Chapter 1

Overview and Descriptive Statistics

What is statistics and why it is important?

The discipline of statistics teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.

Without the presence of uncertainty and variation, there would be little need for statistical methods or statisticians.

Its ultimate goal is translating data into knowledge and understanding of the world around us.

Statistics is the art (and science) of learning from data. It is used to do inference or prediction based on the collected data.

Statistics is very important for the decision making in industry and all other areas.

Examples:

1. Weather forecasts: Computer models are built using statistics that compare prior weather conditions with current weather to predict future weather.

2. Political forecasting: Election forecasting models usually combine hundreds of opinion polls with historical and demographic information to calculate the odds.

3. Auto and home insurance: The premium rate that an insurance company charges you is based upon statistics from all drivers or homeowners in you area.

4. Biology and public health: Statisticians are helping find the important genes, called markers, which influence the whole network of genes. These markers can be used to predict disease risk.

5. Stock market: Stock analysts also use statistical and computational models to forecast what is happening in the economy and to predict the stock trend in the future.

6. Quality Testing: Companies make thousands of products every day and they want to make sure that good quality items are sold. But a company can't test each and every time that they sell a product to you, the consumer. So the company uses statistics to test just a few, called a sample, of what they make. If the sample passes quality tests, then the company assumes that all the items made in the group are good.

1.1 Populations, Samples, and Processes

Table 1: Student information collected from STAT 155 class.

	Gender	Height(in)	IQ
Student 1	M	70	110
Student 2	M	73	121
Student 3	F	65	108
Student 4	M	78	135
Student 5	F	63	115
Student 6	F	68	138
Student 7	F	72	113
Student 8	M	76	123
Student 9	M	68	117
Student 10	F	66	140

Introduction of Basic Terms

- **Population** – the entire collection of objects whose properties are to be analyzed in a particular study.

- **Sample** – a subset of the population.

- **Variable** – any characteristic of interest for each object in a population or a sample.
  We shall initially denote variables by lowercase letters from the end of our alphabet.
  x = Gender
  y = Height
  z = IQ

- **Observation** – the set of measurements obtained for a particular object.

- **Parameter** – a numerical value summarizing the population data.

- **Statistic** – a numerical value summarizing the sample data.

Types of Variables

- **Qualitative/Categorical variables** use labels or names to identify an attribute of an element. Each data value belongs to one of a set of categories. *Arithmetic operations, such as addition and averaging, are NOT meaningful for data resulting from a categorical variable.*

- **Quantitative/Numerical variables** use numeric values that represent different magnitudes of the variable. *Arithmetic operations, such as addition and averaging, are meaningful for data resulting from a numerical variable.*

- **Discrete variable** – a quantitative variable that can take on only a finite or at most a countably infinite number of values. Intuitively, a discrete variable can assume values corresponding to isolated points along a line interval. That is, there is a gap between any two values.

- **Continuous variable** – a quantitative variable that can assume an uncountable number of values. Intuitively, a continuous variable can assume any value along a line interval, including every possible value between any two values.

Univariate, Bivariate and Multivariate data

- **Univariate data** – observations on a single variable
- **Bivariate data** – observations on two variables
- **Multivariate data** – observations on more than one variable

Branches of Statistics

- **Descriptive Statistics** – summarize and present important features of the data in a form that is easy for a reader to understand. These summaries may be tabular, graphical, or numerical.
- **Inferential statistics** – use techniques for generalizing from a sample to a population.

The measurements we make of a variable vary from object to object.

Likewise, results of descriptive and inferential statistics vary, depending on the sample chosen.

The study of variability is a key part of statistics.

Collecting Data

Statistics deals not only with the organization and analysis of data once it has been collected but also with the development of techniques for collecting data.

Sample survey is the process of collecting data on a sample. **Census** is the process of collecting data on the entire population.

It is important to obtain good, representative data. Inferences are made based on statistics obtained from the data. If data are not properly collected, an investigator may not be able to answer the questions under consideration with a reasonable degree of confidence.

With simple random sampling, each subset of objects of the specified size in the population has the same chance of being the sample. This is desirable, because then the sample tends to be a good reflection of the population.

Sometimes alternative sampling methods can be used to make the selection process easier, to obtain extra information, or to increase the degree of confidence in conclusions.

One such method, stratified sampling, entails separating the population units into nonoverlapping groups and taking a sample from each one using simple random sampling.

Frequently, a convenience sample is obtained by selecting objects without systematic randomization.

1.2 Pictorial and Tabular Methods in Descriptive Statistics

Descriptive statistics can be divided into two general subject areas. In section 1.2, we consider representing a data set using visual techniques.

- stem-and-leaf displays
- dotplots
- histograms

In sections 1.3 and 1.4, we will develop some numerical summary measures for data sets.

Notation

n = sample size (number of observations in the sample)

Data values occurring in a sample are symbolically represented by $x_1, x_2, x_3, \cdots, x_n$.

Stem-and-Leaf Display

It's a quick way to obtain an informative visual representation of a numerical data set.

Steps for Constructing a Stem-and-Leaf Display

1. Select one or more leading digits for the stem values. The trailing digits become the leaves.
2. List possible stem values in a vertical column.

3. Record the leaf for every observation beside the corresponding stem value.
4. Indicate the units for stems and leaves someplace in the display.

A stem-and-leaf display conveys information about the following aspects of the data:

- identification of a typical or representative value
- extent of spread about the typical value
- presence of any gaps in the data
- extent of symmetry in the distribution of values
- number and location of peaks
- presence of any outlying values

Dotplots

A dotplot is an attractive summary of numerical data when the data set is reasonable small or there are relatively few distinct data values.

Each observation is represented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically.

As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

Shapes of Distributions

Uniform

Unimodal Bimodal Multimodal

Mound-shaped, Bell-shaped, Symmetrical

Positively Skewed, Right Skewed

Negatively Skewed, Left Skewed

Exercise 1.12 The accompanying specific gravity values for various wood types used in construction appeared in the article Bolted Connection Design Values Based on European Yield Model (*J. of Structural Engr.*, 1993: 2169-2186):

.31 .35 .36 .36 .37 .38 .40 .40 .40
.41 .41 .42 .42 .42 .42 .42 .43 .44
.45 .46 .46 .47 .48 .48 .48 .51 .54
.54 .55 .58 .62 .66 .66 .67 .68 .75

Construct a stem-and-leaf display using repeated stems and comment on any interesting features of the display. Also, construct a dotplot.

Histograms

Constructing a Histogram for **Discrete** Data

1. Determine the frequency and relative frequency of each value of x.
2. Mark possible x values on a horizontal scale.
3. Above each value, draw a rectangle whose height is the relative frequency (or frequency) of that value.

Note: The same method can be applied to ***qualitative/categorical*** *data too.*

Consider data consisting of observations on a **discrete variable** x:

- The **frequency** of any particular x value is the number of times that value occurs in the data set.

- The **relative frequency** of a value is the fraction or proportion of times the value occurs.

  $$\text{relative frequency of a value} = \frac{\text{\# of times the value occurs}}{\text{\# of observations in the data set}}$$

- A **frequency distribution** is a tabulation of the frequencies and/or relative frequencies.

Constructing a Histogram for **Continuous** Data: **Equal** Class Widths

1. Subdivide the measurement axis into a suitable number of equal-width **class intervals** or **classes**. Each observation is contained in exactly one class. An observation on a boundary is placed in the interval to the right of the boundary.
2. Determine the frequency and relative frequency for each class.
3. Mark the class boundaries on a horizontal measurement axis.
4. Above each class interval, draw a rectangle whose height is the corresponding relative frequency (or frequency).

*Note: The same method can be applied to **discrete** data too.*

There are no hard-and-fast rules concerning either the number of classes or choice of classes themselves. Between 5 and 20 classes will be satisfactory for most data sets. Generally, the larger the number of observations in a data set, the more classes should be used. A reasonable rule of thumb is

$$\text{\# of classes} \approx \sqrt{\text{\# of observations}}$$

Equal-width classes may not be a sensible choice if a data set “stretches out” to one side or the other – there are some regions of the measurement scale that have a high concentration of data values and other parts where data is quite sparse.

In that case, Using a small number of equal-width classes results in almost all observations falling in just one or two of the classes. If a large number of equal-width classes are used, many classes will have zero frequency.

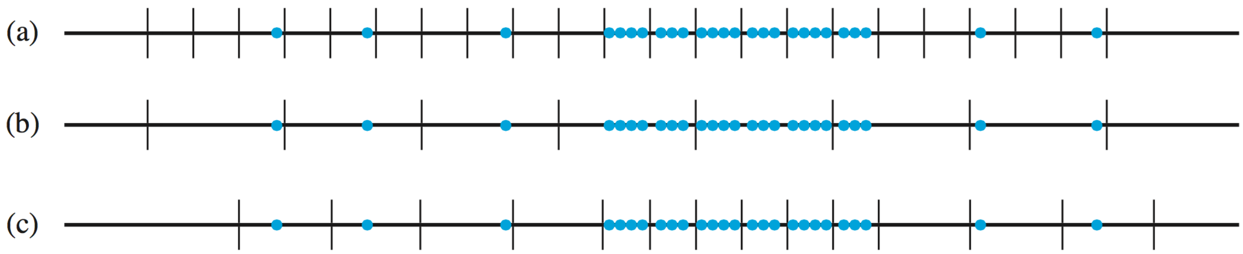


Figure 1.9 Selecting class intervals for "varying density" data: (a) many short equal-width intervals; (b) a few wide equal-width intervals; (c) unequal-width intervals

A sound choice is to use a few wider intervals near extreme observations and narrower intervals in the region of high concentration.

Constructing a Histogram for **Continuous** Data: **Unequal** Class Widths

1. Subdividing the measurement axis to a reasonable number of unequal-width class intervals or classes.
2. Determining frequencies and relative frequencies.
3. Calculate the height of each rectangle

$$\text{rectangle height} = \frac{\text{relative frequency of the class}}{\text{class width}}$$

*Note: The same method can be applied to **discrete** data too.*

The resulting rectangle heights are usually called **densities**, and the vertical scale is the **density scale**.

This prescription will also work when class widths are equal.

A density histogram has one interesting property. Multiplying both sides of the formula for density by the class width gives

$$\begin{aligned}\text{relative frequency} &= \text{(class width)(density)}\\ &= \text{(rectangle width)(rectangle height)}\\ &= \text{rectangle area}\end{aligned}$$

That is, the area of each rectangle is the relative frequency of the corresponding class.

Furthermore, since the sum of relative frequencies should be 1, the total area of all rectangles in a density histogram is 1.

Exercise 1.27 The paper "Study on the Life Distribution of Microdrills" (*J. of Engr. Manufacture*, 2002: 301305) reported the following observations, listed in increasing order, on drill lifetime (number of holes that a drill machines before it breaks) when holes were drilled in a certain brass alloy.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

(a) Construct a relative frequency histogram based on the equal-width class intervals 0$-$ $<$50, 50$-$ $<$100, 100$-$ $<$150, $\cdots$, and comment on features of the histogram.

(b) Construct a histogram of the natural logarithms of the lifetime observations, and comment on interesting characteristics.

(c) What proportion of the lifetime observations in this sample are less than 100? What proportion of the observations are at least 200?

1.3 Measures of Location

- mean
- median
- quartiles
- trimmed mean

Mean – the arithmetic average of the data. Mean is calculated as the sum of all observations divided by the number of observations.

Sample mean $\overline{x}$ of observations $x_1, x_2, x_3, \cdots, x_n$ is:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For reporting $\overline{x}$, it is recommended to use a decimal accuracy of one digit more than the accuracy of the x_i's.

The arithmetic mean is the most widely used measure of central location. However, it is oversensitive to extreme/outlying values and must be used with caution.

Median – the middle value.

Sample median $\widetilde{x}$ is the middle sorted observation. That is, we want a value such that half of the data is smaller than it and half is greater than it.

Steps to find the sample median:

1. Rank the n observations from smallest to largest.

2. If n is <u>odd</u>, median equals to the middle value,
   $\widetilde{x} = (\frac{n+1}{2})^{\text{th}}$ ordered value;
   If n is <u>even</u>, there are two middle values whose average equals the median,
   $\widetilde{x}$ = average of $(\frac{n}{2})^{\text{th}}$ and $(\frac{n+1}{2})^{\text{th}}$ ordered values.

Unlike the mean, the median is insensitive to extreme/outlying values.

In many samples, the relationship between the arithmetic mean and the sample median can be used to assess the shape of a distribution.

Positively/Right Skewed

Symmetrical

Negatively/Left Skewed

Example: In 1994, the Major League Baseball Players association claimed that the median salary for a baseball player was $450,000. The owners reported that the mean salary was $1,168,263. These numbers tell different stories about baseball salaries.

Quartiles – values of the variable that separate a ranked data set into 4 equal parts.

