) by definition.
- $$\begin{aligned} p(-1) &= P\{X = \square 1\} \\ &= P\{\text{losing one unit}\} \\ &= P\{N \text{ does not appear in the 3 rollings}\} \\ &= P\{Y = 0\} \\ &= C(3, 0)(1/6)^0(5/6)^3 && \text{(by Fact 4.6)} \\ &= 125/216. \end{aligned}$$
- $$\begin{aligned} p(+1) &= P\{X = +1\} \\ &= P\{\text{winning one unit}\} \\ &= P\{N \text{ appears once in the 3 rollings}\} \\ &= P\{Y = 1\} \\ &= C(3, 1)(1/6)^1(5/6)^2 && \text{(by Fact 4.6)} \\ &= 75/216. \end{aligned}$$
- Similarly,
  $$\begin{aligned} p(+2) &= P\{X = +2\} \\ &= P\{Y = 2\} \\ &= C(3, 2)(1/6)^2(5/6)^1 && \text{(by Fact 4.6)} \\ &= 15/216. \end{aligned}$$
- $$\begin{aligned} p(+3) &= P\{X = +3\} \\ &= P\{Y = 3\} \\ &= C(3, 3)(1/6)^3(5/6)^0 && \text{(by Fact 4.6)} \\ &= 1/216. \end{aligned}$$
- To determine if the game is fair, we may compute $E[X]$ (as we did in Example 4.8) to see if its value is zero:

$$E[X] = \sum_{x:p(x)>0} xp(x_i) \qquad \text{(by the definition of mean)}$$

$$= \sum_{x:p(x)>0} xP\{X = x_i\} \qquad \text{(by the definition of pmf)}$$

= (□1)×(125/216) + 1×(75/216) + 2×(15/216) + 3×(1/216)
= □17/216.

- This result means that in the long run, the player loses 17 units per every 216 games, or equivalently, loses □17/216 units in each game on the average.
- So the game is *unfair*!

4. Properties of binomial random variables

- ***Fact 4.7***

  If *X* is a binomial random variable with parameter (*n*, *p*), then

$$E[X] = np;$$
$$\text{Var}(X) = np(1 \;\square\; p).$$

Proof: see the reference book.

If we run n trials, where the probability of success for each single trial is p, what is the probability of exactly k successes?

$$\frac{1-p}{1} \quad \frac{p}{2} \quad \frac{p}{3} \quad \frac{1-p}{4} \quad \frac{p}{5} \;\ldots\ldots\; \frac{\_}{n}$$

k slots where prob. success is p , n-k slots where prob. failure is $1-p$
Thus, the probability of obtaining a specific configuration as denoted above is $p^k(1\text{-}p)^{n\text{-}k}$. From here, we must ask ourselves, how many configurations lead to exactly k successes. The answer to this question is simply, "the number of ways to choose k slots out of the n slots above. This is $\binom{n}{k}$. Thus, we must add $p^k(1\text{-}p)^{n\text{-}k}$ with itself exactly $\binom{n}{k}$ times.
This leads to the following answer to the given question:

$$\binom{n}{k} p^k (1-p)^{n-k}$$

We can also define a discrete random variable based on a binomial distribution. We can simply allow the variable to equal the number of successes of running a binomial trial n times. We then separately calculate the probability of obtaining 0 successes, 1 success, etc. , n successes. Here is a concrete example with n = 3 and p = 1/3:

X = 0, with probability $\binom{3}{0}(\frac{1}{3})^0(\frac{2}{3})^3 = \frac{8}{27}$

X = 1, with probability $\binom{3}{1}(\frac{1}{3})^1(\frac{2}{3})^2 = \frac{12}{27}$

X = 2, with probability $\binom{3}{2}(\frac{1}{3})^2(\frac{2}{3})^1 = \frac{6}{27}$

X = 3, with probability $\binom{3}{3}(\frac{1}{3})^3(\frac{2}{3})^0 = \frac{1}{27}$

We can calculate that E(X) = $1(\frac{12}{27}) + 2(\frac{6}{27}) + 3(\frac{1}{27}) = 1$.

Why can we leave at the term when X = 0? Also, why is this value in tune with our intuitive idea of what we should expect? We can formally prove this intuitive notion, namely that for a binomial distribution X, E(X) = np.

H. The Poisson Random Variable

1. Definition of Poisson random variables

♦ ***Definition 4.11***

A random variable *X* taking on one of the values 0, 1, 2, ..., is said to be a Poisson random variable with parameter □ if for some □ > 0, its pmf is of the following form:

$$p(i) = P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!} \qquad \forall\ i = 0, 1, 2, ... \qquad (4.3)$$

(Note: Poisson is pronounced as /pwason/.)

♦ *A comment:*

The Poisson random variable has a lot of applications because it may be used as an approximation of the binomial random variable with parameters (*n*, *p*) when *n* is large and *p* is small enough so that *np* is a moderate value. See the following fact.

2. Approximation of binomial random variables with Poisson random variables

♦ ***Fact 4.8***

When □ = *np* is moderate, we have

$$P\{X = i\} \approx e^{-\lambda}\frac{\lambda^i}{i!} \quad \forall\ i = 1, 2, ..., n$$

where *X* is a binomial random variable with parameters (*n, p*).

(Note: "≈" means "approximately equals.")

Proof: see the reference book.

- ***The meaning of approximation indicated by Fact 4.8 above ---***

  If n independent trials are performed with each resulting in a success with probability p and a failure with probability $1 - p$, then when n is large and p small enough to make np moderate, the number of successes occurring is approximately a Poisson random variable with parameter □ = np.

- ***Applications of the Poisson random variable ---***

  There are a lot of the Poisson random variables:

  - No. of misprints on a page of a book.
  - No. of people in a community living to the age of 100.
  - No. of wrong telephone numbers that are dialed in a day.
  - ....

  (Note: the abbreviation "No." means "the number of," and is equivalent to "#" which we have used before.)

  Why? Because the above numbers of various objects or peoples are all binomial random variables which may be approximated by the Poisson random variable.

Example 4.14

Suppose that the probability that an item produced by a certain machine will be defective is 0.1. Find the probability that a sample of 10 items will contain at most 1 defective item.

Solution:

- According to the binomial random variable, the desired probability for 0 or 1 defective item is

$$
\begin{aligned}
P\{X \le 1\} &= P\{X = 0\} + P\{X = 1\} \\
&= C(10, 0)(0.1)^0(0.9)^{10} + C(10, 1)(0.1)^1(0.9)^9 \\
&= 0.7361.
\end{aligned}
$$

- *Poisson approximation* using $P\{X = i\} \approx e^{-\lambda}\dfrac{\lambda^i}{i!}$ with □ = $np = 10\times 0.1 = 1$ is

$$P\{X \le 1\} = P\{X = 0\} + P\{X = 1\} = e^{-1}\frac{1^0}{0!} + e^{-1}\frac{1^1}{1!} = 2e^{□1} = 0.7358$$

  which is close to 0.7361 computed above!

