2.1 Sampling and Estimation

2.1.1 The need for sampling

In many cases it is not possible to obtain information about all Members of a population, for the following reasons:

(1) The collecting of the information may destroy the sample, e.g. testing fireworks or electric fuses.
(2) The population may be infinite, e.g. the measurements of a physical constant such as g using a particular apparatus.
(3) It may be impracticable to make a measurement for every member of the population, e.g. measuring the length of ants of a particular species.
(4) Even if a measurement could be made for each member of a population, considerations of time and expense usually dictate otherwise.

2.1.2 Random sampling

In order for a sample to be representative of the whole population each member of the population must have an equal chance of being chose. A sample chosen in this way is called a **random sample**.

The simplest method of selecting a random sample is by using a table of **random numbers** (see Table A3). Such tables are now normally compiled electronically but could be made using any device which gives the digits 0 to 9 with equal probability. Such a device is shown if Figure 12.1. It is a prism whose cross-section is a decagon and whose faces are labelled 0 to 9.

Suppose we wish to select two days at random from the month of August using a random number table. Each day is allocated a number 01, 02, 03, etc., up to 31. Note that each day must have the same number of digits in its number so that each number has an equal probability of being chosen. Any starting position can be chosen in Table A3. Suppose we obtain the pairs of digits: (46), (51), 06, (59), (60), 16. The numbers shown in brackets do not correspond to any members of our population and are rejected. Pairs of digits are taken until we have sufficient to give a sample of the required size. In this case the random sample will consist of August 6^{th} and August 16^{th}.

To reduce the amount of numbers which has to be rejected, we can allocate more than one number to each member of the population, on a cyclic basis, thus;

August 1st 01, 32, 64

August 2nd 02, 33, 65 etc.

Each member of the population must have the same number of numbers allocated to it. Using this method the first two random numbers we obtained before, i.e. 46 and 51, correspond to August 15th and August 20th.

2.1.3 Periodic sampling

This is a method of sampling where every nth member of the population is chosen. It is quicker and easier than using random numbers and might be appropriate for, say, selecting names from an electoral register. In some situations it is not suitable, since, for example, choosing every tenth item produced by machine might coincide with a periodicity of the machine.

2.1.4 Stratified random sampling

As its name implies, this method involves dividing the population into strata. A random sample is then selected from each stratum. The size of each sample is in proportion to the size of the stratum from which it is taken. The advantage of stratified random sampling is that the accuracy of the mean is greater than for a stratified sample of the same size. Sometimes, however the differences between the strata may be so great that calculation of a mean for the whole population may seem inappropriate.

2.1.5 Drawing a random sample from a discrete distribution

Random numbers can be used to simulate the drawing of a sample from a given distribution. Suppose we wish to choose a random sample of five from a Binomial distribution for which $p = 1/3$, $n = 3$.

The possible values of the random variable and the corresponding probabilities are given in Table 2.1. We can use random numbers to draw a sample if we assign numbers so that the probability of selecting $x = 0$ is 8/27 etc. these are shown in the third column of Table 2.1, using a cyclic method as before to use as many digits as possible. From the table of random numbers we obtain the pairs of digits.

(94), 68, 81, (97), 25, 39, 68

And the corresponding values of r are 1, 3, 2, 1, 1

Other discrete distributions can be treated similar way. Table 2.2 shows the method for a Poisson distribution with - = 0.3. In this case the probabilities have to be rounded off, here (arbitrarily) to four decimal places, and so the values of the variable > 5 have been

Table 2.1 *choosing a random sample from a Binomial distribution*

X	P(X = x)	Random digits
0	$(⅔)^3 = 8/27$	01 – 08, 28 – 35, 55 - 62
1	$3(⅔)^2(⅓) = 12/27$	09 – 20, 36 – 47, 63 - 74
2	$3(⅔)(⅓)^2 = 6/27$	21 – 26, 48 – 53, 75 - 80
3	$(⅓)^3 = 1/27$	27 53 81

Table 2.2 drawing *a random sample from a Poisson distribution*

X	P(X = x)	Cumulative	Random numbers
0	0.7408	0.7408	0000-7407
1	0.2222	0.9630	7408-9629
2	0.0333	0.9963	9630-9962
3	0.0033	0.9996	9963-9995
4	0.0003	0.9999	9996-9998
>5	0.0001	1.0000	9999

Grouped together. The random numbers start from 0000 so that they all have four digits. If from the table of random numbers we obtain, for example, the four digit number 7452, the corresponding value of the variable is 1.

2.1.6 Drawing a random sample from a continuous distribution

Suppose we wish to simulate drawing a random sample from a Normal distribution. In this case the variate is continuous and so can be selected with varying degrees of accuracy. The probability of getting a value of the standard deviate below a certain value is found from normal Table. For example the probability of a standard deviate z where $z < 0.65$ is 0.7422 and this probability could be represented by assigning it the random numbers 0000 to 7421. Similarly, the probability that $z < 0.64$ is 0.7389 and could be represented by the random numbers 0000 – 7388. Then the probability $0.64 < z < 0.65$ would be represented by the random numbers 7389 to 7421. This suggests that to select a random sample from a Normal distribution we can first select four-figure random numbers and then convert them to F(z), the area under the standardized Normal probability distribution, by putting a decimal point in front. From Table A1 the corresponding value of z is found and hence x, the value of the variable. This is indeed the method used apart from an important proviso. The four-figure random numbers are 0000 to 9999 and if we take the corresponding value of F(z) as 0.0000 to 0.9999, the values of z will not be symmetrically distributed about their mean, 0. To avoid this we add 0.00005 to each F(z) giving a range of 0.00005 to 0.999 95 which is symmetrically about the mean value of F(z), i.e. 0.5 and gives values of z symmetrical about 0.

Example 1

Select a random sample of four values of the variable from a Normal distribution mean 10, s.d. 2. (The measurements should be correct to 1 decimal place.)

The method is set out in Table 2.3.

Four 4-digit random numbers are taken from Table A3. They are converted to values of F(z) adding a decimal point and 0.00005. From Table A1 the range in which Z lies is found and the values of z are converted to values of x using $Z = (x - u)/o$ with $u = 10$, $0 = 2$. (if necessary the range of Z can be reduced by using interpolation in Table A.1) Correct to one decimal place the four randomly chosen values of X are 6.6, 12.8, 14.3 and 11.1.

Table 2.3 *Drawing a random sample from a Normal distribution*

Random numbers	F(x)	Range of Z	Range of Z
0452	0.04525	-1.70 <z < - 1.69	6.60 < x < 6.62
9197	0.91975	1.40 <z < 1.41	12.80 < x < 12.82
9847	0.98475	2.16<z< - 1.41	12.80 < x < 12.82

A similar method can be used for other continuous distributions. Random numbers are taken and converted to a value of the cumulative distribution function in the same way as they were for the Normal distribution. Then the corresponding value of the variable is found using the c.d.f. For example, a continuous Uniform distribution over the range 1 to 3 has probability density function

$F(x) = ½ \quad 1 < x < 3$ and cumulative distribution function

$F(x) = ½ \quad (x - 1)$

Using the random number 6432 gives 0.643 25 for the value of F(x) so that $25 = ½ (x - 1)$

Giving $x = 2{,}2865$

2.1.7 A Biological Example

A botanist wishes to investigate the flora in a field which measures 20 m square and proposes to do so by selecting at random five areas, each one metre square and studying these areas in detail. To select the areas the field is divided into a grid of 1 metre square, as shown in Figure 12.2. Each square is defined by a pair of two-figure numbers, e.g. the square in the top left-hand corner. As before numbers can be allocated on a cyclic basis so that,

