Introduction to Statistics

Objectives

At the completion of this section you should be able to:

- identify and understand various types and applications of statistics
- understand the types of work undertaken by statisticians
- understand statistics in economics and commerce
- be aware of publications in statistics
- understand the role of the Australian Bureau of Statistics
- recognise various types of data

10.1 The meaning of statistics

- *Statistics* plays an important role in many facets of human endeavour and arises from numerous problems in science and technology
- Statistics is often thought to be simply a collection of data presented in the form of tables, graphs and diagrams
- Statistics, in its broad role, is the science of decision making in the face of uncertainty
- It is the scientific method that enables us to make decisions as responsibly as possible

10.1 The meaning of statistics (cont)

- The decision-making process consists of the following steps:
  1. collect pertinent information that is as reliable as possible
  2. select the parts of the available information that are most helpful to making rational decisions
  3. make the actual decisions as sensibly as possible
  4. perceive the risks entailed in the particular decision made, and evaluate the corresponding risks of alternative actions

10.1 The meaning of statistics (cont)

- Statistics can be broadly split into two categories:
  1. Descriptive Statistics
  2. Statistical Inference

- *Descriptive Statistics*
  - The collection, condensation and display of information
  - The meaningful presentation of data such that its characteristics can be effectively observed

10.1 The meaning of statistics (cont)

- Statistical inference
  - Relates to decision making
  - Leads to future action rather than an inspection of the past
  - Relates to all types of decisions, but especially the following:
    1. determining whether any apparent characteristics of a situation are really unusual or whether they could have happened by chance
    2. estimating the values of unknown numerical quantities and determining the reliability of those estimates
    3. using past occurrences to attempt to predict the future

10.1 The meaning of statistics (cont)

- *Data collection*
  - The worth of both descriptive and inferential statistics in any given situation depends on the worth of the available data
  - Collection of reliable data is a fundamental requirement of statistical endeavour
  - This is often included under the heading ‘experimental design’

10.1 The meaning of statistics (cont)

- *Variability*
  - People vary, plants vary, animals vary, weather conditions vary, stockmarket shares vary, the demand for goods varies, profits vary
  - It is this variation that affects the reliability of any information obtained from any set of individuals
  - In statistical inference it must often be decided how much variation is 'acceptable'
  - It is the *amount* of variation from what is expected that must be considered, and this question is best answered by means of an appropriate statistical test

10.2 What is a statistician?

- A *statistician* can broadly be defined as a person who can collect data, present them, analyse them and draw inferences from them
- The advent of fast and efficient computers has enabled statistical operations to be carried out rapidly, making statistical analysis more economical and readily available
- Statisticians may serve in a consulting capacity for individuals or organisations, or they may lecture, undertake research or supervise others who are working on projects

10.3 Statistics in business

- There are many additional problems in business not normally encountered elsewhere. These include:
  - the sheer complexity of the interactions in economic systems
  - the extent to which systems change over time owing to demographic factors, technological and social changes and the political system
  - the incidence of single events that can cause sudden drastic changes
  - the difficulty of defining economic quantities in terms precise enough to enable them to be measured, and the difficulties of obtaining the measurements themselves

10.3 Statistics in business (cont)

- Statistical activity in business requires an extremely sound knowledge of the general principles and techniques of decision making
- In particular, it requires:
  1. techniques that enable the detection of trends and cycles
  2. the distinction of genuine structural changes from the results of random fluctuation
  3. methods to enable the best extrapolation from the past and present to the future, together with an assessment of the possible errors that can result from such extrapolation

10.4 Publications in statistics

- There are many journals throughout the world that publish statistical research
- Some examples of each type of publication:
  - General statistical journals
    - *Applied Statistics*
    - *International Statistical Review*
    - *Journal of the American Statistical Association*
    - *The Statistician*

10.4 Publications in statistics (cont)

- Journals in specific areas
  - *American Journal of Epidemiology*
  - *Biometrika*
  - *Biometrics*
  - *Statistics in Medicine*
  - *Transportation Research*
- Australian journals
  - *Australian and New Zealand Journal of Statistics*
  - *Bulletin of the Operations Research Society of Australia*
  - *Journal of the Australian Mathematical Society (Series A)*

10.5 The Australian Bureau of Statistics (ABS)

- The main source of 'official' Australian statistical data is the *Australian Bureau of Statistics (ABS)*
- It issues about 1600 monthly, quarterly, annual and irregular publications each year covering almost every field of economic and social activity
- These include descriptive matter dealing with Australia's government, international relations, defence, climate, physiography, culture and environment
- The ABS has been delivering services online for some years and these are continually being enhanced

10.5 The Australian Bureau of Statistics - ABS (cont)

- There are many free statistics available online (http://www.abs.gov.au)
- *For example*:
  - census information—a new and exciting range of products including *QuickStats, StaiMaps*, *Census Tables* and *Community Profiles*
  - information about the ABS
  - a schedule of future releases
  - career information
  - teaching resources for educators
  - statistics training