3. The mean and variance of a Poisson random variable

- ***Fact 4.9***

  If X is a Poisson random variable with parameter □, then

$$E[X] = □;$$
$$\text{Var}(X) = □.$$

Proof: see the reference book.

CHAPTER 5
PROBABILITY DISTRIBUTIONS 2
(CONTINUOUS RANDOM VARIABLES)

A. Introduction

Recall – A random variable x is called **continuous** if the possible values of x are all real values in some interval.

To describe the probability distribution of a continuous random variable we use a **probability density function** $p(x)$**.**

Calculating Probabilities Using the Probability Density Function

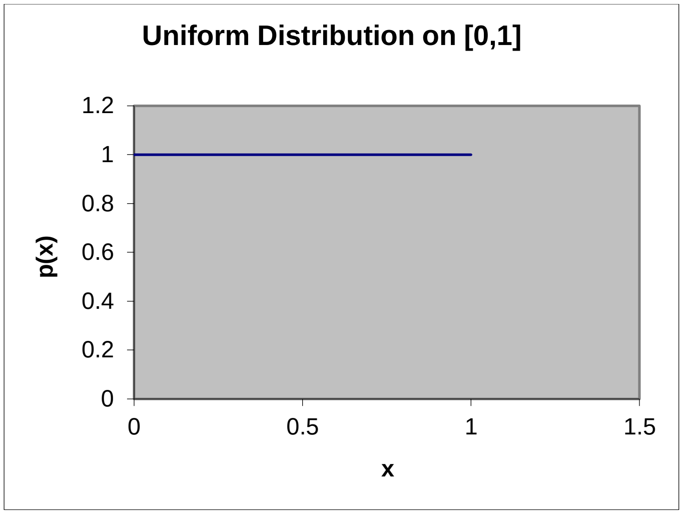
If x is a continuous random variable, then the probability that the value of x will fall between the values a and b is given by the area of the region lying below the graph of $p(x)$ and above the x-axis between a and b.

Note: For any probability density function:

- $p(x) \geq 0$ for all x
- The total area under the graph of $p(x)$ must be 1.

Example: The Continuous Uniform Distribution.

If the random variable x is limited to having values between 0 and 1, then the function $p(x) = 1$ is a possible probability density function for x. From the graph of $p(x)$ shown on the next slide, we se that the area below $p(x)$ between $x = 0$ and $x = 1$ is equal to 1 since the area is simply a 1×1 square.


Uniform Distribution on [0,1]
1.2
1
0.8
0.6
0.4
0.2
0
p(x)
0
0.5
1
1.5
x

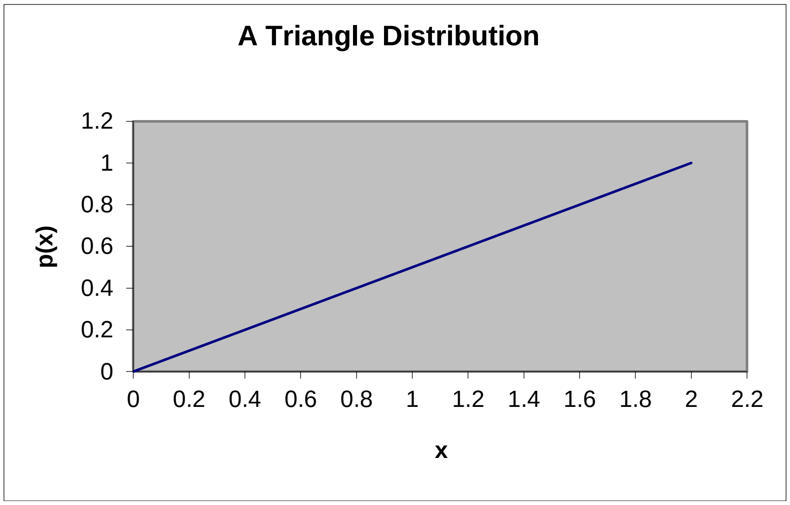
Observe that the probability that *x* lies between .3 and .7 can also be calculated from the graph. The area below $p(x)$ between .3 and .7 forms a rectangle of width .4 and height 1, so the area is $.4\times 1=.4$ and so this is the probability that *x* will assume such a value.

Exercises:

- Find $P(.15\le x\le .75)$
- Find $P\left(\frac{1}{4}\le x\le \frac{2}{3}\right)$

Example: A triangle distribution.

If *x* is a random variable whose values lie between 0 and 2, then the function $p(x)=\frac{x}{2}$ is a possible probability density function for *x*. Observe that the area under the graph of $p(x)$ is a triangle of base 2 and height 1, so the area is $A=\frac{1}{2}\times 2\times 1=1$.



Exercises:

- Find $P(x\le 1)$
- Find $P(x\ge 1.5)$
- Find $P(.5\le x\le 1.5)$

B. Normal Distributions

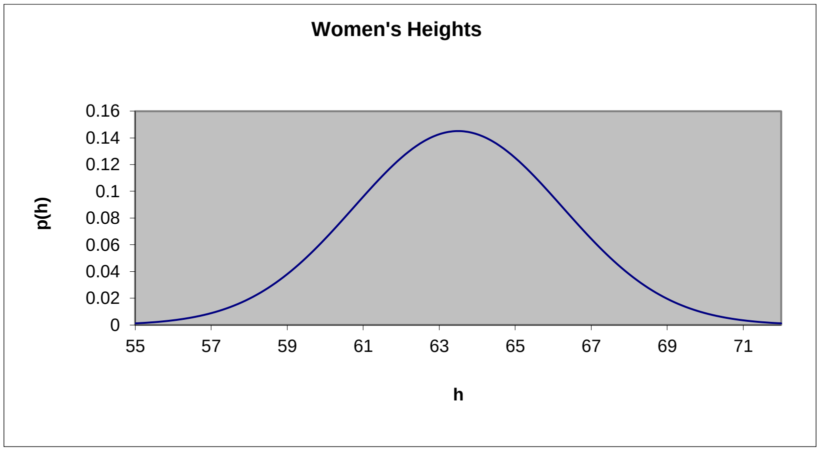
The most commonly used continuous random variables in statistics are **normal random variables**. A continuous random variable *x* is normally distributed if the possible values of *x* are all real numbers and if the probability density function for *x* is given by:

