SIAST Palliser Campus

Mathematics
STAT 201

Lecture Notes and Examples

Unit 3
Inferential Statistics

by Blake Friesen and Robert G. Petry

published by the Department of Mathematics, SIAST Palliser Campus

All numerical data in this document should be considered fabricated unless a source is cited directly. Comments in the transparent copy of the document may contain references indicating inspiration behind some of the data.

History

- Original document produced in 2011 entitled "Inferential Statistics" written by principal authors Blake Friesen and Robert G. Petry with contributions from Mike Decorby. Published by the Department of Mathematics, SIAST Palliser Campus. A transparent copy of this document is available via *http://www.campioncollege.ca/about-us/faculty-listing/dr-robert-petry*

Contents

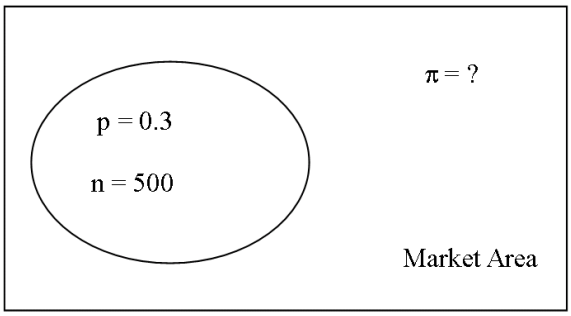
1 Sampling Distributions

1.1 Introduction

The object of statistical sampling is to estimate the value of a population parameter. Sample statistics provide the information for making the estimates. To be a useful estimate of the parameter, each estimate must satisfy two conditions:

1. It must provide a degree of **precision**. (How close do we think we are?)
2. It must provide a level of **confidence**. (What is the probability we are that close?)

Example:
A market analyst examined a sample of 500 households in a market area. The research shows that 30% of the households surveyed plan to buy a new vehicle in the next four years. The analyst has concluded that if this percentage were projected to the whole market area as an estimate of the proportion of households in the market area planning to buy a new vehicle in the next four years, this value would be within 4% of the correct value in 95 samples out of 100.



In this example the degree of precision is 4%; the level of confidence is 95%.

Sampling always produces estimates. It is unlikely that the sample value obtained exactly equals the population value to be estimated. Why are samples used instead of gathering all of the population data?

The term **sampling error** is used to represent the difference between the population parameter and the observed sample statistic.

Example:
In the previous market study on buying intentions of homeowners, suppose that 35% of the households in the market area actually intend to buy a new vehicle. Our sample results based on 500 households showed a proportion of 30%. In this case the sampling error is:

$$E = p - \pi = 0.30 - 0.35 = -.05 = -5\% \ .$$

This means that we are 5% too low in our estimate.

The sample has to be selected in such a way that the sample statistic calculated is not **biased** with respect to the population parameter.

There are many different ways of drawing a sample from a population. Some of these methods are referred to as probability sampling and some are referred to as nonprobability samples. Each of these methods of sampling has specific uses. Probability methods of drawing samples ensure that the estimate of the parameter used is unbiased.

A **probability sample** is one in which each member of the population has a nonzero probability of being included in the sample.

A **simple random sample** is a probability sample in which each item in the population has the same probability of being included in the sample.

A procedure like drawing names from a hat guarantees a simple random sample. Before selections can be made like this, each item in the population must be identified in order to give it a chance of being selected in the sample. Rather than draw items from a hat, statisticians simulate the process by constructing a **sampling frame** and using a table of random numbers to draw from the sample frame. A small set of random[1] digits is given below:

56599	62463	25114	61055	45618	73993	60743
85197	30682	77780	08002	57545	96111	23842
58835	10840	63210	56254	73053	09915	32766
13532	75531	83167	22578	12146	51981	73807
64682	84233	72523	21601	57214	52660	19001
12013	26749	84512	64112	28201	27741	84974
44599	13230	91202	95529	98096	05285	47421
35584	27278	03848	34905	85168	65804	68606
13434	46602	48712	58533	53769	48494	87451
47749	64307	69180	85351	02772	97869	93451
32188	88035	98488	61690	70573	78592	68315
93405	81747	44672	91838	87334	35692	47561
40334	93806	58107	33344	45968	82226	11441
79081	35905	94398	94027	40312	91620	41141
72754	96584	17577	34772	53925	52740	97393

Example:
Construct a sampling frame of a class of 35 students and use the table of random numbers to draw a simple random sample of size 5.

Solution:
Construct a sampling frame for the class by assigning each class member a two digit number starting at 01 and sequentially ending with the last class member, student 35. Read down the first two digits of the first column selecting the first five numbers that appear in sequence that are in the range from 01 to 35 and rejecting those numbers outside of the range.

[1]Most computer languages include a random number generator so this procedure can be done by machine. Modern calculators also have a random number button to produce such numbers. Where do computers and calculators get such random numbers? Computers are deterministic so there are algorithms which take a *seed* number which you provide which they then manipulate repeatedly by a function to produce pseudo-random numbers, namely digits that appear random and hopefully are equally likely. Additionally computers will often keep an *entropy pool* which contains random numbers linked to random events such as when you last turned on your computer, or touched your mouse and keyboard. More recently yet, some computer chip designers have been able to make quantum random number generators that exploit the inherent probabilities underpinning reality to produce true random numbers.

In many applied sampling situations, all the items in the population are not available. Only the sample observations are known.

Example:
A sample of 10 families from a new large suburb are surveyed regarding the balance owing on their house mortgages. The following balances were observed: ($)

145,000	280,000	282,000	290,000	350,000
358,000	402,000	466,000	664,000	714,000

1. What is the estimate of the population mean?
2. If another sample of size 10 were drawn, do you think that the same sample estimate of μ would occur?
3. What is an estimate of the population standard deviation?
4. If another sample of size 10 were drawn, do you think that the same sample estimate of σ would occur?
5. Draw a sketch of the distribution of balances in the community based on a normal curve assuming the estimates of the parameters are accurate.

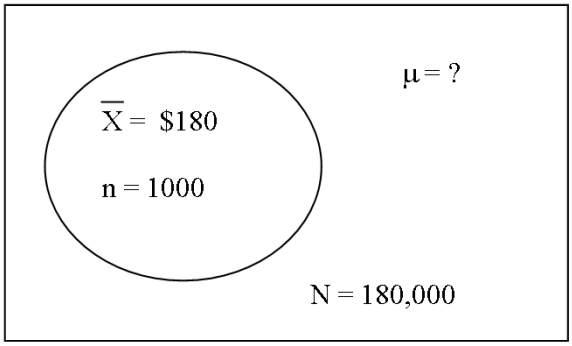
6. Without the assumption of normality is the sketch accurate?
7. Where is a balance of $450,000 located in the distribution?
8. Is there a lot or a little variation in balances?
9. What is the sampling error in the estimate of μ if we know the population mean, μ, is $396,000?

10. Why is it not always possible when sampling to find the exact sampling error in a sampling procedure as was done in the last question?

1.2 The Central Limit Theorem

In an applied sampling problem, usually one sample is drawn for purposes of determining the population parameter. In theory, it is possible to draw many more samples.

Example:
In Saskatchewan there are about 180,000 residents in the age group between 15 and 25 years of age. A market analyst would like to make projections about this group based on a sample of size 1000 individuals. Imagine that the parameter of interest is related to the sporting industry and concerns the average expenditure per individual on sporting equipment per year.



A list of 1000 expenditures is drawn without replacement from 180,000 expenditures. To have 100% confidence that we have the exact value of μ, all 180,000 values would have to be identified. The value that we observe in the sample depends on which values are drawn in the sample. If we sample without replacement, there are ${}_{180,000}C_{1000}$ different samples that could be obtained each with its own mean. If we sample with replacement then there are $180,000^{1000}$ possible samples. The size of this combination or power is too large for purposes of visualizing all the different possible sample possibilities. The sampling error in our estimate depends on which sample we happen to select.

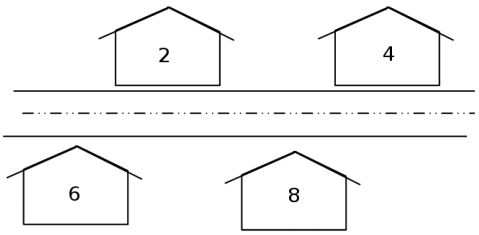
The previous example illustrates the following points which will be true when we try to estimate any population parameter with a sample statistic:

- There are usually a large number of samples that are possible (${}_NC_n$ without replacement, N^n with replacement).
- Each sample will generate a sample statistic (e.g. $\bar{X}$) which will estimate the population parameter (e.g. μ).
- Each such sample statistic will be in error; it will differ from the population parameter by its sampling error E. The sample statistics in general do not equal each other and therefore neither will their errors.

These points make it clear that for an individual sample mean, $\bar{X}$, to be useful one has to have some idea how close it is to the true population mean, μ. To analyze the problem we consider a **sampling distribution**. Suppose we took all possible samples of size n from the population and measured the sample statistic (e.g. $\bar{X}$). The frequency of these mean values would themselves form a distribution.

The above example is too large to study the sampling distribution of the mean so let us consider a simple example where the total number of possible samples is manageable. This situation is too small to actually use sampling techniques to analyze but we will use this example because it is small enough to visualize all of the possible samples that can occur.

Example:
Suppose a social scientist drives through a small village consisting of 4 households. Each household has a different number of occupants specified by the number inside the graphic below.[2] Imagine this to be a sampling situation where the scientist can only observe some of the households. The scientist is interested in the mean number of occupants per household in the village. Suppose 2 households are selected.



How many samples of size 2 can the analyst draw from this village? This depends upon how the sampling is done.

- If sampling is done **without replacement**, the answer is ${}_4C_2$. Confirm that ${}_4C_2 = 6$.
- If sampling is done **with replacement**, the answer is $4^2 = 16$.

In this type of problem sampling would likely be done without replacement but in problems where the sample size is a small fraction of the population the calculations can be done assuming sampling is done with replacement and the population looks the same from one draw to the next. For our purposes we will sample with replacement.[3]

The following table shows the 16 possible samples based on sampling with replacement. (Place the mean to the right of the sample for each possibility).

Sample	Sample Mean, $\bar{X}$	$\bar{X}^2$
(2,2)		
(2,4)		
(2,6)		
(2,8)		
(4,2)		
(4,4)		
(4,6)		
(4,8)		
(6,2)		
(6,4)		
(6,6)		
(6,8)		
(8,2)		
(8,4)		
(8,6)		
(8,8)		
	$\sum \bar{X} =$	$\sum \bar{X}^2 =$

The above table shows that when drawing random samples, the sample mean is itself a variable.[4] This collection of observations for the sample mean is called the **sampling distribution of the mean**.

Questions on this Sampling Distribution Example

1. On your calculator, calculate the population mean and standard deviation of the measured values (2, 4, 6, and 8).

   $\mu =$ $\sigma =$

2. Calculate the mean of the sample mean column.

   $\mu_{\bar{X}} = \dfrac{\sum \bar{X}}{16} =$

3. How does the mean of all sample means, $\mu_{\bar{X}}$ compare to the mean of the X values in the population, μ?

4. If the third sample is the one randomly chosen, what is the sampling error in estimating μ?

   $E = \bar{X} - \mu =$

5. Calculate the standard deviation of the sample mean column.

   $\sigma_{\bar{X}} = \sqrt{\dfrac{\sum \bar{X}^2}{16} - \left(\dfrac{\sum \bar{X}}{16}\right)^2} =$

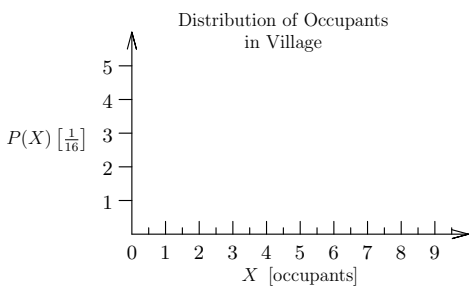
6. How does the standard deviation of the $\bar{X}$ values, $\sigma_{\bar{X}}$, compare to the standard deviation of the X values in the population, σ?

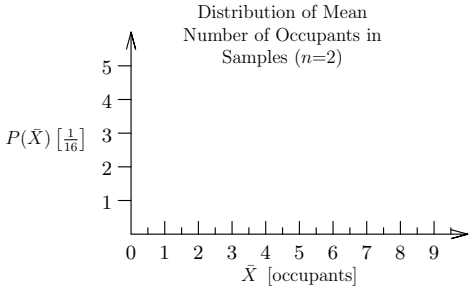
7. Complete the following tables for the probability distribution of X and $\bar{X}$.

X	P(X)
2	/4
4	/4
6	/4
8	/4
	$\sum P(X) =$

$\bar{X}$	P($\bar{X}$)
2	/16
3	/16
4	/16
5	/16
6	/16
7	/16
8	/16
	$\sum P(\bar{X}) =$

8. Now make histograms for these two probability distributions. Add a frequency polygon to your distribution of $P(\bar{X})$.





9. What is the shape of the distribution of sample means?

[2] In this example having a different number of occupants in each household will make our visualization of the samples easier, but it is unnecessary. Try to think how you would approach the analysis if two of the houses had the same number of occupants.

[3] See problem 5 on page 13 for an analysis of the problem sampling without replacement.

The behaviour of the distribution of the mean is summarized by the **Central Limit Theorem (C.L.T.)**. It states the following:

1. The **shape** of the distribution of sample means approaches a **normal curve**, as the size of the sample grows larger when sampling from a large population. For sample sizes of $\boxed{n \geq 30}$ the distribution of $\bar{X}$ will be approximately normal.

   ** Look at the shape of the frequency polygon in your diagram on the right, it is symmetrical but it does not have enough points to produce a smooth bell shape. If our sample size had been 30 or more rather than 2 the shape would be very close to a bell curve. **

2. The mean of the sampling distribution is the same as that of the population.

$$\mu_{\bar{X}} = \frac{\sum \bar{X}}{{}_N C_n \text{ or } N^n} = \mu$$

   ** Examine your two graphs. Their balance points are identical. **

3. The standard deviation of the sampling distribution, called the **standard error of the mean**, is less than the standard deviation in the population. If we are sampling from a distribution that is very large compared to the size of the sample ($n/N < .05$) or we are sampling with replacement as above, the standard error is found by:

$$\boxed{\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}}$$

   ** Verify this holds for the standard deviations you found above. **

As n becomes large relative to N, the standard error becomes small due to the $\sqrt{n}$ in the denominator which also gets larger. A small standard error means that little variation can be expected from one sample mean to the next. Any sample mean we observe for purposes of estimating μ is expected to have little sampling error and so is a good estimate.

Example:
Suppose a population variable has a standard deviation σ of 12 cm and a mean μ of 115 cm. Random samples are drawn from this population for the purpose of analyzing the expected variability in sample means from 115 cm. Compute the standard error of the mean, $\sigma_{\bar{X}}$, if samples of size 50, 100, 500, and 5000 are drawn from the population. Assume we are sampling without replacement but that the population is large with respect to the sample size, $n/N < .05$.