10.5 The Australian Bureau of Statistics - ABS (cont)

Coding of ABS publications

- Each ABS publication has a five-digit code according to the ABS catalogue numbering system
- In this code:

1. The *first* digit defines the broad subject group
2. The *second* digit defines a subgroup of the broad subject group
3. The *third* and *fourth* digits are serial numbers that are permanently allocated to publications within categories formed by the first two digits
4. The *fifth* digit, after the decimal point, represents the state or territory covered by the statistics of the publication

10.6 Types of data

- Statistical data can be obtained in a number of ways. These include:
  - *measurements* using an instrument, such as a device for measuring blood pressure or a simple ruler
  - *counts*, for example where the number of employees absent each day or the number of traffic accidents occurring each year is recorded
- *Rank data* arise in situations where, for example, individuals or objects are ranked according to some criterion, e.g.
  - bank customers could be ranked according to risk when requesting a loan
  - athletes could be ranked according to physical fitness

10.6 Types of data (cont)

- *Categorical data*, or *classification data*, occur where observations are placed into categories, e.g.
  - gender
  - eye colour
  - country of birth
- Analysis of data is described as
  - *quantitative* when it deals with numerical quantities
  - *qualitative* when it deals with non-numerical descriptions

10.6 Types of data (cont)

- *Primary data* are information collected by the person or organisation that will be using the information
  - This includes the case of a researcher who wants to undertake statistical analysis involving information that has not yet been collected
- *Secondary data* are information *already* collected by someone else
  - In this case the researcher has little or no control over the design or method of data collection, which may turn out to be unsatisfactory since it may be inappropriate

Introduction to Statistics Summary

- We looked at identifying and understanding various types and applications of statistics
- We tried to understand the types of work undertaken by a statistician
- We also looked at understanding statistics in economics and commerce
- We became aware of publications about statistics

+

Measures of Central Tendency

Objectives

At the completion of this section you should be able to:

- calculate the mode, median and mean from grouped and ungrouped data
- calculate quartiles, deciles, percentiles and fractiles
- calculate and interpret the geometric mean
- determine the significance of the skewness of a distribution

12.1 Introduction

- It is more convenient to describe a set of numbers by using a *single number*
- Calculating a single number is one of the most frequently encountered methods of condensing data
- The term *measure of central tendency* describes the general idea of a typical value (or average)
- An *average* is simply any single figure that is representative of many numbers
- There are three types of averages: Mean (most common), Median and Mode

12.2 The mode

- The *mode* is the number that occurs most frequently in a set of numbers
- Data with just a single mode are called *unimodal*, while if there are two modes the data are said to be *bimodal*
- The mode is often unreliable as a central measure

Example:

Find the modes of the following data sets:

3, 6, 4, 12, 5, 7, 9, 3, 5, 1, 5

Solution:

The value with the highest frequency is 5 (which occurs 3 times). Hence the mode is *Mo* = 5

12.2 The mode (cont)

- *Calculation of the mode from a frequency distribution*

  The observation with the largest frequency is the mode

  Example:

  A group of 13 real estate agents were asked how many houses they had sold in the past month. Find the mode

Number of houses sold	f
0	2
1	5
2	6
Total	**13**

The observation with the largest frequency (i.e. qty 6) is 2. Hence the mode of these data is 2

12.2 The mode (cont)

- *Calculation of the mode from a grouped frequency distribution*
  - it is not possible to calculate the *exact* value of the mode of the original data from a grouped frequency distribution
  - the class interval with the largest frequency is called the *modal class*

$$Mo \approx L + \frac{d_1}{d_1 + d_2}(i)$$

where:

L = the real lower limit of the modal class

d_1 = the frequency of the modal class minus the frequency of the previous class

d_2 = the frequency of the modal class minus the frequency of the next class above the modal class

i = the length of the class interval of the modal class

12.3 The median

- The *median* is the middle observation in a set
- 50% of the data have a value *less than* the median, and 50% of the data have a value *greater than* the median
- *Calculation of the median from raw data*

Let n = the number of observations

If n is *odd*, the median is $\frac{n+1}{2}$

If n is *even*, the median is the mean of the $\frac{n}{2}$th observation and the $\left(\frac{n}{2}+1\right)$th observation

12.3 The median (cont)

- *Calculation of the median from a frequency distribution*
  - this involves constructing an extra column (*cf*) in which the frequencies are cumulated

Number of pieces	Frequency (f)	Cumulative frequency (cf)
1	10	10
2	12	22
3	16	38
		$\sum f = 38$

  - since *n* is even, the median is the average of the 16th and 17th observations
  - from the *cf* column, the median is 2

12.3 The median (cont)

- *Calculation of the median from a grouped frequency distribution*
  - it is possible to make an *estimate* of the median
  - the class interval that contains the median is called the *median class*

$$\tilde{x} = L + \frac{\left(\frac{n}{2} + C\right)}{f}(i)$$

where:

$\tilde{x}$ = the median

L = the real lower limit of the median class

n = Σf = the total number of observations in the set

C = the cumulative frequency in the class immediately *before* the median class

f = the frequency of the median class

i = the length of the real class interval of the median class

12.4 The arithmetic mean

- The *arithmetic mean* is defined as the sum of the observations divided by the number of observations

$$\bar{x} = \frac{\sum x}{n}$$

where:

$\bar{x}$ = the arithmetic mean calculated from a *sample* (pronounced '*x*-bar')

Σx = the sum of the observations

n = the number of observations in the sample

– the symbol for the arithmetic mean calculated from a *population* is the Greek letter μ

12.4 The arithmetic mean (cont)

- *Use of an arbitrary origin*
  - calculations can be simplified by first removing numbers that have no bearing on a calculation, then restoring them at the end
  - for example, the mean of 1002, 1004 and 1009 is clearly the mean of 2, 4 and 9 with 1000 added (i.e. 5 + 1000 = 1005)
  - if the differences from the arbitrary origin are recorded as *d* then

$$\bar{x} = arbitrary\ \ origin + \frac{\sum d}{n}$$

12.4 The arithmetic mean (cont)

- *Calculation of the mean from a frequency distribution*
  - it is useful to be able to calculate a mean directly from a frequency table
  - the calculation of the mean is found from the formula:

$$\bar{x} = \frac{\sum fx}{\sum f}$$

where:

Σf = the sum of the frequencies

Σfx = the sum of each observation multiplied by its frequency

12.4 The arithmetic mean (cont)

- *Calculation of the mean from a grouped frequency distribution*
  - the mean can only be *estimated* from a grouped frequency distribution
  - assume that the observations are spread evenly throughout each class interval

$$\bar{x} = \frac{\sum fm}{\sum f}$$

where:

Σfm = the sum of the midpoint of a class interval and that class interval's frequency

Σf = the sum of the frequencies

12.4 The arithmetic mean (cont)

- *Weighted means*
  - *weighted arithmetic mean or weighted mean* is calculated by assigning *weights* (or measures of relative importance) to the observations to be averaged
  - if observation *x* is assigned weight *w*, the formula for the weighted mean is:

$$\bar{x} = \frac{\sum xw}{\sum w}$$

  - the weights are usually expressed as percentages or fractions

12.5 Quartiles

- Quartiles divide data into four equal parts
  - First quartile—Q_1
    - 25% of observations are below Q_1 and 75% above Q_1
    - also called the lower quartile
  - Second quartile—Q_2
    - 50% of observations are below Q_2 and 50% above Q_2
    - this is also the median
  - Third quartile—Q_3
    - 75% of observations are below Q_3 and 25% above Q_3
    - also called the upper quartile

12.5 Quartiles (cont)

- There are several methods used to calculate Quartiles. The textbook has adopted the following practice:
  - first, sort the observations in order, lowest to highest
  - divide the number of observations (*n*) by 4 (i.e. a quarter)
  - if the answer is ***not*** a whole number, then
    - round the number ***up*** (i.e 2.25, 2.5 or 2.75 would ***all*** become 3)
    - this "rounded up" number is used to count from the ***lower*** end of the ordered observations to give the 1st Quartile value and from the ***upper*** end of the ordered observations to give the 3rd Quartile value

5. Quartiles (cont)

- if the answer ***is*** a whole number, then
  - count from the ***lower*** end of the ordered observations and the 1st Quartile is halfway between this observation's value and the next observation's value
  - by counting from the ***upper*** end of the ordered observations the 3rd Quartile value is halfway between this observation's value and the next observation's value

12.5 Quartiles (cont)

- *Calculating quartiles*

Example:

- The sorted observations are: 25, 29, 31, 39, 43, 48, 52, 63, 66, 90
- Find the first and third quartile

Solution:

- The number of observations n = 10
- Since $\frac{n}{4} = \frac{10}{4} = 2.5$ is ***not*** a whole number, we round up to 3. Therefore:
- 1st quartile = 3rd observation from the lower end = 31
- 3rd quartile = 3rd observation from the upper end = 63

12.6 Deciles, percentiles & fractiles

- Further division of a distribution into a number of equal parts is sometimes used; the most common of these are *deciles, percentiles and fractiles*
- *Deciles* divide the sorted data into 10 sections
- *Percentiles* divide the distribution into 100 sections
- Instead of using a percentile we could refer to a *fractile*
  - i.e. the 30th percentile is the 0.30 fractile

12.6 Deciles, percentiles & fractiles

- Further division of a distribution into a number of equal parts is sometimes used; the most common of these are *deciles, percentiles and fractiles*
- *Deciles* divide the sorted data into 10 sections
- *Percentiles* divide the distribution into 100 sections
- Instead of using a percentile we could refer to a *fractile*
  - i.e. the 30th percentile is the 0.30 fractile