$$p(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$

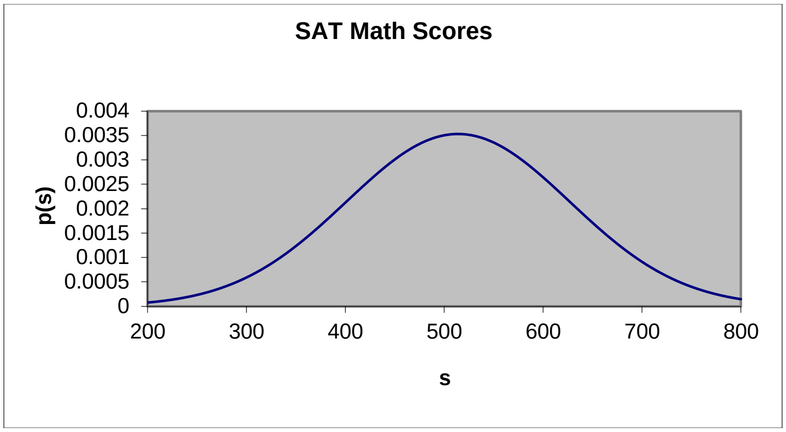
Where the constants μ and σ are the desired mean and standard deviation of *x*.

The graph of the function above is a symmetric bell-shaped distribution. Many quantities measured in everyday life have a distribution which closely matches that of a normal random variable.

Example: Suppose a woman is chosen at random from the population of Indonesian women 19-29 years old. If the random variable *h* represents the height of the woman in inches, then *h* is approximately normally distributed with a mean of $\mu = 63.5$ and a standard deviation of $\sigma = 2.75$ inches. The graph of this distribution is as shown below.



Example: SAT Physics Scores. Suppose a taker of the 2000 SAT Mathematics Exam is chosen randomly. If the random variable *s* represents the taker's score on the exam, then *s* is approximately normally distributed with a mean of $\mu = 514$ and a standard deviation of $\sigma = 113$. The graph of this distribution is as shown below.

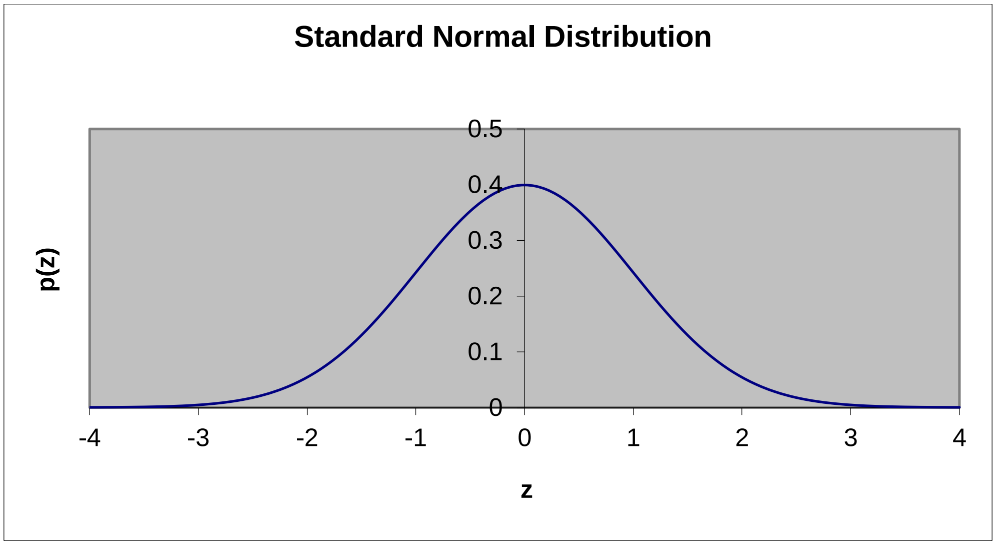


Note: The *inflection points* of these graphs lie exactly one standard deviation from the mean.

C. The Standard Normal Distribution

To find probabilities for a normally distributed random variable, we need to be able to calculate the areas under the graph of the normal distribution.

The table in the front of your book gives the area calculations for a special normal distribution, the **Standard Normal Distribution** which has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$. The graph of this distribution appears below.



Note: The table gives the area under the curve to the **left** of the value *z*. Other types of areas can be found by combining several of the areas as shown in the next example.

D. Calculating Probabilities Using the Standard Normal Table

To find the area under the standard normal curve to the left of a given value of *z*, we look up the ones and tenths values of *z* in the column at the right and the hundredths value in the row at the top of the table. The area (probability) is then the value in the table at that column and row.

Example: If *z* is a continuous random variable with the standard normal distribution, then by using the Standard Normal Table we can see that:

$$P(z \le -.47) = .3192 \text{ and } P(z \le 1.28) = .8997$$

$$P(z \ge -.47) = 1 - .3192 = .6808 \text{ and } \qquad P(z \ge 1.28) = 1 - .8997 = .1003$$

$$P(-.47 \le z \le 1.28) = .8997 - .3192 = .5805$$

Exercises: Compute the following probabilities for a random variable *z* with the standard normal distribution.

- $P(z \le 0.88)$
- $P(z \ge -1.96)$
- $P(-.32 \le z \le 1.10)$
- $P(z \le -1.22 \text{ or } z \ge 1.22)$

E. Calculating Probabilities for a Normal Random Variable

If *x* is a normally distributed random variable with mean μ and standard deviation σ, then the **z-scores** of the values for the random variable have the standard normal distribution. That is the random variable *z* defined by:

$$z = \frac{x - \mu}{\sigma}$$

is normally distributed with $\mu = 0$ and $\sigma = 1$.

Therefore, any interval for the variable *x* can be written as an interval for the z-score *z* and then the probability found by using the Standard Normal Table.

Example: The height *h* (in inches) of a randomly selected woman is approximately normally distributed with a mean of $\mu = 63.5$ and a standard deviation of $\sigma = 2.75$ inches. To calculate the probability that a woman is less than 63 inches tall, we first find the z-score for 63 inches:

$$z = \frac{63 - 63.5}{2.75} = \frac{-0.5}{2.75} = -0.18$$

Thus $P(h \le 63) = P(z \le -0.18)$.

Using the Standard Normal Table, we see that $P(z \le -0.18) = .4286$. So the probability that a randomly chosen woman's height is less than 63 inches is also .4286 or equivalently, 42.86% of women are less than 63 inches tall.

Exercises: Using the information from the women's height example above, Calculate:

- $P(h \ge 65)$
- $P(60 \le h \le 70)$
- $P(63.5 \le h \le 72)$

F. Calculating Values Using the Standard Normal Table

The Standard Normal Table can be used to find percentiles for variables which are normally distributed.