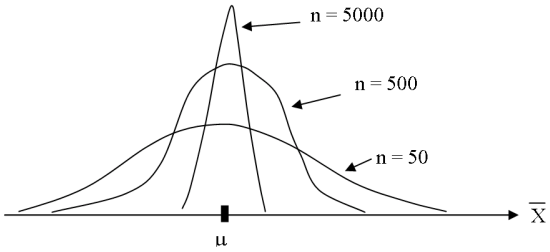
Solution:
Since all of our sample sizes are 30 or more, the C.L.T. states that the samples will have means that are normally distributed. The fact that

$$\frac{n}{N} = \frac{n}{\text{“large”}} \approx 0 < .05$$

means we can use the formula above for the standard error of the mean.

n	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (cm)
50	$\frac{12}{\sqrt{50}} = 1.70$
100	$\frac{12}{\sqrt{100}} = 1.20$
500	$\frac{12}{\sqrt{500}} = 0.54$
5000	$\frac{12}{\sqrt{5000}} = 0.17$

If we plotted several of these distributions they would look as follows:



As n becomes large relative to N, the standard error becomes small. A small standard error means that little variation can be expected from one sample mean to the next. The larger sample size ensures that any sample mean we measure is much more likely to lie near the unknown population mean we are trying to estimate. In other words, any sample mean we observe for purposes of estimating μ is expected to have little sampling error. This is why we prefer larger samples.

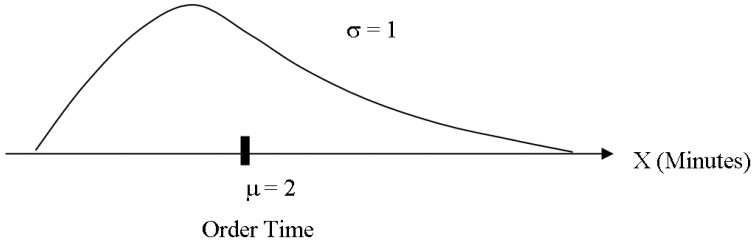
1.3 The Central Limit Theorem Applied

To aid in understanding the C.L.T. we can solve conceptual problems involving the likelihood of finding a given sample statistic if the population parameters are known. In practice the latter is not usually the case – we are taking a sample to estimate an unknown population parameter. We will see in later sections how we can use the C.L.T. to our advantage in this estimation and for hypothesis testing. For now, however, consider the following simpler conceptual problem.

Example:
A fast food service has a very large clientele. On the average it takes 2 minutes with a standard deviation of 1 minute to take an order from a customer. Suppose a simple random sample of 100 customers is observed, what is the probability that the average time $\bar{X}$ it takes to take a customer's order within the sample is:

1. Between 2 and 2.1 minutes?
2. Less than 1.8 minutes?
3. Over 3 minutes?

Solution:
Notice that there is no information about the shape of the distribution of times to take a customer's order among all of the clientele. It may be skewed or bimodal. It could look like this:



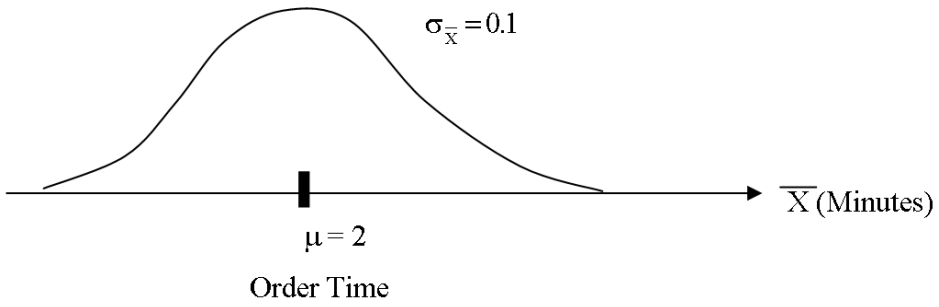
Since $n = 100 \geq 30$ the C.L.T. guarantees us that if all possible samples of size 100 are taken from this population of times, the distribution of mean times from sample to sample will be a normal curve. The mean and standard deviation of this sampling distribution are known. The mean of the sampling distribution is the same as that of the population, namely

$$\mu_{\bar{X}} = \mu = 2 \text{ min} .$$

The standard deviation of the sampling distribution (the standard error of the mean) is less than the standard deviation of the population.[4]

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1 \text{ min}}{\sqrt{100}} = 0.1 \text{ min}$$

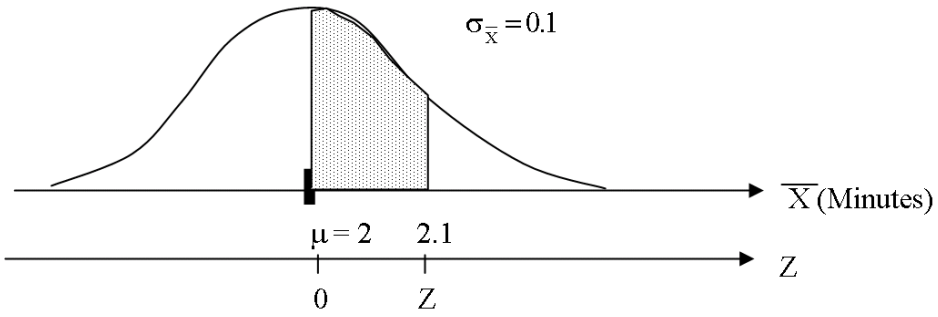
The sampling distribution of mean times would look like this:



Since the shape, mean, and standard deviation of the sampling distribution curve are known, the probability associated with an interval under this curve can be found by using a table of areas under the normal curve, where Z on the sampling curve is found by:

$$\boxed{Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}}$$

1. For example, to find $P(2 \text{ min} < \bar{X} < 2.1 \text{ min})$, find the Z values for $\bar{X}$'s of 2 and 2.1 minutes. Since 2 minutes is the mean, its Z value is 0.00 .



If $\bar{X} = 2.1$ minutes then $Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{2.1 - 2}{0.1} = 1.00$.

Look up the area between centre and Z = 1.00 in the table. The answer is 0.3413.

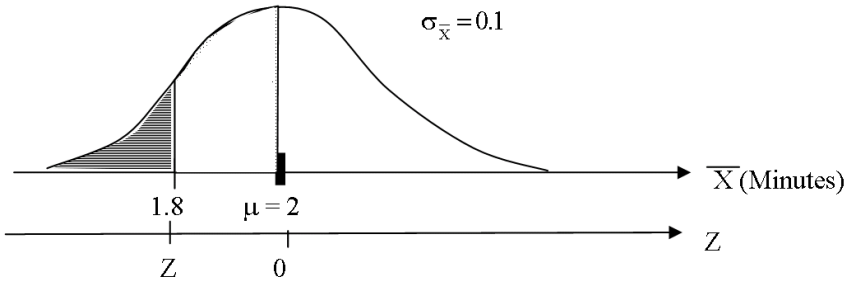
The proper use of probability symbols in this case is:

$$P(2 \text{ min} < \bar{X} < 2.1 \text{ min}) = P(0.00 < Z < 1.00) = 0.3413$$

Caution: A common error is to use the population standard deviation rather than the standard error for the standard deviation in the Z value calculation.

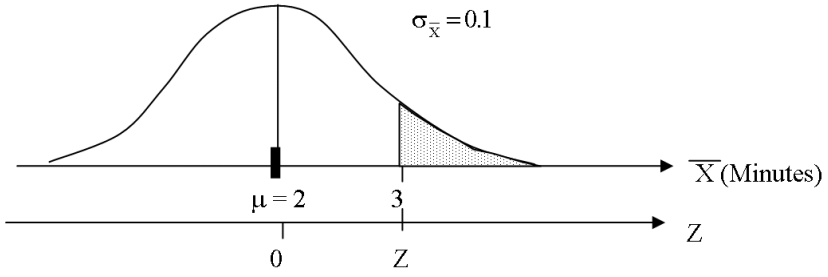
Compute the answers to parts 2 and 3 by drawing the diagrams, shading in the correct region and doing the calculation. Satisfy yourself that they should be as follows.

2.



$$P(\bar{X} < 1.8 \text{ min}) = P(Z < -2.00) = 0.0228$$

3.



$$P(3 \text{ min} < \bar{X}) = P(10.00 < Z) = 0.0000$$

Notice that if we had just asked the probability that a single order (X) was longer than 3 minutes the answer would have been, if the underlying population variable X were normally distributed,

$$P(3 \text{ min} < X) = P(1.00 < Z) = 0.1587 \ .$$

The result here is significantly different than that of the sample mean due to the use of the standard deviation when calculating Z.

[4]The finite correction factor (see Section 1.4.1) is not required here because it is stated that the population is very large so the sample is of insignificant size ($< 5\%$) relative to the population and the *F.C.F.* would be very close to 1. Note that the underlying population here is the orders, not the customers so a large clientele implies a large number of orders. Even a pub with only 20 regular customers would still have a large number of orders (effectively infinite) over time.

1.4 Extending the Central Limit Theorem

The central limit theorem describes the behaviour of the distribution of sample means. In the previous examples, it has been demonstrated that the sample mean is itself a random variable. When one sample is drawn from a population, the only information that is known about the population mean is that the mean of this sample is one of the possible sample means. If the conditions of the C.L.T. are met, then this sample mean will lie on a normal distribution curve that has the population mean as its centre.

The central limit theorem as stated requires a sample size of $n \geq 30$. When sampling without replacement we also have required a population that is much larger ("infinite") compared to the sample size ($n/N < .05$). If these conditions do not hold then some adjustments must be made about some of the statements in the central limit theorem.

1.4.1 Finite Populations

If the population cannot be considered to be infinite, that is when the **sample size constitutes 5% or more of the population**, there will be less variation among sample means than can be expected by the C.L.T. when we are sampling **without replacement**. In this case the **finite correction factor (*F.C.F.*)** is applied to the standard error of the mean,

$$\boxed{\sigma_{\bar{x}} = (F.C.F.) \cdot \frac{\sigma}{\sqrt{n}}},$$

where

$$\boxed{F.C.F. = \sqrt{\frac{N-n}{N-1}}}.$$

Since the $F.C.F. \leq 1$, its inclusion makes the standard error smaller.

Example:
Repeat the last example ($\sigma = 12$ cm, $\mu = 115$ cm) but now with a population of size $N = 10,000$. Compute the standard error of the mean if samples of size 50, 100, 500, and 5000 are drawn without replacement from the population.

Solution:

n	n/N	$F.C.F. = \sqrt{\frac{N-n}{N-1}}$	$\sigma_{\bar{x}} = (F.C.F.) \cdot \frac{\sigma}{\sqrt{n}}$ (cm)
50	.005	$\sqrt{\frac{9950}{9999}} = .9975$	$\sqrt{\frac{9950}{9999}} \cdot \frac{12}{\sqrt{50}} = 1.69$
100	.010	$\sqrt{\frac{9900}{9999}} = .9950$	$\sqrt{\frac{9900}{9999}} \cdot \frac{12}{\sqrt{100}} = 1.19$
500	$\boxed{.050}$	$\sqrt{\frac{9500}{9999}} = \boxed{.9747}$	$\sqrt{\frac{9500}{9999}} \cdot \frac{12}{\sqrt{500}} = \boxed{0.52}$
5000	$\boxed{.500}$	$\sqrt{\frac{5000}{9999}} = \boxed{.7071}$	$\sqrt{\frac{5000}{9999}} \cdot \frac{12}{\sqrt{5000}} = \boxed{0.12}$

For n small relative to the size of N, that is for $n/N < .05$, the size of the $F.C.F.$ is approximately 1, and as a result makes very little difference to the value of the standard error from the last example. In these two cases ($n = 50$, $n = 100$) the population is effectively infinite and we can neglect the $F.C.F.$. For the cases where $n/N \geq .05$ the $F.C.F.$ differs significantly from 1 and introduces a significant modification to the standard error in the mean as can be seen by comparing the entries (boxed) for $n = 500$ and $n = 5000$ to their previous values. In these cases the finite size of the population is noticeable.

Note that had we sampled with replacement, no *F.C.F.* would have been required regardless of the sample size. **We will always assume we are sampling without replacement unless stated otherwise.** Therefore test to see if $n/N < .05$. If it is not, the *F.C.F.* is required.

1.4.2 Small Samples

The sampling distribution of the mean is always symmetrical in shape regardless of the size of the sample or the shape of the population from which the sample is drawn. If the sample size is thirty or more and the population is infinite the distribution of sample means will always be normal regardless of the shape of the population distribution from which the sample is drawn. However sometimes large samples are very expensive or impossible to get.

If the sample size is less than 30, the sampling distribution of the mean is no longer normal. However if the variable X itself is known to have a normal distribution then the sample mean $\bar{X}$ also will be normally distributed. However, for reasons that will be discussed in Section 2.4 the small sample size problems we will be interested in ultimately require the use of what is called a ***t* distribution**.

If the C.L.T.'s conditions are met (infinite population, large sample), one could wonder why we do not use the finite correction factor and a t-distribution even then if the results are more accurate. In principle we could and some do. However, in this course the neglect of the *F.C.F.* for infinite populations and the use of a normal distribution for large sample size will be used both because it is easier and also to give a better understanding of what is going on.

Assignment:

For any conceptual problem involving application of the C.L.T. :

⇒ Draw a Venn diagram with labeled parameters and statistics.
⇒ Draw a sampling curve with appropriate labels and the region that answers the question shaded.
⇒ Use the proper symbols for sampling distributions and probability.

1. The average income in a country of 2,500,000 people is \$9600 with a standard deviation of \$640. A simple random sample of 1024 of the citizens of this country is selected and their average income is computed.
   (a) Draw a diagram of the sampling distribution of the mean.
   (b) What is the probability of observing a sample average within \$50 of the population average?
   (c) What is the probability of observing a sample average that is more than \$60 higher than the population average?
2. The average weight of a vehicle traveling on a municipal road is 2500 kg with a standard deviation of 700 kg. A simple random sample of 40 vehicles traveling the road is selected.
   (a) Draw a diagram of the sampling distribution of the mean.
   (b) What is the probability of observing a sample mean within 200 kg of the average weight?
   (c) What is the probability of observing a sample mean that differs from the average weight by more than 500 kg?
3. A personnel test is administered to 500 applicants. A simple random sample of 100 scores is selected from the group of all test scores. The average test score for all people writing the test is 75% with a standard deviation of 10%.
   (a) What is the probability the sample mean will be above 76%?
   (b) Between what two values would the average for the sample fall in 95 cases out of 100?

   ** Caution: How does n compare with N and what are the implications? **
4. What happens to the standard error in the mean, $\sigma_{\bar{x}}$, when
   (a) The sample size is just one, $n = 1$?
   (b) The sample is the whole population, $n = N$?

   Explain your answers.
5. Repeat the example of sampling two of the four houses in the village for the number of occupants, but now sampling without replacement.
   (a) What are the ${}_4C_2 = 6$ possible samples?
   (b) What is the mean of the sample means, $\mu_{\bar{X}}$? Compare it to the population mean, μ.
   (c) What is the standard error of the mean, $\sigma_{\bar{x}}$? Show that it satisfies the equation for $\sigma_{\bar{x}}$ of the C.L.T. if the finite correction factor is used.
   (d) Which of the conclusions of the C.L.T. are not true in this example?