12.7 The geometric mean

- When dealing with quantities that change over a period, we would like to know the mean rate of change
- Examples include
  - the mean growth rate of savings over several years
  - the mean ratios of prices from one year to the next
- Geometric mean of *n* observations x_1, x_2, x_3,…x_n is given by:

$$Geometric\ mean = \sqrt[n]{x_1\ x_2\ x_3 \ldots x_n}$$

12.8 Skewness

- The *skewness* of a distribution is measured by comparing the relative positions of the mean, median and mode
- Distribution is *symmetrical*

  Mean = Median = Mode
- Distribution *skewed right*

  Median lies between mode and mean, and mode is less than mean
- Distribution *skewed left*

  Median lies between mode and mean, and mode is greater than mean

12.8 Skewness (cont)

- There are two measures commonly associated with the shapes of a distribution—*kurtosis* and *skewness*
- *Kurtosis* is the degree to which a distribution is peaked
- The kurtosis for a normal distribution is zero
- If the kurtosis is *sharper* than a normal distribution, the kurtosis is *positive*
- If it is *flatter* than a normal distribution, the kurtosis is *negative*

+ Measures of Central Tendency Summary

- We have looked at calculating the mode, median and mean from grouped and ungrouped data
- We also looked at calculating quartiles, deciles, percentiles and fractiles
- We have discussed calculating and interpreting the geometric mean
- We determined the significance of the skewness of a distribution

+

Confidence Intervals

Objectives

- At the completion of this topic you should be able to:
  - calculate estimates and their standard errors
  - calculate confidence intervals for the population mean
  - calculate confidence intervals for the population proportion

16.8 Confidence intervals for an unknown population mean

- *Point estimates*
  - a single estimate of an unknown population mean can be obtained from a random sample
  - different random samples give different values of the mean
  - a single estimate is referred to as a *point estimate*
  - *accuracy depends on:*
    - variability of data in the population
    - size of the random sample

16.8 Confidence intervals for an unknown population mean (cont)

- *The standard error of the mean*

standard error of the mean provides the precise measure of accuracy of a point estimate of the mean

$$\text{Standard error of the mean} = \frac{\sigma}{\sqrt{n}}$$

where:

σ = population standard deviation

n = size of random sample

the value of σ can be replaced by the sample standard deviation, s

$$\text{Standard error of the mean} \approx \frac{s}{\sqrt{n}}$$

16.8 Confidence intervals for an unknown population mean (cont)

- *The meaning of confidence intervals for μ*
  - instead of providing a single point estimate for μ, we consider a *range of values* or an *interval* within which the value of μ may lie
  - be able to provide a *probability* or *level of confidence* that this interval does indeed contain the true value of μ
  - common probabilities to use are 0.90, 0.95 and 0.99
  - confidence intervals for μ are based on the values from random samples

16.8 Confidence intervals for an unknown population mean (cont)

- *Construction of confidence intervals for μ*
  - Suppose that we have the following information:
    - the population has a normal distribution.
    - the population standard deviation (σ) is known.
    - the sample size *(n)* can be of *any* size
  - then a 95% confidence interval for *μ* is:

$$\left(x - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \right)$$

where:

$\bar{x}$ = the mean of the sample

σ = the population standard deviation

n = the size of the sample

+ 16.9 Confidence intervals for an unknown population proportion

- *Point estimates*
  - to give a *single* estimate of an unknown population proportion (π), use the value of the proportion (p) obtained from a random sample taken from that population
  - a single estimate of π is referred to as a *point estimate*
  - point estimates are particularly important in survey work, for example, to get an idea of how the population sampled feels about a certain issue
  - how accurate a point estimate is depends on two factors:
    - how *variable* the data are in the population
    - the *size* of the random sample used to make the estimate

16.9 Confidence intervals for an unknown population proportion (cont)

- *The standard error of the proportion*
  - the precise measure of accuracy of a point estimate (*p*) of a population proportion is provided by the *standard error of the proportion*
  - the formal definition is:

$$\text{Standard error of the proportion} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

where:

π = the population proportion

n = the size of the random sample

the value of π can be replaced by the *sample* standard proportion, *p*

$$\text{Standard error of the proportion} = \sqrt{\frac{p(1-p)}{n}}$$

16.9 Confidence intervals for an unknown population proportion (cont)

- *The meaning of confidence intervals for π*
  - instead of providing a single point estimate for π, we consider a *range of values* or an *interval* within which the value of π may lie
  - be able to provide a *probability* or *level of confidence* that this interval does indeed contain the true value of π
  - common probabilities to use are 0.90, 0.95 and 0.99, although any probability could be used
  - confidence intervals for π are based on the values from random samples