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0																				
1																				
2																				
3																				
4																				
5																				
6																				
7																				

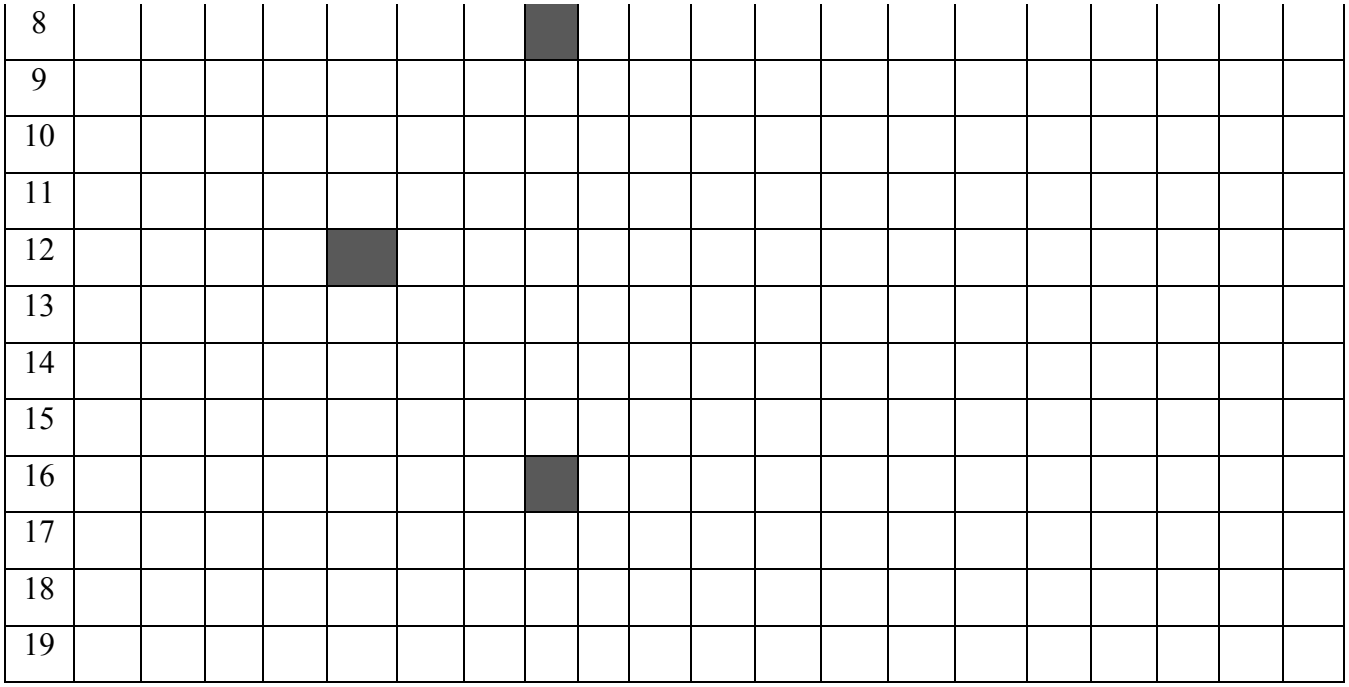


Figure 2.2. Diagram to illustrate Section 12.7

for example, 22, 42, 62, 82 all correspond 02. Using the random number table gives five sets of four digits: 67 28, 96 25, 68 36, 24 72, 03 85

The corresponding squares are 07 08, 16 05, 08 16, 04 12, 03 05

They are shown shaded in Figure 2.2.

2.1.8 Practical Sampling

In discussing and comparing sampling schemes the following criteria should be borne in mind:

(i) the randomness of the sample,

(ii) time,

(iii)cost,

(iv)Convenience to the person being questioned.

Consider, for example, some of the ways in which a survey might be made in a small town to find out whether parents of children under five consider the nursery school facilities satisfactory. A truly random sample is one in which all of the parents have an equal chance of being chosen. This could be achieved by selecting names from the electoral register using random numbers and interviewing those chosen. This has the disadvantages that (i) the time and expense involved in travelling to peoples' homes would be considerable, especially since more than one visit would be required if they were out, and (ii) a large number of those chosen would not have children under five and further names would have to be chosen to

replace them. The latter disadvantage could be overcome by selecting the sample from a list of parents with children under five, possibly obtainable from the Local Health Authority. The sampling could be stratified by dividing the parents into groups according to which district of the town they live in so that different income (and possibly ethnic) groups are fairly represented.

An attractive way of overcoming the first disadvantage mentioned above, i.e. the time and expense involved in traveling, would seem to be offered by selecting names at random from the telephone directory. This method is not satisfactory since (i) many people, generally those less well off, do not have a phone, and (ii) people are suspicious of being phoned by strangers.

One of the simplest ways of obtaining a sample would be to stop people with small children in the town centre. This method is quick and cheap but has the disadvantage that the sample is not necessarily random. In practice, however, a compromise may have to be made between obtaining a random sample and considerations of time, cost and convenience.

2.1.9Sampling distribution

2.1.9.1 Sampling

The main purpose of taking a sample is to obtain information about the parameters of the population from which the sample is drawn. For example, the mean of a sample gives us an estimate of the mean of the population.

If we took another sample from the population and calculated *its* mean we should be most unlikely to obtain the same value as we did for the first sample. In fact, if we continued taking samples and calculating their means, these means would have a frequency distribution of their own. If we consider all possible values that the mean can take when all possible random samples (of a given size) are drawn from the population then we can form the probability distribution of the sample mean. The random variable defined by this probability distribution, in this case the sample mean, is called an estimator. The probability distribution of an estimator is called its **sampling distribution.**

In this chapter we look at sampling distributions, in particular that of the mean, and the properties which an estimator needs to give a ‘good’ estimate of a population parameter.

Sampling distribution

2.1.9.2 Estimation Theory

Introduction

Estimation, in statistics, is any of numerous procedures used to calculate the value of some property of a population from observations of a sample drawn from the population. There are two types of estimates: point and interval. A point estimate is a value of a sample statistic that is used as a single estimate of a population parameter. An interval estimate defines a range within which the value of the property can be expected (with a specified degree of confidence) to fall. Interval estimates of population parameters are called confidence intervals

Some important statistics

Central Tendency in the sample

If $X_1, X_2, ..., X_n$ represents a random sample of size n, then the sample mean is defined to be

statistic: $\overline{X} = \frac{\sum_1 X_i n}{n}$,

- $\overline{X}$ is a statistic because it is a function of the random sample $X_1, X_2, ..., X_n$
- $\overline{X}$ has same unit of $X_1, X_2, ..., X_n$
- $\overline{X}$ measures the central tendency in the sample (location)

Variability in the Sample

Definition: If $X_1, X_2, ..., X_n$ represents a random sample of size n, then the sample variance is defined to be the statistic

$$s^2 = \frac{\sum_1^n (x-\bar{x})^2}{n-1} = \frac{\sum_1^n x^2 - n\bar{x}}{n-1} = \frac{n\sum x^2 - \left(\sum_1^n x\right)^2}{n(n-1)}$$

- s^2 is a statistic because it is a function of the random sample $X_1, X_2, ..., X_n$
- s^2 measures the variability in the sample

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{1}^{n}(x-\bar{x})^2}{n-1}}$$

2..9.3 Sampling Distribution

Each sample give a different estimate of the population mean. In order to find out how close an estimate will be to the population parameter. We discuss the sampling distribution of the mean and find out how it relates to the population distribution.

Example 2: Take a sample of size 2 drawn from a population { 1,2,3} with replacement. Estimate the population mean and variance?