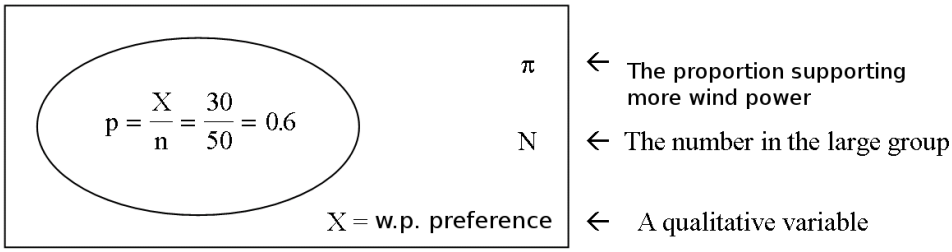
1.5 Sampling Distribution of Proportions

Every sample statistic has a sampling distribution. That is, for any given sample statistic, there is bound to be some variation observed from sample to sample in the sample statistic. The C.L.T. describes the behaviour of the variation among sample means. The distribution of sample modes, sample medians or sample standard deviations could be analyzed in a similar manner. Each of these distributions would have the same attributes related to shape, centre and variability.

Another statistic of importance whose value varies from sample to sample is the sample proportion p. The symbol for the population parameter is π.

Example:
Suppose a simple random sample of 50 people is selected from a large group of people. It is found that 30 people in the sample are in favour of more wind power generation. Estimate the proportion of the population in favour of more wind power generation.

Solution:



Is it likely that $p = \pi$? If we took another sample of size 50 we are unlikely to observe the same p value?

Large samples are always used to estimate population proportions. For samples of fixed size n, the distribution of the sample proportion has the following three features.

1. The distribution of p approaches a normal curve as n increases. The normal approximation is valid provided $n\pi > 5$ and $n(1-\pi) > 5$.

2. The mean of all sample proportions is the population proportion, π:

$$\mu_p = \pi$$

3. The standard deviation of sample proportions, called the **standard error of proportion**, is found by[5]

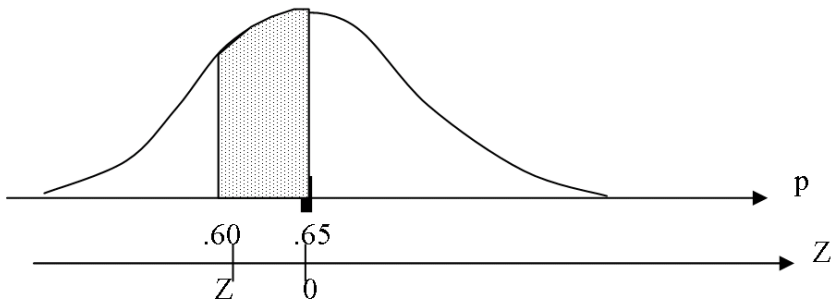
$$\boxed{\sigma_p = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}}$$

[5]Note:

- The similarity to the standard deviation formula of the binomial probability distribution is no coincidence. The result arises from the fact that the binomial distribution is approximately normal under the constraints given above. The connection between a binomial probability and a sample proportion is demonstrated in problem 2 on page 16 .
- Some use $n\pi > 10$ and $n(1-\pi) > 10$ for the realm of validity. Once again see problem 2 for an example of the magnitude of error introduced near the cutoff.

Example:
In the previous example, suppose that π, the proportion of the population who support more wind power, is 0.65. What is the probability of finding a sample proportion, p, between 0.60 and 0.65?

Solution:
Since we have that $n\pi = (50)(.65) = 32.5 > 5$ and $n(1-\pi) = (50)(.35) = 17.5 > 5$, p should be approximately normally distributed. Draw the sampling distribution curve for the statistic.



The mean of the sampling distribution curve is the same as the population proportion 0.65 but the standard error or proportion is:

$$\sigma_p = \sqrt{\frac{\pi \cdot (1-\pi)}{n}} = \sqrt{\frac{(0.65)(.35)}{50}} = 0.0675$$

Compute the Z value under the normal curve that corresponds to a p value of 0.60 and 0.65. For $p = 0.60$,

$$Z = \frac{p-\pi}{\sigma_p} = \frac{.60-.65}{0.0675} = -0.74$$

The area between centre and that Z value in the table is 0.2704. The answer written in proper probability symbols is:

$$P(0.60 < p < 0.65) = P(-0.74 < Z < 0) = 0.2704$$

**Caution: A common confusion here is calculating the value of the standard deviation of this sampling distribution using p rather than π.

For the rest of the unit we will be dealing with problems involving either sample means, $\bar{X}$, or sample proportions, p. To identify the type of problem that you are working with, it is helpful to note that in sample mean problems the underlying variable X is always quantitative, since it is something that has to be averaged. For a proportion problem, the underlying variable X is qualitative since we are calculating the proportion of the population that satisfies some criterion. As such the values of the variable are either X = true (satisfies the criterion) or X = false (does not satisfy the criterion). To simplify the notation when dealing with proportions as done in the last problem we will tend to write $p = \frac{X}{n}$ where here X means the number of sample elements satisfying the criterion. As such our usage of X here corresponds to the count appearing in the binomial distribution. Finally note that as a proportion, it must be the case that p be dimensionless and satisfy $0 \leq p \leq 1$.

Assignment:

1. Currently the ruling political party in a large country is supported by 45% of the voters. If a simple random sample of 1024 voters is chosen in this country, what is the probability that the proportion in the sample in favour of the ruling party differs by more than 2% from the population proportion?

2. A large box of batteries contains 20% that are defective. A simple random sample of 30 batteries is chosen. What is the probability of finding from 3 to 6 batteries in the sample defective

   (a) Using the normal approximation for a sample proportion?

   (b) Using an exact binomial probability formula calculation?

   Your answers should differ considerably. One reason is that we are approximating the discrete binomial probability distribution with a continuous distribution. For instance if we wanted to evaluate the probability of exactly $X = 3$ defective batteries we would need to evaluate the area under the continuous normal curve for the proportion range corresponding to 2.5 to 3.5 defective batteries.

   (c) To improve on our approximation, redo part (a) using the proportions for 2.5 to 6.5 defective batteries instead and compare this with the actual answer in (b). Modifying intervals in this ways is referred to as introducing a *continuous correction factor*.

2 Point Estimates and Confidence Intervals

2.1 Confidence Intervals for the Population Mean (Large Samples)

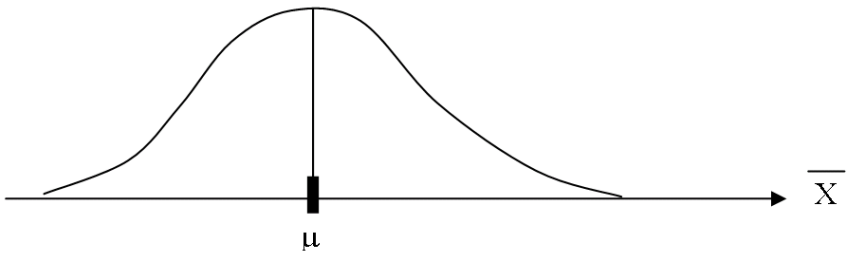
The previous sections have demonstrated the theoretical working of the C.L.T. for sample means and the equivalent theorem for sample proportions. We will modify this slightly to apply the theory in a practical setting.

In all of the sampling theory developed so far, the assumption was made that the population parameters were known. For example, to compute $\mu_{\bar{x}}$ or $\sigma_{\bar{x}}$, the assumption is made that μ and $\bar{X}$ are known. If this were the case, there would be no need to sample because one would never use an estimate if an exact value were available.

In practice only one sample is selected. This means that we have exactly one point known to us on the whole sampling distribution curve. The exact position of this point on that curve is also unknown to us.

Example:
A simple random sample of 50 wage earners from a population of 10,000 is selected. The average wage in the sample is \$42,000 with a standard deviation of \$5000. Suppose we wish to estimate μ. If another sample of size 50 was chosen, is it likely that we would see the same sample statistics?

Solution:
Because the sample size exceeds 30, the only thing known here is:

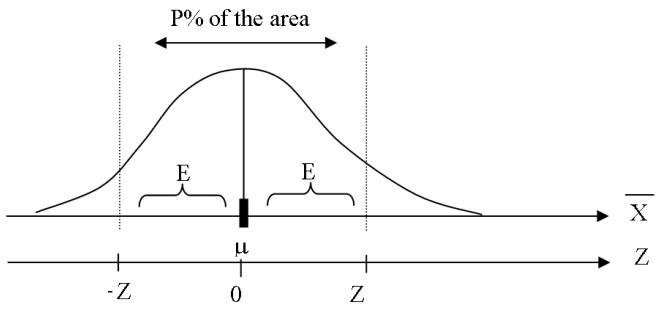


We do not know whether to place the $\bar{X}$ we observed to the right or to the left of μ. The best we can say is that $\mu \approx \bar{X} = \$42,000$.

We call sample statistics such as these **point estimates** of the parameter because they represent only one of the many possible points on the distribution curve.

Parameter	Point Estimate
μ	X
σ	s
π	p

The problem with point estimates is that there is no way to assess how precise they are or how much confidence to place in them. To satisfy these criteria, we construct **interval estimates**.



The above diagram shows that P% of all sample means are within a distance of $\boxed{E = Z \cdot \sigma_{\bar{X}}}$ of the mean. In probability symbols:

$$\boxed{P\left(\left[\bar{X} - E\right] < \mu < \left[\bar{X} + E\right]\right) = P\%}$$

This probability event is called a confidence interval estimate for the population mean. This interval estimate $\left[\bar{X} - E\right] < \mu < \left[\bar{X} + E\right]$ has a **precision** of E and a **degree of confidence** of $P\%$.

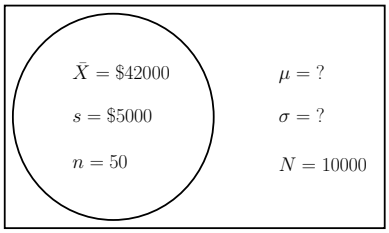
Example:
Using the previous sampling problem on incomes, construct a 95% confidence interval for μ.

Solution:
Keeping in mind that the goal is to fill in the blanks of the confidence interval statement,

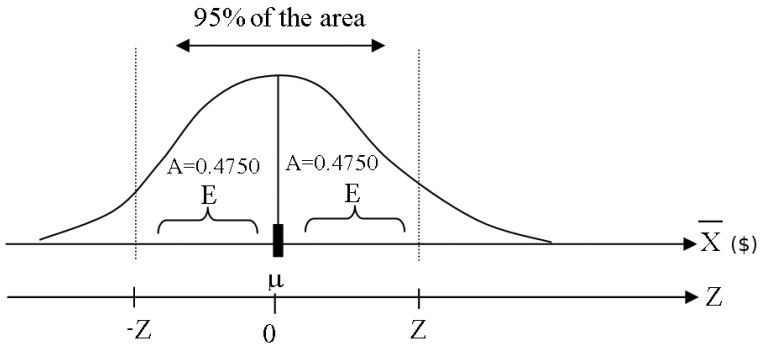
$$P\left(\left[\bar{X} - E\right] < \mu < \left[\bar{X} + E\right]\right) = P\% \ ,$$

proceed with the following steps for any confidence interval:

Step 1) Identify all given information with symbols. This is conveniently done with a Venn diagram:



Step 2) Draw a diagram of the sampling distribution. Since $n = 50 \geq 30$ the distribution of $\bar{X}$ is approximately normal by the C.L.T. .



Step 3) Calculate the Z-value. Use the confidence, here $P = 95\%$, to find the area required for the table,

$$A = \frac{.95}{2} = 0.4750 \ ,$$

and the corresponding Z-value,

$$Z = 1.96 \ ,$$

from the table.

Step 4) Calculate the standard error. To calculate the precision E, we require the standard error of the mean, $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$, but σ is unknown so replace it by its point estimate, s. We designate the approximate standard error of the mean as

$$\boxed{s_{\bar{X}} = \frac{s}{\sqrt{n}}} = \frac{\$5000}{\sqrt{50}} = \$707.11$$

** Note that at this step we used the fact that $\frac{n}{N} = \frac{50}{10000} = .005 < .05$ so that, assuming we were sampling without replacement, we could ignore a *F.C.F.* in these formulae. If $\frac{n}{N} > .05$ we would have to introduce an *F.C.F.* into the standard error here.

Step 5) Calculate the precision (maximum sampling error).

$$\boxed{E = Z \cdot s_{\bar{X}}} = (1.96)(\$707.11) = \$1385.94$$

Step 6) Make the confidence interval statement.

$$\begin{aligned} P\left(\left[\bar{X} - E\right] \ &< \ \mu \ < \ \left[\bar{X} + E\right]\right) = P\% \\ P\left(\left[\$42,000 - \$1385.94\right] \ &< \ \mu \ < \ \left[\$42,000 + \$1385.94\right]\right) = 95\% \\ P(\$40,614.06 \ &< \ \mu \ < \ \$43,385.94) = 95\% \end{aligned}$$

In words there is a .95 probability that the population mean μ is between \$40,614.06 and \$43,385.94.

Notes:

1. Some terms that are used in connection with confidence intervals are:

   Lower Limit \$40,614.06

   Upper Limit \$43,385.94

   Precision or Maximum Sampling Error \$1385.94

   Confidence Coefficient 95%

2. The reason we will always write a confidence interval statement which includes the interval along with the confidence is because the interval **depends on the confidence coefficient chosen**. To see this, repeat the above calculation at a 99% level of confidence.

   Step 1) Identify all given information with symbols on a Venn Diagram.

Step 2) Draw a diagram of the sampling distribution.

Step 3) Calculate the Z-value.

Step 4) Calculate the standard error.

Step 5) Calculate the precision (maximum sampling error).

Step 6) Make the confidence interval statement.

3. One does not get something for nothing when creating confidence intervals with the same raw information. Comparing our 99% confidence interval with the 95% confidence interval we see that if we want to be more confident about catching the population mean we have to have a wider interval (larger E). If you wanted a narrower interval with the same level of confidence what would you have to do?

4. When creating a confidence interval we choose the level of confidence. Usually we want to be fairly confident we have caught the population mean in the interval.[6] Standard confidence coefficients are in the following table, along with their corresponding Z-value for normally distributed sample statistics.

Confidence, P	Z
80%	1.282
90%	1.645
95%	1.960
98%	2.326
99%	2.576
99.9%	3.291

   These may be verified up to two digits after the decimal from the normal distribution table.

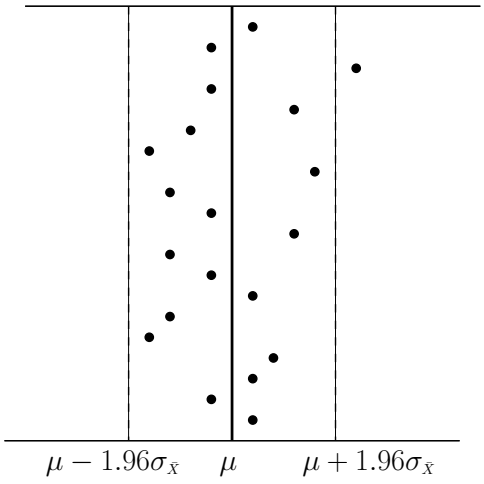
5. We stated that the generic confidence interval for the sample mean is

$$P\left(\left[\bar{X} - E\right] < \mu < \left[\bar{X} + E\right]\right) = P\% \ ,$$

   but looking at the sampling distribution it ought to read

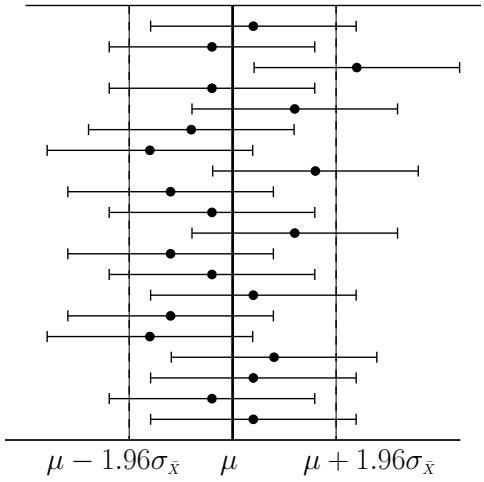
$$P\left(\left[\mu - E\right] < \bar{X} < \left[\mu + E\right]\right) = P\% \ \ !$$

   What gives? To see why we can exchange the population and sample means consider the following diagram illustrating means of 20 possible samples. The 95% confidence interval around the population mean is indicated and as expected there is one sample out of 20 (the third) which does not lie in this interval while 19/20 or 95% lie within the interval.