Order n observations from smallest to largest and separate the smallest half from the largest half; the median is included in both halves if n is odd.

- Q_1 **lower quartile** or **lower fourth** – the median of the smallest half of the data
- Q_2 – the median of the data
- Q_3 **upper quartile** or **upper fourth** – the median of the largest half of the data

Quartiles will be used to construct boxplot in Section 1.4.

The mean is quite sensitive to a single outlier, whereas the median is impervious to many outliers.

The mean and the median are at opposite extremes of the same family of measures. The mean is the average of all the data, whereas the median results from eliminating all but the middle one or two values and then averaging.

That is, the mean involves trimming 0% from each end of the sample, whereas for the median the maximum possible amount is trimmed from each end.

Trimmed mean $\overline{x}_{tr(100\alpha)}$ – a compromise between the mean $\overline{x}$ and the median $\widetilde{x}$ in which the smallest $100\alpha\%$ and the largest $100\alpha\%$ of the data are eliminated and the average is computed from what is left over.

A trimmed mean with a moderate trimming percentage – someplace between 5% and 25% – will yield a measure of center that is neither as sensitive to outliers as is the mean nor as insensitive as the median.

If the desired trimming percentage is $100\alpha\%$ and $n\alpha$ is not an integer, the trimmed mean must be calculated by interpolation using the appropriate weighted average.

Exercise 1.40 Compute the sample mean, the sample median, 25% trimmed mean, and 10% trimmed mean for the lifetime data given in Exercise 1.27, and compare these measures.

11	14	20	23	31	36	39	44	47	50
59	61	65	67	68	71	74	76	78	79
81	84	85	89	91	93	96	99	101	104
105	105	112	118	123	136	139	141	148	158
161	168	184	206	248	263	289	322	388	513

Categorical Data and Sample Proportions

Consider sampling a population that consists of only two categories. If we let x denote the number in the sample falling in category 1, then the number in category 2 is $n - x$. The relative frequency or **sample proportion** in category 1 is x/n and the sample proportion in category 2 is $1 - x/n$.

Focus attention on a particular category and code the sample results so that a 1 is recorded for an observation in the category and a 0 for an observation not in the category. Then the sample proportion of observations in the category is the sample mean of the sequence of 1s and 0s.

Thus, a sample mean can be used to summarize the results of a categorical sample.

These remarks also apply to situations in which categories are defined by grouping values in a numerical sample or population (e.g., we might be interested in knowing whether individuals have owned their present automobile for at least 5 years, rather than studying the exact length of ownership).

---

	sample statistic	population parameter
mean		
median		
proportion		

Exercise 1.41 A sample of n = 10 automobiles was selected, and each was subjected to a 5-mph crash test. Denoting a car with no visible damage by S (for success) and a car with such damage by F, results were as follows:

S S F S S S F F S S

(a) What is the value of the sample proportion of successes x/n?

(b) Replace each S with a 1 and each F with a 0. Then calculate $\overline{x}$ for this numerically coded sample. How does $\overline{x}$ compare to x/n?

(c) Suppose it is decided to include 15 more cars in the experiment. How many of these would have to be S's to give x/n = .80 for the entire sample of 25 cars?

1.4 Measures of Variability

- range
- variance
- standard deviation

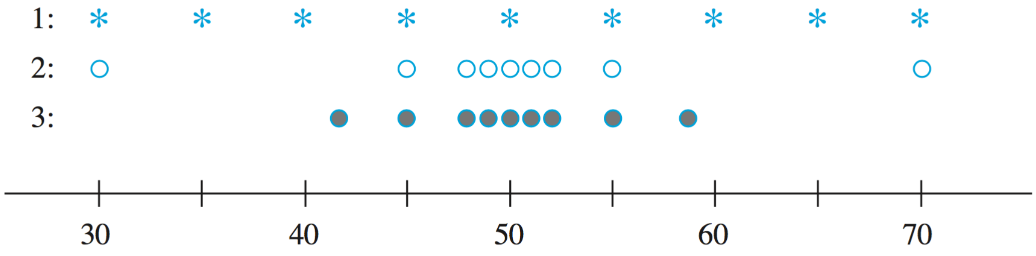


Figure 1.19 Samples with identical measures of center but different amounts of variability

Range – The difference between the largest and smallest sample values.

A defect of the range is that it depends on only the two most extreme observations and disregards the positions of the remaining $n-2$ values.

Our primary measures of variability involve the **deviations** from the mean, $x_i - \overline{x}$, for $i = 1, 2, \cdots, n$.

A deviation will be positive if the observation is larger than the mean and negative if the observation is smaller than the mean.

Note that sum of deviations $= \sum(x_i - \overline{x}) = 0$.

Sample variance – the average of the squared deviations from the mean.

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

Why $n-1$*?*

It is customary to refer to s^2 as being based on $n-1$ **degrees of freedom** (df). This terminology results from the fact that although s^2 is based on n deviations, these sum to zero. Therefore, only $n-1$ of the deviations are freely determined.

An alternative computational formula for the numerator of s^2 is

$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = \sum x_i^2 - n\overline{x}^2$$

Sample standard deviation – the (positive) square root of the sample variance.

$$s = \sqrt{s^2}$$

The unit of s is the same as the unit of each of the x_is, and is more easily interpreted than the variance. The bigger s is, the more spread out the data is.

	sample statistic	population parameter
variance		
standard deviation		

Sample statistics are **point estimates** of corresponding population parameters. We'll learn point estimation in Chapter 6.

Some Properties of the Variance and Standard Deviation

- Change the origin:
  If $y_i = x_i + c$, for $i = 1, 2, \cdots, n$, then $s_y^2 = s_x^2$ and $s_y = s_x$.
- Change the scale:
  If $y_i = cx_i$, for $i = 1, 2, \cdots, n$, then $s_y^2 = c^2 s_x^2$ and $s_y = cs_x$.

Therefore, adding a constant to each data value does not change the sample variance; whereas, multiplying each data value by a constant results in a new sample variance that is equal to the old one multiplied by the square of the constant.

Exercise 1.45 The value of Young's modulus (GPa) was determined for cast plates consisting of certain intermetallic substrates resulting in the following sample observations ("Strength of Modulus of a Molybdenum-Coated Ti-25Al-10Nb-3U-1Mo Intermetallic," *J. of Materials Engr. And Performance*, 1997: 46-50):

116.4 115.9 114.6 115.2 115.8

(a) Calculate $\overline{x}$ and the deviations from the mean.

(b) Use the deviations calculated in part (a) to obtain s^2 and s.

(c) Calculate s^2 by using the computational (shortcut) formula for S_{xx}.

(d) Subtract 100 from each observation to obtain a sample of transformed values. Now calculate the sample variance of these transformed values can compare it to s^2 for the original data.

Boxplot – a pictorial summary that describes the following most prominent features of the data

- center
- spread
- the extent and nature of any departure from symmetry
- identification of "outliers"

Steps for Constructing a Boxplot that Shows Outliers

1. Draw a horizontal measurement scale.
2. Draw vertical lines at the lower fourth/quartile Q_1, the median $\widetilde{x}$, and the upper fourth/quartile Q_3. Enclose these vertical lines in a box.
3. Compute the **fourth spread** f_s or the **interquartile range (IQR)**. IQR $= f_s =$ upper fourth $-$ lower fourth $= Q_3 - Q_1$
4. Detect outliers.
   Any observation farther than $1.5f_s$ beyond the closest fourth is an **outlier**. That is, an outlying value is less than $Q_1 - 1.5f_s$ or greater than $Q_3 + 1.5f_s$.
   An outlier is **extreme** if it is more than $3f_s$ beyond the nearest fourth, and it is **mild** otherwise.
5. Draw a whisker from Q_1 to the smallest data value that is larger than the lower fence and a whisker from Q_3 to the largest data value that is smaller than the upper fence.
6. Represent each mild outlier by a closed circle and each extreme outlier by an open circle.

Outliers distort both the mean and the standard deviation, since neither is resistant. Statistical inference based a set of data that contains outliers could be flawed.

Exercise 1.56 The following data on distilled alcohol content (%) for a sample of 35 port wines was extracted from the article “A Method for the Estimation of Alcohol in Fortified Wines Using Hydrometer Baume and Refractometer Brix” (*Amer. J. Enol. Vitic.*, 2006: 486490). Each value is an average of two duplicate measurements.

16.35	18.85	16.20	17.75	19.58	17.73	22.75	23.78	23.25
19.08	19.62	19.20	20.05	17.85	19.17	19.48	20.00	19.97
17.48	17.15	19.07	19.90	18.68	18.82	19.03	19.45	19.37
19.20	18.00	19.60	19.33	21.22	19.50	15.30	22.25	

Construct a boxplot that shows outliers. Describe/summarize the data.

Distribution Shape Based upon Boxplot

If the median is near the center of the box and each of the horizontal lines is of approximately equal length, then the distribution is roughly symmetric.

If the median is to the left of the center of the box or the right line is substantially longer than the left line, the distribution is positively/right skewed.

If the median is to the right of the center of the box or the left line is substantially longer than the right line, the distribution is negatively/left skewed.

Comparative Boxplots

A comparative or side-by-side boxplot is a very effective way of revealing similarities and differences between two or more data sets consisting of observations on the same variable.

Chapter 2

Probability

Probability allows us to make the inferential jump from a sample to a population.

In this chapter, probability is defined and some rules for working with probabilities are introduced.

The term **probability** refers to the study of randomness and uncertainty.

The discipline of probability provides methods for quantifying the chances, or likelihoods, associated with the various outcomes.

2.1 Sample Spaces and Events

Experiment – an activity or process that whose outcome is subject to uncertainty. That is, experiment lead to random outcomes.

Sample Space – the set of all possible outcomes of an experiment, denoted by S.

Event – any collection (subset) of outcomes of interest denoted by a capital letter. Event is a subset of the sample space. Events can be **simple** (consists of exactly one outcome) or **compound** (consists of more than one outcomes).

Venn diagram – a graphical representation of events that is very useful for illustrating logical relations.

Complement of an event A – denoted by A', is the event that A does not occur.
$A' = \{x \mid x \notin A\}$

Union of two events A and B – denoted by $A \cup B$, is the event that either A or B occurs, or they both occur. That is, at least one of the events occur.
$A \cup B = \{x \mid x \in A \text{ or } x \in B\}$

Intersection of two events A and B – denoted by $A \cap B$, is the event that both A and B occur simultaneously.
$A \cap B = \{x \mid x \in A$ and $x \in B\}$

Null event – denoted by $\emptyset$, consists of no outcomes.

Two events A and B are said to be **mutually exclusive** or **disjoint** if $A \cap B = \emptyset$ so that they have no outcomes in common.

$A \cap A$ = $A \cup A$ =

$A \cap S$ = $A \cup S$ =

$A \cap \emptyset$ = $A \cup \emptyset$ =

$A \cap A'$ = $A \cup A'$ =

Exercise 2.4 Each of a sample of four home mortgages is classified as fixed rate (F) or variable rate (V).

(a) What are the 16 outcomes in S?

(b) Which outcomes are in the event that exactly three of the selected mortgages are fixed rate?

(c) Which outcomes are in the event that all four mortgages are of the same type?

(d) Which outcomes are in the event that at most one of the four is a variable-rate mortgage?

(e) What is the union of the events in parts (c) and (d), and what is the intersection of these two events?

(f) What are the union and intersection of the two events in parts (b) and (c)?

2.2 Axioms, Interpretations, and Properties of Probability

Given an experiment and a sample space S, the objective of probability is to assign to each event A a number $P(A)$, called the **probability** of the event A, which will give a precise measure of the chance that A will occur.

Law of Large Numbers – the relative frequency of the number of times that an outcome occurs when an experiment is replicated over and over again approaches the theoretical probability of the outcome.

The probability should satisfy the following **three axioms (basic properties)**:

Axiom 1 For any event A, $P(A) \geq 0$.

Axiom 2 $P(S) = 1$.

Axiom 3 If $A_1, A_2, \cdots$, is an infinite collection of disjoint events, then

$$P\left(A_1 \cup A_2 \cup A_3 \cup \cdots\right) = \sum_{i=1}^{\infty} P(A_i).$$

Properties of Probability

- $P(\emptyset) = 0$

- $A_1, \cdots, A_k$ are disjoint events, then

$$P(A_1 \cup A_2 \cup A_3 \cdots A_k) = \sum_{i=1}^{k} P(A_i)$$

- $P(A) + P(A') = 1$

- $0 \le P(A) \le 1$

- For any two events A and B,
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  $\Rightarrow$ If events A and B are disjoint events, then
  $P(A \cup B) = P(A) + P(B)$.

- For any three events A, B, and C,

$$\begin{aligned} P(A \cup B \cup C) =& P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C) \end{aligned}$$

Example: Suppose that A and B are two events. $P(A) = 0.8$, $P(B) = 0.7$.