Solution; We know $\mu = \sum_{i}^{n} xp(x)$ and $\sigma^2 = \sum_{i}^{n} X^2 P(X) - \left[\sum_{i}^{n} XP(X)\right]^2$

$$\mu = \frac{\sum_{i}^{n} X}{n} \quad \text{and } \sigma^2 = \frac{\sum_{i}^{n}(X-\bar{X})^2}{n} \quad \sigma^2 = \frac{\sum_{i}^{n} X^2}{n} - \left[\bar{X}\right]^2$$

Possible samples are: (1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), (3,3). The table below shows the expected values of X.

Samples	**1**	**2**	**3**
1	**1**	$3/2$	**2**
2	$3/2$	**2**	$5/2$
3	**2**	$5/2$	**3**

$$\mu = \sum_{i}^{n} XP(X) = 2$$

$$\mu_{\bar{X}} = \sum_{i}^{n} \bar{X}^2 P(\bar{X}) = 2$$

$$Var(\bar{X}) = \sum \bar{X}^2 P(\bar{X}) = \frac{1}{3}$$

Definition: The probability distribution of a statistic is called a sampling distribution.

Theorem: If μ all possible random samples of size n are drawn (with replacement) from a pop, with mean μ and s.d σ , Then the means of the samples have a probability distribution known as a sampling distribution of means, with means μ and s.d $\sigma/\sqrt{n}$. The standard deviation of the sampling distribution of the means is known as the standard error(s.e.) of the mean.

Proof

We know $E(\bar{X}) = \mu$ and $E(\bar{X})$ is the mean of the sampling distribution of the mean. The variance of the sample distribution is Var($\bar{X}$),

$$\mathrm{V}(\bar{X}) = Var\left(\frac{\sum_i^n X}{n}\right)$$

$$\begin{aligned}
&= \frac{1}{n^2} Var\left\{\sum_i^n X\right\} \\
&= \frac{1}{n^2}\left\{\mathrm{var}(X_1) + \mathrm{var}(X_2) + ... + \mathrm{var}(X_n)\right\} \\
&= \frac{1}{n^2} n\sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned} \qquad \text{(A)}$$

Therefore, the sd of the sampling distribution of the mean as $\frac{\sigma}{\sqrt{n}}$.

If the sampling is without replacement and the pop size is N then (A) becomes

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

If n is less than N , equation (A) holds. In this course we consider n less than N.

Note that whether the sampling is without replacement $Var(\bar{X})$ decrease as the sample size increases. The larger the sample size the closer the sample mean is likely to be the population mean.

Assignment 1:

1. A large population consist of equal numbers of the digits 1 and 3.

(a) Find the mean and variance of this population

(b) Find the probability distribution of the mean of samples size three taken from this population and verify that this mean mean is equal to the population mean and its variance is equal to one-third of the population variance.

2. The discrete random variable J has the distribution given below.

(a) Find the mean μ and variance σ^2 of the distribution. Random samples size two are taken from the distribution. By considering all possible samples, obtain the probability distribution of the mean of such samples. Verify that this distribution has mean μ and variance $\frac{1}{2}\sigma$.

(b) What would be the mean and variance of the distribution of the mean of random samples of size three fro the original distribution?

j	-2	-1	0	1	2
P(J = j)	0.1	0.2	0.4	0.2	0.1

2.1.9.4 Unbiased Estimator

Sample estimates are either biased or unbiased and our interest is study how we can show that this estimate is unbiased. We begin by showing that the sampling distribution 0f the mean is an unbiased estimate of the population mean. That is:

$E(X) = \mu$

Proof

Let $\overline{X} = \frac{\sum_i^n X_i}{n}$, hence,

$$E(\overline{X}) = E\left[\frac{\sum_i^n X_i}{n}\right] = \frac{1}{n}\sum_i^n E(X_i) = \frac{1}{n}\left[E(X_1) + E(X_1) + ... + E(X_n)\right]$$

We know that $E(X) = \mu$

$$E(\bar{X}) = \frac{1}{n}[\mu + \mu + ... + \mu]$$
$$= \frac{1}{n} n\mu \qquad 1$$
$$E(\bar{X}) = \mu$$

Therefore, the sample estimate is an unbiased estimate of the population mean.

However, we might expect the random variable $\sum_{i}^{n} \frac{(X_i - \mu)^2}{n}$ to be an unbiased estimator of σ^2 and this can be proved as follows by finding its expected value:

$$E\left[\sum_{1}^{n} \frac{(X_i - \mu)^2}{n}\right] = E\left[\frac{(X_1 - \mu)^2}{n} + \frac{(X_2 - \mu)}{n} + ... + \frac{(X_n - \mu)^2}{n}\right]$$
$$= E\left[\frac{(X_1 - \mu)^2}{n}\right] + E\left[\frac{(X_2 - \mu)^2}{n}\right] + ... + E\left[\frac{(X_n - \mu)^2}{n}\right] \qquad \text{(B)}$$
$$= \frac{1}{n}\sigma^2 + \frac{1}{n}\sigma^2 + ... + \frac{1}{n}\sigma^2$$
$$= \sigma^2$$

Unfortunately we are not usually in the position of requiring an estimate of σ when we know μ. In most cases we have only an estimate of μ ie $\bar{X}$. Using $\bar{X}$ we can calculate the s.d., s, of the sample, from the formula

$$S^2 = \sum_{1}^{n} \frac{(X_i - \bar{X})^2}{n}$$

However, we cannot use $S^2 = \sum_{1}^{n} \frac{(X_i - \bar{X})^2}{n}$ to give an unbiased estimate of variance, since the sum of the squares of the deviations of the $x_i's$ from $\bar{X}$ is less than the sum of the squares of the deviations from μ and consequently S^2 underestimate σ^2.

We can find an unbiased estimateor of σ^2 as follows;-

$$\sigma^2 = E\left\{\sum_i^n \frac{(X_i-\mu)^2}{n}\right\}$$
$$= E\left\{\sum_1^n \frac{\left[(X_i-\overline{X})-(\mu-\overline{X})\right]^2}{n}\right\}$$
$$= E\left\{\sum_1^n \frac{\left[(X_i-\overline{X})^2-2(X_i-\overline{X})(\mu-\overline{X})+(\mu-\overline{X})^2\right]}{n}\right\}$$
$$= E\left\{\sum_i^n \frac{(X_i-\overline{X})^2}{n}-2(\mu-\overline{X})\sum_1^n \frac{(X_i-\overline{X})}{n}\right\}+n.\frac{(\mu-\overline{X})^2}{n}$$

The second term is zero, since

$$\sum_1^n (X_i-\overline{X}) = \sum_1^n X_i - \sum_1^n \overline{X} = n\overline{X}-n\overline{X} = 0$$

So we have

$$\sigma^2 = E\left\{\sum_1^n \frac{(X_i-\overline{X})^2}{n}\right\}+E\left\{\mu-\overline{X})^2\right\} \qquad \text{(C)}$$

The third term is $E(S^2)$ where S^2 is the random variable $\sum_i^n \frac{(X_i-\overline{X})^2}{n}$. The second term, $E\left\{(\mu-\overline{X})^2\right\}$ or $E\left\{(\overline{X}-\mu)^2\right\}$, is $Var(\overline{X})$ which was shown in thr proof (B) above to be $\frac{\sigma^2}{n}$.