   In the next diagram we draw the confidence intervals centred on each sample mean.

[6] A level of confidence of 30% would not be very useful information since the probability of being outside the stated interval would be 70%!



One notices that 19 of the 20 sample confidence intervals (95%) contain the true population mean μ while only one (again the third) does not. As such for a sample chosen at random we have a 95% chance of containing the actual population mean within the confidence interval centred on the sample mean which is the statement we wanted to make. Note however that when drawing the sampling distribution, the variable is the sample statistic $\bar{X}$ not the population parameter μ. The latter is a fixed, if unknown, constant.

Assignment:

1. Mercury levels were tested in a sample of 36 Walleye at Threepoint Lake in northern Manitoba in 2005 and found to have a mean of .300 ppm (parts per million) with a standard deviation of .040 ppm. Assuming the Walleye population in the lake is significantly larger, construct a 98% confidence interval for the average concentration of mercury in Walleye in that lake for that year.

2. To estimate the average amount owing in a large journal of receivables, a random sample of 100 entries is selected from the entries. The average amount owing in the sample is \$512 with a standard deviation of \$58. Construct a 90% confidence interval for the average amount of an entry in the journal.

3. A community has 500 family units. An estimate of the average amount spent per household winterizing a family dwelling is required. 80 families are randomly selected. The average amount in the sample spent on winterizing a dwelling is \$85 with a standard deviation of \$15.

   (a) Construct a 95% confidence interval for the average amount of money spent on winterizing a dwelling. ** Caution: Examine n. **

   (b) Repeat (a) but for a 99% confidence interval.

4. A taxi company wishes to obtain an estimate of an average fare. A random sample of 20 customers shows that the average fare is \$19.50 with a standard deviation of \$4.80. Why can our knowledge of the C.L.T. not be used to construct a 90% confidence interval or any other confidence interval in this case?

5. A training seminar has been offered many times over the course of the life of a business to many of its employees. At the end of the seminar an exam is administered to measure the knowledge level of an employee at the end of the seminar. An estimate is to be obtained of the average score on the test from the large bank of test scores of all employees who have taken the test. A random sample of 180 scores from the files of the company show the average in the sample to be 82% with a standard deviation of 6%.

   (a) Construct a 99% confidence interval for the average of all test scores.

   (b) Repeat part (a) but now assume that our sample of 180 was taken out of a population of only 1000 exams.

2.2 Confidence Intervals for the Population Proportion

The same theory holds for sample proportions as for sample means as was done on the previous example with the exception that the standard error of the sampling distribution is the standard error of proportion. Ultimately we wish to make a confidence interval statement of the form

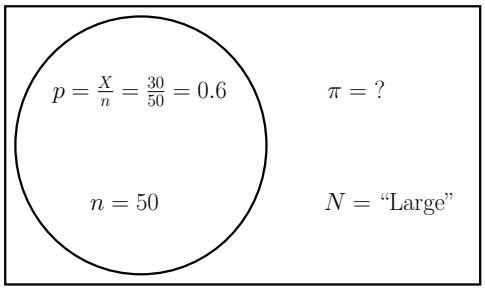
$$\boxed{P\left([p-E] < \pi < [p+E]\right) = P\%}\ ,$$

from which the steps below logically follow. The maximum sampling error (precision) now depends on the standard error of proportion, $\boxed{E = Z \cdot \sigma_p}$.

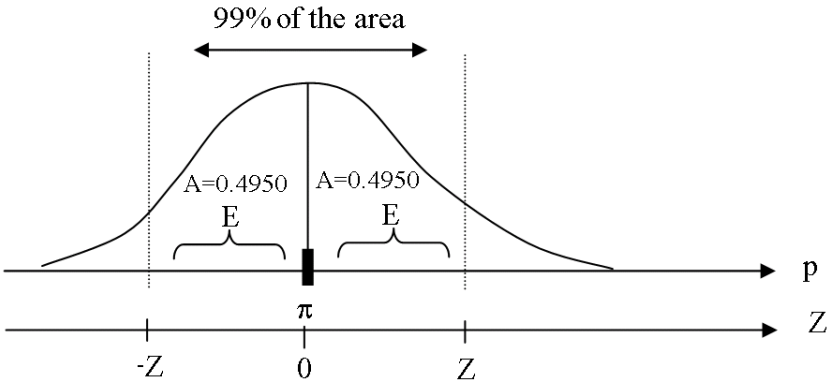
Example:
A previous problem dealt with a sample of size 50 randomly selected from a large group for the purpose of estimating the proportion of the group who support more wind power generation. Construct a 99% confidence interval for the proportion in the population who support more wind power if it is observed that there were 30 supporters in the sample.

Solution:

Step 1) Draw a Venn diagram labelling the information known. It is also convenient to calculate the sample proportion here.



Step 2) Construct a sampling distribution diagram. Since we are now assuming we do not know what the true population proportion is, we can use the the point estimate of $\pi \approx p = .6$ to confirm that, $n\pi \approx (50)(.6) = 30 > 5$ and $n(1-\pi) \approx (50)(.4) = 20 > 5$ so the distribution of sample p should be normal.



Step 3) Calculate the Z-value. To trap 99% of all sample proportions about the centre, we require an area of

$$A = \frac{.99}{2} = .4950\ ,$$

so from the tables (selecting the value midway between 2.57 and 2.58) we need to go out 2.575 standard deviation from the centre, that is

$$Z = 2.575 \ .$$

Step 4) Calculate the standard error. The standard error of proportion, in theory is found by $\sigma_p = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$. Since π is unknown in this case, π is replaced by its point estimate, p, and the approximate standard error of proportion, designated by s_p, is replaced by

$$\boxed{s_p = \sqrt{\frac{p \cdot (1-p)}{n}}} = \sqrt{\frac{(0.60)(0.40)}{50}} = 0.069 \ .$$

Note that at this step we had no need for a *F.C.F.* because $\frac{n}{N} = 50/\text{“Large”} \approx 0 < 0.05$. Had this not been the case we would have had to include a *F.C.F.* in our standard error formula.

Step 5) Calculate the precision (maximum sampling error).

$$\boxed{E = Z \cdot s_p} = (2.575)(0.069) = 0.18$$

Step 6) Make the confidence interval statement.

$$\begin{aligned} P\left([p-E] \ < \ \pi \ < \ [p+E]\right) &= P\% \\ P\left([.6-0.18] \ < \ \pi \ < \ [.6+0.18]\right) &= 99\% \\ P(.42 \ < \ \pi \ < \ .78) &= 99\% \end{aligned}$$

In words there is a .99 probability that the population proportion π is between .42 and .78 .

Assignment:

1. The Saskatchewan Ministry of the Environment has a surveillance program to monitor the incidence of Chronic Wasting Disease (CWD). Between 1997 and 2007 around 34,000 cervids (which include deer, elk, and moose) in Saskatchewan have been tested for CWD with 197 testing positive.

   (a) Will the sample proportion p of infected animals be normally distributed?

   (b) Construct a 90% confidence interval for the proportion of cervids in Saskatchewan with CWD. (Assume the actual population size of cervids in Saskatchewan is in the millions.)

   (c) According to the Ministry of Environment (MOE) website:

   > "The CWD surveillance program is based primarily on the testing of hunter-killed animals and to a lesser extent on the testing of cervids submitted through the passive surveillance of sick or dead animals in the province. In areas where CWD was detected, MOE implemented herd reduction programs, primarily through increased hunting quotas and reduced hunting licence fees, in an attempt to increase hunter harvest to reduce deer densities and limit the spread of CWD."

   Does this affect the conclusion in (b)? If so, how?

2. A credit union branch has 650 members some of which have more than one type of account. A random sample of 50 of the members shows that 26 have a secondary account.

   (a) What is a point estimate for the proportion of members who have a secondary account at the credit union?

   (b) Construct a 99% confidence interval for the proportion of members who have a secondary account at the credit union.

   ** Hint: Note size of n and N.

3. A poll of 1000 people taken from a large population asking their voting intention gave the following results:

Response	Number
Pink Party	560
Blue Party	320
Green Party	90
Undecided	30
	1000

   (a) Construct a 95% confidence interval for the proportion of people who support the Pink Party.

   (b) Based on your result in (a) can we conclude that the majority (more than half) of the population support the Pink Party?

   (c) If you wanted to be 99% confident that the Pink Party had majority support would your result in (b) still be true?

2.3 Sample Size Determination

In all of the previous problems, it was assumed that a sample had already been taken and an estimate was to be based upon this sample. In a practical situation, a sample must first be selected before any of the sample statistics can be computed. The size of the sample cannot just arbitrarily be selected. There are three factors that determine the size of the sample to be selected both from a statistical point of view and from an economic point of view.

From an economic point of view, the sampling size turns out to be costly because it involves the time, resources, and funds of the firm. The method of choosing the sample should minimize these costs while at the same time guaranteeing the statistical reliability and precision of the estimate provided.

From a statistical point of view, the size of the sample depends upon three things:

1. The amount of precision required in the estimate.
2. The degree of confidence to be placed on the estimate.
3. The amount of variation in the population.

Recall that:

- In estimating **means**, the maximum possible sampling error (precision) is found by:

$$E = Z \cdot \sigma_{\bar{X}} = Z \cdot \frac{\sigma}{\sqrt{n}}$$

  Solve this equation for n:

$$\boxed{n = \left[\frac{Z \cdot \sigma}{E}\right]^2}$$

- In estimating **proportions**, the maximum possible sampling error (precision) is found by:

$$E = Z \cdot \sigma_{p} = Z \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

  Solve this equation for n:

$$\boxed{n = \pi \cdot (1-\pi) \cdot \left[\frac{Z}{E}\right]^2}$$

Notice that in both the case of estimating means and of estimating proportions, the formula developed requires the value of a parameter in the population. The parameters in the population are unknown to us so an estimate of these parameters must be available to us before the sample is selected. This is often done by using a small pilot survey or in the case of proportions the worst case can be assumed, namely where, $\pi = 0.5$. The latter maximizes the function $\pi \cdot (1-\pi)$. Note the sample size n must be an integer and to be conservative we always round up regardless of the decimal fraction.

Example:
A survey is planned to determine the amount of time computer-owning adults in Regina spend on their home computers during a week. A pilot survey of such adults found that 8 hours were spent on average with a standard deviation of 3 hours. If we want to estimate the actual average weekly home computer use to within 1/4 of an hour, how many should be surveyed if we want to be 95% confident of the result? (Answer: $n = 554$ computer-owning Regina adults)

Solution:

Example:
In a previous example a sample of 50 was taken from a large population and it was found that 30 were supporters of more wind power generation. The 99% confidence interval for the proportion of support based on this small sample was quite wide, namely
$([.6 - 0.18] < \pi < [.6 + 0.18]) = (.42 < \pi < .78)$.

1. How large a sample should be taken if we desire to know the true proportion to within .05, still at a 99% level of confidence. Use the small sample proportion as an estimate of π. (Answer: $n = 637$ people, if $Z = 2.575$ is used.)
2. How large a sample should be used if no estimate of π had been available? (Answer: $n = 664$ people)

Solution:

1.

2.

Assignment:

1. A manufacturing company produces a component for a product on an assembly line. The company wishes to know the average time taken for the component to travel from one workstation on the line to the next. By taking a handful of such measurements a worker has estimated that the mean time taken is 202.0 seconds with a standard deviation of 5.0 seconds. What sample size should be used in the actual analysis in order to be 95% certain that the mean transfer time is estimated to within 1.0 seconds in the final confidence interval?
2. What sample size should a government official employ if he wishes to estimate the proportion of citizens of Quebec who wish to secede from Canada to become a sovereign country to within 1% at a 99% level of confidence. (The last Quebec referendum in 1995 was 49.42% in favour of leaving Canada.)
3. In order to assess the cost of commuting daily between Regina and Moose Jaw the average mileage of carpool vehicles is required. A pilot survey of six such vehicles found the following mileages:

   $$6.8,\ 8.1,\ 8.5,\ 9.4,\ 12.3,\ 13.6 \quad \text{(km/litre)}$$

   If the average mileage is desired to be known to within .5 km/litre at a 90% level of confidence, how many car mileages should be measured in the actual survey?
4. In an attempt to assess the status of women in nineteenth century Paris, a researcher wishes to know the proportion of burial plots occupied by women in the Père Lachaise cemetery, a large affluent necropolis in Paris containing over 300,000 burial plots.
   (a) If the researcher wishes to know the fraction of female burials to within 4% at a 95% level of confidence, how many nineteenth century burial plots will she need to examine in the cemetery?
   (b) In an attempt to lessen the number of plots requiring examination, the researcher chooses 20 nineteenth century plots at random, looks at the names, and determines that 6 are women. How large a sample size would she need now to achieve a confidence interval with the requirements given in (a)?
   (c) The researcher decides that this is still a larger sample size than she wishes to take. What other options does she have to reduce the sample size required?

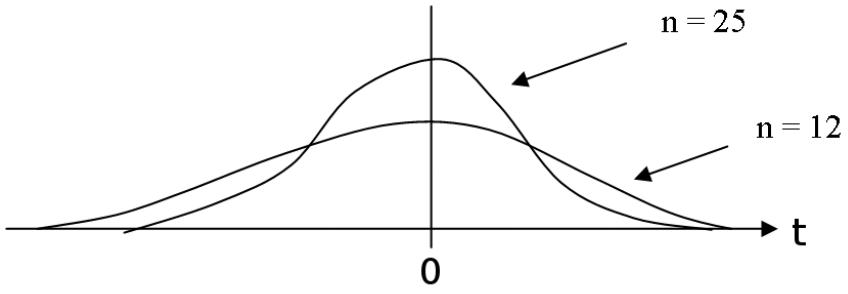
2.4 The Sampling Distribution of the Mean Using Small Samples

In some instances it is not practical to take the large sample ($n \geq 30$) required for the C.L.T. to be applied. (Name some!) However, if a random variable X is **known to be approximately normally distributed** (something not required by the C.L.T.) it can be shown theoretically that both $\bar{X}$ and $Z = (\bar{X} - \mu)/\sigma_{\bar{X}} = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ are normally distributed even for small samples ($n < 30$).[7] The problem that arises for small samples, however, is that, as before, we typically do not know the population standard deviation σ but wish to approximate it by the sample standard deviation s . Now the test statistic, call it t, defined in the same way as was done for Z,

$$\boxed{t = \frac{\bar{X} - \mu}{s_{\bar{X}}}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} \ ,$$

is no longer normally distributed as it was before for large samples. It has what is called a **t distribution**. Essentially when $n < 30$ the variation of s about σ causes the above value to depart from normality. By definition, the t-value measures the number of standard deviations from the mean of a given value, just as with the Z-score.