Example: To find the score which marks the 80th percentile for SAT Math Scores, we use the fact that SAT Math scores *s* are approximately normally distributed with $\mu = 514$ and $\sigma = 113$. From the Standard Normal Table, the z-score for which closest to 80 percent of values lie to the left is 0.84 which corresponds to a probability of .7995. The SAT score which corresponds to a z-score of 0.84 can be found by solving $0.84 = \frac{s - 514}{113}$ for *s*. This yields $s = 608.92$. So a score of 609 is better than 80% of all other test scores.

Exercises: For the normal distribution above:

- Find P_{35}.

- If a person scores in the top 5% of test scores, what is the minimum score they could have received?

- If a person scores in the bottom 10% of test scores, what is the maximum score they could have received?

G. The Central Limit Theorem

The Central Limit Theorem shows why normal distributions are so common and so useful. Essentially, it says that if a large sample is drawn, the sample averages for any random variable have a normal distribution. More specifically:

Central Limit Theorem – If the following conditions are true:

- *x* is a random variable with a known mean μ and known standard deviation σ
- A random sample of *n* values of the random variable *x* is drawn
- Either *x* is normally distributed or $n \geq 30$

Then for the random variable $\overline{x}$ which represents the sample mean for our sample of *n* values, the following are also true:

- The distribution of $\overline{x}$ is approximately normal with greater values of *n* giving a closer approximation.
- The mean of $\overline{x}$ denoted by $\mu_{\overline{x}}$ is equal to the mean of *x*: $\mu_{\overline{x}} = \mu$
- The standard deviation of $\overline{x}$ denoted by $\sigma_{\overline{x}}$ is equal to the standard deviation of x divided by the square root of the sample size: $\sigma_{\overline{x}} = \sigma/\sqrt{n}$

Example: If the random variable *h* represents the height of a randomly selected woman, then *h* is normally distributed with a mean of $\mu = 63.5$ inches and a standard deviation of $\sigma = 2.75$ inches. If random samples of 16 women are selected, then the Central Limit Theorem can be applied.

The sample mean $\overline{h}$ is a normally distributed random variable with $\mu_{\overline{h}} = 63.5$ and

$$\sigma_{\overline{h}} = 2.75/\sqrt{16} = 0.6875$$

We can now calculate the probability of selecting a sample of 16 women whose average height is more than 65 inches.

We are interested in $P(\bar{h} \geq 65)$. Converting 65 to a
z-score we have:

$$z = \frac{\bar{h} - \mu_{\bar{h}}}{\sigma_{\bar{h}}} = \frac{65 - 63.5}{0.6875} = \frac{1.5}{0.6875} = 2.18$$

The probability in the Standard Normal Table for this value is .9854. Thus $P(\bar{h} \geq 65) = 1 - .9857 = .0143$. So there is only a 1.43% chance that we would choose 16 women at random with an average height of at least 65".

Note that $P(h \geq 65) = .2912$ was calculated in one of the exercises above. It is far more likely to find a single woman who is taller than 65" than a group of 16 women whose average height is more than 65".

Exercises:

- Find $P(\bar{h} \leq 64.5)$
- Find $P(63 \leq \bar{h} \leq 64.5)$

Example: If *w* represents the weight of an American man chosen at random, then *w* is a continuous random variable with a right-skewed distribution with a mean of $\mu = 180$ lbs. and a standard deviation of $\sigma = 20$ lbs. If a random sample of 43 men is selected, and $\bar{w}$ is the average weight of the 43 men, find the following probabilities:

- Find $P(175 \leq \bar{w} \leq 185)$
- Find $P(170 \leq \bar{w} \leq 190)$
- Find $P(\bar{w} \geq 162)$

Example: Consider the lottery example introduced in Chapter 4 Notes. An instant lottery ticket is purchased for $2. The possible prizes are $0, $2, $20, $200, and $1000. Let *Z* be the random variable representing the amount won on the ticket, and suppose *Z* has the following distribution:

Z	0	2	20	200	1000
$P(Z)$	.7489	.2	.05	.001	.0001

We determined the mean of *Z* to be $\mu = \$1.70$, and the standard deviation of *Z* to be $\sigma = \$12.57$.

If a player purchases 1000 random tickets, then we apply the Central Limit Theorem to $\bar{Z}$ the average amount won per ticket. $\mu_{\bar{Z}} = \$1.70$ and $\sigma_{\bar{Z}} = 12.57\big/\sqrt{1000} \approx \$0.3975$.

We can now determine the probability that the player gains money. In order for the player to win money on the 1000 tickets, $\bar{Z}$ must exceed \$2. To find $P(\bar{Z} \geq 2)$ we calculate the z-score for $\bar{Z} = 2$:

$$z = \frac{\bar{Z} - \mu_{\bar{Z}}}{\sigma_{\bar{Z}}} = \frac{2 - 1.7}{.3975} = \frac{.3}{.3975} = .75$$

The standard normal table gives a value of .7734 for the
z-score .75, and so: $P(\bar{Z} \geq 2) = 1 - .7734 = .2266$.

Exercise: Repeat the above with 10,000 tickets.

CHAPTER 6
POINT ESTIMATION OF PARAMETERS

A. Introduction

Last week you became familiar with the normal distribution. We now **estimate** the parameters of a normally distributed population by analysing a sample taken from it. In this lecture we will be concentrating on the estimation of percentages and means of populations but do note that any population parameter can be estimated from a sample.

1. Sampling

Sampling theory takes a whole lecture on its own! Since any result produced from the sample can be used to estimate the corresponding result for the population it is absolutely essential that the sample taken is as representative as possible of that population. Common sense rightly suggests that the larger the sample the more representative it is likely to be but also the more expensive it is to take and analyse. A random sample is ideal for statistical analysis but, for various reasons, other methods also have been devised for when this ideal is not feasible. We will not study sampling in this lecture but just give a list of the main methods below.

- Simple Random Sampling
- Systematic Sampling
- Stratified Random Sampling
- Multistage Sampling
- Cluster Sampling
- Quota Sampling

It is usually neither possible nor practical to examine every member of a population so we use the data from a sample, taken from the same population, to estimate the 'something' we need to know about the population itself. The sample will not provide us with the exact 'truth' but it is the best we can do. We also use our knowledge of samples to estimate limits within which we can expect the 'truth' about the population to lie and state how confident we are about this estimation. In other words instead of claiming that the mean cost of buying a small house is, say, exactly £75 000 we say that it lies between £70 000 and £80 000.