16.9 Confidence intervals for an unknown population proportion (cont)

- *Construction of confidence intervals for π*
  - we are able to construct an *interval estimate* for the true value of an unknown *population proportion* π
  - the *confidence interval* for *π* is of the form:

$$\left(p - z\sqrt{\frac{p(1-p)}{n}},\ p + z\sqrt{\frac{p(1-p)}{n}} \right)$$

where the value of z is chosen as follows:

- $z = 1.645$ for a 90% confidence interval
- $z = 1.96$ for a 95% confidence interval
- $z = 2.58$ for a 99% confidence interval

Confidence Intervals Summary

- We calculated estimates and their standard errors
- We calculated confidence intervals for the population mean
- We calculated confidence intervals for the population proportion

+

Measures of Variation

Objectives

At the completion of this section you should be able to:

- calculate common measures of variation (including the range, interquartile range, mean deviation and standard deviation) from grouped and ungrouped data
- calculate and interpret the coefficient of variation

13.1 Introduction

- A measure of central tendency in itself is not sufficient to describe a set of data adequately
- A measure of *variation* (or *dispersion)* of the data is usually required
- This measure gives an indication of the internal *spread* of the data—that is, the extent to which data items vary from one another or from a central point
- Some reasons for requiring a measure of dispersion of a set of data:
  - as an indication of the *reliability* of the average value
  - to assist in *controlling* unwanted variation

13.2 The range

- The simplest measure of variation is the *range*
- It is the difference between the largest and smallest values in a set of data

Range = largest observation – smallest observation

- Examples of uses of range include:
  - temperature fluctuations on a given day
  - movement of share prices
  - acceptable range of systolic and diastolic blood pressures

13.2 The range (cont)

- Range is considered primitive as it considers only the extreme values, which may not be useful indicators of the bulk of the population
- Extreme values, called *outliers*, may often result from errors of measurement
- *Outliers* are defined as values that are inconsistent with the rest of the data
- Although the range is the quickest and easiest measure of variation to calculate, its should be interpreted with some caution

13.3 The interquartile range (midspread)

- Measures the range of the middle 50% of the values only
- Is defined as the difference between the upper and lower quartiles

Interquartile range = upper quartile – lower quartile

$$= Q_3 - Q_1$$

- May be calculated from grouped frequency distributions that contain open-ended class intervals
- It is usually only used with a large number of observations

13.5 The standard deviation

- The most commonly used measure of variation is the *standard deviation*
- It takes into account every observation and measures the 'average deviation' of observations from mean
- It works with squares of residuals, not absolute values, therefore it is easier to use in further calculations
- The values of the mean deviation and standard deviation should be reasonably close, since they are both measuring the variation of the observations from their mean

13.5 The standard deviation (cont)

- *Population standard deviation*
  - *Uses squares* of the residuals, which will eliminate the effect of the signs, since squares of numbers cannot be negative

    Step 1: Find the sum of the squares of the residuals

    Step 2: Find their mean

    Step 3: Take the *square root* of this mean

$$\text{Standard deviation} = \sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{N}}$$

    Where N = the size of the population

    The *square* of the population standard deviation σ is called the *population variance* $Variance = \sigma^2$

13.5 The standard deviation (cont)

- *Sample standard deviation*
  - It is rare to calculate the value of σ since populations are usually very large
  - It is far more likely that the *sample standard deviation* (denoted by *s*) will be needed

$$\text{Sample standard deviation} = s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

  - *where*: $(n - 1)$ is the number of observations in the sample

13.5 The standard deviation (cont)

- *Important points about the standard deviation*
  - the standard deviation cannot be negative
  - the standard deviation of a set of data is zero if, and only if, the observations are of equal value
  - the standard deviation can *never* exceed the range of the data
  - the more scattered the data, the greater the standard deviation
  - the *square* of the standard deviation is called the *variance*

13.6 The coefficient of variation

- This is a measure of relative variability used to:
  - measure changes that have occurred in a population over time
  - compare variability of two populations that are expressed in different units of measurement
- It is expressed as a percentage rather than in terms of the units of the particular data

13.6 The coefficient of variation (cont)

- The formula for the coefficient of variation (V) is:

$$V = 100\left(\frac{s}{\bar{x}}\right)\%$$

where: $\bar{x}$ = the mean of the sample

s = the standard deviation of the sample

13.6 The coefficient of variation (cont)

- *Example:*

  Calculate the coefficient of variation for the price of 400 g cans of pet food, given that the mean is 81 cents and s = 6.77 cents. Interpret the results

- *Solution:*

$$V = 100\left(\frac{s}{\bar{x}}\right)\%$$

$$= 100\left(\frac{6.77}{81}\right)\%$$

$$= 8.36\%$$

This means that the standard deviation of the price of a 400g can of pet food is 8.36% of the mean price