The shape of the t distribution depends upon the sample size n. The larger the sample, the closer the shape is to a normal curve. The smaller the size of the sample, the more the distribution tends to be dispersed from the centre and the flatter the sampling distribution curve.



The total area under any of these curves, is, as usual for a continuous probability distribution, equal to one. Rather than label the different t curves by their associated sample size n, statisticians use a related number called the **number of degrees of freedom**, defined by

$$\boxed{df = n - 1} \ .$$

This is the same quantity that is the denominator in the sample variance formula:

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

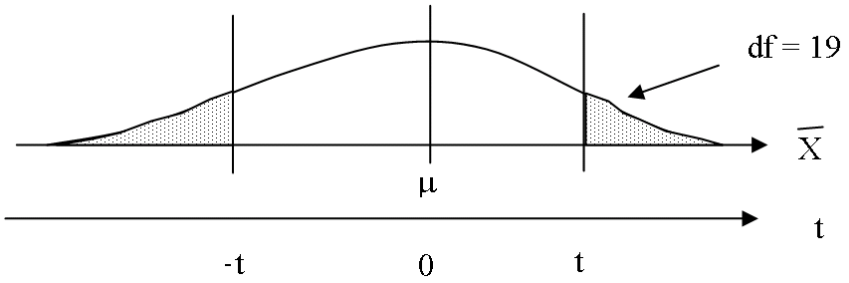
Essentially one has a different t distribution for each different value of df (or n), and we could make an area under the curve table for each one as was done for the normal (Z) distribution. In practice, since we are only interested in applying the t distribution to find confidence intervals (and later for hypothesis testing) this is unnecessary; we do not need 30 tables. Recall only a handful of numbers off the normal distribution table were required for these problems. We will grab the same set of numbers for each t distribution and put them on a single table. Furthermore our t table will illustrate the area under the t distribution in a different format than the Z table; we will use the area in the tails rather

[7] Here we neglected the *F.C.F.* in $\sigma_{\bar{X}}$ since we will assume a small sample always satisfies $n/N < .05$ as n is small (< 30).

than the area in the centre. Finally, whereas the normal table has the area (probability) inside the table, since we will be interested in doing reverse lookups (using area to get t), it is the t-value which is inside the table. See the t distribution table. Notice that the values for the Z distribution are given at the bottom ($df = \infty$).[8]

Example:
Find the t-value based on a sampling distribution with samples of size 20 that places 5% of the area under the curve in two tails. Draw a sketch of the sampling curve.

Solution:
If $n = 20$ then $df = n - 1 = 20 - 1 = 19$.



The total shaded region in two tails is 0.05 . Read down the appropriate column to find $t = 2.093$.

Confidence intervals for the mean are done exactly as before but now we need to be careful to identify which t distribution is the correct one to find t by using df as just shown. Also since we are using a t distribution rather than the normal distribution our maximum sampling error (precision) for our confidence interval is $\boxed{E = t \cdot s_{\bar{x}}}$. Finally we note that the assumption that X be normal is not overly restrictive as many distributions are normal. Moreover the t statistic is **robust** which means that its distribution remains largely unchanged even if the underlying distribution is not normal. In practice if X has a distribution that merely has a mound in the centre this analysis will still typically work.

Example:
A simple random sample of 10 small cars is tested for fuel economy. Here are the results:

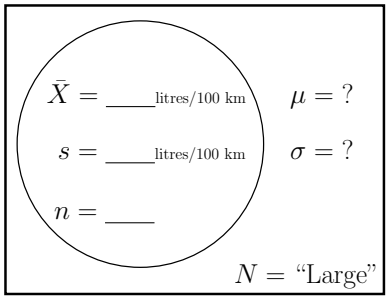
5.3, 6.2, 8.2, 6.1, 5.9, 7.3, 9.7, 5.5, 10.3, 7.5 (litres/100 km)

Construct a 98% confidence interval for the mean fuel economy of small cars based upon this evidence if the fuel economy of cars is known to be approximately normally distributed.

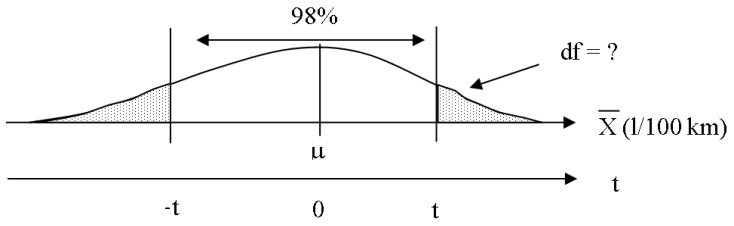
Solution:
We follow the same steps as before, guided by what we need to fill in the confidence interval statement.

[8]The ∞ symbols means infinity, or "very large" df. Recall as n gets really large (≥ 30) the t distribution converges to the normal curve.

Step 1) Identify all given information with symbols on a Venn Diagram. Compute $\bar{X}$ and s using your calculator and the appropriate statistical keys.



Step 2) Draw a diagram of the sampling distribution for samples of this size. Since $n = 10$ is less than thirty and X is known to be normal we will need a t distribution.[9]



Step 3) Calculate the t-value.

- Decide on the number of degrees of freedom, $df =$ _____.
- Find the total shaded area in two tails, Area $= 1 -$ _____ $=$ _____.
- The t-value from the tables is $t =$ _____.

Step 4) Calculate the standard error.

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} =$$

Step 5) Compute the precision (maximum sampling error).

$$E = t \cdot s_{\bar{X}} =$$

Step 6) Make the confidence interval statement.

$$P\left(\left[\bar{X} - E\right] \;<\; \mu \;<\; \left[\bar{X} + E\right]\right) = P\%$$

$$P\left([\ \_\_\_\_\_ - \_\_\_\_\_\ ] \;<\; \mu \;<\; [\ \_\_\_\_\_ + \_\_\_\_\_\ ]\ \right) = \_\_\_\_\_\ \%$$

$$P(\ \_\_\_\_\_ \;<\; \mu \;<\; \_\_\_\_\_\ ) = \_\_\_\_\_\ \%$$

[9]The perceptive student should note that in drawing this distribution the two curves are different since $\bar{X}$ has a normal distribution while the statistic we are manipulating $t = (\bar{X} - \mu)/s_{\bar{X}}$ has a t distribution. For our purposes we are only sketching a lump-shaped distribution in either case. The point of the diagram is to indicate important areas and the fact that $\bar{X}$ values and the dimensionless t values are interrelated.

Assignment:

Note in the confidence interval problems below that the "fun" part of using small samples is that the instructor can ask the student to calculate the descriptive statistics required ($\bar{X}$ and s) since there are few enough data elements (less than 30) that you can do this yourself on a calculator. After that step is taken, however, these problems are no more complicated that the large sample ones except use of the t distribution is required.

1. Find t-values required to solve:
   (a) a 98% confidence interval with sample size 15.
   (b) a 90% confidence interval with sample size 28.
   (c) a 99% confidence interval with sample size 6.
   (d) a 99.9% confidence interval with sample size 560.

2. According to Environment Canada, the average water discharge for the month of May of the South Saskatchewan River at Medicine Hat for the years 2000-2009 was measured to be (m^3/s):

69	51	64	272	87
82	238	294	317	89

   (a) If we could assume that the discharges are normally distributed, what would be the 95% confidence interval for the mean average water discharge for the month of May of the South Saskatchewan River at Medicine Hat.
   (b) If the underlying distribution is not approximately normally distributed, what could we do to construct the confidence interval?

3. The distributions of purchases at a snack food outlet is thought to be normal. A random sample of 12 purchases shows the following amounts ($):

0.75	1.78	4.25	0.50	2.34	3.25
0.98	2.22	1.75	1.50	0.88	4.15

   (a) Construct a 90% confidence interval for the average value of a purchase at the outlet.
   (b) A taxation officer observes the outlet over the course of a week and notices that 800 customers have entered the cash register line of the outlet. He seizes the tapes from the cash register and adds the total value of all purchases over this time. If purchases have been correctly accounted for, between what two values should the total of all purchases lie?

2.5 Sampling and Confidence Interval Summary Exercise

1. Suppose that there are currently 1,200,000 unemployed members of Canada's labour force. The mean length of time that a person is unemployed is 12 weeks with a standard deviation of 4 weeks. The distribution of time-spent unemployed is approximately normal.

   (a) A person is randomly chosen from among those unemployed in the labour force. What is the probability that the person has been unemployed in excess of 18 weeks?

   (b) Suppose that a simple random sample of 100 people is selected from those unemployed in the labour force. What is the probability that the mean length of time unemployed for the people in the sample exceeds 18 weeks?

   (c) Suppose we do not know what the mean or standard deviation in the labour force is but we do know that a simple random sample of 1000 unemployed workers yields an average of 11.80 weeks with a standard deviation of 3.75 weeks.

      i. Based on the 99% confidence coefficient, what is the maximum possible sampling error?

      ii. What is the confidence interval for the mean length of time spent unemployed by a person in the labour force?

   (d) Recalculate your answers to part (c) based upon a sample size of 20 rather than a sample size of 1000 using the same point estimates.

   (e) Recalculate the answers to part (c) based upon a sample size of 150,000 rather than a sample size of 1000 using the same point estimates. ** Check! What is n/N ? **

   (f) A small pilot survey was selected from the 1,200,000 unemployed. It was found that the standard deviation in the population was 4.2 weeks. How large should the random sample be in order to estimate the population mean based upon a 92% level of confidence if we wish to be within 0.5 weeks in our estimate?

2. A company has over 5000 salespeople spread across Canada. A random survey of business travel for these people showed that on the basis of 200 of these salespeople, the average distance traveled was 35,000 km with a standard deviation of 2,500 km. Construct a 98% confidence interval for the average travel of the 5000 salespeople.

3. A company wishes to purchase a compact vehicle to obtain the best fuel economy for urban deliveries. They chose 10 compact vehicles from sales lots and drove them under similar conditions. The average fuel economy achieved was 7.2 1/100 km with a standard deviation of 1.2 1/100 km. On the basis of the sample results, construct a 90% confidence interval for the average fuel economy one can expect from compact cars driven under these conditions.

4. In 2011 the Saskatchewan government decided not to hold a province-wide vote on whether to introduce daylight savings time (DST) arguing that in a recent government poll, 66 per cent of Saskatchewan residents opposed switching to DST, while only 27 per cent were in favour. Seven per cent had no opinion. Further details of the poll were as follows:

   > "The province-wide poll of 1,012 Saskatchewan residents was conducted by Fast Consulting between Jan. 10 and Jan. 24. It has a margin of error of 3.1%, 19 times out of 20."

   (a) Calculate the maximum sampling error for the proportion of the population in favour of switching to DST at the 95% level of confidence. Why is there a discrepancy between your answer and the 3.1% reported?

   (b) Find the confidence interval for the proportion in favour of switching to the DST at a 99.9% level of confidence. (You can find the appropriate Z-value on the bottom of the t table with $df = \infty$.) State your conclusion using the same statistical language of the quotation. Would the 99.9% confidence interval have been more useful to report than the 95%?

5. A shaker box contains 10,000 beads of which some are white and some are coloured. A random sample of 500 of these beads are chosen. 100 are found to be coloured. Construct a 99% confidence interval for the proportion of beads in the box that are coloured. ** Caution, watch the sample size. **

6. To test the durability of a new paint for white center lines, a highway department painted test strips across heavily traveled roads in eight different locations, and electronic counters showed that they deteriorated after having been crossed by (to the nearest hundred):

   142,600 167,800 136,500 108,300 126,400 133,700 162,000 149,400

   Construct a 95% confidence interval for the average amount of traffic (vehicles) the new white paint can withstand before it deteriorates. Assume that the data are approximately normally distributed.

7. A political party is interested in its current status among the electorate. A random sample of voters is to be polled regarding their current political preference. How many voters should be included in the poll so that the party will be off in its estimate from the true proportion by no more than 2% with a 95% degree of confidence.

8. An auditor is checking the accounts receivable journal of a firm. There are 10,000 entries in the journal. The auditor wishes to know the average value of an entry to the journal to within $40. How many entries should be included in the sample to have a 99% degree of confidence in the answer if he estimates the standard deviation of entries to be about $800?

3 Hypothesis Testing

3.1 The Logic Of Hypotheses Testing

3.1.1 The Null and Alternative Hypotheses

Statisticians use a very precise definition of the word "hypothesis". From a statistical point of view a hypothesis is an assumption made about a population parameter.

Grammatically, a hypothesis is always phrased as a declarative sentence. A declarative sentence is one, which has a truth-value. Hypotheses are either true or false but not both.

Example:
The following are examples of **Declarative Sentences**:

- It is sunny outside.
- The Passenger Pigeon is extinct.

The following are examples of sentences which are **Not Declarative**:

- Hi, how are you?
- Have a nice day!

From the above examples, we can see that a declarative sentence is either true or false but other types of sentences have no truth-value.

Hypothesis testing is a decision-making procedure that uses sample evidence to test the truthfulness of a statement about a population parameter.

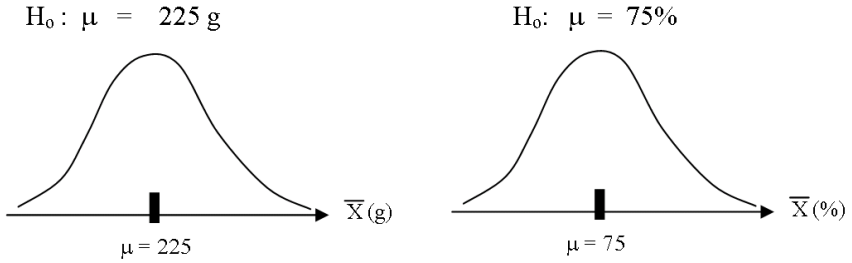
Example:
The following are examples of some hypotheses we may want to test:

1. The net contents of a can of tuna fish coming off of the production line averages 225 g.
2. Historically, the average mark scored by applicants on a personnel test is 75%.
3. The proportion of overweight children in Canada is 26% .
4. Most widgets produced in Canadian factories are within 5 grams of the average weight.

Sampling theory is used in hypothesis testing for the purpose of avoiding bias in making the decision. It is an important part of the process that the decision be arrived at based upon an initial impartial point of view. It has been said that statistics can be used to prove any position that you wish to take in a discussion. That can be the case if statistics procedures are used incorrectly.

The decision to be tested is called the **null hypothesis**, $\mathbf{H_0}$. In the null hypothesis, the parameter to be tested is always set equal to some assumed value. This assumed value is the centre of the sampling distribution curve on which we will base our decision. The word *null* means *no difference* in statistics. It is another way of saying, *equal to*.

Example:
The null hypotheses from examples 1 and 2 above would imply the following sampling distributions:



The **alternative hypothesis**, $\mathbf{H_a}$, is a hypothesis that will be accepted if the statistical evidence leads to the conclusion that the null hypothesis is probably false. There are three different arguments that can be put up in opposition to the statement that a parameter is equal to a certain value. We could conclude that it is *less than*, *greater than* or simply *not equal to* that value depending on the circumstances. The alternative hypothesis will always have one of these three mathematical relation symbols: $<$, $>$, or $\neq$.

Example:
Possible alternative hypotheses for the previous example 1 are:

$$H_a : \mu < 225 \text{ g} \qquad H_a : \mu > 225 \text{ g} \qquad H_a : \mu \neq 225 \text{ g}$$

The first step in decision-making is to formulate these two hypotheses so that we can gather some sample evidence and use it to examine how the sample observations compare to the assumed population parameter value. When we make our decision, it will either be *the evidence suggests that we reject* H_0 *and accept* H_a or it will be *there is no evidence to reject* H_0.