Thus

$$\sigma^2 = E(S^2)+\frac{\sigma^2}{n}$$

Rearranging,

$$\sigma^2 = \frac{n}{n-1}E(S^2) = E\left\{\frac{n}{n-1}S^2\right\} \qquad \text{(D)}$$

Substituting for S

$$\sigma^2 = E\left\{\left(\frac{n}{n-1}\right)\left(\sum_1^n \frac{(X_i-\overline{X})^2}{n}\right)\right\}$$

$$\sigma^2 = E\left\{\sum_1^n \frac{(X_i - \overline{X})^2}{n-1}\right\} \qquad \text{(E)}$$

Equation (E) gives us an unbiased estimator of variance, which we shal denote by $\hat{S}^2$, where

$$\hat{S}^2 = \sum_1^n \frac{(X_i - \overline{X})^2}{n-1} \qquad \text{(F)}$$

And $\hat{S}$ and S are related by

$$\hat{S} = \sqrt{\left(\frac{n}{n-1}\right)S} \qquad \text{(G)}$$

The term n – 1 in the denominator of equation (F) is referred to as the number of degrees of freedom and is given the symbol ν. The reason for this name is as follows: If we knew μ then the ariance could be calculated from n independent deviations, using $\sum_1^n \frac{(X_i - \overline{X})^2}{n}$.

If instead we measure the deviations from $\overline{X}$, then we have only n – 1 independent deviationa, as when (n – 1) deviations are given the last deviation can be deduced using the fact that $\sum_1^n (X_i - \overline{X}) = 0$. One degree of freedom has been lost since only n – 1 of the deviations can be varied independently.

Equation (G) shows that for large n, there is only a small error in takin S^2 rather than $\hat{S}^2$ as an estimator of σ^2.

Exercise: Five measurments of the volume of acid required in a ttitration are: 25.1, 25.2,, 25.2, 25.0, 25.5 cm^3 .Use these results to obtain estimates for the mean and s.d. of the volume of acid required.

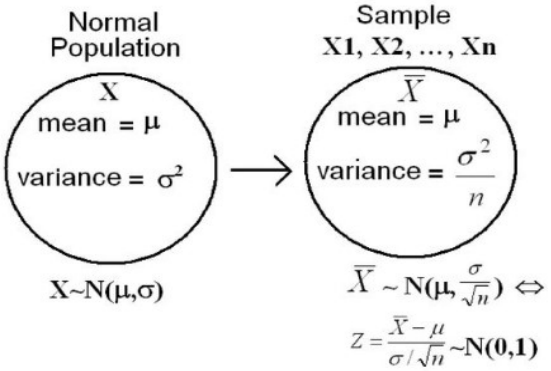
Example 3: If $X_1, X_2, ..., X_n$ represents a random sample of size n, then the probability distribution of $\overline{X}$ is called the sampling distribution of the sample $\overline{X}$.

If $X_1, X_2, ..., X_n$ is a random sample of size n taken from a normal distribution with mean μ and variance σ^2 ie $N(\mu, \sigma)$, then the sample mean $\overline{X}$ has a normal distribution with mean

$E(\overline{X}) = \mu_X = \mu$ and the variance $Var(\overline{X}) = \frac{\sigma^2}{n}$.

If $X_1, X_2, ..., X_n$ is a random sample of size n from $N(\mu, \sigma)$, the

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



Theorem: (Central Limit Theorem):

If $X_1, X_2, ..., X_n$ is a random sample of size n from any distribution(population) with mean μ and finite variance σ^2, then, if the sample size n is large, the random variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Is approximately standard normal variable ie,

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = N(0,1) \text{ approximately}$$

- $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(),1) \Leftrightarrow \overline{X} \approx N(\mu, \frac{\sigma}{\sqrt{n}})$
- We consider n large when n >=30
- For large sample size n , $\overline{X}$ has approximately a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, ie, $\overline{X} \approx N(\mu, \frac{\sigma}{\sqrt{n}})$ approximately.
- The sampling distribution of is used for inferences about the population mean μ.

Example 4

What is the probability that the mean of 100 digits taken from a random number table is greater than 5.0?

Solution. Random numbers form a discrete Uniform distribution.

By symmetry mean = 4.5

The variance for a Uniform distribution where the variable takes the values 1, 2, 3, . . . n was shown to be

$\frac{1}{12}(n^2 - 1)$. The numbers 0, 1, 2, . . . , 9 will have the same

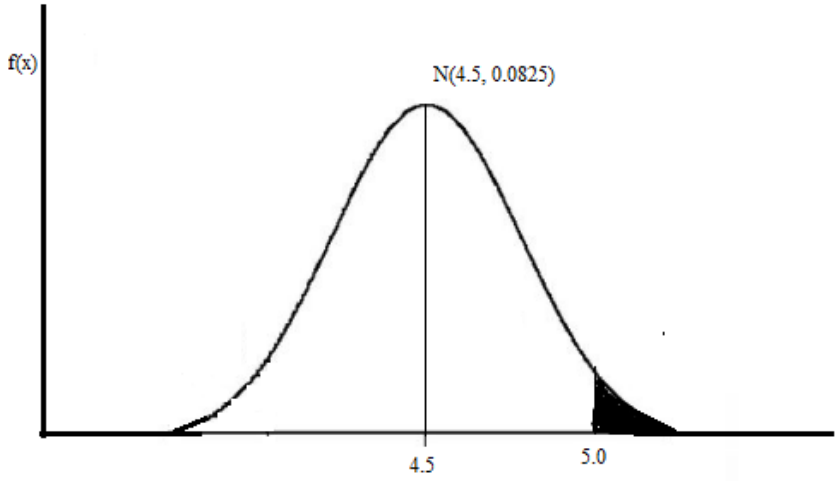


Figure : Graph to illustrate Example 3

Variance as 1, 2, 3, . . . , 10, since a change of origin does not alter the variance. Thus

Variance of random numbers = $\frac{1}{12}(10^2 - 1) = 8.25$

$= 8.25$

Using the relation $\sigma^2 = E(S^2) + \frac{\sigma^2}{n}$, the variance of the sampling distribution of the mean of 100 random digits is $\frac{\sigma^2}{n} = \frac{8.25}{100} = 0.0825$, and the mean of the sample distribution of the mean is 4.5. the central limit theorem tells us that the sampling distribution of the mean is Normal as shown in Figure above the s.d. of this distribution is $\sqrt{0.0825} = 0.287$. The probability that the mean of 100 members is greater than 5 is given by the shaded area.

$$z = \frac{5.0 - 4.5}{0.28} = 1.74$$

Required probability = P(Z > 1.74)

$= 1 - P(Z < 1.74)$

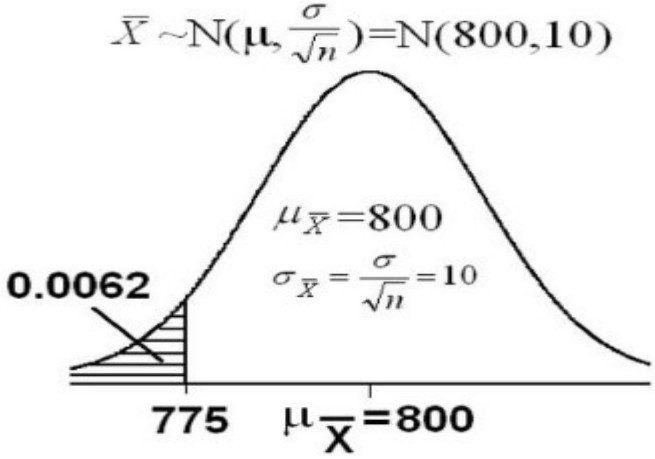
$= 1 - 0.9591$

$= 0.0409$

Example 5: An electric firm manufactures light bulbs that have a length of ife that is approximately normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Solution: Let X be the length of life of randomly selected bulbs. Given that $\mu = 800$ hours and $\sigma = 40$. That mean X ¬ $N(800,40)$, and the sample size n = 16 bulbs.