2. Types of Parameter estimates

These two types of **estimate** of a **population parameter** are referred to as:

- Point estimate - one particular value;
- Interval estimate - an interval centred on the point estimate.

3. Point Estimates of Population Parameters

From the sample, a value is calculated which serves as a point estimate for the population parameter of interest.

a. The best estimate of the population **percentage**, π, is the sample percentage, p.

b. The best estimate of the unknown population **mean**, μ, is the sample mean, $\bar{x} = \frac{\sum x}{n}$

This estimate of μ is often written $\hat{\mu}$ and referred to as 'mu hat'.

c. The best estimate of the unknown population **standard deviation**, σ, is the sample standard deviation s, where:

$s = \sqrt{\frac{\sum(x-\bar{x})^2}{(n-1)}}$ This is obtained from the $x\sigma_{n-1}$ key on the calculator.

N.B. $s = \sqrt{\frac{\sum(x-\bar{x})^2}{(n)}}$ from $x\sigma_n$ is **not** used as it underestimates the value of σ

4. Interval Estimate of Population Parameter (Confidence interval)

Sometimes it is more useful to quote two limits between which the parameter is expected to lie, together with the probability of it lying in that range. The limits are called the **confidence limits** and the interval between them the **confidence interval.**

The width of the confidence interval depends on three sensible factors:

a. the degree of confidence we wish to have in it,
i.e. the probability of it including the 'truth', e.g. 95%;

b. the size of the sample, n;

c. the amount of variation among the members of the sample, e.g. for means this the standard deviation, s.

The **confidence interval** is therefore an interval centred on the point estimate, in this case either a percentage or a mean, within which we expect the population parameter to lie. The width of the interval is dependent on the confidence we need to have that it does in fact include the population parameter, the size of the sample, n, and its standard deviation, s, if estimating means. These last two parameters are used to calculate the **standard error**, s/√n, which is also referred to as the standard deviation of the mean.

The number of standard errors included in the interval is found from statistical tables - either the normal or the t-table. Always use the normal tables for percentages which need large samples. For means the choice of table depends on the sample size and the population standard deviation:

	Population standard deviation

Sample size	Known: standard error = $\frac{\sigma}{\sqrt{n}}$	Unknown: standard error = $\frac{s}{\sqrt{n}}$
Large	Normal tables	Normal tables
Small	Normal tables	t-tables

5. Interpretation of Confidence intervals

How do we interpret a confidence interval? If 100 similar samples were taken and analysed then, for a 95% confidence interval, we are confident that 95 of the intervals calculated would include the true population mean. In practice we tend to say that we are 95% confident that our interval includes the true population value. Note that there is only one true value for the population mean, it is the variation between samples which gives the range of confidence intervals.

6. Confidence Intervals for a Percentage or Proportion

The only difference between calculating the interval for percentages or for proportions is that the former total 100 and the latter total 1. This difference is reflected in the formulae used, otherwise the methods are identical. Percentages are probably the more commonly calculated so in Example 2 we will estimate a population percentage.

The confidence interval for a population percentage or a proportion, π, is given by:

$\pi = p \pm z\sqrt{\frac{p(100-p)}{n}}$ for a percentage or $\pi = p \pm z\sqrt{\frac{p(1-p)}{n}}$

for a proportion

where: π is the unknown population percentage or proportion being estimated,

p is the sample percentage or proportion, i.e. the point estimate for π,

z is the appropriate value from the normal tables,

n is the sample size.

The formulae $\sqrt{\frac{p(100-p)}{n}}$ and $\sqrt{\frac{p(1-p)}{n}}$ represent the standard errors of a percentage and a proportion respectively.

The samples must be large, (>30), so that the normal table may be used in the formula.

We therefore estimate the confidence limits as being at z standard errors either side of the sample percentage or proportion. The value of z, from the normal table, depends upon the degree of confidence, e.g. 95%, required. We are prepared to be incorrect in our estimate 5% of the time and confidence intervals are always symmetrical so, in the tables we look for Q to be 5%, two tails.

Example 2

In order to investigate shopping preferences at a supermarket a random sample of 175 shoppers were asked whether they preferred the bread baked in-store or that from the large national bakeries. 112 of those questioned stated that they preferred the bread baked in-store. Find the 95% confidence interval for the percentage of all the store's customers who are likely to prefer in-store baked bread.

The point estimate for the population percentage,π, is p = $\frac{112}{175}\times 100$ = 64%

Use the formula: $\pi \quad = \quad p \quad \pm \quad z\sqrt{\frac{p(100-p)}{n}}$ where p = 64 and n = 175

From the short normal table 95% confidence ⇒ 5%, two tails ⇒ z = 1.96

$$p \quad \pm \quad z\sqrt{\frac{p(100-p)}{n}} \quad \Rightarrow 64 \pm 1.96 \times \sqrt{\frac{64 \times 36}{175}} \quad =$$

the confidence limits for the population percentage, π, are and <

π <

5. Confidence interval for the Population Mean, μ, when the population standard deviation, σ, is known.

Example 3: For the small supermarket as a whole it is known that the standard deviation of the wages for part-time employees is £1.50.

A random sample of 10 employees from the small supermarket gave a mean wage of £4.15 per hour. Assuming the same standard deviation, calculate the 95% confidence interval for the average hourly wage for employees of the small branch and use it to see whether the figure could be the same as for the whole chain of supermarkets which has a mean value of £4.50.

As we actually know the population standard deviation we do not need to estimate it from the sample standard deviation. The normal table can therefore be used to find the number of standard errors in the interval.

Confidence Interval: $\mu = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$ where z comes from the short normal table

$$\mu = \bar{x} \pm z\frac{\sigma}{\sqrt{n}} \Rightarrow 4.15 \pm 1.96 \times \frac{1.50}{\sqrt{10}} =$$

This interval includes the mean, £4.50, for the whole chain so the average hourly wage could be the same for all employees of the small supermarket.

6. Confidence interval for the Population Mean, μ, when the population standard deviation is not known, so needs estimating from the sample standard deviation, s.

The population standard deviation is unknown, so the t-table, must be used to compensate for the probable error in estimating its value from the sample standard deviation.

Example 4: Find the 99% confidence interval for the mean value of all the invoices, in Example 1, sent out by the small supermarket branch. If the average invoice value for the whole chain is £38.50, is the small supermarket in line with the rest of the branches?

Confidence Interval: $\mu = \bar{x} \pm t\frac{s}{\sqrt{n}}$ where the value of t comes from the table of 'percentage points of the t-distribution' using n - 1 degrees of freedom ($\nu = n - 1$)

From Example 1 $\bar{x}$ = £32.92, s = £7.12, n = 20. degrees of freedom = (20 – 1) = 19; 99% confidence; from tables t = 2.87

$$\mu = \bar{x} \pm t\frac{s}{\sqrt{n}} =$$

This interval does not include £38.50, so the small branch is out of line with the rest.