3.1.2 Type I and Type II Errors

The Risk of Making A Wrong Decision Based On Incomplete Information

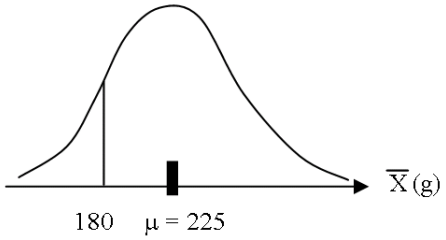
One of the objectives of decision-making is to minimize the risk of making the wrong decision. Here is a diagram that illustrates the pitfalls that can occur when making a decision based on sample evidence.

		Decision Made About H_0	
		Accept	**Reject**
Unknown State of H_0	**True**	Correct	Incorrect *(Type I error)*
	False	Incorrect *(Type II error)*	Correct

Sample evidence drawn from the population is used to decide the truthfulness of the null hypothesis about the population parameter. If the sample statistic deviates too far from the assumed population parameter, we say the difference is **statistically significant** and we reject H_0 . The probability of

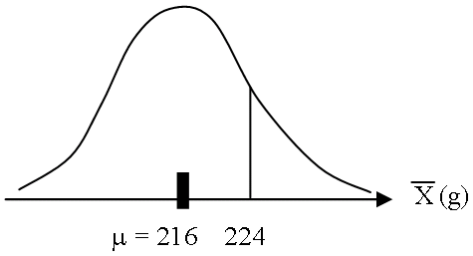
rejecting H_0 when it is true is called the **α risk**. This risk, also called the **level of significance**, is set arbitrarily low. Common values of α are 0.01, 0.05, and 0.10. These numbers represent the probability of making a **Type I error** in decision-making.

Example:
Use the previous example 1 situation and a sample of 50 cans of tuna fish. Suppose the average weight in the sample is 180 grams. We might reject the null hypothesis that the average weight of all cans of tuna fish is 225 grams because there is such a large difference between the values. It is possible just due to the random nature of sampling that we picked the 50 cans that had the lightest weight among all cans packaged and the true average is in fact 225 grams for all cans but that would be very unlikely.



The **β risk** is the probability of **accepting a false null hypothesis** on the basis of sample evidence. This type of error is called a **Type II error** in hypothesis testing.

Example:
Use the previous example 1 situation and a sample of 50 cans of tuna fish. Suppose the average weight in the sample is 224 grams. We might accept the null hypothesis that the average weight of all cans of tuna fish is 225 grams because there is such a small difference between the values. The average weight of all cans might in fact be 216 grams and through random selection we just happened to pick 50 cans that were very heavy relative to the distribution.



The interrelationship between the errors may be understood with a law metaphor. Suppose we are wishing to the evaluate the guilt of an accused person standing trial. In an English court of law the defendant is innocent until proven guilty. Let the null hypothesis H_0 be the hypothesis that the defendant is innocent. The alternative hypothesis H_a is that the defendant is guilty. In this case a Type I error (α), rejecting the null hypothesis when it is true, corresponds to finding the defendant guilty when he is innocent. This is considered a more serious error legally than a Type II error (β), accepting the null hypothesis when it is false, which corresponds to finding the defendant innocent when he is guilty. For purposes of this course, we need to be able to describe what a Type II error is. We will not deal with calculating the probability of making this type of error.

3.1.3 Evaluating the Evidence

Evaluation of the evidence is always the last consideration in making a statistical decision. The strategy for how a decision is to be made is set forth and then the sample evidence is gathered from the population based on random sampling techniques. The evidence is always weighed against the theoretical assumptions in order to judge the probability of observing these sample results using the theoretical statistical model.

Example:
In the previous example, the null hypothesis is that the average weight of all cans of tuna fish is 225 grams. We may decide to reject this only if the sample observation is in the bottom 1% of possible observations for the sample mean. The risk of rejecting a true null hypothesis would then be 1%, a very small risk. Having established this process, we would then go and examine a random selection of 50 cans of tuna fish.

Assignment:

Formulate the null and alternative hypotheses for the following situations. (This is the first step of any hypothesis test.)

1. An ambulance service is considering replacing its ambulances with new equipment. If \$52.00 is the average weekly maintenance cost of one of the old ambulances and μ is the average weekly maintenance cost it can expect for one of the new ones, what hypotheses should be used if the service wants to buy the ambulances only if it can be shown that this will significantly reduce the average weekly maintenance cost?
   - H_0:
   - H_a:
2. What hypotheses should the ambulance service use in problem 1 if the ambulance service is anxious to buy the new ambulances (which have some other nice features) unless it can be shown that they will significantly increase the average weekly maintenance cost?
   - H_0:
   - H_a:
3. A major Canadian brewery has held an 18% share of the market. However, because of an increased marketing effort, company officials believe the brewery's market share is now greater than 18%. What hypotheses should the company officials use if they want to prove that their market share has significantly increased?
   - H_0:
   - H_a:
4. What hypotheses should a quality improvement team use if it wants to make sure that wire leads are meeting a specification of 10 cm in length?
   - H_0:
   - H_a:

3.2 Testing Hypotheses About a Single Mean

Our first type of hypothesis testing involves a single parameter, We begin by testing a population mean, μ.

The theory that was presented in Section 3.1 is formulated into a formal six step procedure. All tests of hypothesis follow this process.

Example:
A promotional agency markets chocolate almonds through service clubs and charities. The agency claims that the packages contain 350 g. Evaluate this claim at the 1% level of significance if in a random sample of 49 boxes a mean of 338 grams with a standard deviation of 28 grams was measured.

Solution:

Step 1) Formulate the null and alternative hypotheses.

- $\mathbf{H_0 : \mu = 350\ g}$ $\Leftarrow$ Remember the null hypothesis must always have an "=" sign.
- $\mathbf{H_a : \mu < 350\ g}$ $\Leftarrow$ The only point of contention is when the consumer is short changed.

Step 2) State the level of significance.

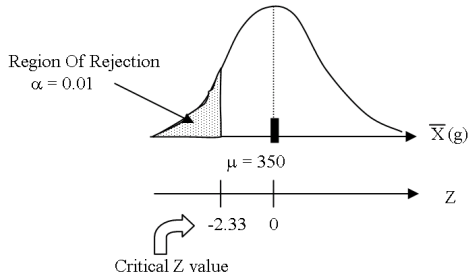
$\boldsymbol{\alpha = 0.01}$ $\Leftarrow$ This is some low value, here given as 1%.

Step 3) Determine the test statistic.

We will examine a large sample of boxes and determine their average weight so a ***Z*-test** applies because the distribution of the mean for large samples is a normal curve. If we were confident that the population of weights of all boxes was itself normally distributed and we examined a small sample of boxes, then a t test would apply.

Step 4) Establish a decision rule.

Draw a **sampling distribution curve** and determine the **critical Z value** which identifies the region in which the null hypothesis is rejected.



Here since $\alpha = .01$ is the tail area on the left side of a normal curve we find $Z_{\text{critical}} = -2.33$.

This is called a **one-tail test** because the region of rejection falls in one tail. If the alternative hypothesis has a "$\neq$" sign, there will be two tails of rejection and the probability of rejection (the significance) would be split equally between them.

A useful trick to identify the region of rejection is to look at the alternative hypothesis and consider the inequality as pointing to the region. (So $<$ means left, $>$ means right, and $\neq$ means left and right.)

Step 5) Evaluate the evidence.

Given the above data the estimated standard error in the mean is

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \frac{28 \text{ g}}{\sqrt{49}} = 4 \text{ g}$$

(No *F.C.F.* is required in calculating $s_{\bar{X}}$ since the population size N of all boxes produced by the factory is much larger than the sample so $n/N = 49/\text{Large} \approx 0 < .05$.)

The calculated Z-score for our measured sample mean is:

$$Z = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{338 - 350}{4} = -3.00$$

Step 6) State your decision.

Since the Z-value lies in the region of rejection, the evidence suggests, at a 1% level of significance, that we reject H_0 and accept H_a. In words, the mean weight is less than 350 grams.

Assignment:

The objective of this exercise is to gain familiarity with the formal testing procedure for hypothesis testing. Make sure to follow the process as outlined in the example.

1. An accountant reads a summary about a large journal which states that the average value of a journal entry is $560. If she takes a sample of 50 entries and finds a mean of $600 with a standard deviation of $80, evaluate the claim at a 10% level of significance.

2. Repeat the previous question, but now the accountant takes a sample size of only 20 and finds a mean of $595 with a standard deviation of $100. What assumption do we have to make about the population of entries before we can test the hypothesis?

3. A merchant sells butter which is advertised as weighing 454 grams. Establish a testing procedure to test the merchant's claim based on large samples and a level of significance of 5%. Evaluate his claim using sample evidence which shows that a random sample of 60 bricks averaged 447 grams with a standard deviation of 10 grams.

4. What is the probability of committing a Type I error in problem 3?

5. What would it mean to commit a Type II error in problem 3 in terms of the words of the problem?

6. An astronomer observes a group of over one thousand stars that appear to be clustered together in a small area of the sky. She believes that rather than being due to the random chance of stars being lined up along her line of sight, the stars appear clustered because they are actually near each other and likely have a common formation history. To test her theory she measures the *metallicity*[10]of several stars in the group with the following results (dex):

-0.50	0.00	0.02	0.06	0.08	0.09	0.11	0.12
0.14	0.19	0.19	0.20	0.22	0.25	0.30	0.38

   The astronomer consults a journal article that shows that stars in a larger region of the sky in which the cluster lies have a metallicity that is normally distributed with a mean of -.301 . Test the the claim, at a level of significance of 1%, that the cluster stars have this mean metallicity. What is the likely explanation of the star with metallicity -0.50?

7. According to the Canadian Mortgage and Housing Corporation the average rent for a two-bedroom apartment in October, 2010 in Regina was $881. Suspecting that the average rent for such dwellings has increased, a renter surveys 10 two-bedroom apartments and finds the following monthly rents ($):

750	800	810	880	910
925	990	1000	1010	1030

   Evaluate the claim, at the 5% level of significance using this data that the rent is still $881.

[10]Astronomers consider *metals* to be any element other than hydrogen and helium! After the Big Bang, the universe was primarily hydrogen and helium, metals only forming later in stars, such as in supernova events. A high metallicity therefore suggests a later generation star, such as our sun, which has formed out of such metal-enriched matter. Metallicity is written as [Fe/H], which represents the logarithm of the ratio of a star's iron abundance, easily identified in the star's light spectrum, compared to that of the Sun. A star with metallicity 0 has the same iron abundance as our sun, while metallicity of 1 would be 10 times solar, and -1 would be $1/10^{\text{th}}$ solar.

3.3 Testing Hypotheses About a Single Proportion

We next consider tests involving a single population proportion, π. For our purposes sample sizes must be large, as discussed in Section 1.5, so that we may assume that sample proportions are normally distributed. The six step hypothesis testing procedure applies here as well.

Example:
A government has a policy to provide funding to political parties with more than 4% support. A new political party emerges claiming that they have the required popular support. Assuming the government will evaluate such a claim with a poll, devise a statistical test at the 5% level of significance to determine whether the new party has the required support.

Solution:

Step 1) Formulate the null and alternative hypotheses.

- $H_0 : \pi = 0.04$ $\Leftarrow$ The new party support is 4%.
- $H_a : \pi < 0.04$ $\Leftarrow$ The government official doubts they have achieved this level of support.

Step 2) State the level of significance.

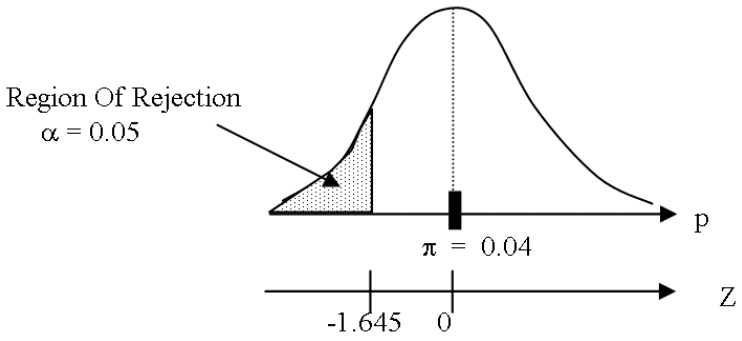
$\alpha = 0.05$ $\Leftarrow$ As directed in the problem.

Step 3) Determine the test statistic.

The official plans on using a sample of size $n = 500$ for the hypothesis test so, assuming the null hypothesis is true, $n\pi = (500)(.04) = 20$ and $n(1-\pi) = (500)(.96) = 480$ which are both greater than 5. We can therefore assume that p will be approximately normal normally distributed and thus a Z test applies. (For this course in tests about proportions this will always be the case because large size samples will always be used.)

Step 4) Establish a decision rule.

Draw a sampling distribution curve and determine the critical Z value which identifies the region in which the null hypothesis is rejected.



Here for the given area, $Z_{\text{critical}} = -1.645$.

Step 5) Evaluate the evidence.

To evaluate the claim, a government official does a poll of 500 individuals and find the new party has the support of 15 individuals.

The observed sample proportion is: $p = \dfrac{X}{n} = \dfrac{15}{500} = 0.03$

The standard error of proportion is: $\sigma_p = \sqrt{\frac{\pi \cdot (1-\pi)}{n}} = \sqrt{\frac{(0.04) \cdot (0.96)}{500}} = 0.008764$

(No *F.C.F.* was required to calculate σ_p as the population of eligible voters is assumed much larger than 500.)

The calculated Z value is: $\boxed{Z = \frac{p - \pi}{\sigma_p}} = \frac{0.03 - 0.04}{0.008764} = -1.141$

Note that we have used the population parameter π of the null hypothesis in the calculation of the standard error since we are assuming the null hypothesis is true and knowing π is sufficient to calculate σ_p .

Step 6) State your decision.

Since the Z value does not lie in the region of rejection we fail to reject the null hypothesis H_0. At a level of significance of 5% there is no evidence to suggest the new party does not have the required 4% support.

A critic of the official who analyzed the support of the new party questions how it is that the official is concluding the new party has the required 4% support when the poll only indicated 3%. What could the official have done to make a more convincing evaluation of the support?

Assignment:

1. A television network sells advertising based on the proportion of the large metropolitan area it reaches that will watch a given program. A company buys advertising during a particular show with the expectation that 40% of households will be watching the program. After the program airs the company samples 100 households and finds that only 30 watched the show. If the 10% level of significance is used, can the null hypothesis that 40% of the households would watch the program be rejected?

2. In 2008, 68% of Canadian women were unable to identify that the leading cause of death among women was stroke and heart disease. From 2008 onward the Heart and Stroke Foundation ran *The Heart Truth* campaign to raise awareness of the severity of the issue and the warning signs associated with stroke and heart disease. In 2011 if a Harris-Decima poll of 1,013 Canadian women found 53% of Canadian women unable to identify their leading cause of death, can we conclude the proportion of women unable to identify their leading cause of death decreased over the three year interim period at the 2% level of significance?

3. A counselor claims that at least 30 percent of the students attending a local college hold a part-time job. Doubting the claim, a college official does a random sample of 75 students finds that 15 have part-time jobs. Is there adequate evidence to suggest that the counselor is incorrect in her claim at the 5% level of significance. The college has approximately 2000 students.