13.7 Remarks

- Among the more important characteristics of the standard deviation are:
  - It is the most frequently used measure of variation, and because of its mathematical properties it has widespread use in problems involving statistical inference
  - If the mean cannot be calculated, neither can the standard deviation
  - Its value is affected by the value of every observation in the data
  - If the data have a number of extreme values, the value of the standard deviation may be distorted so as not to be a good ‘representative’ measure of variation

Measures of variation Summary

- Among the more important characteristics of the standard deviation are:
  - It is the most frequently used measure of variation, and because of its mathematical properties it has widespread use in problems involving statistical inference
  - If the mean cannot be calculated, neither can the standard deviation
  - Its value is affected by the value of every observation in the data
  - If the data have a number of extreme values, the value of the standard deviation may be distorted so as not to be a good 'representative' measure of variation

+

Hypothesis Testing

Objectives

At the completion of this topic you should be able to:

- understand the principles of statistical inference
- formulate null and alternative hypotheses
- understand one-tailed and two-tailed tests
- understand type I and type II errors
- understand test statistics
- understand the significance level of a test

+

Hypothesis Testing (Cont)

Objectives (cont)

- understand & calculate critical values
- understand the regions of acceptance and rejection
- calculate and interpret a one-sample z-test statistic
- calculate and interpret a one-sample t-test statistic
- calculate and interpret a paired t-test statistic
- calculate and interpret a two-sample t-test statistic
- understand and calculate p-values

21.1 Statistical Inference

- One of the major roles of statisticians in practice is to draw conclusions from a set of data
- This process is known as statistical inference
- We can put a probability on whether a conclusion is correct ‘beyond reasonable doubt’
- The major question to be answered is whether any difference between samples, or between a sample and a population, has occurred simply as a result of natural variation or because of a real difference between the two

21.1 Statistical Inference (cont)

- In this decision-making process we usually undertake a series of steps, which can include the following:
  1. collecting the data
  2. summarising the data
  3. setting up an hypothesis (i.e. a claim or theory), which is to be tested
  4. calculating the probability of obtaining a sample such as the one we have if the hypothesis is true
  5. either accepting or rejecting the hypothesis

21.1 Statistical Inference (cont)

- The null hypothesis
  - a technique for dealing with these problems begins with the formulation of an hypothesis
  - the null hypothesis is a statement that nothing unusual has occurred. The notation is H_o
  - the alternative hypothesis states that something unusual has occurred. The notation is H_1 or H_A
  - together they may be written in the form:

    H_o: (statement) vs. H_1: (alternative statement)

21.1 Statistical Inference (cont)

- The null hypothesis (cont)
  - Question: Is the population mean, μ, equal to a specified value?

    Hypotheses:

    H_o : The population mean, μ, is equal to the specified value.

    vs.

    H_1 : The population mean, μ, is not equal to the specified value
  - Another way of expressing the null and alternative hypotheses is in the form of symbols, e.g.

$$H_0: \mu = \mu_0$$

21.1 Statistical Inference (cont)

- The alternative hypothesis

  The alternative hypothesis may be classified as two-tailed or one-tailed

  – two-tailed test (two-sided alternative)

    - we do the test with no preconceived notion that the true value of μ is either above or below the hypothesised value of μ_o
    - the alternative hypothesis is written:

$$H_1 : \mu \neq \mu o$$

21.1 Statistical Inference (cont)

- The alternative hypothesis (cont)
  - One-tailed test (one-sided alternative)
    - we do the test with a strong conviction that, if H_o is not true, it is clear that μ is either greater than μ_o or less than μ_o
    - the alternative hypothesis is written as

      $$H_1 : \mu > \mu_o$$

      if we feel that the only possible alternative is that μ is greater than μ_o, or as

      $$H_1 : \mu < \mu_o$$

      if we feel that the only possible alternative is that μ is less than μ_o

21.1 Statistical Inference (cont)

- Significance level
  - by chance or whether an unusual event has taken place
  - after the appropriate hypotheses have been formulated, we must decide upon the significance level (or α -level) of the test
  - the most common significance level used is 5%, commonly written as α = 0.05
  - this level represents the borderline probability between whether an event (or sample) has ensued, which occurs less than 5% of the time, is considered unusual

21.2 Errors

- There are two possible errors in making a conclusion about a null hypothesis
  - Type I errors occur when you reject H_o (i.e. conclude that it is false) when H_o is really true. The probability of making a type I error is, in fact, equal to α, the significance level of the test
  - Type II errors occur when you accept H_o (i.e. conclude that it is true) when H_o is really false. The probability of making a type II error is denoted by the Greek letter β (beta)

21.3 The one-sample z-test

- Deals with the case of a single sample being chosen from a population and the question of whether that particular sample might be consistent with the rest of the population
- To answer this question, we construct a test statistic according to a particular formula
- It is important that the correct test be used, since the use of an incorrect test statistic can lead to an erroneous conclusion

21.3 The one-sample z-test (cont)

- Information required in calculation:
  - the size (n) of the sample
  - the mean $(\overline{x})$ of the sample
  - the standard deviation (s) of the sample
- Other information of interest might include:
  - Does the population have a normal distribution?
  - Is the population's standard deviation (σ) known?
  - Is the sample size (n) large?