7. Comparison of Means using 'Overlap' in Confidence Intervals

We are going to extend the previous method to see if two populations could have the same mean or, alternatively, if two samples could have come from the same population as judged by their means. We assume that the standard deviations of both populations are the same.
If, for example, a supermarket chain wished to see if a local advertising campaign was successful or not they could take a sample of customer invoices before the campaign and another after the campaign and calculate confidence intervals for the mean spending of all customers at both times. If the intervals were found to overlap the means could be the same so the campaign might have had no effect. If, on the other hand, they were quite separate with the later sample giving the higher interval then the campaign must have been effective in increasing sales.

Example 5

The till slips of supermarket customers were sampled both before and after an advertising campaign and the results were analysed with the following results:

Before: $\bar{x}$ = £37.60, s = £6.70, n = 25
After: $\bar{x}$ = £41.78, s = £5.30, n = 25

Has the advertising campaign been successful in increasing the mean spending of all the supermarket customers? Calculate two 95% confidence intervals and compare the results.

For both, n = 25 so 24 degrees of freedom giving t = 2.06 for 5%, 2-tails.

<u>Before</u>: $\mu_B = \bar{x}_B \pm t\frac{s_B}{\sqrt{n_B}} \Rightarrow 37.60 \pm 2.06 \times \frac{6.70}{\sqrt{25}} = 37.60 \pm 2.76$

$$£34.84 < \mu < £40.36$$

<u>After</u>: $\mu_A = \bar{x}_A \pm t\frac{s_A}{\sqrt{n_A}} \Rightarrow$

Interpretation: The sample mean had risen considerably but, because the confidence intervals overlap, the mean values for **all** the sales may lie in the common ground. There may be no difference between the two means so the advertising campaign has
not been proved to be successful.

We shall improve on this method in the next example.

8. Confidence Intervals for Paired Data

If two measures are taken from each case, i.e. 'before' and 'after', in every instance then the 'change' or 'difference' for each case can be calculated and a confidence interval for the mean of the 'changes' calculated. If the data can be 'paired', i.e. it is not independent, then this method should be used as a smaller interval is produced for the same percentage confidence giving a more precise estimate.

Confidence Interval:

$\mu_d = \bar{x}_d \pm t\frac{s_d}{\sqrt{n_d}}$ $\bar{x}_d$, s_d and n_d refer to the calculated differences.

Example 6

The supermarket statistician realised that there was a considerable range in the spending power of its customers. Even though the overall spending seemed to have increased the high spenders still spent more than the low spenders and that the individual increases would show a smaller spread. In other words these two populations, 'before' and 'after', are not independent.

Before the next advertising campaign at the supermarket, he took a random sample of 10 customers, A to J, and collected their till slips. After the campaign, slips from the same 10 customers were collected and both sets of data recorded. Using the paired data, has there been any mean change at a 95% confidence level?

	A	B	C	D	E	F	G	H	I	J
Before	42.30	55.76	32.29	10.23	15.79	46.50	32.30	78.65	32.20	15.90

After	43.09	59.20	31.76	20.78	19.50	50.67	37.32	77.80	37.39	17.24

We first need to calculate the differences. The direction doesn't matter but it seems sensible to take the earlier amounts away from the later ones to find the changes:

	A	B	C	D	E	F	G	H	I	J
Diffs.	0.79	3.44	-0.53	10.55	3.71	4.17	5.02	-0.85	5.19	1.34

We can now forget the original two data sets and just work with the differences.

From the differences $\bar{x}_d = £3.28$, $s_d = £3.37$ and $n_d = 10$

$$\underline{\text{95\% C.I.}}\ \mu_d \quad = \quad \bar{x}_d \pm t\frac{s_d}{\sqrt{n_d}} \quad \Rightarrow \quad 3.28 \pm 2.26 \times \frac{3.37}{\sqrt{10}} \quad \Rightarrow \quad 3.28 \pm 2.41$$

$$£0.87 < \mu_d < £5.69$$

This interval does not include zero so the possibility of 'no change' has been eliminated. Because the mean amount spent after the campaign is greater than that before it, there has been a significant increase in spending.

In general: If the confidence interval includes zero, i.e. it changes from negative to positive, then there is a possibility that no change has taken place and the original situation has remained unchanged. If both limits have the same sign, as above, then zero is excluded and some change must have taken place.

When interpreting this change look carefully at the direction of the change as this will depend on your order of subtraction. In the above case it is obvious that $\bar{x}_d$ is positive so there has been an increase in the average spending.

B. Properties of Point Estimators and Methods of Estimation

An estimator $\hat{\theta}$ for a target parameter θ is a function of the random variables and therefore it is itself a random variable.

Consequently an estimator has a probability distribution which we call the *sampling distribution* of the estimator.

We noted before that if $E\left(\hat{\theta}\right)=\theta$, the estimator is unbiased.

1. **Relative Efficiency**

   We know that usually it is possible to obtain more than one unbiased estimator for the same target parameter. If we have two unbiased estimators we prefer the one with the smaller variance.

   The definition of the relative efficiency:

   $$eff\left(\hat{\theta}_1,\hat{\theta}_2\right)=\frac{V\left(\hat{\theta}_1\right)}{\hat{\theta}_1\left(\hat{\theta}_2\right)}$$

   How do you interpret it?

2. **Consistency**

   Suppose that a coin, which has probability *p* of resulting in heads, is tossed *n* times. If *p* is unknown, the sample proportion, Y/n, is an estimator of *p*.
   What happens to it if *n* increases?
   Intuitively, as *n* increases, the increases, Y/n should get closer to *p*.

3. **Definition**

   $\hat{\theta}_n$ is a *consistent estimator* of θ if for any $\varepsilon>0$

   $$\lim_{n\to\infty}P\left(\left|\hat{\theta}_n-\theta\right|\le\varepsilon\right)=1$$
   $$or$$
   $$\lim_{n\to\infty}P\left(\left|\hat{\theta}_n-\theta\right|>\varepsilon\right)=0$$

   Theorem 6.1

   An unbiased estimator $\hat{\theta}_n$ for θ is a consistent estimator of θ if

   $$\lim_{n\to\infty}V\left(\hat{\theta}_n\right)=0$$

   Previously we considered $\overline{Y}_1-\overline{Y}_2$ as an intuitive estimator for $\mu_1-\mu_2$. The next theorem will be useful in establishing the consistency of such estimators.