(a) Is it possible that $P(A \cap B) = 0.1$?

(b) What is the smallest possible value for $P(A \cap B)$?

(c) Is it possible that $P(A \cap B) = 0.77$?

Exercise 2.12 Consider randomly selecting a student at a certain university, and let A denote the event that the selected individual has a Visa credit card and B be the analogous event for a MasterCard. Suppose that $P(A) = .5$, $P(B) = .4$, and $P(A \cap B) = .25$.

(a) Compute the probability that the selected individual has at least one of the two types of cards (i.e., the probability of the event $A \cup B$).

(b) What is the probability that the selected individual has neither type of card?

(c) Describe, in terms of A and B, the event that the selected student has a Visa card but not a MasterCard, and then calculate the probability of this event.

When the sample space S is either finite or “countably infinite”, the probability of any event A is computed by adding together the individual probabilities for each outcome in A.

If $A = \{E_1, E_2, \cdots, E_k\}$, then $P(A) = \sum_{i=1}^{k} P(E_i)$.

When all the outcomes in the sample space S are equally likely to happen,

$$P(A) = \frac{\text{number of outcomes within A}}{\text{total number of outcomes}} = \frac{N(A)}{N}.$$

Example There are 2 black balls and 2 white balls in a box. Suppose you close your eyes and randomly draw 2 balls from the box. What is the probability that you get exactly one black and one white balls?

2.3 Counting Techniques

Product Rule for Ordered Pairs

If the first element of an ordered pair can be selected in n_1 ways, and for each of these n_1 ways the second element of the pair can be selected in n_2 ways, then the number of pairs is $n_1 n_2$.

Example Suppose that a box contains 7 red balls and 5 blue balls. If two balls are drawn out at random, how many possible ordered pairs are there? What is the probability that the first one is red and the second one is blue?

General Product Rule for k-Tuples

Suppose a set consists of ordered collections of k elements (k-tuples) and that there are n_1 possible choices for the first element; for each choice of the first element, there are n_2 possible choices of the second element; $\ldots$; for each possible choice of the first $k-1$ elements, there are n_k possible choices of the kth element. Then there are a total of $n_1 n_2 \cdots n_k$ possible k-tuples.

Example Suppose a package of 30 Milk Chocolate M&M candies contain 9 brown, 10 green and 11 yellow M&M candies. If you randomly choose 3 candies, what's the probability of getting all three colors?

Permutations and Combinations

Permutation – order does matter

An ordered subset is called a **permutation**. The number of permutations of size k that can be formed from the n individuals or objects in a group will be denoted by $P_{k,n}$.

$$P_{k,n} = \frac{n!}{(n-k)!}$$

Combination – order does NOT matter

An unordered subset is called a **combination**. The number of permutations of size k that can be formed from the n individuals or objects in a group will be denoted by $C_{k,n}$, or $\binom{n}{k}$.

$$C_{k,n} = \binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$$

Example: Suppose we have 18 people competing in a game. How many ways can we award a 1st, 2nd and 3rd place prize among the 18 contestants?

Example: Suppose we need to select a committee of size 3 from a group of 18 people. How many ways can the committee of size 3 be selected?

Exercise 2.38 A box in a certain supply room contains four 40-W light bulbs, five 60-W bulbs, and six 75-W bulbs. Suppose that three bulbs are randomly selected.

(a) What is the probability that exactly two of the selected bulbs are rated 75 W?

(b) What is the probability that all three of the selected bulbs have the same rating?

(c) What is the probability that one bulb of each type is selected?

(d) Suppose now that bulbs are to be selected one by one until a 75-W bulb is found. What is the probability that it is necessary to examine at least six bulbs?

Exercise 2.43 In five-card poker, a straight consists of five cards with adjacent denominations (e.g., 9 of clubs, 10 of hearts, jack of hearts, queen of spades, and king of clubs). Assuming that aces can be high or low, and you are dealt a five-card hand.

(a) what is the probability that it will be a straight with high card 10?

(b) What is the probability that it will be a straight?

(c) What is the probability that it will be a straight flush (all cards in the same suit)?

2.4 Conditional Probability

Example The probability of observing an even number (event A) on a toss of a fair die is 0.5, where $S = \{1, 2, 3, 4, 5, 6\}$ and $A = \{2, 4, 6\}$. Suppose we're given the information that on a particular throw of the die the result was a number less than or equal to 3 (event B), where $B = \{1, 2, 3\}$. Would the probability of observing an even number on that throw of the die still be equal to 0.5?

For any two events A and B with $P(B) > 0$, the **conditional probability of A given B has occurred** is defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Multiplication Rule for $P(A \cap B)$

$$P(A \cap B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

Exercise 2.47 Return to the credit card scenario of Exercise 2.12, where $A = \{\text{Visa}\}$, $B = \{\text{MasterCard}\}$, $P(A) = .5$, $P(B) = .4$, and $P(A \cap B) = .25$. Calculate and interpret each of the following probabilities .

(a) $P(B \mid A)$ (b) $P(B' \mid A)$

(c) $P(A \mid B)$ (d) $P(A' \mid B)$

(e) Given that the selected individual has at least one card, what is the probability that he or she has a Visa card?

Exercise 2.56 For any events A and B with $P(B) > 0$, show that
$P(A \mid B) + P(A' \mid B) = 1$.

Exercise 2.57 If $P(B \mid A) > P(B)$, show that $P(B'|A) < P(B')$.

A set of events $A_1, A_2, \cdots, A_k$ is **exhaustive** if at least one of the events must occur.

$$A_1 \cup A_2 \cup \cdots \cup A_k = \bigcup_{i=1}^{k} A_i = S$$

Law of Total Probability

Let $A_1, A_2, \cdots, A_k$ be mutually exclusive and exhaustive events. The unconditional probability of B, $P(B)$, can then be written as a weighted average of the conditional probabilities of B given A_i, $P(B \mid A_i)$, as follows:

$$P(B) = \sum_{i=1}^{k} P(B \mid A_i) P(A_i)$$

Bayes' Theorem

Let $A_1, A_2, \cdots, A_k$ be a collection of k mutually exclusive and exhaustive events with **prior probabilities** $P(A_i)$ ($i = 1, \cdots, k$). Then for any other event B for which $P(B) > 0$, the **posterior probability** of A_j given that B has occurred is

$$P(A_j \mid B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B \mid A_j)P(A_j)}{\sum_{i=1}^{k} P(B \mid A_i)P(A_i)},$$

for $j = 1, \cdots, k$.

⇒ For any two events A and B with $P(B) > 0$,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A')}$$

Exercise 2.59 At a certain gas station, 40% of the customers use regular gas (A_1), 35% use plus gas (A_2), and 25% use premium (A_3). Of those customers using regular gas, only 30% fill their tanks (event B). Of those customers using plus, 60% fill their tanks, whereas of those using premium, 50% fill their tanks.

(a) What is the probability that the next customer will request plus gas and fill the tank ($A_2 \cap B$)?

(b) What is the probability that the next customer fills the tank?

(c) If the next customer fills the tank, what is the probability that regular gas is requested? Plus? Premium?

2.5 Independence

Two events A and B are said to be **independent** (denoted by $A \perp B$) if $P(A \mid B) = P(A)$ and are **dependent** otherwise.

$P(A \mid B) = P(A) \Longleftrightarrow P(B \mid A) = P(B)$

Multiplication Rule for $P(A \cap B)$

A and B are **independent** if and only if (iff) $P(A \cap B) = P(A)P(B)$

More generally, events $A_1, \cdots, A_n$ are **mutually independent** if for every k $(k = 2, \cdots, n)$ and every subset of indices $i_1, \cdots, i_k$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k})$$

Example: If A and B are disjoint with $P(A) > 0$ and $P(B) > 0$, are they independent?

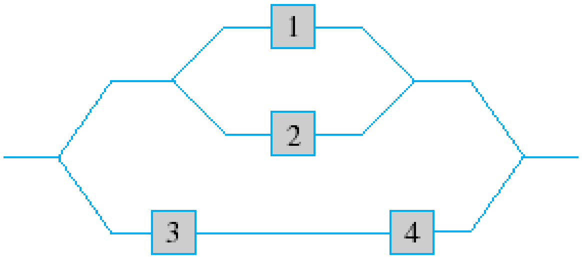
Exercise 2.73: If A and B are independent events, show that A' and B are also independent.

Exercise 2.74 Suppose that the proportions of blood phenotypes in a particular population are as follows:

A	B	AB	O
.42	.10	.04	.44

Assuming that the phenotypes of two randomly selected individuals are independent of one another, what is the probability that both phenotypes are O? What is the probability that the phenotypes of two randomly selected individuals match?

Exercise 2.80 Consider the system of components connected as in the accompanying picture. Components 1 and 2 are connected in parallel, so that subsystem works iff either 1 or 2 works; since 3 and 4 are connected in series, that subsystem works iff both 3 and 4 work. If components work independently of one another and P(component works) = .9, calculate P(system works).



Chapter 3

Discrete Random Variables and Probability Distributions

3.1 Random Variables

Random variable (rv) – is obtained by assigning a numerical value to each outcome in the sample space S of a particular experiment. In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers. Random variables are denoted by uppercase letters such as X and Y.

- A **discrete random variable** is a random variable that takes either a finite number of possible values or at most a countably infinite number of possible values.

- A **continuous random variable** is a random variable that takes an infinite number of possible values that is not countable. For any possible value c, $P(X = c) = 0$.

In this chapter, we examine the basic properties and discuss the most important examples of discrete variables. Chapter 4 focuses on continuous random variables.

Exercise 3.7 For each random variable defined here, describe the set of possible values for the variable, and state whether the variable is discrete.

(a) X = the number of unbroken eggs in a randomly chosen standard egg carton

(b) Y = the number of students on a class list for a particular course who are absent on the first day of classes

(c) U = the number of times a duffer has to swing at a golf ball before hitting it

(d) X = the length of a randomly selected rattlesnake

(e) Z = the amount of royalties earned from the sale of a first edition of 10,000 textbooks

(f) Y = the pH of a randomly chosen soil sample

(g) X = the tension (psi) at which a randomly selected tennis racket has been strung

(h) X = the total number of coin tosses required for three individuals to obtain a match (HHH or TTT)

3.2 Probability Distributions for Discrete Random Variables

Probability distribution – A table, graph, or mathematical formula that provides the possible values of a random variable X and their corresponding probabilities.

The **probability distribution** or **probability mass function (pmf)** of a discrete variable X is defined for every number x by
$p(x) = P(X = x) = P(\text{all } s \in S; X(s) = x)$,
which is the sum of the probabilities of all sample points in S that are assigned the value x.

- $0 \leq p(x) \leq 1$
- $\sum_x p(x) = 1$

The **cumulative distribution function (cdf)** $F(x)$ of a discrete random variable X is defined for every number x by
$F(x) = P(X \leq x) = \sum\limits_{y: y \leq x} p(y)$.

- $0 \leq F(x) \leq 1$
- $F(a) = 0$, for $a < x_{\min}$, $x_{\min}$ is the smallest possible X value
- $F(b) = 1$, for $b \geq x_{\max}$, $x_{\max}$ is the largest possible X value

Example A fair coin is thrown three times and the sequence of heads (H) and tails (T) is recorded. The sample space is $S = \{$HHH,HHT,HTT,HTH,TTT, TTH,THH,THT$\}$. Define random variable X as the total number of heads. Find probability mass function (pmf) $p(x)$ and construct a table, a **line graph** and a **probability histogram** to represent the pmf.
Also, find the cumulative distribution function (cdf) $F(x)$. Use both table and **step graph** to describe the cdf.

The pmf $p(x)$ can be recovered by calculating the jump value of $F(x)$ at x, i.e., $p(x) = P(X = x) = F(x) - F(x^-)$, where x^- represents the largest possible X value that is strictly less than x.

For any two numbers x_1 and x_2 with $x_1 \leq x_2$, we have $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1^-)$.

Exercise 3.24 An insurance company offers its policyholders a number of different premium payment options. For a randomly selected policyholder, let X = the number of months between successive payments. The cdf of X is as follows:

$$F(x) = \begin{cases} 0 & x < 1 \\ .30 & 1 \leq x < 3 \\ .40 & 3 \leq x < 4 \\ .45 & 4 \leq x < 6 \\ .60 & 6 \leq x < 12 \\ 1 & 12 \leq x \end{cases}$$

(a) What is the pmf of X?

(b) Using just the cdf, compute $P(3 \leq X \leq 6)$ and $P(4 \leq X)$.

3.3 Expected Values

If X is a discrete random variable with pmf $p(x)$, then the **expected value** or **mean value** of X, denoted by $E(X)$ or μ_X or just μ, is

$$E(X) = \mu_X = \sum_x xP(X = x) = \sum_x xp(x)$$

and for any real-valued function $h(X)$, its expected value is

$$E[h(X)] = \mu_{h(X)} = \sum_x h(x)P(X = x) = \sum_x h(x)p(x)$$

Example A fair coin is thrown three times and the sequence of heads (H) and tails (T) is recorded.