4. In an effort to cut costs a computer chip manufacturer replaces a semiconducting material in an existing component with a cheaper material. The manufacturer wishes to test the effect, if any, of such a change. With the old material the failure rate of the component was 7%. If in a batch of 200 of the new components there are 16 failures, test the claim that the new failure rate is still 7% at the 2% level of significance.

5. To meet a government regulation a gasoline retailer claims to contain at least 6% ethanol by volume in its gasoline. A government regulator wishes to test this claim at the 5% level of significance. Would she use a proportion hypothesis test? Explain your answer.

3.4 Review Exercises On One Parameter Hypothesis Testing

** Use the formal six step procedure for hypothesis testing to do the following problems. **

1. A building products association sets standards for building sizes of products used in building construction. To set the standard for door heights, the association is interested in knowing if the average height of adult males in North America has increased from its previous value of 178 cm. A sample of 121 males shows an average height of 181 cm with a standard deviation of 5cm. Can an increase in the average height be supported at the 5% level of significance?

2. A trucking firm claims that the weights of trucks it sends over a municipal road is 25,000 kg. Test the claim at the 1% level of significance if a sample of 100 random weighings of the trucks shows a mean weight of 28,000 kg with a standard deviation of 3000 kg.

3. It is thought that half of all winners of major prizes in raffles held at summer fairs are males. A random sample of 100 prize winners shows that 40 are male. At the 10% level of significance can we conclude that less than half of all prize winners are male?

4. A drink dispenser is supposed to be set to fill an average of 300 ml of beverage in a cup. Using a 1% level of significance, should the drink dispenser be adjusted if a random sample of 8 cups filled showed the following volumes (ml):

   280 290 321 315 302 275 318 302

   (Assume drink dispensations are approximately normally distributed.)

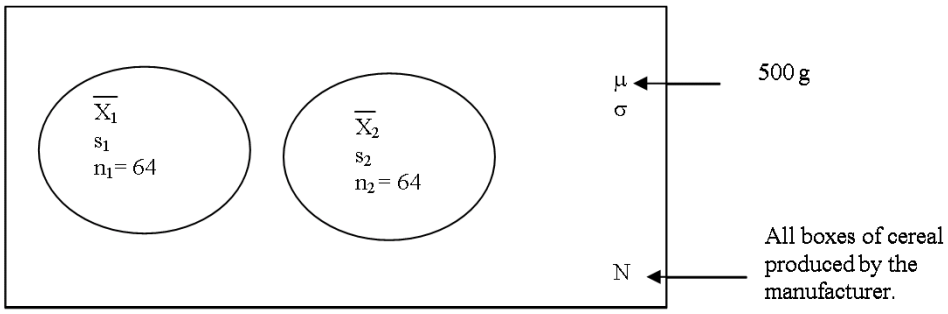
5. A large manufacturing company investigated the service it received from suppliers and discovered that, in the past, 68% of all materials shipments were received on time. However, the company recently installed a just-in-time system in which suppliers are linked more closely to the manufacturing process. A random sample of 110 deliveries since the just-in-time system was installed reveals that 81 deliveries were on time. Use this sample information to test whether the proportion of on time deliveries has significantly increased at the 5% level.

3.5 Testing Hypotheses About the Difference Between Means

Sometimes we are interested in testing whether two means are different. For instance, imagine that a breakfast cereal manufacturer produces the same boxes of cereal at two different plants. Suppose the manufacturer suspects that the mean weight of a box coming from plant A actually differs from the mean weight of a box at plant B. Obviously one cannot test every box of cereal produced by each plant to calculate their respective mean weights to see if they are the same, so a sample would be taken at each and its mean value calculated. These sample means would have some error from their true means, so a difference in their values would not necessarily imply a difference in the population means of the different plants. How can we resolve this? We start by making our null hypothesis that the two populations really are the same, with same mean and standard deviation. If that were the case our two samples would be just two of many possible samples from the one large population and we can ask how the difference of two such means will be distributed. If our measured difference in that scenario appears extraordinary then we will reject the null hypothesis.

In order to test a hypotheses about the difference in population means, it is therefore necessary to analyze the distribution of the difference between sample means drawn from the same population.

Example:
Suppose that two different samplers each select a random sample of 64 boxes of a popular breakfast cereal. The stated net contents on the box is 500g.



Each sampler has the possibility of observing ${}_NC_{64}$ different sample means. The C.L.T. describes the behaviour of each of these distributions. If we were able to list the joint possibilities for the samplers and calculate the difference in their results we might see a table such as this:

$\overline{X}_1$	$\overline{X}_2$	$\overline{X}_1 - \overline{X}_2$
512	502	+10
490	504	-14
497	492	+5
---	---	---
---	---	---

The distribution of the difference between sample means

There will be the same number of observable possible differences as there are individual possible paired samples. The difference of sample means $\bar{X}_1 - \bar{X}_2$ is itself a variable over the population comprised of paired samples.

3.5.1 Large Independent Samples

As just demonstrated, the difference of sample means is itself a variable over the population comprised of paired samples. The distribution of the difference between sample means $\bar{X}_1 - \bar{X}_2$ from the same population has these three characteristics:[11]

1. If the sample size is large and the samples are drawn independently, the shape of the distribution will be very close to normal.
2. The mean difference will be zero.
$$\mu_{\bar{X}_1 - \bar{X}_2} = 0$$
3. The standard error of the difference between means **in theory** is found by this formula:
$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

   In practice, the standard deviation of the population, σ, is unknown. In the case of large samples each sample standard deviation is used as an estimate of the population value. In practice the standard error formula therefore is:
$$\boxed{s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Example:
Management at a large company wishes to know if there is a significant increase in the average sales performance between salespeople who receive a certain type of training over sales people who do not receive the training. We wish to test this at the 5% level of significance based on large independent samples.

Solution:

Step 1) Formulate the null and alternative hypotheses.

The null hypothesis would take the view that the populations of sales people were the same. That is that there should be no difference between the means in the populations. The hypotheses statements must be formulated as differences so that we can use the sampling distribution for the test. Let the subscript T refer to trained and NT to not trained.

- $\mathrm{H}_0 : \mu_T - \mu_{NT} = 0$
- $\mathrm{H}_\mathrm{a} : \mu_T - \mu_{NT} > 0 \;\Leftarrow$ We suspect $\mu_T > \mu_{NT}$ which implies this.

Step 2) State the level of significance.

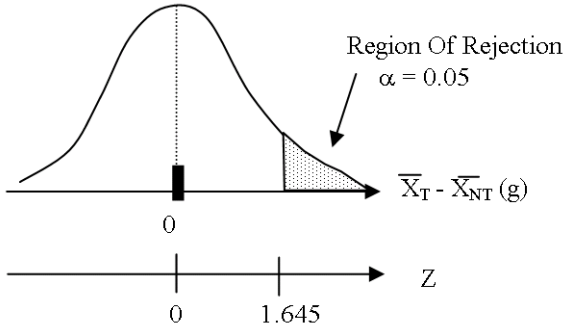
Use an $\alpha = 0.05$ as directed in the problem.

Step 3) Determine the test statistic.

A Z test applies because the decision will be based on large, independent samples.

[11] The sum (difference) of two normal variables X and Y is itself normally distributed with a mean which is the sum (difference) of the means, $\mu_{X \pm Y} = \mu_X \pm \mu_Y$, and a variance satisfying $\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$. In our case the two variables are $\bar{X}_1$ and $\bar{X}_2$. Since our samples are of large size, these are, by the C.L.T., normally distributed with means both equalling μ and standard deviations approximated by the sample standard errors in the mean, $s_1/\sqrt{n_1}$ and $s_2/\sqrt{n_2}$, from which our results above follow. It is to be noted that the general property, namely that the variance of the sum or difference of two normal variables is added in quadrature, is also useful in combining error estimates in calculations.

Step 4) Establish a decision rule.



For the given tail area we have $Z_{\text{critical}} = 1.645$.

Step 5) Evaluate the evidence.

Suppose 60 salespeople were trained and compared to 80 salespeople who were not trained. The sample results are summarized in the following table:

	Mean Weekly Sales ($)	Standard Deviation ($)	Number In Group
Trained Group	5000	800	60
Untrained Group	4700	1100	80

The observed difference in means is:

$$\bar{X}_T - \bar{X}_{NT} = \$5000 - \$4700 = \$300$$

To test the significance of this difference we must see where it lies on the sampling distribution curve.

$$s_{\bar{X}_T - \bar{X}_{NT}} = \sqrt{\frac{\bar{X}_T^2}{n_T} + \frac{\bar{X}_{NT}^2}{n_{NT}}} = \sqrt{\frac{800^2}{60} + \frac{1100^2}{80}} = \$160.60$$

The Z value for this sampling distribution is calculated by:

$$\boxed{Z = \frac{\left(\bar{X}_T - \bar{X}_{NT}\right) - 0}{s_{\bar{X}_T - \bar{X}_{NT}}}} = \frac{300 - 0}{160.60} = +1.87$$

Step 6) State your decision.

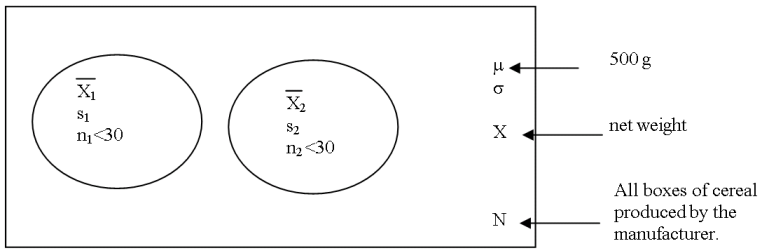
Because the observed Z value lies in the region of rejection the decision would be to reject H_0: Based on the evidence, at a 5% level of significance, the average weekly sales for trained salespeople is higher than the average weekly sales for untrained salespeople.

(If we had a calculated Z of a value like +0.87, the decision would be: There is no evidence to suggest a difference in the average weekly sales between trained and untrained salespeople.)

3.5.2 Small Independent Samples

In certain cases it is too costly or time consuming to draw large samples in order to test assertions about the difference between population means. If smaller sample sizes are used, there is another sampling distribution used in the decision rule.

Example:
Take the case of the two samplers of cereal boxes from the Section 3.5 and let us suppose that they drew samples less than 30 in size.

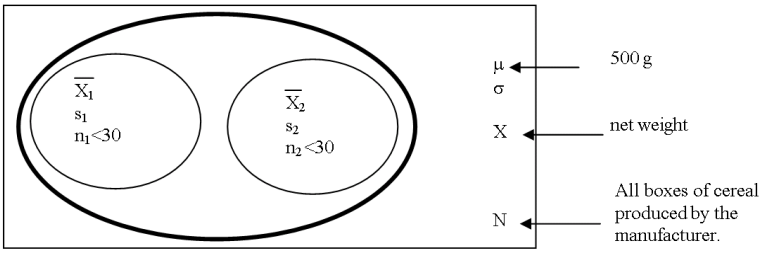


In this case the difference of sample means will not have a normal distribution because the sample sizes are not large enough. However, we do know from Section 2.4 that if the underlying distribution of the weights, X, is normal then the sample means will each have a t distribution. It turns out that this will also be the case for the difference of the sample means.

If we were able to draw the distribution curve of the net contents of all the boxes manufactured and it resulted in a bell curve, the distribution of the difference between sample means from the same population has these three characteristics:

1. When the sample size is small and the samples are drawn independently, the shape of the distribution will be a t curve with
$$\boxed{df = n_1 + n_2 - 2}$$
degrees of freedom.
2. The mean difference will be zero.
$$\mu_{\bar{X}_1 - \bar{X}_2} = 0$$
3. The standard error of the difference between means **in theory** is found by this formula:
$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

In practice, the standard deviation of the population, σ, is unknown. In the case of small samples, the variability in each sample is pooled into one large sample to obtain an estimate of the population standard deviation. The pooled estimate of σ is s_{Pool}.



Recall from page 30 that the variance of a sample can be defined as:

$$s^2 = \frac{\sum \left(X - \bar{X}\right)^2}{df}$$

The pooled estimate of variance, s^2_{Pool} is defined using the above definition:

$$s^2_{\text{Pool}} = \frac{\sum \left(X - \bar{X}_1\right)^2 + \sum \left(X - \bar{X}_2\right)^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2}$$

Substitute the pooled estimate into the theoretical standard error formula estimate

$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{s^2_{\text{Pool}}}{n_1} + \frac{s^2_{\text{Pool}}}{n_2}} = \sqrt{s^2_{\text{Pool}} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]},$$

to obtain

$$\boxed{s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}$$

Example:
Platinum is a very expensive metal. A purchaser wishes to know if there is any significant difference in the mean weight of a platinum item supplied by Supplier A versus those supplied by supplier B. It is assumed that the weight of platinum parts supplied by all suppliers is distributed normally. Because of the expense involved, a small random sample of parts will be tested from each supplier. Test the assumption that there is no difference in the weights of parts supplied by these two suppliers at the 10% level of significance if 12 items are tested from each supplier.

Solution:

Step 1) Formulate the null and alternative hypotheses.

- $\text{H}_0 : \mu_A - \mu_B = 0$
- $\text{H}_\text{a} : \mu_A - \mu_B \neq 0$ $\Leftarrow$ We suspect $\mu_A \neq \mu_B$ which implies this.

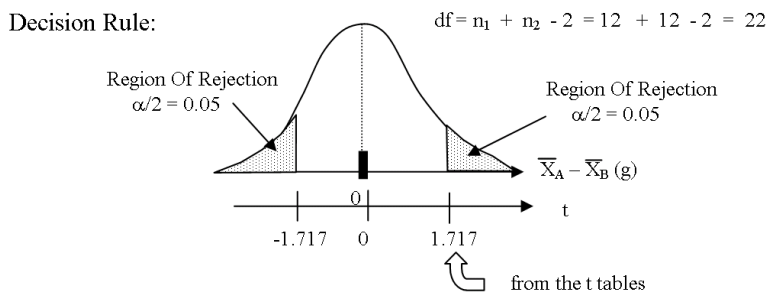
Step 2) State the level of significance.

Use an $\alpha = 0.10$ as directed in the problem.

Step 3) Determine the test statistic.

A t test applies because the decision will be based on small, independent samples drawn from a normal population.

Step 4) Establish a decision rule.



For our given tail area and using a t distribution we have $t_{\text{critical}} = \pm 1.717$.

Step 5) Evaluate the evidence.

The sample results are summarized in the following table:

Supplier	Mean Weight (g)	Standard Deviation (g)	Number
A	17	2	12
B	20	4	12

The observed difference in means is:

$$\bar{X}_A - \bar{X}_B = 17 \text{ g} - 20 \text{ g} = -3 \text{ g}$$

To test the significance of this difference we must see where it lies on the sampling distribution curve.

Calculate the standard error of the sampling distribution.

$$\begin{aligned} s_{\bar{X}_A - \bar{X}_B} &= \sqrt{\frac{(n_A - 1) \cdot s_A^2 + (n_B - 1) \cdot s_B^2}{n_A + n_B - 2} \cdot \left[\frac{1}{n_A} + \frac{1}{n_B}\right]} \\ &= \sqrt{\frac{(12 - 1) \cdot 2^2 + (12 - 1) \cdot 4^2}{12 + 12 - 2} \cdot \left[\frac{1}{12} + \frac{1}{12}\right]} \\ &= 1.29 \text{ g} \end{aligned}$$

Calculate the t value for the observed sample difference.

$$\boxed{t = \frac{\left(\bar{X}_A - \bar{X}_B\right) - 0}{s_{\bar{X}_A - \bar{X}_B}}} = \frac{-3 - 0}{1.29} = -2.32$$

Step 6) State your decision.