   Theorem 6.2

   Suppose that $\hat{\theta}_n$ converges in probability to θ and that $\hat{\theta}_n\,'$ converges in probability to θ'.

a. $\hat{\theta}_n + \hat{\theta}_n{}'$ converges in probability to $\theta + \theta'$

b. $\hat{\theta}_n \times \hat{\theta}_n{}'$ converges in probability to $\theta \times \theta'$

c. $\hat{\theta}_n / \hat{\theta}_n{}'$ converges in probability to θ / θ', provided that $\theta' \neq 0$.

d. If g(.) is a real-valued function that is continuous at θ, then $g\left(\hat{\theta}_n\right)$ converges in probability to $g\left(\theta\right)$.

We considered large-sample confidence intervals for some parameters or practical interest. In particular, if $Y_1, Y_2, \ldots, Y_n$ is a random sample form any distribution with mean μ and variance σ^2, we established that

$$\bar{Y} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Is a valid large sample confidence interval with confidence coefficient $\approx (1-\alpha)$. If sample size is large and σ^2 is unknown, it is recommended to substitute S for σ. The following theorem provides the theoretical justification.

Theorem 6.3

Suppose that U_n has a distribution that converges to a standard normal distribution as $n \to \infty$. If W_n converges in probability to 1, then the distribution function of U_n / W_n converges to a standard normal distribution function.

4. Sufficiency

Up to this point we have chosen estimators on the basis of intuition. We have shown that $\bar{Y}$ and S^2 are unbiased estimators of μ and σ^2. Are we loosing any information about our target parameters relying on these statistics?

In this section we present methods for finding statistics, that summarizes all the information about target parameters. Such statistics are said to have the property of *sufficiency*, or they are called *sufficient statistics.*

“Good” estimators are (or can be made to be) functions of any sufficient statistic.

To illustrate let us consider the outcomes of *n* trials of binomial experiment, $X_1, X_2, \ldots, X_n$, where

$$X_i = \begin{cases} 1, & \text{if the } i\text{th trial is a success} \\ 0, & \text{if the } i\text{th trial is a failure} \end{cases}$$

If *p* is a probability of success on any trial then, for $i = 1, 2, \ldots, n$,

$$X_i = \begin{cases} 1, & \text{with probability p} \\ 0, & \text{with probability q=1-p} \end{cases}$$

Suppose we are given

$$Y = \sum_{i=1}^{n} X_i$$

The number of successes among the *n* trials. If we know the value of *Y*, can we gain any further information about *p* by looking at other functions of $X_1, X_2, \ldots, X_n$?

One way to answer this question is to look at the conditional distribution of $X_1, X_2, \ldots, X_n$ given *Y*:

$$P\left(X_1 = x_1, \ldots, X_n = x_n \mid Y = y\right)$$
$$= \frac{P\left(X_1 = x_1, \ldots, X_n = x_n, Y = y\right)}{P\left(Y = y\right)}$$

The numerator is 0 unless $Y = \sum_{i=1}^{n} x_i$, and it is the probability of an independent sequence of 0s and 1s with the total sum of *y* 1s and (*n*-*y*) 0s if $Y = \sum_{i=1}^{n} X_i$. Also the denominator is the binomial probability of exactly *y* success in *n* trials. Therefore, if $y = 1, 2, \ldots, n,$

$$P\left(X_1 = x_1, \ldots, X_n = x_n \mid Y = y\right)$$
$$= \begin{cases} \dfrac{p^y (1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \dfrac{1}{\binom{n}{y}} & \text{if } \sum_{i=1}^{n} x_i = y \\ 0 & \text{otherwise} \end{cases}$$

It is important to note that the conditional distribution of $X_1, X_2, \ldots, X_n$ given Y does not depend upon *p*. That is, once Y is known, no other function of $X_1, X_2, \ldots, X_n$ will shed additional light on the possible value of *p*. Therefore statistics Y is said to be *sufficient* for *p*.

Definition 9.3

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from a probability distribution with unknown parameter θ. Then the statistics is said to be sufficient for θ if the conditional distribution of $Y_1, Y_2, \ldots, Y_n$ given *U* does not depend on θ.

This definition tells us how to check whether statistic is sufficient, but it really does not tell us how to find a sufficient statistic. We will move towards it in two steps.

First we will define the concept of *likelihood*.
Recall that in the discrete case the joint distribution of discrete random variables $Y_1, Y_2, \ldots, Y_n$ is given by a probability function of

$p(y_1, y_2, \ldots, y_n)$.

If this joint probability depends explicitly on the value of a parameter θ, we write it as

$p(y_1, y_2, \ldots, y_n \mid \theta)$.

This function gives the probability or *likelihood* of observing the event $(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$ given parameter θ.

In the continuous case we will write the joint density function

$f(y_1, y_2, \ldots, y_n)$.

It will be convenient to have a single name.

Definition 9.4
Let $y_1, y_2, \ldots, y_n$ be sample observations taken on correspondent random variables $Y_1, Y_2, \ldots, Y_n$ whose distribution depends on parameter θ.
Then, if $Y_1, Y_2, \ldots, Y_n$ are discrete random variables, the *likelihood of the sample*, $L(y_1, y_2, \ldots, y_n \mid \theta)$ is defined to be the joint probability of $y_1, y_2, \ldots, y_n$.
If $Y_1, Y_2, \ldots, Y_n$ are continuous random variables,. The likelihood $L(y_1, y_2, \ldots, y_n \mid \theta)$ is defined to be the joint density evaluated at $y_1, y_2, \ldots, y_n$.

Theorem 9.4
Let *U* be a statistic based on the random sample $Y_1, Y_2, \ldots, Y_n$. Then *U* is *sufficient statistic* for the estimation of a parameter θ if and only if the likelihood $L(y_1, y_2, \ldots, y_n \mid \theta)$ can be factored into two nonnegative functions

$$L(y_1, y_2, \ldots, y_n \mid \theta) = g(u, \theta) \times h(y_1, y_2, \ldots, y_n)$$

Where $g(u,\theta)$ is a function only of *u* and θ, and $h(y_1, y_2, \ldots, y_n)$ is not a function of θ.

C. The Rao-Blackwell Theorem

In words:

If

a. $\hat{\theta}$ is an unbiased estimator for θ

and

b. *U* is a statistic that is sufficient for θ,

then there is a function of *U* that is

i) an unbiased estimator for θ

ii) which has *no larger* variance than $\hat{\theta}$

Definition

A sufficient statistic for a parameter θ is called a *minimal sufficient statistic* if it can be expressed as a function of any sufficient statistic for θ.

How to find it?