(a) Let random variable X = the total number of heads. Find $E(X)$, which is the expected number of heads you will observe.

(b) If you win \$2 every time obseving a head and have to pay \$1 every time observing a tail. What is the expected profit of this game?

Example Suppose that a player randomly draws a card from a pack of cards, and wins \$15 if an Ace, King, Queen, or Jack is obtained, and otherwise wins the face value of the card in dollars. Would you pay \$9 to play this game?

Let X have pmf $p(x)$ with mean μ. Then the **variance** of X, denoted by $V(X)$ or σ_X^2 or just σ^2, is defined by

$$\begin{aligned} V(X) = \sigma^2 = E[(X-\mu)^2] \qquad & = \sum_x (x-\mu)^2 p(x) \\ = E(X^2) - [E(X)]^2 \qquad & = \Big[\sum_x x^2 p(x)\Big] - \mu^2 \end{aligned}$$

The **standard deviation (SD)** of X is $\sigma = \sqrt{\sigma^2}$.

For any real-valued function $h(X)$, the variance of $h(X)$ is

$$V[h(X)] = \sigma^2_{h(X)} = E\{h(x) - E[h(x)]\}^2 = \sum_x \{h(x) - E[h(x)]\}^2 p(x)$$

Example An insurance company will sell you a \$10,000 term life policy for an annual premium of \$300. Based on a period life table from the U.S. government, the probability that you will survive the coming year is 0.999, what is the expected gain and the variance of the gain for the insurance company for the coming year?

Properties of the Expected Value

- If X is a random variable and $Y = aX + b$ for some constants a and b, then $E(Y) = E(aX + b) = aE(X) + b$.

  $\Rightarrow E(aX) = aE(X)$

  $\Rightarrow E(b) = b$

- For any random variables X and Y, $E(X + Y) = E(X) + E(Y)$.

  $\Rightarrow E\Big(\sum_i X_i\Big) = \sum_i E(X_i)$

Properties of the Variance

- If X is a random variable and $Y = aX + b$ for some constants a and b, then $V(Y) = V(aX + b) = a^2V(X)$.

  $\Rightarrow V(aX) = a^2V(X)$

  $\Rightarrow V(X + b) = V(X)$

  $\Rightarrow V(b) = 0$

- If random variables X and Y are independent, then
  $V(X + Y) = V(X) + V(Y)$
  $V(X - Y) = V(X) + V(Y)$

Example Given a random variable X with $E(X) = \mu$ and $V(X) = \sigma^2$, let $Y = \frac{X-\mu}{\sigma}$. Find $E(Y)$ and $V(Y)$.

Exercise 3.32 An appliance dealer sells three different models of upright freezers having 13.5, 15.9, and 19.1 cubic feet of storage space, respectively. Let X = the amount of storage space purchased by the next customer to buy a freezer. Suppose that X has pmf

x	$p(x)$
13.5	.2
15.9	.5
19.1	.3

(a) Compute $E(X)$, $E(X^2)$, and $V(X)$.

(b) If the price of a freezer having capacity X cubic feet is $25X - 8.5$, what is the expected price paid by the next customer to buy a freezer?

(c) What is the variance of the price $25X - 8.5$ paid by the next customer?

(d) Suppose that although the rated capacity of a freezer is X, the actual capacity is $h(X) = X - .01X^2$. What is the expected actual capacity of the freezer purchased by the next customer?

3.4 The Binomial Probability Distribution

Bernoulli random variable – a random variable whose only possible values are 0 and 1, with probabilities $1 - p$ and p, respectively. Its pmf is thus

$$\begin{aligned} p(1) &= P(X = 1) = p \\ p(0) &= P(X = 0) = 1 - p = q\ . \end{aligned}$$

The mean and variance of the Bernoulli rv are

$$\begin{aligned} E(X) &= p \\ V(X) &= p(1-p) = pq. \end{aligned}$$

Note: Suppose an experiment, which results in a “success” with probability p and a “failure” with probability $1 - p$, is performed. The random variable defined to be 1 or 0 according to whether the experiment was a “success” or a “failure” is a Bernoulli random variable.

Suppose for rv X its pmf $p(x)$ depends on a quantity that can be assigned any one of a number of possible values, with each different value determining a different probability distribution. Such a quantity is called a **parameter** of the distribution.

The collection of all probability distributions for different values of the parameter is called a **family** of probability distributions.

Bernoulli random variable deals only with one single experiment resulting in a "success" with probability p and a "failure" with probability $1-p$.

If we repeat it n times, then the total number of successes, X, is a **binomial random variable**, with parameters n and p.
We write $X \sim Bin(n,p)$.
Its pmf is

$$b(x;n,p) = P(X=x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0,1,\cdots,n \\ 0 & \text{otherwise} \end{cases}$$

Its cdf is

$$B(x;n,p) = P(X \le x) = \sum_{y=0}^{x} \binom{n}{y} p^y (1-p)^{n-y}, \quad x = 0,1,\cdots,n$$

And its mean and variance are

$$\begin{aligned} E(X) &= np \\ V(X) &= np(1-p) = npq \end{aligned}$$

Example It is known that screws produced by a certain company will be defective with probability .01 independently of each other. The company sells the screws in packages of 10 and offers a money-back guarantee that at most 1 of the 10 screws is defective. What proportion of packages sold by this company would need to be refunded?

Exercise 3.54 A particular type of tennis racket comes in a mid size version and an oversize version. Sixty percent of all customers at a certain store want the oversize version.

(a) Among ten randomly selected customers who want this type of racket, what is the probability that at least six want the oversize version?

(b) Among ten randomly selected customers, what is the probability that the number who want the oversize version is within 1 standard deviation of the mean value?

(c) The store currently has seven rackets of each version. What is the probability that all of the next ten customers who want this racket can get the version they want from current stock?

3.5 Hypergeometric and Negative Binomial Distributions

Suppose the population or set to be sampled consists of N individuals, objects, or elements (a finite population).

Each individual can be characterized as a success (S) or a failure (F), and there are M successes in the population.

A sample of n individuals is selected without replacement such that each subset of size n is equally likely to be chosen.

The random variable of interest, X = the number of S's in the sample, is a **hypergeometric random variable** with parameters n, M, and N.
Its pmf is

$$h(x; n, M, N) = P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}},$$

for any integer x satisfying $\max(0, n-N+M) \leq x \leq \min(n, M)$.

And its mean and variance are

$$E(X) = n \cdot \frac{M}{N}$$
$$V(X) = \left(\frac{N-n}{N-1}\right) \cdot n \cdot \frac{M}{N} \cdot \left(1 - \frac{M}{N}\right)$$

Exercise 3.68 An electronics store has received a shipment of 20 table radios that have connections for an iPod or iPhone. Twelve of these have two slots (so they can accommodate both devices), and the other eight have a single slot. Suppose that six of the 20 radios are randomly selected to be stored under a shelf where the radios are displayed, and the remaining ones are placed in a storeroom. Let X = the number among the radios stored under the display shelf that have two slots.

(a) What kind of a distribution does X have (name and values of all parameters)?

(b) Compute $P(X = 2)$, $P(X \leq 2)$, and $P(X \geq 2)$.

(c) Calculate the mean value and standard deviation of X.

Suppose that independent Bernoulli trials, each with probability p of being a success, are performed until a total number of r successes occurs, where r is a specified positive integer.

The random variable X = the number of failures that precede the rth success, is called a **negative binomial random variable** with parameters p and r.
We write $X \sim NB(r, p)$.

Its pmf is:

$$nb(x; r, p) = P(X = x) = \binom{x+r-1}{r-1} p^r (1-p)^x, \quad x = 0, 1, 2, \cdots$$

And its mean and variance are

$$E(X) = \frac{r(1-p)}{p}$$
$$V(X) = \frac{r(1-p)}{p^2}$$

Example When a fisherman catches a fish, if it is young with a probability of 0.2, the fisherman returns the fish to the water. On the other hand, an adult fish will be kept. Suppose the fisherman sets a goal of 10 adult fish.

(a) What is the expected number of young fish caught by the fisherman before the 10th adult fish is caught?

(b) What is the expected number of all fish the fisherman needs to catch in order to reach the goal?

Exercise 3.75 Suppose that $p = P(\text{male birth}) = .5$. A couple wishes to have exactly two female children in their family. They will have children until this condition is fulfilled.

(a) What is the probability that the family has x male children?

(b) What is the probability that the family has four children?

(c) What is the probability that the family has at most four children?

(d) How many male children would you expect this family to have? How many children would you expect this family to have?

Geometric rv as a special case of Negative Binomial rv

Suppose we perform the independent Bernoulli trials until one success occurs. The random variable X = the number of failures and Y = the number of trials (i.e., $Y = X + 1$), are called **geometric random variables** with parameter p.

That is, $X \sim Geometric(p)$ is equivalent to $X \sim NB(1, p)$.

We have the geometric pmfs,

$$\begin{aligned} g(x; p) &= P(X = x) = p(1-p)^x, & x &= 0, 1, 2, \cdots \\ g(y; p) &= P(Y = y) = p(1-p)^{y-1}, & y &= 1, 2, 3, \cdots \end{aligned}$$

And the means and variances are

$$\begin{aligned} E(X) &= \frac{1-p}{p} & E(Y) &= \frac{1}{p} \\ V(X) &= \frac{1-p}{p^2} & V(Y) &= \frac{1-p}{p^2} \end{aligned}$$

Example When a fisherman catches a fish, if it is young with a probability of 0.2, the fisherman returns the fish to the water. On the other hand, an adult fish will be kept.

(a) What is the expected number of fish caught by the fisherman until the first adult fish is caught?

(b) What is the probability that the fifth fish caught is the first young fish?

(c) If the fisherman catches five fish, what is the probability that there are exactly one young fish?

3.6 The Poisson Probability Distribution

A discrete random variable X is said to have a **Poisson distribution** with parameter $\mu(\mu > 0)$ if the pmf of X is,

$$p(x;\mu) = P(X = x) = \frac{e^{-\mu}\mu^{x}}{x!}, \quad x = 0, 1, 2, \ldots$$

The mean and variance of the Poisson rv X are

$$E(X) = \mu$$
$$V(X) = \mu$$

Poisson distribution deals with counting the number of times an event occurs in a given interval (time, space, volume, etc.). X = # of occurrences of some event over a given interval.

Poisson distribution can be used to model the number visits to a particular website, the number of pulses of some sort recorded by a counter, the number of accidents in an industrial facility, the number of particles emitted by a radioactive source, or the number of births during a given day.

General assumptions of the Poisson distribution

1. The probability that an event occurs in a given interval of time is proportional to the length of the interval.
2. Events do not happen simultaneously in a sufficient small interval.
3. What happens in one subinterval is independent of what happens in any other non-overlap subinterval.

Exercise 3.86 The number of people arriving for treatment at an emergency room can be modeled by a Poisson process with a rate parameter of five per hour.

(a) What is the probability that exactly four arrivals occur during a particular hour?

(b) What is the probability that at least four people arrive during a particular hour?

(c) How many people do you expect to arrive during a 45-min period?

Poisson Approximation of Binomial

A binomial random variable with parameters (n, p) can be approximated by a Poisson random variable with $\mu = np$, when n is large enough $(n > 50)$ and p is small enough so that np approaches a moderate value ($np < 5$).

Theoretically, $b(x; n, p) \xrightarrow[p \to 0]{n \to \infty} p(x; \mu = np).$

$$X \sim Bin(n, p) \overset{n \to \infty}{\underset{p \to 0}{\Longrightarrow}} X \sim Poisson(\mu = np)$$

Exercise 3.84 Suppose that only .10% of all computers of a certain type experience CPU failure during the warranty period. Consider a sample of 10,000 computers.

(a) What are the expected value and standard deviation of the number of computers in the sample that have the defect?

(b) What is the (approximate) probability that more than 10 sampled computers have the defect?

(c) What is the (approximate) probability that no sampled computers have the defect?

Chapter 4

Continuous Random Variables and Probability Distributions

- A **discrete random variable** is a random variable that takes either a finite number of possible values or at most a countably infinite number of possible values.

- A **continuous random variable** is a random variable that takes an infinite number of possible values that is not countable. For any possible value c, $P(X = c) = 0$.

In this chapter, we introduce basic properties of continuous random variables and discuss important examples of continuous random variables and their probability distributions.

4.1 Probability Density Functions

Let X be a continuous random variable (rv), then the **probability distribution** or **probability density function (pdf)** of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

That is, the probability that X falls in the interval $[a, b]$ is the area under the density function between a and b.

$P(a \leq X \leq b)$ = $P(a < X \leq b)$ = $P(a \leq X < b)$ = $P(a < X < b)$

The graph of $f(x)$ is open referred to as the **density curve**.