So we have $\mu_{\bar{X}} = \mu = 800$ hours.



$$\sigma_{\bar{\mu}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{16}} = 10$$

Therefore,

$$\bar{X} \approx N(\mu, \frac{\sigma}{\sqrt{n}}) = N(800,10) \Leftrightarrow Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} =$$

$$Z = p\left(\frac{\bar{X} - 800}{10} < \frac{775 - 800}{10}\right)$$

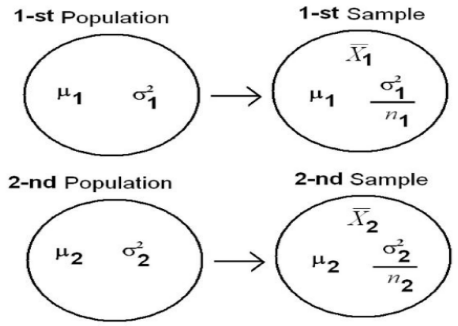
$$Z = p\left(Z < \frac{775 - 800}{10}\right)$$

$= p(Z < -2.50$

0.0062

2.1.9.5 Sampling Distribution of the difference between two means

Statistical analyses are very often concerned with the difference between means. A typical example is an experiment designed to compare the mean of a control group with the mean of an experimental group. Inferential statistics used in the analysis of this type of experiment depend on the sampling distribution of the difference between means.



The sampling distribution of the difference between means can be thought of as the distribution that would result if we repeated the following three steps over and over again:

(1) Sample n_1 scores from Population 1and n_2 scores from Population 2,

(2) Compute the means of the two samples (M1 and M2),

(3) Compute the difference between means M1 - M2. The distribution of the differences between means is the sampling distribution of the difference between means.

As you might expect, the mean of the sampling distribution of the mean is:

$$\mu_{M_1-M_2} = \mu_1 - \mu_2$$

Which says that the mean of the distribution of differences between sample means is equal to the difference between population means.

Suppose that we have two populations:

- 1st Population with μ_1 and variance σ^2_1
- 2nd Population with μ_2 and variance σ^2_2
- We are interested in comparing μ_1 and μ_2 or equivalently, making inferences about $\mu_1 - \mu_2$
- We independently select a random sample of size n_1 from the 1st Population and another random sample of size n_2 from the 2nd Population:
- Let $\overline{X}_1$ be the sample mean of the 1st sample,
- Let $\overline{X}_2$ be the sample mean of the 1st sample

The sampling distribution of $\mu_1 - \mu_2$ is used to make inference about $\mu_1 - \mu_2$.

Theorem: If n_1 and n_2 are large, then the sampling distribution of $\overline{X}_1 - \overline{X}_2$ is approximately normal with mean

$E(\overline{X}_1 - \overline{X}_2) = \mu_{X_1} - \mu_{X_2} = \mu_1 - \mu_2$ and the variance of

$Var(\overline{X}_1 - \overline{X}_2) = \sigma^2{}_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\varpi_2^2}{n_2}$, that is

$$\overline{X}_1 - \overline{X}_2 \approx N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$Z = \frac{(\overline{X}_1 = \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(),1)$$

Note that

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\sigma^2_{\overline{X}_1 - \overline{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \neq \sqrt{\frac{\sigma_1^2}{n_1}} + \sqrt{\frac{\sigma_2^2}{n_2}}$$

Example 6: The television picture tubes of manufacturer A have a mean lifetime of 6.5years and standard deviation of of 0.9 years, while those of manufacturer B have a mean lifetime of 6 years and standard deviation of 0.8 years. What is the probability that a random sample of 36 tubes from manufacturer A will have a mean lifetime that is at least 1 year more than the mean lifetime of a random sample of 49 tubes from manufacturer B?

Solution

Population A	**Population B**
$\mu_1 = 6.5$	$\mu_2 = 6$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

We need to find the probability that the mean lifetime of manufacturer A is at least 1 year more than the mean lifetime of manufacturer B which is $p(\overline{X}_1 > \overline{X}_2 + 1)$

The sample distribution of $\overline{X}_1 - \overline{X}_2$ is

$$\overline{X}_1 - \overline{X}_2 = N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

$$E(\vec{X}_1 - \overline{X}_2) = \mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2 = 6.5 - 6.0 = 0.5$$

$$Var(\overline{X}_1 - \overline{X}_2) = \sigma_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(0.9)^2}{36} + \frac{(0.8)2}{49} = 0.03556$$

$$\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n}} = \sqrt{0.03556} = 0.189$$

Therefore, $\overline{X}_1 - \overline{X}_2 \approx N(0.5, 0.189)$

Recall $Z = \dfrac{(\overline{X}_1 = \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \approx N(),1)$

$$= P(\overline{X}_1 \geq \overline{X}_2 + 1) = P(\overline{X}_1 - \overline{X}_2 \geq 1)$$

$$= P\left(\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \geq \frac{1 - (\mu_1 - \mu_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \right)$$

$$= P(Z \geq \frac{1 - 0.5}{0.189}$$

$$= P(Z \geq 2.65) = 1 - P(Z < 2.65) = 1 - 0.9960 = 0.0040$$

Example 7

Assume there are two species of green beings on Mars. The mean height of Species 1 is 32 while the mean height of Species 2 is 22. The variances of the two species are 60 and 70 respectively and the heights of both species are normally distributed. You randomly sample 10 members of Species 1 and 14 members of Species 2. What is the probability that the mean of the 10 members of Species 1 will exceed the mean of the 14 members of Species 2 by 5 or more? Without doing any calculations, you probably know that the probability is pretty high since the difference in population means is 10. But what exactly is the probability?

First, let's determine the sampling distribution of the difference between means. Using the formulas above, the mean is

$$\mu_{M_1-M_2} = 32 - 22 = 10$$

The standard error is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{60}{10} + \frac{70}{14}} = 3.317$$

The sampling distribution is shown in Figure 1. Notice that it is normally distributed with a mean of 10 and a standard deviation of 3.317. The area above 5 is shaded blue.

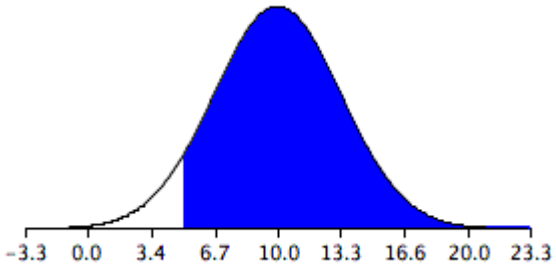


Figure 1

The last step is to determine the area that is shaded blue. Using either a Z table, the area can be determined to be 0.934. Thus the probability that the mean of the sample from Species 2 will exceed the mean of the sample from Species 1 by 5 or more.

As shown below, the formula for the standard error of the difference between means is much simpler if the sample sizes and the population variances are equal. Since the variances and samples sizes are the same, there is no need to use the subscripts 1 and 2 to differentiate these terms.

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

This simplified version of the formula can be used for the following problem:

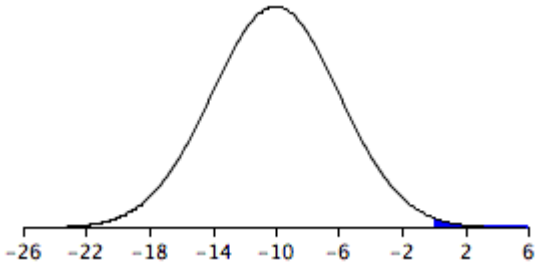
Example 8

The mean height of 15-year olds boys (in cm) is 175 and the variance is 64. For girls, the mean is 165 and the variance is 64. If eight boys and eight girls were sampled, what is the probability that the mean height of the sample of girls would be higher than the mean height of the boys? In other words, what is the probability that the mean height of girls minus the mean height of boys is greater than 0?