1. *Typically* factorization criterion
2. Lehman and Scheffe method:

Consider the ratio of likelihoods evaluated at two points, $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$:

$$\frac{L(x_1, x_2, \ldots, x_n \mid \theta)}{L(y_1, y_2, \ldots, y_n \mid \theta)}$$

Many times it is possible to find a find a function

$$g(x_1, x_2, \ldots, x_n)$$

Such that this ratio is free of the unknown parameter θ if and only if:

$$g(x_1, x_2, \ldots, x_n) = g(y_1, y_2, \ldots, y_n)$$

If such a function can be found, then $g(Y_1, Y_2, \ldots, Y_n)$ is a minimal sufficient statistic for θ.

MVUE = a minimum variance unbiased estimator

D. The Method of Moments

i) Sample moments should provide good estimates of the corresponding population moments.

ii) Because the population moments are functions of population parameters, we can use i) to get these parameters

Formal Definition:

Choose as estimates those values of the parameters that are solutions of the equations $\mu_k^{'} = m_k^{'}$, for $k = 1, 2, ..., t$, where t is the number of parameters to be estimated.

Example 9.11

A random sample $Y_1, Y_2, ..., Y_n$ is selected from a population in which Y_i possesses a uniform density function over the interval $(0, \theta)$ where θ is unknown. Use the method of moments to estimate θ.

Solution

The value of $\mu_1^{'}$ for a uniform random variable is

$$\mu_1^{'} = \mu = \frac{\theta}{2}$$

The corresponding first sample moment is

$$m_1^{'} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$$

From which:

$$\mu_1^{'} = \frac{\theta}{2} = \overline{Y}$$

Thus,

$$\hat{\theta} = 2\overline{Y}$$

E. Method of Maximum Likelihood

Suppose that the likelihood function depends on *k* parameters, $\theta_1, \theta_2, ..., \theta_k$. Choose as estimates those values of the parameters that maximize the likelihood

$$L(y_1, y_2, ..., y_n \mid \theta_1, \theta_2, ..., \theta_k)$$

Example 9.14

Binomial experiment with *n* trials resulted in observations $(y_1, y_2, ..., y_n)$, where y_i=1 if trial is successful and 0 otherwise. Find the maximum likelihood estimator of *p*, the probability of success.

Solution

$$L(p) = L(y_1, y_2, ..., y_n \mid p) = p^y (1-p)^{n-y},$$

where $y = \sum_{i=1}^{n} y_i$

Now we want to maximize it with respect to *p*. Since $\ln\left[L(p)\right]$ is a monotonically increasing function of $L(p)$, the value of p which maximizes both functions will be the same.

$\ln\left[L(p)\right] = y\ln(p) + (n-y)\ln(1-p)$

FOC: $\dfrac{y}{p} - \dfrac{n-y}{1-p} = 0$

From which $\hat{y} = \dfrac{y}{n}$

Summary

Moment estimators are consistent but generally not very efficient.

MLEs are consistent and, if adjusted to be unbiased, often lead to minimum variance estimators.

MLEs – a popular method of estimation

the density function of *X* and suppose that *U* = *h*(*X*). Show that *U* is minimally sufficient for *a* if the following condition holds:

f(*x* | *a*) / *f*(*y* | *a*) does not depend on *a* if and only if *h*(*x*) = *h*(*y*).

E. The Factorization Theorem

The definition precisely captures the intuitive notion of sufficiency given above, but can be difficult to apply. We must know in advance a candidate statistic *U*, and then we must be able to compute the conditional distribution of *X* given *U*. The *factorization theorem* given in the next exercise frequently allows the identification of a sufficient statistic from the form of the density function of *X*.

3. Let *f*(*x* | *a*) denote the density function of *X*. Show that *U* = *h*(*X*) is sufficient for *a* if and only if there exist functions *G*(*u* | *a*) and *r*(*x*) such that
   f(*x* | *a*) = *G*[*h*(*x*) | *a*] *r*(*x*) for *x* in *S* and *a* in *A*.
   As the notation indicates, *r* depends only on the data *x* and not on the parameter *a*.
4. Show that if *U* and *V* are equivalent statistics and *U* is sufficient for *a* then *V* is sufficient for *a*.
5. Suppose that the distribution of *X* is a *k*-parameter exponential families with the natural statistic *h*(*X*). Show that *h*(*X*) is sufficient for *a*.
   Because of this result, *h*(*X*) is referred to as the *natural sufficient statistic* for the exponential family.
6. Suppose that X_1, X_2, ..., X_n is a random sample of size *n* from the normal distribution with mean μ in *R* and variance $d^2 > 0$.
   a. Show that ($X_1 + X_2 + \cdots + X_n$, $X_1^2 + X_2^2 + \cdots + X_n^2$) is sufficient for (μ, d^2),
   b. Show that (*M*, S^2) is sufficient for (μ, d^2) where *M* is the sample mean and S^2 is the sample variance. *Hint*: Use part (a) and equivalence.
7. Suppose that X_1, X_2, ..., X_n is a random sample of size *n* from the Poisson distribution with parameter *a* > 0. Show that $X_1 + X_2 + \cdots + X_n$ is sufficient for *a* where

8. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the gamma distribution with shape parameter $k > 0$ and scale parameter $b > 0$.

a. Show that $(X_1 + X_2 + \cdots + X_n, X_1 X_2 \cdots X_n)$ is sufficient for (k, b).

b. Show that (M, U) is sufficient for (k, b) where M is the (arithmetic) sample mean and U is the geometric sample mean. *Hint*: Use part (a) and equivalence.

9. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the beta distribution with parameters $a > 0$ and $b > 0$. Show that (U, V) is sufficient for (a, b) where

$U = X_1 X_2 \cdots X_n$, $V = (1 - X_1)(1 - X_2) \cdots (1 - X_n)$.

10. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the uniform distribution on the interval $[0, a]$ where $a > 0$. Show that $X_{(n)}$ (the n'th order statistic) is sufficient for a.

12. Show that if U and V are equivalent statistics and U is minimally sufficient for a then V is minimally sufficient for a.

13. Suppose that the distribution of X is a k-parameter exponential family with natural sufficient statistic $U = h(X)$. Show that U is a minimally sufficient for a. *Hint*: Recall that j is the smallest integer such that X is a j-parameter exponential family.

14. Show that the sufficient statistics given above for the Bernoulli, Poisson, normal, gamma, and beta families are minimally sufficient for the given parameters.

15. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the uniform distribution on the interval $[a, a + 1]$ where $a > 0$. Show that $(X_{(1)}, X_{(n)})$ is minimally sufficient for a.

In the last exercise, note that we have a single parameter, but the minimally statistics is a vector of dimension 2.