Any legitimate probability density function (pdf) $f(x)$, should satisfy the following conditions:

1. $f(x) \geq 0$, for all x
2. $\int_{-\infty}^{\infty} f(x)dx =$ area under the entire density curve $= 1$
   The rule of total probability holds.

For any possible value c, $P(X = c) = 0$

- $P(X = c) = P(c \leq X \leq c) = \int_c^c f(x)dx = 0$
- $P(c - \dfrac{\epsilon}{2} \leq X \leq c + \dfrac{\epsilon}{2}) = \int_{c-\frac{\epsilon}{2}}^{c+\frac{\epsilon}{2}} f(x)dx \approx \epsilon f(c)$

So the probability that X falls in an interval of length ϵ around the point c is approximately $\epsilon f(c)$. Thus $f(c)$ is a measure of how likely it is that the random variable will be near c.

Exercise 4.3 The error involved in making a certain measurement is a continuous random variable X with pdf

$$f(x) = \begin{cases} .09375(4 - x^2) & -2 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

(a) Sketch the graph of $f(x)$.

(b) Compute $P(X > 0)$.

(c) Compute $P(-1 < X < 1)$.

(d) Compute $P(X < -.5$ or $X > .5)$.

4.2 Cumulative Distribution Functions and Expected Values

The **cumulative distribution function (cdf)** $F(x)$ for a continuous random variable X is defined for every number x by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

Thus, $F(x)$ is the area under the density curve to the left of x. $F(x)$ is often referred to as the left-tail probability.

From $F(x)$ to $f(x)$ by the fundamental theorem of calculus

If X is a continuous rv with pdf $f(x)$ and cdf $F(x)$, then at every x at which the derivative $F'(x)$ exists,

$$F'(x) = f(x)$$

Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$.

- $F(x) \to 0$, as $x \to -\infty$
- $F(x) \to 1$, as $x \to \infty$
- $F(x)$ is a monotonic non-decreasing functions of x
- $F(x)$ does not need to be smooth, but **is continuous**

- For any number a, $P(X > a) = 1 - F(a)$

- For any two number a and b with $a \leq b$,
$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$$
$$= \int_a^b f(x)dx = F(b) - F(a)$$

Percentiles of a Continuous Distribution

Let p be a number between 0 and 1. The **(100p)th percentile** of the distribution of a continuous rv X, denoted by $\eta(p)$, is defined by

$$p = F(\eta(p)) = P(X \leq \eta(p)) = \int_{-\infty}^{\eta(p)} f(y)dy$$

The median of a continuous distribution, denoted by $\widetilde{\mu}$, is the 50th percentile, so $\widetilde{\mu}$ satisfies $F(\widetilde{\mu}) = .5$.
That is, half the area under the density curve is to the left of $\widetilde{\mu}$ and half is to the right of $\widetilde{\mu}$.
If the pdf is symmetric, the median is the point of symmetry.

Exercise 4.12 The cdf for X = measurement error in Exercise 4.3 is

$$F(x) = \begin{cases} 0 & x < -2 \\ \frac{1}{2} + \frac{3}{32}\left(4x - \frac{x^3}{3}\right) & -2 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$

(a) Compute $P(X < 0)$.

(b) Compute $P(-1 < X < 1)$.

(c) Compute $P(.5 < X)$.

(d) Verify that $f(x)$ is as given in Exercise 4.3 by obtaining $F'(x)$.

(e) Verify that $\widetilde{\mu} = 0$.

Expected Values

If X is a continuous random variable with pdf $f(x)$, then its **expected value** or **mean value** is

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

For any function $h(X)$,

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f(x) dx$$

If X is a continuous random variable with mean value μ, then the **variance** of X, denoted by $V(X)$, is defined by

$$\begin{aligned}\sigma_X^2 = V(X) = E[(X-\mu)^2] &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx \\ = E(X^2) - [E(X)]^2 &= \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2\end{aligned}$$

The **standard deviation (SD)** of X is $\sigma_X = \sqrt{V(X)}$.

Properties

- $E(aX+b) = aE(X) + b$
- $V(aX+b) = a^2V(X)$.
- $E(X+Y) = E(X) + E(Y)$
- $V(X \pm Y) = V(X) + V(Y)$, if X and Y are independent.

Exercise 4.13 “Time headway” in traffic flow is the elapsed time between the time that one car finishes passing a fixed point and the instant that the next car begins to pass that point. Let X = the time headway between two randomly chosen consecutive cars (sec.) in a traffic environment. The distribution of time headway has the form

$$f(x) = \begin{cases} \frac{k}{x^4} & x > 1 \\ 0 & x \leq 1 \end{cases}$$

(a) Determine the value of k for which $f(x)$ is a legitimate pdf.

(b) Obtain the cumulative distribution function cdf $F(x)$.

(c) Use the cdf $F(x)$ to determine the probability that headway exceeds 2 sec and also the probability that headway is between 2 and 3 sec.

(d) Obtain the mean value and the standard deviation of headway.

(e) What is the probability that headway is within 1 standard deviation of the mean value?

A continuous rv X is said to have a **uniform distribution** on the interval $[A, B]$, if its pdf is

$$f(x; A, B) = \begin{cases} \dfrac{1}{B-A} & A \leq x \leq B \\ 0 & \text{otherwise} \end{cases}$$

We write $X \sim U(A, B)$.

Its mean and variance are

$$E(X) = \frac{A+B}{2} \qquad\qquad V(X) = \frac{(B-A)^2}{12}$$

Example Suppose Buses arrive at a specific bus stop at 15-minute intervals. If a passenger arrives at the bus stop at a random time, then X = waiting time in minutes, is uniformly distributed between 0 and 15.

(a) Compute $P(X = 5)$, $P(3 < X < 8)$, and $P(-1 < X < 5)$.

(b) For a satisfying $0 < a < a + 5 < 10$, compute $P(a < X < a + 5)$.

(c) Compute $E(X)$ and $V(X)$.

(d) For the next 10 passengers, what's the probability that exactly five of them need to wait less than five minutes?

4.3 The Normal Distribution

The normal or “bell-shaped” distribution is the cornerstone of most methods of estimation and hypothesis testing developed in the rest of this course.

A continuous random variable X is said to have a **normal distribution** with parameters μ and σ, where $-\infty < \mu < \infty$ and $\sigma > 0$, if the pdf of X is given by

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty$$

We write $X \sim N(\mu, \sigma^2)$.

Its mean and variance are

$$E(X) = \mu \qquad\qquad V(X) = \sigma^2$$

If $X \sim N(\mu, \sigma^2)$, then

- X has a bell-shaped probability distribution.
- The probability distribution is perfectly symmetric and centered at its mean μ.
- Its spread is determined by σ.
- There are an infinitely large number of normal curves, one for each pair of μ and σ.
- The total area under any normal curve is 1.

Specially, $N(\mu = 0, \sigma^2 = 1)$ is called **standard normal distribution**. A random variable having a standard normal distribution is called a **standard normal random variable** and will be denoted by Z. That is, $Z \sim N(0, 1)$. The pdf of Z is

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$$

The cdf of Z is denoted by $\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} f(y; 0, 1)dy$.

Appendix Table A.3 gives $\Phi(z)$, the area under the standard normal density curve to the left of z.

If $Z \sim N(0, 1)$, then

- $P(Z \le -x) = P(Z \ge x)$
- $\Phi(-x) = 1 - \Phi(x)$
- If $X = \mu + \sigma Z$, then $X \sim N(\mu, \sigma^2)$

Standardizing Normal Distributions

If $X \sim N(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$, then $Z \sim N(0, 1)$.

- $F(X) = \Phi(Z)$
- $P(a \le X \le b) = F(b) - F(a) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$
- [(100p)th percentile of X] = $\mu + \sigma \times$ [(100p)th percentile of Z]

4 Typical Problems about Normal Distribution

- Compute $P($about $Z)$
- Find z-values
- Compute $P($about $X)$
- Find x-values

Compute $P($about $Z)$ – Use Table A.3 to find $\Phi(z)$

- $P(Z \leq b) = \Phi(b)$
- $P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$
- $P(Z \geq a) = 1 - \Phi(a)$

Example Suppose $Z \sim N(0, 1)$, find the following probabilities

(a) $P(Z < 1)$

(b) $P(-2 \leq Z \leq 2)$

(c) $P(0.38 \leq Z \leq 1.25)$

(d) $P(Z \geq -3)$

Find z-values

1. Find the z-value's left-tail probability $\Phi(z)$
2. Search for the z-value in Table A.3

Example: Find a value of the standard normal random variable Z, call it z_0, such that

(a) $P(Z \geq z_0) = 0.2090$

(b) $P(Z < z_0) = 0.7090$

(c) $P(-z_0 \leq Z < z_0) = 0.8472$

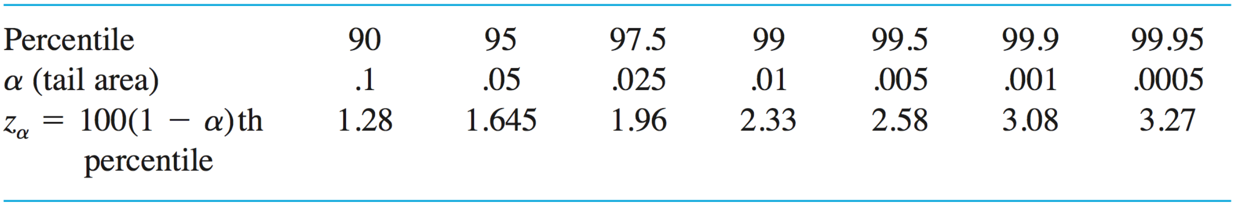
In statistical inference, we will need the values on the horizontal z axis that capture certain small tail areas under the standard normal curve.

Notation: For $0 \leq \alpha \leq 1$, the $100(1-\alpha)$th percentile of the distribution is denoted by z_α, so that

$$\Phi(z_\alpha) = 1 - \alpha \quad \Rightarrow \quad P(Z \geq z_\alpha) = \alpha$$

The percentiles z_α are often referred to as the **“z critical values”**. That is, z_α denotes the value on the z axis for which α of the area under the z curve lies to the right of z_α.

Table 4.1 Standard Normal Percentiles and Critical Values



Percentile	90	95	97.5	99	99.5	99.9	99.95
α (tail area)	.1	.05	.025	.01	.005	.001	.0005
$z_\alpha = 100(1-\alpha)$th percentile	1.28	1.645	1.96	2.33	2.58	3.08	3.27

Example: Compute $P(|Z| \leq z_{\alpha/2})$.

Example: Find $z_{0.025}$ and $z_{0.2}$.

Compute $P(\text{about } X)$, where $X \sim N(\mu, \sigma^2)$

1. $Z = \frac{X-\mu}{\sigma}$, i.e., standardizing X
2. Compute $P(\text{about } Z)$

Example Suppose $X \sim N(3, 9)$, Compute

(a) $P(3 < X < 6)$

(b) $P(X > 0)$

(c) $P(|X - 3| > 6)$

Example Suppose $X \sim N(\mu, \sigma^2)$, find $P(\mu - c\,\sigma \leq X \leq \mu + c\,\sigma)$ where

(a) $c = 1$

(b) $c = 2$

(c) $c = 3$

Find x-values – Think Percentiles!

1. Find the desired z-value such at $F(x) = \Phi(z)$
2. Convert z to x using $x = \mu + \sigma z$.

Example Suppose $X \sim N(9, 4)$, find

(a) $\eta(.75)$

(b) $\eta(.08)$

Example Suppose $X \sim N(6, 16)$, find a value of the normal random variable X, call it x_0, such that

(a) $P(X \leq x_0) = 0.8531$

(b) $P(X > x_0) = 0.025$

(c) $P(X > x_0) = 0.95$

Exercise 4.35 Suppose the diameter at breast height (in.) of trees of a certain type is normally distributed with μ = 8.8 and σ = 2.8.

(a) What is the probability that the diameter of a randomly selected tree will be at least 10 in.? Will exceed 10 in.?

(b) What is the probability that the diameter of a randomly selected tree will exceed 20 in.?

(c) What is the probability that the diameter of a randomly selected tree will be between 5 and 10 in.?

(d) What value c is such that the interval ($8.8 - c$, $8.8 + c$) includes 98% of all diameter values?

(e) If four trees are independently selected, what is the probability that at least one has a diameter exceeding 10 in.?

The normal distribution is often used as an approximation to the distribution of values in a discrete population. In such situations, extra care should be taken to ensure that probabilities are computed in an accurate manner.

Normal Approximation of Binomial Distribution

Let $X \sim Bin(n,p)$ be a binomial rv based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has approximately a normal distribution $N\left(\mu = np, \sigma^2 = np(1-p)\right)$.