As before, the problem can be solved in terms of the sampling distribution of the difference between means (girls - boys). The mean of the distribution is 165 - 175 = -10. The standard deviation of the distribution is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{2\sigma^2}{n}} = \sqrt{\frac{(2)(64)}{8}} = 4$$

A graph of the distribution is shown in Figure 2. It is clear that it is unlikely that the mean height for girls would be higher than the mean height for boys since in the population boys are quite a bit taller. Nonetheless it is not inconceivable that the girls' mean could be higher than the boys' mean.



A difference between means of 0 or higher is a difference of 10/4 = 2.5 standard deviations above the mean of -10. The probability of a score 2.5 or more standard deviations above the mean is 0.0062.

Example 9

For boys, the average number of absences in the first grade is 15 with a standard deviation of 7; for girls, the average number of absences is 10 with a standard deviation of 6.

In a nationwide survey, suppose 100 boys and 50 girls are sampled. What is the probability that the male sample will have *at most* three more days of absences than the female sample?

Solution

The correct answer is B. The solution involves three or four steps, depending on whether you work directly with raw scores or z-scores. The "raw score" solution appears below:

- Find the mean difference (male absences minus female absences) in the population.

$$\mu_d = \mu_1 - \mu_2 = 15 - 10 = 5$$

- Find the standard deviation of the difference.

$$\sigma_d = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
$$= \sqrt{\frac{7^2}{100} + \frac{6^2}{50}} = 1.1$$

Find the probability. This problem requires us to find the probability that the average number of absences in the boy sample minus the average number of absences in the girl sample is less than 3. To find this probability, we use Normal Distribution table.

$$Z = \frac{3-5}{1.1} = -\frac{2}{1.1} = -1.818$$

$$P(Z < -1.818) = 0.034$$

Specifically, we enter the following inputs: 3, for the normal random variable; 5, for the mean; and 1.1, for the standard deviation. We find that the probability of the mean difference (male absences minus female absences) being 3 or less is about 0.035.

Thus, the probability that the difference between samples will be no more than 3 days is 0.035.

Example 10

Assume there are two species of green beings on Mars. The mean height of Species 1 is 32 while the mean height of Species 2 is 22. The variances of the two species are 60 and 70 respectively and the heights of both species are normally distributed. You randomly sample 10 members of Species 1 and 14 members of Species 2. What is the probability that the mean of the 10 members of Species 1 will exceed the mean of the 14 members of Species 2 by 5 or more? Without doing any calculations, you probably know that the probability is pretty high since the difference in population means is 10. But what exactly is the probability?

First, let's determine the sampling distribution of the difference between means. Using the formulas above, the mean is

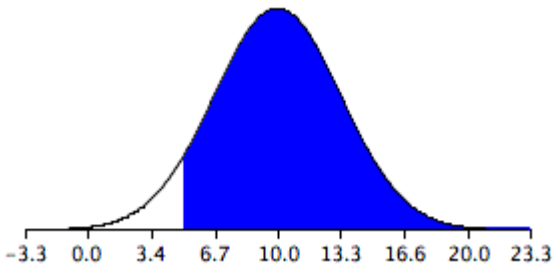
$$\mu_{M_1-M_2} = 32 - 22 = 10$$

The standard error is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{60}{10} + \frac{70}{14}} = 3.317$$

The sampling distribution is shown in Figure 1. Notice that it is normally distributed with a mean of 10 and a standard deviation of 3.317. The area above 5 is shaded blue.

Figure 1. The sampling distribution of the difference between means.



The last step is to determine the area that is shaded blue. Using either a Z table, the area can be determined to be 0.934. Thus the probability that the mean of the sample from Species 2 will exceed the mean of the sample from Species 1 by 5 or more.

As shown below, the formula for the standard error of the difference between means is much simpler if the sample sizes and the population variances are equal. Since the variances and samples sizes are the same, there is no need to use the subscripts 1 and 2 to differentiate these terms.

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$$

This simplified version of the formula can be used for the following problem:

$Z = \frac{5-10}{3.317} = -1.507$ Use the standard normal table to find P(Z>-1.507)

Example 11

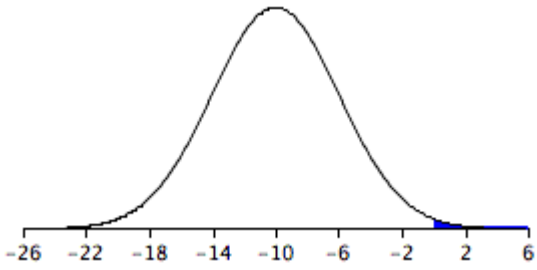
The mean height of 15-year olds boys (in cm) is 175 and the variance is 64. For girls, the mean is 165 and the variance is 64. If eight boys and eight girls were sampled, what is the probability that the mean height of the sample of girls would be higher than the mean height of the boys? In other words, what is the probability that the mean height of girls minus the mean height of boys is greater than 0?

As before, the problem can be solved in terms of the sampling distribution of the difference between means (girls - boys). The mean of the distribution is 165 - 175 = -10. The standard deviation of the distribution is:

$$\sigma_{M_1-M_2} = \sqrt{\frac{2\sigma^2}{n}} = \sqrt{\frac{(2)(64)}{8}} = 4$$

A graph of the distribution is shown in Figure 2. It is clear that it is unlikely that the mean height for girls would be higher than the mean height for boys since in the population boys are quite a bit taller. Nonetheless it is not inconceivable that the girls' mean could be higher than the boys' mean.

Figure 2. Sampling distribution of the difference between mean heights.



A difference between means of 0 or higher is a difference of 10/4 = 2.5 standard deviations above the mean of -10. The probability of a score 2.5 or more standard deviations above the mean is 0.0062.

2.1.9.6 Sampling Distribution of the Sample Proportion

The distribution of the sample proportion approximates a normal distribution under the following 3 conditions.

Over the years the values of the conditions have changed. The examples that follow in the remaining lessons will use the first set of conditions at 5. If any set of the two conditions listed below are satisfied, the sampling distribution of the sample proportion is...

- approximately normal
- with mean, $\mu = p$
- standard deviation [standard error], $\sigma = \sqrt{\frac{\hat{p}(1-p)}{n}}$

If the sampling distribution of $\hat{p}$ is approximately normal, we can convert a sample proportion to a z-score using the following formula:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

We can apply this theory to find probabilities involving sample proportions.

Example. 12

Suppose it is known that 43% of Zambian own an iPhone. If a random sample of 50 Zambians were surveyed,

(a) What is the probability that the proportion of the sample who owned an iPhone is between 45% and 50%?

Solution

For this problem, we know p = 0.43, n = 50. First, we should check our conditions for the sampling distribution of the sample proportion.

np = 50(0.43) = 21.5 and n(1-p) = 50(1-0.43) = 28.5 – both are greater than 5.

Since the conditions are satisfied, $\hat{p}$ will have a sampling distribution that is approximately

Normal with mean $\mu = 0.43$ and standard deviation (standard error) $\sqrt{\frac{0.43(1-0.43)}{50}} \approx 0.07$

$$P(0.45 < \hat{P} < 0.5) = P(\frac{0.45-0.43}{0.07} < \frac{\hat{P}-P}{\sqrt{\frac{\hat{P}(1-P)}{n}}} < \frac{0.5-0.43}{0.07}$$

$$= P(0.286 < Z < 1$$
$$= P(Z < 1) - P(Z < 0.286)$$
$$= 0.8413 - 0.6126$$
$$= 0.2287$$

Therefore, if the true proportion of Americans who own an iPhone is 43%, then there would be a 22.87% chance that we would see a sample proportion between 45% and 50% when the sample size is 50.