21.3 The one-sample z-test (cont)

- There are different cases for the one-sample z-test statistic

***Case I* is a situation where:**

1. the population has a normal distribution and
2. the population standard deviation, σ, is known

***Case II* is a situation where:**

1. the population has any distribution
2. the sample size, n, is large (i.e. at least 25), and
3. the value of σ is known

21.3 The one-sample z-test (cont)

In both these cases we can use a z-test statistic defined by:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

21.3 The one-sample z-test (cont)

***Case III* is a situation where:**

1. the population has *any* distribution
2. the sample size, *n*, is large (i.e. at least 25), and
3. the value of σ is unknown (however, since *n* is large, the value of σ is approximated by the sample standard deviation, *s*)

In this case we can use a *z-test statistic* defined by:

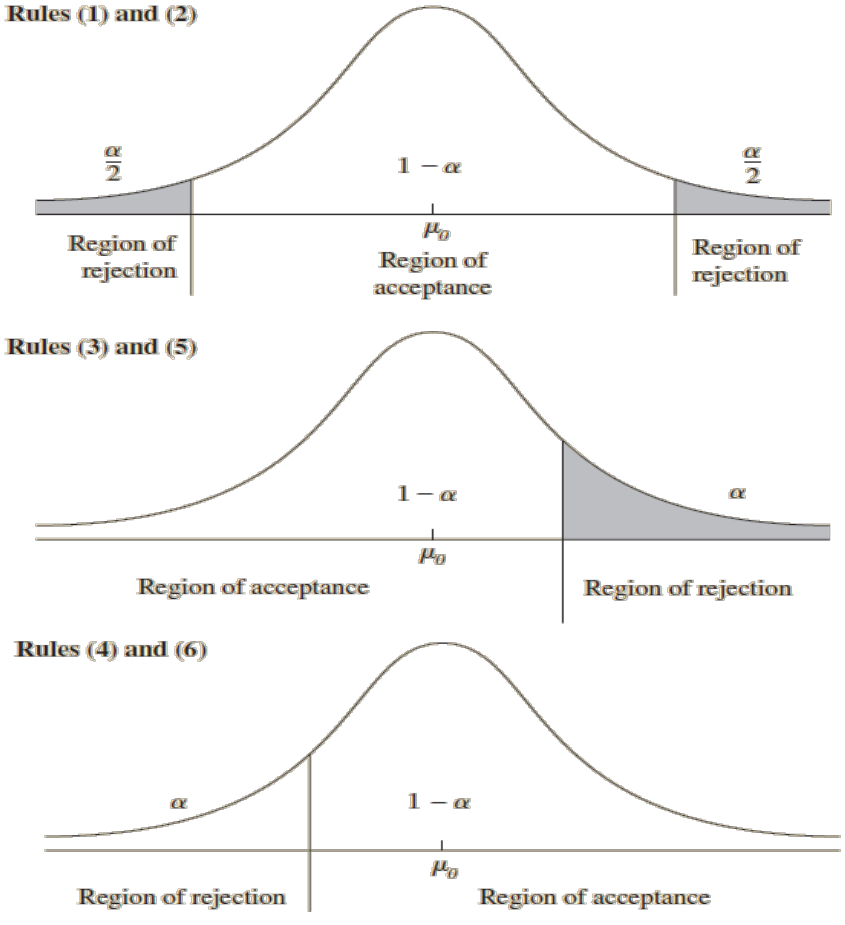
$$z = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

21.3 The one-sample z-test (cont)

- A *critical value* is one that represents the cut-off point for *z*-test statistics in deciding whether we do not reject or do reject H_0. The particular critical value to use depends on two things:
  1. whether we are using a one-sided or two-sided test, and
  2. the significance level used

21.3 The one-sample z-test (cont)

Rules for rejecting or not rejecting H_0 in a one-sample z-test



21.4 The one-sample t-test

- Consider a test that has the same objective as a one-sample *z*-test, that is to determine whether a sample is significantly different from the population as a whole
- However, the one-sample *t*-test has a different set of assumptions

Case IV is a situation where:

1. the population is normally distributed
2. the population standard deviation, s , is unknown and
3. the sample size, n, is small

21.4 The one-sample t-test (cont)

- Then we can use a t-test statistic defined as:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

21.4 The one-sample t-test (cont)

- Unlike the z-test statistic, the t-test statistic has associated with it a quantity called degrees of freedom
- The degrees of freedom are denoted by the Greek letter υ and are defined by $\upsilon = n - 1$
- Degrees of freedom relate to the fact that the sum of all deviations in a sample of size n must add up to zero
- When using a t-test statistic, we must use special critical values in a table