In practice, the approximation is adequate provided that both $np \geq 10$ and $n(1-p) \geq 10$, since there is then enough symmetry in the underlying binomial distribution.

Example Assume that $X \sim Bin(n,p)$. In which of the following cases would it be appropriate to use the normal approximation to the binomial?

(a) n = 100, p = .01

(b) n = 25, p = .6

(c) n = 10, p = .4

Since we are using a continuous probability distribution to approximate probabilities for a discrete probability distribution and for $X \sim N(\mu, \sigma^2)$,

$$P(X = x) = 0, \quad -\infty < x < \infty,$$

and for $X \sim Bin(n, p)$,

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \cdots, n,$$

we must do <u>**continuity correction**</u>.

If $X \sim Bin(n, p)$ can be approximated by $N(\mu, \sigma^2)$, where $\mu = np$ and $\sigma = \sqrt{np(1-p)}$, we have

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \overset{approx.}{\sim} N(0, 1)$$

In particular, for x = a possible value of X,

$$\begin{aligned} P(X \le x) = B(x; n, p) &\approx \begin{pmatrix} \text{area under the normal curve} \\ \text{to the the left of } x + 0.5 \end{pmatrix} \\ &= \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

Example Suppose $X \sim Bin(n,p)$. Show how to do continuity correction when calculating the following probabilities using normal approximation.

(a) $P(X = 10)$

(b) $P(X \leq 12)$

(c) $P(X < 9)$

(d) $P(X \geq 19)$

(e) $P(X > 28)$

(f) $P(2 \leq X \leq 17)$

(g) $P(3 \leq X < 27)$

(h) $P(6 < X \leq 40)$

(i) $P(5 < X < 8)$

Exercise 4.54 Suppose that 10% of all steel shafts produced by a certain process are nonconforming but can be reworked (rather than having to be scrapped). Consider a random sample of 200 shafts, and let X denote the number among these that are nonconforming and can be reworked. What is the (approximate) probability that X is

(a) At most 30?

(b) Less than 30?

(c) Between 15 and 25 (inclusive)?

4.4 The Exponential and Gamma Distributions

There are many practical situations in which the variable of interest to an investigator might have a skewed distribution. One family of distributions that has this property is the gamma family. We first consider a special case, the exponential distribution, and then generalize later in the section.

A continuous rv X is said to have an **exponential distribution** with parameter $\lambda > 0$, if the pdf of X is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The cdf of the exponential rv X is

$$F(x;\lambda) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

And its mean and variance are

$$E(X) = \frac{1}{\lambda} \qquad V(X) = \frac{1}{\lambda^2}$$

Relationship between Poisson and Exponential distributions

Suppose that the number of events occurring in any time interval of length t has a **Poisson** distribution with parameter at (where a, the rate of the event process, is the expected number of events occurring in 1 unit of time) and that numbers of occurrences in nonoverlapping intervals are independent of one another. Then the distribution of elapsed time between the occurrence of two successive events is **exponential** with parameter $\lambda = a$.

The exponential distribution is frequently used as a model for the distribution of elapsed time between the occurrence of successive events, such as customers arriving at a service facility or calls coming in to a switchboard.

Exponential distribution is **memoryless**

$$P(X \geq t + t_0 \mid X \geq t_0) = P(X \geq t)$$

where X is interpreted as the waiting time for an event to occur relative to some initial time.

The above relation implies that, if X is conditioned on a failure to observe the event over some initial period of time t_0, the conditional probability that the remaining waiting time is greater than or equals to t, i.e., $P(X \geq t + t_0 \mid X \geq t_0)$, is the same as the original unconditional probability that the waiting time is great than or equals to t, i.e., $P(X \geq t)$.

In other words, the distribution of remaining waiting time is independent of the starting time. Thus, the exponential distribution is often used to model the waiting time until some specific event occurs. For example, the amount of time from now until an earthquake occurs, or the survival time of patients with certain disease.

Exercise 4.59 Let X = the time between two successive arrivals at the drive-up window of a local bank. If X has an exponential distribution with $\lambda = 1$, compute the following:

(a) The expected time between two successive arrivals

(b) The standard deviation of the time between successive arrivals

(c) $P(X \leq 4)$

(d) $P(2 \leq X \leq 5)$

Gamma Distribution

To define the family of gamma distributions, we first need to introduce a function that plays an important role in many branches of mathematics.

For $\alpha > 0$, the **gamma function** $\Gamma(\alpha)$ is defined by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$$

Properties of Gamma Function

- For any $\alpha > 1$, $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$
- For any positive integer n, $\Gamma(n) = (n-1)!$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

A continuous random variable X is said to have a **gamma distribution** if the pdf of X is

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where the parameters α and β satisfy $\alpha > 0$,$\beta > 0$.

And its mean and variance are

$$E(X) = \alpha\beta \qquad V(X) = \alpha\beta^2$$

The **standard gamma distribution** has $\beta = 1$, so the pdf of a standard gamma rv X is

$$f(x;\alpha) = \begin{cases} \dfrac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

And the cdf of the standard gamma rv X is

$$\gamma(x,\alpha) = F(x;\alpha) = \int_0^x \frac{y^{\alpha-1}e^{-y}}{\Gamma(\alpha)} dy \qquad x > 0$$

$\gamma(x,\alpha)$ is also called **incomplete gamma function** (Table A.4).

If $X \sim Gamma(\alpha,\beta)$, then the cdf of X is

$$F(x;\alpha,\beta) = P(X \leq x) = \gamma(x/\beta,\alpha)$$

The exponential distribution is a special case of gamma distribution by taking $\alpha = 1$ and $\beta = \frac{1}{\lambda}$.

$$X \sim Exp(\lambda) \Longleftrightarrow X \sim Gamma(\alpha = 1, \beta = \frac{1}{\lambda})$$

Exercise 4.66 Suppose the time spent by a randomly selected student who uses a terminal connected to a local time-sharing computer facility has a gamma distribution with mean 20 min and variance 80 min^2.

(a) What are the values of α and β?

(b) What is the probability that a student uses the terminal for at most 24 min?

(c) What is the probability that a student spends between 20 and 40 min using the terminal?

Chapter 5

Joint Probability Distributions and Random Samples

Many problems in probability and statistics involve several random variables simultaneously. In this chapter, we first discuss probability models for the joint (i.e., simultaneous) behavior of several random variables.

Then later in this chapter, we consider functions of n random variables $X_1, X_2, \cdots, X_n$, focusing especially on their average $\frac{1}{n}\sum_{i=1}^{n} X_i$. We call any such function, itself a random variable, a **(sample) statistic**. Methods from probability are used to obtain information about the distribution of a statistic.

The premier result of this type is the **Central Limit Theorem (CLT)**, the basis for many inferential procedures involving large sample sizes.

5.1 Jointly Distributed Random Variables

Here we start with joint probability distributions for two (discrete or continuous) random variables.

Two Discrete Random Variables

Suppose X and Y are both discrete random variables. The **joint probability mass function** $p(x, y)$ is defined for each pair (x, y) by

$$p(x, y) = P(X = x, Y = y)$$

satisfying $p(x, y) \geq 0$ and $\sum_x \sum_y p(x, y) = 1$.

Now let A be any set consisting of pairs of (x, y) values (e.g., $A = \{(x, y) : x + y = 5\}$ or $\{(x, y) : \max(x, y) \leq 3\}$).
Then the probability $P[(X, Y) \in A]$ is obtained by summing the joint pmf over pairs in A:

$$P[(X, Y) \in A] = \sum_{(x,y) \in A}\sum p(x, y)$$

The **marginal probability mass function** of X and Y, denoted by $p_X(x)$ and $p_Y(y)$, respectively, are given by

$$p_X(x) = P(X = x) = \sum_y p(x, y) \quad \text{for each possible value } x$$

$$p_Y(y) = P(Y = y) = \sum_x p(x, y) \quad \text{for each possible value } y$$

Exercise 5.2 When an automobile is stopped by a roving safety patrol, each tire is checked for tire wear, and each headlight is checked to see whether it is properly aimed. Let X denote the number of headlights that need adjustment, and let Y denote the number of defective tires.

(a) If X and Y are independent with $p_X(0) = .5$, $p_X(1) = .3$, $p_X(2) = .2$, and $p_Y(0) = .6$, $p_Y(1) = .1$, $p_Y(2) = p_Y(3) = .05$, $p_Y(4) = .2$, display the joint pmf of (X, Y) in a joint probability table.

(b) Compute $P(X \leq 1$ and $Y \leq 1)$ from the joint probability table, and verify that it equals the product $P(X \leq 1) \cdot P(Y \leq 1)$.

(c) What is $P(X + Y = 0)$, the probability of no violations?

(d) Compute $P(X + Y \leq 1)$.

Two Continuous Random Variables

Let X and Y be continuous random variables. A **joint probability density function** $f(x, y)$ for these two variables is a function satisfying $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$.
Then for any two-dimensional set A

$$P[(X, Y) \in A] = \iint\limits_{A} f(x, y)\, dx\, dy$$

In particular, if A is the two-dimensional rectangle $\{(x, y) : a \leq x \leq b, c \leq y \leq d\}$, then

$$P[(X, Y) \in A] = P(a \leq X \leq b, c \leq Y \leq d) = \int_{a}^{b} \int_{c}^{d} f(x, y)\, dy\, dx$$

The **marginal probability density function** of X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{for } -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad \text{for } -\infty < y < \infty$$

Two random variables X and Y are said to be **independent** if for every pair of x and y values

$$p(x, y) = p_X(x) \cdot p_Y(y) \quad \text{when } X \text{ and } Y \text{ are discrete}$$

or

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad \text{when } X \text{ and } Y \text{ are continuous}$$

If the above condition is not satisfied for all (x, y), then X and Y are said to be **dependent**.

If X and Y are two discrete random variables with joint pmf $p(x, y)$ and marginal X pmf $p_X(x) > 0$, then **conditional probability mass function of Y given that $X = x$** is

$$p_{Y|X}(y \mid x) = P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{p(x, y)}{p_X(x)},$$

for any possible value y.

If X and Y are two continuous random variables with joint pdf $f(x, y)$ and marginal X pdf $f_X(x)$, then for any X value x for which $f_X(x) > 0$, the **conditional probability density function of Y given that $X = x$** is

$$f_{Y|X}(y \mid x) = \frac{f(x, y)}{f_X(x)} \qquad \text{for } -\infty < y < \infty$$

Exercises 5.9 & 5.19 Each front tire on a particular type of vehicle is supposed to be filled to a pressure of 26 psi. Suppose the actual air pressure in each tire is a random variable – X for the right tire and Y for the left tire, with joint pdf

$$f(x,y) = \begin{cases} K(x^2+y^2) & 20 \le x \le 30,\ 20 \le y \le 30 \\ 0 & \text{otherwise} \end{cases}$$

(a) What is the value of K?

(b) What is the probability that both tires are underfilled?

(c) What is the probability that the difference in air pressure between the two tires is at most 2 psi?

(d) Determine the (marginal) distribution of air pressure in the right tire.

(e) Are X and Y independent rv's?

(f) Determine the conditional pdf of Y given that $X = x$ and the conditional pdf of X given that $Y = y$.

(g) If the pressure in the right tire is found to be 22 psi, what is the probability that the left tire has a pressure of at least 25 psi? Compare this to $P(Y \geq 25)$.

(h) If the pressure in the right tire is found to be 22 psi, what is the expected pressure in the left tire, and what is the standard deviation of pressure in this tire?

5.2 Expected Values, Covariance, and Correlation

Let X and Y be jointly distributed random variables with pmf $p(x,y)$ or pdf $f(x,y)$ according to whether the variables are discrete or continuous. Then the **expected value of a function $h(X,Y)$**, denoted by $E[h(X,Y)]$ or $\mu_{h(X,Y)}$, is given by

$$E[h(X,Y)] = \begin{cases} \sum\limits_{x}\sum\limits_{y} h(x,y)p(x,y) & X,Y \text{ discrete} \\ \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} h(x,y)f(x,y)\,dx\,dy & X,Y \text{ continuous} \end{cases}$$

When two random variables X and Y are not independent, it is frequently of interest to assess how strongly they are related to one another. The extent to which two random variables vary together (co-vary) can be measured by their covariance.

The **covariance** between two random variables X and Y is

$$\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
$$= \begin{cases} \sum\limits_{x}\sum\limits_{y}(x-\mu_x)(y-\mu_y)p(x,y) & X,Y \text{ discrete} \\ \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y)f(x,y)\,dx\,dy & X,Y \text{ continuous} \end{cases}$$

Shortcut formula for Cov(X,Y)

$$\text{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - \mu_X\mu_Y$$

Interpreting the Covariance

- positive covariance – indicates that higher than average values of one variable tend to be paired with higher than average values of the other variable.
- negative covariance – indicates that higher than average values of one variable tend to be paired with lower than average values of the other variable.
- zero covariance – If the two random variables are independent, the covariance will be zero. However, a covariance of zero does not necessarily mean that the variables are independent. A nonlinear relationship can exist that still would result in a covariance value of zero.