(b) If a random sample of size of seventy five was surveyed, what is the probability we would find more than 50% of Zambians with an iPhone?

Solution: First, check our conditions: $np = 75(0.43)$ and $n(1-n) = 75(1-0.43)$ are both greater than five. The sampling distribution of the sample proportion is approximately

Normal with Mean, Standard deviation $\sqrt{\frac{\hat{P}(1-P)}{n}} = \sqrt{\frac{0.43(1-0.43)}{75}} \approx 0.05717$

$$P(\hat{P} > 0.5) = \left(\frac{\hat{P}}{\sqrt{\frac{\hat{P}(1-P)}{n}}} > \frac{0.5 - 0.43}{\sqrt{\frac{0.43(1-0.43)}{75}}} \right)$$

$$\approx P(Z > 1.22)$$

$$= 1 - P(X < 1.22)$$

$$= 1 - 0.8888$$

$$= 0.1112$$

Therefore, there is a 11.1% chance to get a sample proportion of 50% or higher in a sample size of 75

(c) Suppose it is known that 43% of Zambian own an iPhone. If a random sample of 50 Zambians were surveyed,

What is the probability that the proportion of the sample who owned an iPhone is between 45% and 50%?

Solution

For this problem, we know p = 0.43, n = 50. First, we should check our conditions for the sampling distribution of the sample proportion.

np = 50(0.43) = 21.5 and n(1-p) = 50(1-0.43) = 28.5 – both are greater than 5.

Since the conditions are satisfied, $\hat{p}$ will have a sampling distribution that is approximately

Normal with mean $\mu = 0.43$ and standard deviation (standard error) $\sqrt{\frac{0.43(1-0.43)}{50}} \approx 0.07$

$$P(0.45 < \hat{P} < 0.5) = P(\frac{0.45 - 0.43}{0.07} < \frac{\hat{P} - P}{\sqrt{\frac{\hat{P}(1-P)}{n}}} < \frac{0.5 - 0.43}{0.07}$$

$$= P(0.286 < Z < 1$$

$$= P(Z < 1) - P(Z < 0.286)$$

$$= 0.8413 - 0.6126$$

$$= 0.2287$$

Therefore, if the true proportion of Americans who own an iPhone is 43%, then there would be a 22.87% chance that we would see a sample proportion between 45% and 50% when the sample size is 50.

(c) If a random sample of size of seventy five was surveyed, what is the probability we would find more than 50% of Zambians with an iPhone?

Solution: First, check our conditions: $np = 75(0.43)$ and $n(1-n) = 75(1-0.43)$ are both greater than five. The sampling distribution of the sample proportion is approximately Normal with Mean, Standard deviation $\sqrt{\frac{\hat{P}(1-P)}{n}} = \sqrt{\frac{0.43(1-0.43)}{75}} \approx 0.05717$

$$P(\hat{P} > 0.5) = \left(\frac{\hat{P}}{\sqrt{\frac{\hat{P}(1-P)}{n}}} > \frac{0.5-0.43}{\sqrt{\frac{0.43(1-0.43)}{75}}} \right)$$

$$\approx P(Z > 1.22)$$
$$= 1 - P(X < 1.22)$$
$$= 1 - 0.8888$$
$$= 0.1112$$

Therefore, there is a 11.1% chance to get a sample proportion of 50% or higher in a sample size of 75

2.1.9.7 t- Distribution:

Recall that , If $X_1, X_2, ..., X_n$ is a random sample of size n taken from a normal distribution with mean μ and variance σ^2 ie $N(\mu, \sigma)$, then

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} = N(0,1)$$

We can only apply this result only when σ^2 is known. If σ^2 is not known, we replace the population variance σ^2 with the sample variance $S^2 = \frac{\sum (X_i - \overline{X})^2}{n-1}$, to have the following statistic

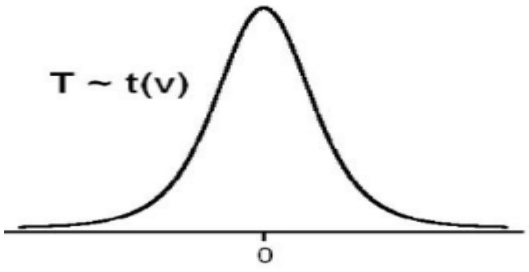
$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

If $X_1, X_2, ..., X_n$ is a random sample of size n taken from a normal distribution with mean μ and variance σ^2 ie $N(\mu, \sigma)$, then

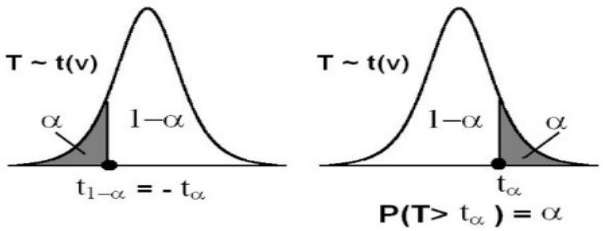
$T = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ has a t- distribution with v = n – 1 degrees of freedom (df), and we write $T \approx t_v$

Note

- t-distribution is a continuous distribution
- The shape of t- distribution is similar to the shape of the standard normal distribution
- We use the t- distribution for small sample size.



Notation:



- t_α = The t-value above which we find an area equal to α, that is $P(T > t_\alpha) = \alpha$
- Since the curve of the pdf of T~ t(ν) is symmetric about 0, we have

Values of t_v are tabulated in the t- distribution table.

Summary

In this Lecture, we learned how to use the Central Limit Theorem to find the sampling distribution for the sample mean and the sample proportion under certain conditions.

We learned that the sampling distributions are centered on the population parameter with variability. All of this theory was built knowing the parameter. Can we use this information in situations where the parameter is unknown? We take our first step into inference and into the "real world" where the population parameters are unknown and need to be estimated.

Confidence Intervals for Variance and Standard Deviation

Introduction

We have learned that estimates of population means can be made from sample means, and confidence intervals can be constructed to better describe those estimates. Similarly, we can estimate a population standard deviation from a sample standard deviation, and when the original population is normally distributed, we can construct confidence intervals of the standard deviation as well. Variances and standard deviations are a very different type of measure than an average, so we can expect some major differences in the way estimates are made.

We know that the population variance formula, when used on a sample, does not give an unbiased estimate of the population variance. In fact, it tends to underestimate the actual population variance. For that reason, there are two formulas for variance, one for a population and one for a sample. The sample variance formula is an unbiased estimator of the population variance. (Unfortunately, the sample standard deviation is still a biased estimator.)

Also, both variance and standard deviation are nonnegative numbers. Since neither can take on a negative value, the domain of the probability distribution for either one is not $(-\infty,\infty)$, thus the normal distribution cannot be the distribution of a variance or a standard deviation. The correct PDF must have a domain of $[0,\infty)$. It can be shown that if the original population of data is normally distributed, then the expression $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $n-1$ degrees of freedom.

The chi-square distribution of the quantity $\frac{(n-1)S^2}{\sigma^2}$ allows us to construct confidence intervals for the variance and the standard deviation (when the original population of data is normally distributed).