21.4 The one-sample t-test (cont)

- This information is given in the “Critical Values for the t-distribution” table (Table 7, page 785 of textbook) with the critical values given to three decimal places
- After the value of the *t*-test statistic is found, the method of either not rejecting or rejecting H_0 is similar to Rules (1) to (6) described for the *z*-test

21.5 Drawing a conclusion

- Once a decision has been made to either reject or not reject H_o, a conclusion should be written in terms of what is the actual problem

21.5 Drawing a conclusion (cont)

- The steps that should be undertaken to perform a one-sample test are:
  1. Set up the null and alternative hypotheses. This includes deciding whether you are using a one-sided or two-sided test
  2. Decide on the significance level
  3. Write down the relevant data
  4. Decide on the test statistic to be used
  5. Calculate the value of the test statistic
  6. Find the relevant critical value and decide whether H_o is to be rejected or not rejected
  7. Draw an appropriate conclusion

21.6 The paired t-test

- Now we will consider problems in which there are two samples that are to be compared with each other
- These are often referred to as *two-sample problems* and there are a number of test statistics available
- In some instances, the two samples have a structure such that the data are *paired*
- A comparison of the values in two samples requires the calculation of a *two-sample test statistic*
- In such cases, the most commonly used test statistic is a *paired t-test*, which takes into account this natural pairing

21.6 The paired t-test (cont)

- The steps involved in the analysis are as follows:

1. Construct the null and alternative (two-tailed) hypotheses.
   H_0 always states that there is *no difference* between the two samples. H_1 always states that there is *some difference* between the two samples.

   In *symbol* form these can be written as:

   $$H_0: \mu_d = 0 \qquad \text{v.} \qquad H_1: \mu_d \neq 0$$

2. Calculate the actual *differences* between the values in the two samples. It is important to retain the minus sign if the subtraction yields a *negative* number
3. Calculate the mean and standard deviation of the values of the differences

21.6 The paired t-test (cont)

4. Calculate the value of the *paired t-test statistic using*:

$$t = \frac{\bar{x}_d - \mu_d}{\frac{s_d}{\sqrt{n}}}$$

where:

n = the number of pairs of differences

$\mu_d = 0$ (according to the null hypothesis)

and with $\upsilon = n - 1$ degrees of freedom

21.6 The paired t-test (cont)

5. Look up the *critical value* in Table 7 for the desired significance level

If | test statistic | > critical value, we *reject* H_0 and the conclusion is that H_1 applies

If | test statistic | < critical value, we *cannot reject* H_0 and the conclusion is that there is no evidence of a significant difference between the two samples

21.7 The two-sample t-test

- This time, suppose that the samples are *not* paired; that is, they are *independent*
- The two samples need not contain the same number of observations
- The most common test statistic used in this situation is a *two-sample t-test* (also known as a *pooled t-test*)

21.7 The two-sample t-test (cont)

- The steps for using a two-sample *t*-test are as follows:

1. Construct the null and alternative (two-tailed) hypotheses.
   H_0 always states that there is *no difference* between the two samples, while H_1 always states that there is *some difference* between the two samples

$$H_0: \mu_1 = \mu_2 \qquad \text{v.} \qquad H_1: \mu_1 \neq \mu_2$$

2. Calculate the following statistics from the samples: the number of observations in Sample 1 and 2, the mean of Sample 1 and 2, the standard deviation of Sample 1 and 2

21.7 The two-sample t-test (cont)

3. Calculate the value of the *pooled standard deviation*

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

4. Calculate the value of the *two-sample t-test statistic*

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

21.7 The two-sample t-test (cont)

5. Look up the critical value in Table 7 for the desired significance level:

If | test statistic | > critical value, we *reject* H_0 and the conclusion is that H_1 applies. Hence, there is a significant difference between the two samples

If | test statistic | < critical value, we *cannot reject* H_0 and the conclusion is that there is no evidence of a significant difference between the two samples

21.8 p-values

- An alternative to using critical values for testing hypotheses is to use a *p*-value approach
- The value of the test statistic is still calculated in the usual way
- In this case we now calculate the probability of obtaining a value *as extreme* as the value of the test statistic if H_0 were true
- Compare the *p*-value with α. Then:
  - If *p*-value > α, we *cannot reject* H_0 at that α-level
  - If *p*-value < α, we *reject* H_0 at that α-level

Hypothesis Testing Summary

- We looked at and understood the principles of statistical inference
- We formulated null and alternative hypotheses
- We understood:
  - one-tailed and two-tailed tests
  - type I and type II errors
  - test statistics
  - the significance level of a test
  - the regions of acceptance and rejection

Hypothesis Testing Summary (cont)

- We understood and calculated critical values
- We calculated and interpreted
  - a one-sample *z*-test statistic
  - a one-sample *t*-test statistic
  - a paired *t*-test statistic
  - a two-sample *t*-test statistic
- We understood and calculated *p*-values