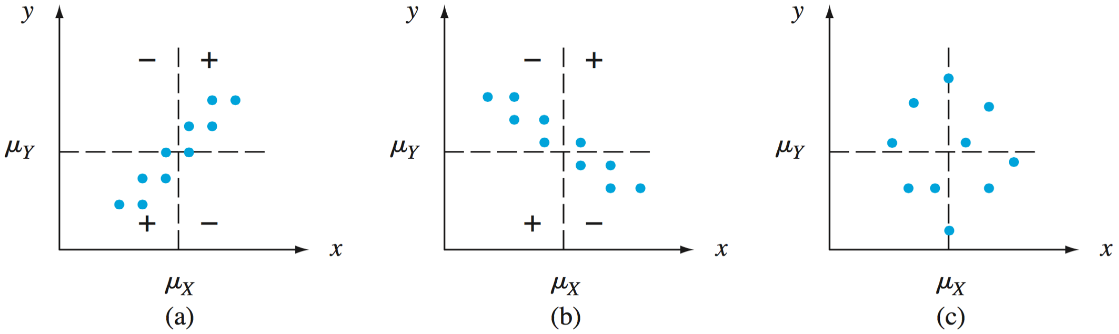


Figure 5.4 $p(x, y) = 1/10$ for each of ten pairs corresponding to indicated points: (a) positive covariance; (b) negative covariance; (c) covariance near zero

Because the value of covariance depends critically on the units of measurement, it is difficult to compare covariances among data sets having different scales. A value that might represent a strong linear relationship for one data set might represent a very weak one in another. The correlation coefficient addresses this issue by creating a dimensionless quantity that facilitates this comparison.

The **correlation coefficient** between two random variables X and Y, denoted by Corr(X, Y) or $\rho_{X,Y}$ or just ρ, is defined by

$$\rho_{X,Y} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Properties about Correlation

- If a and c are either both positive or both negative, Corr$(aX + b, cY + d) =$ Corr(X, Y)
- For any two rv's X and Y, $-1 \leq$ Corr$(X, Y) \leq 1$.
- If X and Y are independent, then $\rho = 0$.
  But $\rho = 0$ does not imply independence.
- $\rho = 1$ or -1 iff $Y = aX + b$ for some numbers a and b with $a \neq 0$.
- Cov$(X, X) = E[(X - \mu_X)^2] = V(X)$ and Corr$(X, X) = 1$.

Exercise 5.31 Given the two jointly distributed rv's X and Y in Exercise 5.9

(a) Compute the covariance between X and Y.

(b) Compute the correlation coefficient ρ for X and Y.

5.3 Statistics and Their Distributions

General Problems of Statistics: A sample of size n is a collection of observations selected from a particular population distribution (with pmf $p(x)$ or pdf $f(x)$, for discrete or continous cases, respectively). The data values recorded, $x_1, \cdots, x_n$, are the observed values of a set of n independent random variables $X_1, \cdots, X_n$, and each has the same probability distribution $p(x)$ or $f(x)$. The general problem is to estimate some unknown quantity about $p(x)$ or $f(x)$ based on the collected data $x_1, \cdots, x_n$.

A **statistic** is any quantity whose value can be calculated from sample data. Prior to obtaining data, there is uncertainty as to what value of any particular statistic will result. Therefore, a statistic is a random variable and will be denoted by an uppercase letter; a lowercase letter is used to represent the calculated or observed value of the statistic.

Any statistic, being a random variable, has a probability distribution. For example, the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has a probability distribution, and the sample variance
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ also has a probability distribution.

The probability distribution of a statistic is sometimes referred to as its **sampling distribution** to emphasize that it describes how the statistic varies in value across all samples that might be selected.

Statistics are random variables; parameters are constants.

Random Samples

The probability distribution of any particular statistic depends not only on the population distribution (normal, uniform, etc.) and the sample size n but also on the method of sampling.

The rv's $X_1, X_2, \cdots, X_n$ are said to form a **random sample** of size n if

- The X_i's are independt rv's.
- Every X_i has the same probability distribution.

In other words, we say that the X_i's are **independent and identically distributed (iid)**.

If sampling is either with replacement or from an infinite population, these conditions are satisfied exactly.

These conditions will be approximately satisfied if sampling is without replacement, yet the sample size n is much smaller than the population size N.

In practice, if $n/N \leq .05$ (at most 5% of the population is sampled), we can proceed as if the X_i's form a random sample. The virtue of this sampling method is that the probability distribution of any statistic can be more easily obtained than for any other sampling method.

Exercise 5.38 There are two traffic lights on a commuters route to and from work. Let X_1 be the number of lights at which the commuter must stop on his way to work, and X_2 be the number of lights at which he must stop when returning from work. Suppose that X_1 and X_2 are independent and each has the same distribution as given in the accompanying table (so that X_1, X_2 is a random sample of size $n = 2$).

x_i	0	1	2
$p(x_i)$	.2	.5	.3

$\mu = 1.1, \sigma^2 = .49$

(a) Determine the pmf of $T_o = X_1 + X_2$.

(b) Calculate μ_{T_o}. How does it relate to μ, the population mean?

(c) Calculate $\sigma^2_{T_o}$. How does it relate to σ^2, the population variance?

(d) Let X_3 and X_4 be the number of lights at which a stop is required when driving to and from work on a second day assumed independent of the first day. With T_o = the sum of all four X_i's, what now are the values of $E(T_o)$ and $V(T_o)$?

(e) Referring back to (d), what are the values of $P(T_o = 8)$ and $P(T_o \geq 7)$?

5.4 The Distribution of the Sample Mean

If $X_1, \cdots, X_n$ are rv's and $a_1, \cdots, a_n$ and b are constants, then

$$E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b$$

If, in addition, $X_1, \cdots, X_n$ are independent, then

$$V(a_1X_1 + \cdots + a_nX_n + b) = a_1^2V(X_1) + \cdots + a_n^2V(X_n)$$

Let $X_1, \cdots, X_n$ be a random sample from a distribution with mean value μ and standard deviation σ. Then

- $\mu_{\overline{X}} = E(\overline{X}) = \mu$
- $\sigma_{\overline{X}}^2 = V(\overline{X}) = \sigma^2/n$ and $\sigma_{\overline{X}} = \sigma/\sqrt{n}$

In addition, with $T_o = X_1 + \cdots + X_n$ (the sample total), we have $E(T_o) = n\mu$, $V(T_o) = n\sigma^2$, and $\sigma_{T_o} = \sqrt{n}\sigma$.

Let $X_1, \cdots, X_n$ be iid rv's and $X_i \sim N(\mu, \sigma^2)$, then for any n, $\overline{X} \sim N(\mu_{\overline{X}} = \mu, \sigma_{\overline{X}}^2 = \sigma^2/n)$.

The Central Limit Theorem (CLT)

Let $X_1, \cdots, X_n$ be a random variables from a distribution with mean μ and variance σ^2.
Then if n is sufficiently large, the distribution of their average $\overline{X}$ can be approximated by a $N(\mu, \sigma^2/n)$ distribution.
Similarly, the distribution of sample total $T_o = X_1 + \cdots + X_n$, can be approximated by a $N(n\mu, n\sigma^2)$ distribution.
The larger the value of n, the better the approximation.

Rule of the Thumb
If $n > 30$, the Central Limit Theorem can be used.

Exercise 5.50 The breaking strength of a rivet has a mean value of 10,000 psi and a standard deviation of 500 psi.

(a) What is the probability that the sample mean breaking strength for a random sample of 40 rivets is between 9900 and 10,200?

(b) If the sample size had been 15 rather than 40, could the probability requested in part (a) be calculated from the given information?

Exercise 5.52 The lifetime of a certain type of battery is normally distributed with mean value 10 hours and standard deviation 1 hour. There are four batteries in a package. What lifetime value is such that the total lifetime of all batteries in a package exceeds that value for only 5% of all packages?

5.5 The Distribution of a Linear Combination

If $X_1, \cdots, X_n$ are independent, normally distributed rv's (with possibly different means and/or variances), then any linear combination of the X_i's also has a normal distribution.

In particular, the difference $X_1 - X_2$ between two independent, normally distributed variables is itself normally distributed with parameters $\mu_{X_1 - X_2} = \mu_{X_1} - \mu_{X_2}$ and $\sigma^2_{X_1 - X_2} = \sigma^2_{X_1} + \sigma^2_{X_2}$.

Exercise 5.62 Manufacture of a certain component requires three different machining operations. Machining time for each operation has a normal distribution, and the three times are independent of one another. The mean values are 15, 30, and 20 min, respectively, and the standard deviations are 1, 2, and 1.5 min, respectively. What is the probability that it takes at most 1 hour of machining time to produce a randomly selected component?

Chapter 6

Point Estimation

Given a parameter of interest, such as a population mean μ or population proportion p, the objective of point estimation is to use a sample to compute a number that represents in some sense a good guess for the true value of the parameter.

6.1 Some General Concepts of Point Estimation

When discussing general concepts and methods of inference, it is convenient to have a generic symbol for the parameter of interest. We will use the Greek letter θ for this purpose.

A **point estimate** of a parameter θ is a single number that can be regarded as a sensible value for θ. A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the **point estimator** of θ, often denoted by $\hat{\theta}$.

Given a random sample $x_1, \cdots, x_n$ from a population with unknown parameter μ and variance σ^2, then

- the sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a point estimate of μ
- the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ is a point estimate of σ^2

If we write $\hat{\theta} = \theta +$ error of estimation
then an accurate estimator would be one resulting in small estimation errors, so that estimated values will be near the true value.

A point estimator $\hat{\theta}$ is said to be an **unbiased estimator** of θ if $E(\hat{\theta}) = \theta$. If $\hat{\theta}$ is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the **bias** of $\hat{\theta}$.

That is, $\hat{\theta}$ is unbiased if its sampling distribution is always centered at the true value of the parameter θ.

When choosing among several different estimators of θ, select one that is unbiased.

Examples of unbiased estimators

If $X_1, \cdots, X_n$ is a random sample from a distribution with mean μ and variance σ^2

- Sample mean $\overline{X}$ is an unbiased estimator of μ.
- If the distribution is continuous and symmetric, sample median $\widetilde{X}$ and any trimmed mean $\overline{X}_{tr(100\alpha)}$ are also unbiased estimators of μ.
- Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator of σ^2.

If $X \sim Bin(n, p)$, then the sample proportion of successes $\hat{p} = \frac{X}{n}$ is an unbiased estimator of p.

Among all estimators of θ that are unbiased, choose the one that has minimum variance. The resulting $\hat{\theta}$ is called the **minimum variance unbiased estimator (MVUE)** of θ.

One of the triumphs of mathematical statistics has been the development of methodology for identifying the MVUE in a wide variety of situations. The most important result of this type for our purposes concerns estimating the mean μ of a normal distribution.

Let $X_1, \cdots, X_n$ be a random sample from a normal distribution with parameters μ and σ.

- Sample mean $\hat{\mu} = \overline{X}$ is the MVUE of μ.
- Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is the MVUE estimator of σ^2.

The **standard error** of an estimator $\hat{\theta}$ is its standard deviation $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. It is the magnitude of a typical or representative deviation between an estimate and the value of θ.

If the standard error itself involves unknown parameters whose values can be estimated, substitution of these estimates into $\sigma_{\hat{\theta}}$ yields the **estimated standard error** ($\hat{\sigma}_{\hat{\theta}}$ or $s_{\hat{\theta}}$) of the estimator.

Exercise 6.2 A sample of 20 students who had recently taken elementary statistics yielded the following information on brand of calculator owned (T = Texas Instruments, H = Hewlett Packard, C = Casio, S = Sharp):

T T H T C T T S C H
S S T H C T T T H T

(a) Estimate the true proportion of all such students who own a Texas Instruments calculator.

(b) Of the 10 students who owned a TI calculator, 4 had graphing calculators. Estimate the proportion of students who do not own a TI graphing calculator.

Exercise 6.10 Using a long rod that has length μ, you are going to lay out a square plot in which the length of each side is μ. Thus the area of the plot will be μ^2. However, you do not know the value of μ, so you decide to make n independent measurements $X_1, X_2, \cdots, X_n$ of the length. Assume that each X_i has mean μ and variance σ^2.

(a) Show that $\overline{X}^2$ is not an unbiased estimator for μ^2.

(b) For what value of k is the estimator $\overline{X}^2 - kS^2$ unbiased for μ^2.

6.2 Methods of Point Estimation

The Method of Moments

Let $X_1, \cdots, X_n$ be a random sample from a pmf or pdf $f(x)$. For $k = 1, 2, \cdots$

- The k**th population moment** or k**th moment of the distribution** is $E(X^k)$.
- The k**th sample moment** is $\frac{1}{n}\sum_{i=1}^{n} X_i^k$.