For a confidence level $1-\alpha$, we will have the inequality $\frac{(n-1)S^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_L^2}$.
Solving this inequality for the population variance σ^2, and then the population standard deviation σ, leads us to the following pair of confidence intervals.

$$\sqrt{\frac{(n-1)S^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi_L^2}}$$

It is worth noting that since the chi-square distribution is not symmetric, we will be obtaining confidence intervals that are not symmetric about the point estimate.

Since variance and standard deviation are statistical quantities of a set of data, can we construct confidence intervals for these quantities?

When products that fit together (such as pipes) are manufactured, it is important to keep the variations of the diameters of the products as small as possible; otherwise, they will not fit together properly and will have to be scrapped. In the manufacture of medicines, the variance and standard deviation of the medication in the pills play an important role in making sure patients receive the proper dosage. For these reasons, confidence intervals for variances and standard deviations are necessary.

Chi-Square Distributions The chi-square distribution must be used to calculate confidence intervals for variances and standard deviations. The chi-square variable is similar to the t variable in that its distribution is a family of curves based on the number of degrees of freedom. The symbol for chi-square is (Greek letter chi, pronounced "ki"). A chi-square variable cannot be negative, and the distributions are skewed to the right.

Chi-Square Distributions. At about 100 degrees of freedom, the chi-square distribution becomes somewhat symmetric. The area under each chi-square distribution is equal to 1.00, or 100%.

The Chi-square Distribution

The point estimate for σ^2 is S^2

The point estimate for σ is S

Where σ^2 and σ are estimate population parameters and S^2 and S are sample statistics.

You can use the Chi-square distribution to construct a confidence interval for the variance and standard deviation. If the random variable x has a normal distribution then the distribution of

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

Form a chi-square distribution for samples of any size > 1

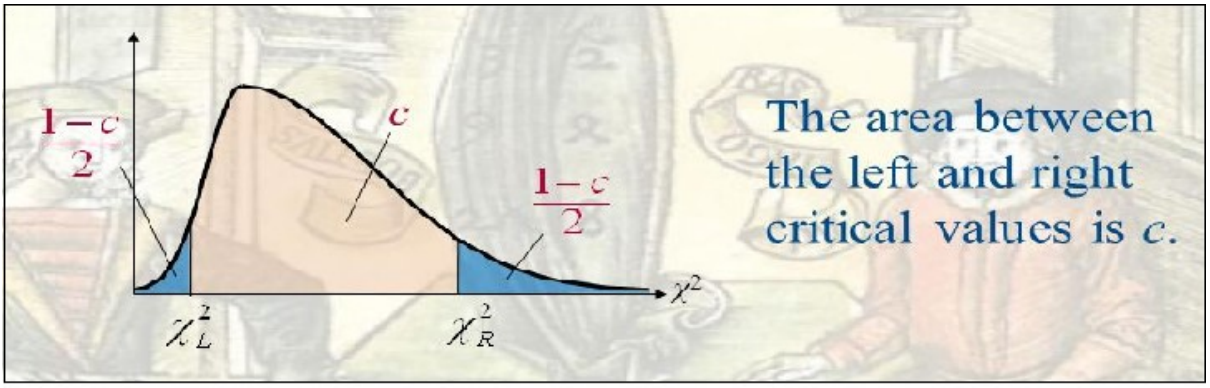
Properties of the chi-square distribution

1. All chi-square values χ^2 are greater than or equal to zero
2. The chi-square distribution is a family of curves, each determined by degrees of freedom. To form a confidence interval for, σ^2 use χ^2 distribution with degrees of freedom equal to one less than the sample size.

   $d.f. = n - 1$ Degrees of freedom
3. The area under each curve of the chi square distribution equals one.

Critical Values for χ^2

- There are two critical values for each level of confidence
- The value χ^2_R represents the right tail critical value
- The value χ^2_L represents the left tail critical value



Example 13: Find the critical values χ^2_R and χ^2_L for a 90% confidence interval when the sample size is 20.

Solution

d.f = n-1 = 20 – 1=19

Each area in the table represents the region under the chi square curve to the right of the critical value

Area to the right of $\chi^2_R = \frac{1-c}{2} = \frac{1-0.90}{2} = 0.05$

Area to the left of $\chi^2_L = \frac{1+c}{2} = \frac{1+0.90}{2} = 0.95$

Solution: Finding critical values for χ^2_R

Table 6: χ^2-Distribution

Degrees of freedom	α									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997

$\chi^2_R = 30.144$ $\chi^2_L = 10.117$

90% of the area under the curve lies between 10.117 and 30.144

Confidence intervals for σ^2 and σ

Confidence interval for σ^2

$$\frac{(n-1)S^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_L^2}$$

Confidence interval for σ

$$\sqrt{\frac{(n-1)S^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi_L^2}}$$

The probability that the confidence intervals contain σ^2 or σ is c.

Confidence intervals for σ^2 and σ

In words	In symbols
Verify that the pop has a normal distribution	
Identify the sample statistics n and the df	d.f. =n - 1
Find the point estimate S^2	$S^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
Find the critical values $\chi^2{}_R$ and $\chi^2{}_L$ that correspond to the given level of confidence c	Use chi square table
Find the left and right endpoints and form the confidence for the pop variance	$\frac{(n-1)S^2}{\chi_R^2} < \sigma \frac{(n-1)S^2}{\chi_L^2}$
Find the left and right endpoints and form the confidence for the pop standard deviation	$\sqrt{\frac{(n-1)S^2}{\chi_R^2}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi_L^2}}$

Example 14: You randomly select and weigh 30 samples of an allergy medicine. The sample standard deviation is 1.20 mg. assuming the weights are normally distributed. Construct 99% CI for the pop variance and SD.

Solution: df = n – 1 = 30 -1 29

Area to the right of $\chi^2_R = \frac{1-c}{2} = \frac{1-0.99}{2} = 0.005$

Area to the left of $\chi^2_L = \frac{1+c}{2} = \frac{1+0.99}{2} = 0.995$

The critical values are: $\chi^2_R = 52.336$ and $\chi^2_L = 13.121$

Solution: Constructing a CI

Left end point : $\frac{(n-1)S^2}{\chi^2_R} = \frac{(30-1)(1.20)^2}{52.336} = 0.80$

Left end point : $\frac{(n-1)S^2}{\chi^2_L} = \frac{(30-1)(1.20)^2}{13.121} = 3.18$

$0.80 < \sigma^2 < 3.18$

With 99% confidence you can say that the population variance is between 0.80 and 3.18mg

Constructing a CI for σ :

$$\sqrt{\frac{(n-1)S^2}{\chi^2_R}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_L}}$$

$$\sqrt{\frac{(30-1)(1.20)^2}{52.336}} < \sigma < \sqrt{\frac{(30-1)(1.20)^2}{13.121}}$$

$$0.89 < \sigma < 1.78$$

With 99% confidence you can say that the population variance is between 0.89 and 1.78mg

Example 15: A statistician chooses 27 randomly selected dates, and when examining the occupancy records of a particular motel for those dates, finds a standard deviation of 5.86 rooms rented. If the number of rooms rented is normally distributed, find the 95% confidence interval for the population standard deviation of the number of rooms rented.

Solution: For a sample size of n=27, we will have df=n−1=26 degrees of freedom. For a 95% confidence interval, we have α=0.05, which gives 2.5% of the area at each end of the chi-square distribution. We find values of $\chi_{0.975}2 = 13.844$ and $\chi^2_{0.025} = 41.923$.

Evaluating $\frac{(n-1)S^2}{\chi^2}$, we obtain 21.297 and 64.492. This leads to the inequalities

$21.297 \leq \sigma^2 \leq 64.492$ for the variance, and taking square roots, $4.615 \leq \sigma \leq 8.031$ for the standard deviation.