Let $X_1, \cdots, X_n$ be a random sample from a distribution with pmf or pdf $p(x; \theta_1, \cdots, \theta_m)$, where $\theta_1, \cdots, \theta_m$ are parameters whose values are unknown. Then the **moment estimators** $\hat{\theta}_1, \cdots, \hat{\theta}_m$ are obtained by equating the first m sample moments to the corresponding first m population moments and solving for $\theta_1, \cdots, \theta_m$.

The basic idea of this method is to equate certain sample characteristics, such as the mean, to the corresponding population expected values. Then solving these equations for unknown parameter values yields the estimators.

Example Suppose $X_1, \cdots, X_n$ is a random sample from a Bernoulli distribution with parameter p. That is, each X_i takes the value 1 with probability p and the value 0 with probability $1 - p$. Find the moment estimator of p. Is the moment estimator unbiased?

Example Suppose $X_1, \cdots, X_n$ is a random sample from a normal distribution with parameters μ and σ. Find the moment estimators of μ and σ^2. Are they unbiased?

Maximum Likelihood Estimation

Let $X_1, \cdots, X_n$ have joint pmf or pdf

$$f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_m) = f(x_1, \cdots, x_n; \mathbf{\Theta})$$

where the parameters $\mathbf{\Theta} = \{\theta_1, \cdots, \theta_m\}$ have unknown values.

When $x_1, \cdots, x_n$ are the observed sample values and $L = f(x_1, \cdots, x_n; \mathbf{\Theta})$ is regarded as a function of $\mathbf{\Theta} = \{\theta_1, \cdots, \theta_m\}$, it is called the **likelihood function**.

The **maximum likelihood estimates (mle's)** $\hat{\mathbf{\Theta}} = \{\hat{\theta}_1, \cdots, \hat{\theta}_m\}$ are those values of the θ_i's that maximize the likelihood function.

When the X_i's are substituted in place of the x_i's, the **maximum likelihood estimators** result.

The likelihood function tells us how likely the observed sample is as a function of the possible parameter values. Maximizing the likelihood gives the parameter values for which the observed sample is most likely to have been generated — that is, the parameter values that “agree most closely” with the observed data.

Notes on finding the mle

1. If $X_1, \cdots, X_n$ is a random sample, i.e. X_i's are independent and identically distributed (iid). Because of independence, the likelihood function L is a product of the individual pmf's or pdf's.

2. Finding $\boldsymbol{\Theta}$ to maximize $\ln(L)$ is equivalent to maximizing L itself. In statistics, taking the logarithm frequently changes a product to a sum, which is easier to work with.

$$\begin{aligned} \ln(xy) &= \ln(x) + \ln(y) \\ \ln(x/y) &= \ln(x) - \ln(y) \\ \ln(x^y) &= y \ln(x) \end{aligned}$$

3. To find the values of θ_i's that maximize $\ln(L)$, we must take the partial derivatives of $\ln(L)$ or with respect to each θ_i, equate them to zero, and solve the equations for θ_i's. This solution is $\hat{\boldsymbol{\Theta}} = \{\hat{\theta}_1, \cdots, \hat{\theta}_m\}$, the mle.

Example Suppose $X_1, \cdots, X_n$ is a random sample from a Bernoulli distribution with parameter p. That is, each X_i takes the value 1 with probability p and the value 0 with probability $1 - p$. Find the mle of p. Is the mle unbiased?

Example Suppose $X_1, \cdots, X_n$ is a random sample from a normal distribution with parameters μ and σ. Find the mle's of μ and σ^2. Are they unbiased?

Exercise 6.22 Let X denote the proportion of allotted time that a randomly selected student spends working on a certain aptitude test. Suppose the pdf of X is

$$f(x;\theta) = \begin{cases} (\theta+1)x^{\theta} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

where $-1 < \theta$. A random sample of ten students yields data $x_1 = .92, x_2 = .79, x_3 = .90, x_4 = .65, x_5 = .86, x_6 = .47, x_7 = .73, x_8 = .97, x_9 = .94$, and $x_{10} = .77$.

(a) Use the method of moments to obtain an estimator of θ, and then compute the estimate for the data.

(b) Obtain the maximum likelihood estimator of θ, and then compute the estimate for the given data.

Chapter 7

Statistical Intervals Based on a Single Sample

A point estimate says nothing about how close it might be to the parameter. An alternative to reporting a single sensible value for the parameter being estimated is to calculate and report an entire interval of plausible values – an **interval estimate** or **confidence interval (CI)**.

A confidence interval is always calculated by first selecting a **confidence level**, which is a measure of the degree of reliability of the interval. The most frequently used confidence levels are 99%, 95%, and 90%. The higher the confidence level, the more strongly we believe that the value of the parameter being estimated lies within the interval.

Information about the precision of an interval estimate is conveyed by the width of the interval. If the confidence level is high and the resulting interval is quite narrow, our knowledge of the value of the parameter is reasonably precise. A very wide confidence interval, however, gives the message that there is a great deal of uncertainty concerning the value of what we are estimating.

7.1 Basic Properties of Confidence Intervals

The basic concepts and properties of confidence intervals (CIs) are most easily introduced by first focusing on a simple, albeit somewhat unrealistic, problem situation. Suppose that the parameter of interest is a population mean μ and that

- The population distribution is normal
- The value of the population standard deviation σ is known

If $X_1 = x_1, \cdots, X_n = x_n$ is a random sample from a normal distribution with mean value μ and standard deviation σ. The interval $\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$ is a **95% confidence interval for μ**. A concise expression for the interval is $\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$, where $-$ gives the left endpoint (lower limit) and $+$ gives the right endpoint (upper limit).

We can also say $\overline{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}$ with 95% confidence.

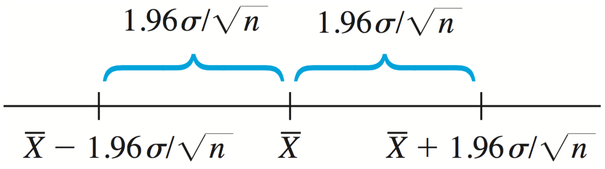

$1.96\sigma/\sqrt{n}$
$1.96\sigma/\sqrt{n}$
$\overline{X} - 1.96\sigma/\sqrt{n}$
$\overline{X}$
$\overline{X} + 1.96\sigma/\sqrt{n}$

Figure 7.2 The random interval (7.4) centered at $\overline{X}$

Interpreting a Confidence Level

If our confidence level is 95% and if we were to draw repeated random samples of the same size from a population and form confidence intervals each time, approximately 95% of the intervals would contain the population parameter.

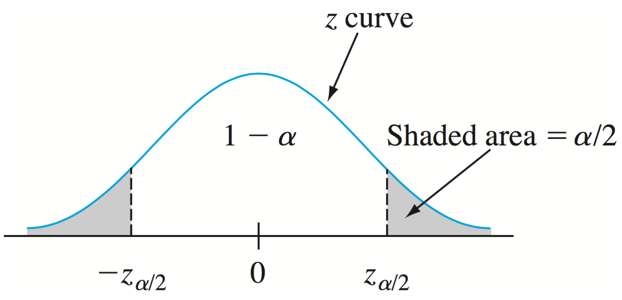


Figure 7.4 $P(-z_{\alpha/2} \le Z < z_{\alpha/2}) = 1 - \alpha$

$1-\alpha$	.80	.85	.90	.95	.98	.99
$z_{\alpha/2}$	1.28	1.44	1.645	1.96	2.33	2.575

A **$100(1-\alpha)$% confidence interval** for the mean μ of a normal population when the value of standard deviation σ is known, is given by

$$\left(\overline{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

or, equivalently $\overline{x} \pm z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$.

The **level of confidence**, $1-\alpha$, in a confidence interval is the proportion of intervals that will contain the parameter being estimated if a large number of repeated samples are obtained.

- $\overline{x}$ always the center of the CI for μ.
- $w = 2z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} =$ the width of the CI.
  - If $z_{\alpha/2}\uparrow$, then $w\uparrow$.
  - If $\sigma\uparrow$, then $w\uparrow$.
  - If $n\uparrow$, then $w\downarrow$.
  - If $(1-\alpha)\uparrow$, then $w\uparrow$.

Sample Size Determination for $100(1-\alpha)$% CI for μ

The sample size necessary for the CI $\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ to have a width w is

$$n = \left(2z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2$$

Note: Always round up when calculating the sample size n!

Exercise 7.5 Assume that the helium porosity (in percentage) of coal samples taken from any particular seam is normally distributed with true standard deviation .75.

(a) Compute a 95% CI for the true average porosity of a certain seam if the average porosity for 20 specimens from the seam was 4.85.

(b) Compute a 98% CI for true average porosity of another seam based on 16 specimens with a sample average porosity of 4.56.

(c) How large a sample size is necessary if the width of the 95% interval is to be .40?

(d) What sample size is necessary to estimate true average porosity to within .2 with 99% confidence?

7.2 Large-Sample Confidence Intervals for a Population Mean and Proportion

The CI for μ given in the section 7.1 assumed that the population distribution is normal with the value of σ known. We now present large-sample CIs whose validity do not require these assumptions.

A Large-Sample CI for Population Mean μ

Let $X_1, \cdots, X_n$ be a random sample from a population having a mean μ and standard deviation σ. Provided that n is sufficiently large, the Central Limit Theorem (CLT) implies that $\overline{X}$ has approximately a normal distribution whatever the nature of the population distribution.

$$\overline{X} \overset{approx.}{\sim} N(\mu, \frac{\sigma^2}{n})$$

In practice, σ is usually unknown, but can be estimated by the unbiased estimator S.

Therefore, the standardized variable

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{approx.}{\sim} N(0,1)$$

This implies that

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

is a **large-sample confidence interval for** μ with confidence level approximately $100(1-\alpha)$%. This formula is valid regardless of the shape of the population distribution.

Sample Size Determination for $100(1-\alpha)$% CI for μ

The sample size necessary for the CI $\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ to have a width w is

$$n = \left(2z_{\alpha/2} \cdot \frac{s}{w}\right)^2$$

Notes: We use s, an unbiased estimate of the unknown σ.

Exercise 7.14 The article "Evaluating Tunnel Kiln Performance" (*Amer. Ceramic Soc. Bull.*, Aug. 1997: 59-63) gave the following summary information for fracture strengths (MPa) of $n = 169$ ceramic bars fired in a particular kiln: $\overline{x} = 89.10$, $\sigma = 3.73$.

(a) Calculate a (two-sided) confidence interval for true average fracture strength using a confidence level of 95%. Does it appear that true average fracture strength has been precisely estimated?

(b) Suppose the investigators had believed *a priori* that the population standard deviation was about 4 MPa. Based on this supposition, how large a sample would have been required to estimate μ to within .5 MPa with 95% confidence?

A Large-Sample CI for Population Proportion p

Let p denote the proportion of "success" in a population. A random sample of size n is selected and X is the number of successes in the sample. If the sample is drawn with replacement, or n/N is sufficiently small, we have $X \sim Bin(n, p)$.

Furthermore, if $np \geq 10$ and $n(1-p) \geq 10$, X has approximately a normal distribution $N\big(\mu = np, \sigma^2 = np(1-p)\big)$.

The natural (and unbiased) estimator for p is $\hat{p} = \frac{X}{n}$, the sample fraction of successes. And $\hat{p}$ also has approximately a normal distribution $N\big(\mu = p, \sigma^2 = p(1-p)/n\big)$.

In practice, p is unknown, we can replace p by $\hat{p}$.

$$\hat{p} \overset{approx.}{\sim} \big(\mu = \hat{p}, \sigma^2 = \hat{p}(1-\hat{p})/n\big)$$

The interval $$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

$$= \Big(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\Big)$$

is called a **confidence interval for a population proportion** p with confidence level approximately $100(1-\alpha)$%.

Sample Size Determination for $100(1-\alpha)$% CI for p

The sample size necessary for the CI $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$ to have a width w is

$$n = \left(\frac{2z_{\alpha/2}}{w}\right)^2 \hat{p}(1-\hat{p})$$

Note: if we don't have a good estimate of p, we just use $\hat{p} = 0.5$. This will guarantee you a large enough sample.

Example According to *Thomson Financial*, through January 25, 2006, the majority of companies reporting profits had beaten estimates. A sample of 162 companies showed 104 beat estimates, 29 matched estimates, and 29 fell short.

(a) What is the unbiased point estimate of the proportion that beat estimates?

(b) Calculate a 95% CI for the proportion that beat estimates.

(c) What sample size is needed if the desired width of a 95% CI is 0.10?

(d) What sample size is needed if the desired width of a 99% CI is 0.10?

(e) What sample size would be required for which the width of a 99% CI to be at most 0.10, irrespective of p.