Because the observed t value lies in the region of rejection the decision would be to reject H_0: Based on the evidence, at a 10% level of significance, there is a difference in the mean weight of the platinum parts supplied by these two suppliers.

Assignment:

1. A researcher wishes to test if there is a difference between the average monthly household income in two large communities. A random sample of 40 households in the first community has a mean of \$1,900 with a standard deviation of \$540. For the second community a sample of 30 households has a mean of \$1,600 with a standard deviation of \$420. Using the five percent level of significance, test the hypothesis that there is no difference between the average monthly household income in the two communities.

2. The Moose Jaw Sporting Goods Company wishes to evaluate the use of advertising to increase sales. To do so they placed an advertisement on television that was shown three times during the live broadcast of the Saskatchewan Roughriders' first away game. The sales for each of the seven days after the ad was placed to be compared with the sales for the seven days immediately before running the ad. The following data representing the total dollar sales each day were collected:

Sales Before Ad (\$)	Sales After Ad (\$)
1,765	2,045
1,543	2,456
2,867	2,590
1,490	1,510
2,800	2,850
1,379	1,255
2,097	2,255

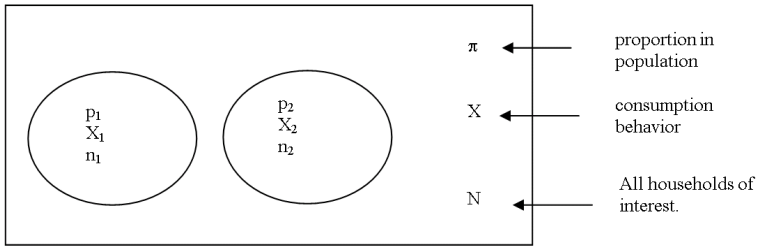
   Based on the sample data, would you conclude that there was a significant increase in sales after the advertising? Test using an alpha level of 0.01. Assume the population of sales is normally distributed.

3. In an attempt to evaluate the hypothesis that a "good night's sleep" gives better performance, a researcher created two groups of subjects. Those with regular sleep (group R) were allowed to sleep up to 7 hours maximum at night, while the second group with enhanced sleep (group E) were allowed up to 9 hours of sleep. After 1 week all individuals in the study were timed to see how long it took each of them to accomplish the same task involving motor skills and coordination. The 40 individuals in the regular sleep group had a mean time of 53.1 seconds with a standard deviation of 8.0 seconds, while the 50 in the enhanced sleep group took an average time of 51.2 seconds with a standard deviation of 9.0 seconds to accomplish the task. Can we concluded that the average time required by a person with enhanced sleep (E) is significantly less than that of a person with regular sleep (R) at the 10% level of significance?

3.6 Testing Hypotheses About the Difference Between Proportions (Large Samples)

As with the comparison of two means, we may also be interested in comparing to proportions. For instance, we might be interested in knowing if two different groups have a different incidence of diabetes. If we could test each member of the two groups to see if they had diabetes we could work out the proportions to see if they are equal or not. However this is too costly so we want, instead, to sample each group and get an estimate of the proportions and compare these. However, due to the sampling error, each such proportion differs from its true group's proportion and thus a difference in sample proportions does not imply a difference in the two actual proportions. Once again we proceed by assuming a null hypothesis that there is no difference between the actual proportions of the two groups. The two samples would then just belong to one larger population with the same proportion. We can then consider how differences between such sample proportions are distributed. If we find that our measured difference is exceptional, we will reject the null hypothesis. We thus need to know how the difference between to sample proportions drawn from the same population is distributed.

Example:
Take the case of two market analysts who sample a population of households to determine the proportion of households that consume more than 2 packages of potato chips per week.



Sampler 1 could find any of ${}_NC_{n_1}$ possible sample proportions while Sampler 2 could find any of ${}_NC_{n_2}$ possible sample proportions. Assuming the sample sizes are large enough these sample proportions will each be normally distributed. If we were able to list the joint possibilities from the samplers and calculate the difference in their results we could see a table like this:

p_1	p_2	$p_1 - p_2$
.32	.29	.03
.22	.26	-.04
.31	.30	.01
- - -	- - -	- - -
- - -	- - -	- - -

The difference between proportions $p_1 - p_2$ is itself a variable over the population of paired samples.

If two samples are drawn from the same population and the sample sizes are sufficiently large so that $np > 5$ and $n(1-p) > 5$ the distribution of the difference between the sample proportions $p_1 - p_2$ has these three characteristics:[12]

1. The difference will be normally distributed.

[12]These results follow from the properties of combining normal variables outlined in the footnote on page 48 and the known standard deviation and mean of the proportion sampling distributions.

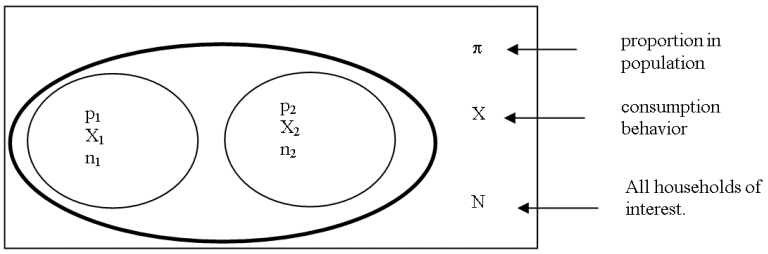
2. The mean difference will be zero:

$$\mu_{p_1-p_2} = 0$$

3. The standard error of the difference between proportions **in theory** is found by this formula:

$$\sigma_{p_1-p_2} = \sqrt{\frac{\pi \cdot (1-\pi)}{n_1} + \frac{\pi \cdot (1-\pi)}{n_2}}$$

In practice, the proportion of the population, π, is unknown. In the case of small samples, the variability in each sample is pooled into one large sample to obtain an estimate of the population proportion.



The pooled estimate of π is p_{Pool} given by

$$\boxed{p_{\text{Pool}} = \frac{X_1 + X_2}{n_1 + n_2}}$$

Substitute the pooled estimate into the standard error formula to obtain:

$$s_{p_1-p_2} = \sqrt{\frac{p_{\text{Pool}} \cdot (1-p_{\text{Pool}})}{n_1} + \frac{p_{\text{Pool}} \cdot (1-p_{\text{Pool}})}{n_2}}\ ,$$

or, as found on the formula sheet,

$$\boxed{s_{p_1-p_2} = \sqrt{p_{\text{Pool}} \cdot (1-p_{\text{Pool}}) \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}}\ .$$

Example:
A biologist believes that during afternoon hours a greater proportion of elk that are active are female than in the evening. Test this claim at the 1% level of significance.

Solution:

Step 1) Formulate the null and alternative hypotheses.

- $\text{H}_0 : \pi_A - \pi_E = 0$ $\Leftarrow$ There is no difference in the proportions (equivalent to $\pi_A = \pi_E$)
- $\text{H}_\text{a} : \pi_A - \pi_E > 0$ $\Leftarrow$ The biologist suspects $\pi_A > \pi_E$ which implies this.

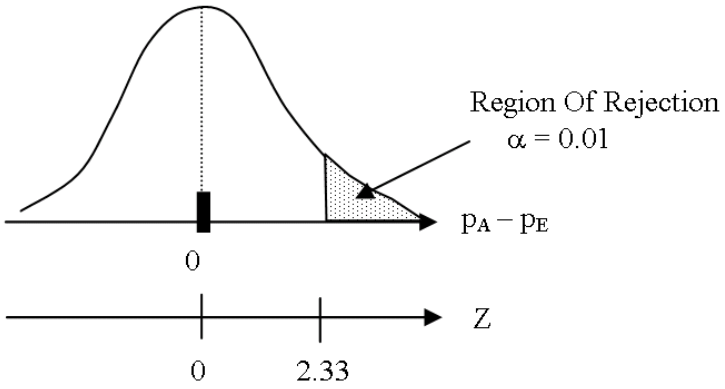
Step 2) State the level of significance.

Use an $\alpha = 0.01$ as directed in the problem.

Step 3) Determine the test statistic.

A Z test applies because the decision will be based on large sample observations of proportion. Specifically we check (see evidence below) that $n_A p_A = (350)(.643) = 225$, $n_A(1-p_A) = (350)(.357) = 125$, $n_E p_E = (200)(.500) = 225$, and $n_E(1-p_E) = (200)(.500) = 100$ are all greater than 5.

Step 4) Establish a decision rule.



For our given tail area and using the normal distribution we have $Z_{\text{critical}} = 2.33$.

Step 5) Evaluate the evidence.

The biologist observes elk that are active in the afternoon and in the evening. The two sample results are summarized in the following table:

Time	Number of Females	Number of Males	Total Number
Afternoon	225	125	350
Evening	100	100	200
Total	325	225	550

The observed difference in proportions is: $p_A - p_E = \dfrac{225}{350} - \dfrac{100}{200} = 0.643 - .500 = 0.143$

To test the significance of this difference we must see where it lies on the sampling distribution curve.

Calculate the pooled estimate of the population proportion.

$$p_{\text{Pool}} = \frac{X_A + X_E}{n_A + n_E} = \frac{225 + 100}{350 + 200} = 0.591$$

Calculate the standard error of the sampling distribution.

$$s_{p_A - p_E} = \sqrt{p_{\text{Pool}} \cdot (1 - p_{\text{Pool}}) \cdot \left[\frac{1}{n_A} + \frac{1}{n_E}\right]} = \sqrt{0.591 \cdot (1 - 0.591) \cdot \left[\frac{1}{350} + \frac{1}{200}\right]} = 0.04358$$

Calculate the Z-value for the observed sample difference.

$$\boxed{Z = \frac{(p_A - p_E) - 0}{s_{p_A - p_E}}} = \frac{0.143 - 0}{0.04358} = 3.28$$

Step 6) State your decision.

Because the observed Z value lies in the region of rejection the decision would be to reject H_0: Based on the evidence, at a 1% level of significance, the proportion of active elk which are female is greater in the afternoon than in the evening.

Assignment:

1. A counselor at a college suspects that a higher proportion of business students holds part-time jobs than do science students. To test the hypothesis, a random sample of 100 business students was taken and 30 had part-time jobs, while a sample of 150 science students found 36 with part-time jobs. The two programs have large enrolments (so use of an *F.C.F.* is not required.) Can the counselor confirm his suspicion at the 5% level of significance?

2. In the 1975 comedy "Monty Python and the Holy Grail", it is suggested that there would be a difference between the ability of an African swallow and a European swallow to carry a coconut over a large distance. To test the claim, an ornithologist (with too many financial resources and no ethics) attaches coconuts to 40 African swallows and 36 European swallows and releases them from a common location. Using transmitters he has inserted in the coconuts the researcher is able to determine that 18 of the African and 9 of the European swallows were able to travel a distance of over 100 km with coconuts intact.

   (a) Can the researcher conclude at the 20% level of significance that there is a difference in the ability of African and European swallows to travel laden with coconuts?

   (b) How would the problem change if one had hypothesized that African swallows were better coconut carriers than European ones (at the 20% level of significance)?

3.7 Review Exercises on Two Parameter Hypothesis Testing

** Use the formal six step procedure for hypothesis testing to do the following problems. **

1. A study of financial planning services showed that 380 of 1160 people who were over 45 years of age preferred an accountant for assistance in their financial planning. Of those 45 years of age and less, 120 of 470 surveyed preferred an accountant for assistance in their financial planning. Using an alpha risk of 1%, is there a difference between the two groups in the proportion who prefer an accountant for financial planning.

2. A firm hires a large number of professional people to carry out the services it offers. At salary negotiation time, the question is always raised as to whether the professionals at this firm earn similar wages on average to professionals in other firms who do similar duties. Can the firm conclude at the 1% level that its professionals earn less based on the following sample results?

Professionals	Average Wage($/yr)	Std Dev($/yr)	Number
Internal	45000	5000	15
External	48000	4000	17

3. An insurance company wishes to know whether foreign cars are more expensive to repair than domestic cars on average. Can their claim be supported at the 5% level if the following sample results are observed?

	Average Repair ($)	Std Dev ($)	Number
Domestic	8000	2000	50
Foreign	9500	3100	60

4 Paired Data Analysis

4.1 Introduction To Paired Data Analysis

Paired data analysis refers to those situations where two data sets are analyzed to see if there are any trends or relationships between the variables.

Example:
Consider the following questions:

1. Is there any relationship between the variation in time spent on training an employee and the performance of salespeople on the job?
2. Is there any relationship between the variation in money a government spends on the military (per capita) and the number of times the government goes to war over the same time period?
3. Is there any relationship between the variation in the amounts of money spent on advertising in an area and the sales volume in that area?
4. Is there any relationship between the variation in the frequency of numbers that are played in a lottery and the relative frequency of numbers that win in the lottery?
5. Is there any relationship between the period of oscillation of a Cepheid variable star and its intrinsic brightness (absolute magnitude)?

Each of the above examples contains two quantities which vary. The variability in these quantities is pictured by way of a mathematical graph. In paired data analysis a theoretical mathematical curve is assumed to occur between variables. When a set of paired data is analyzed by the technique of **linear regression**, the assumption is made that there is a straight line which best describes the relationship between the data sets. We will restrict ourselves to this type of curve fitting.

Example:
A municipal parks department is interested in knowing the relationship between the price of a lawnmower in their fleet of small lawnmowers and the number of breakdowns it has over a five year period. A random sample of 10 lawnmowers are studied. The observations are summarized as follows:

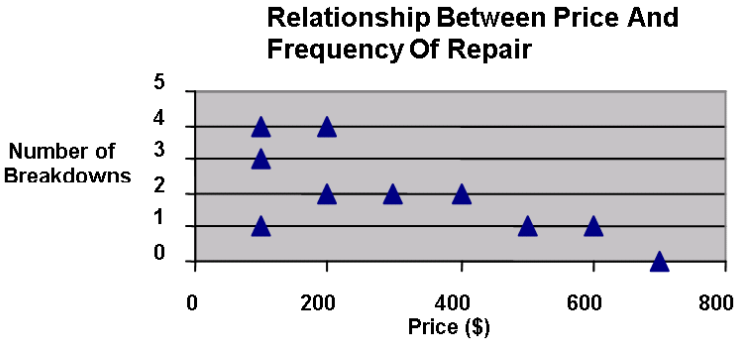
Price($)	600	300	200	500	100	700	100	400	100	200
# of Breakdowns	1	2	2	1	4	0	3	2	1	4

The attributes of each lawnmower are identified by two numbers which combine to make an **ordered pair**. In statistical analysis, as in other disciplines, the ordered pairs are plotted on a **coordinate plane**. The variable plotted on the horizontal axis is called the **independent variable** while the variable that is plotted on the vertical axis is called the **dependent variable**. For purposes of later calculations, the independent variable is always identified by the symbol X while the dependent variable is always identified by the symbol Y. The ordered pairs formed by the paired data are plotted in the form of a **scatter diagram** on the coordinate axes. It is important that the dependent variable be plotted on the vertical axis.

In regression analysis the objective is to predict the value of Y given a value of X. The dependent variable can always be identified because it is the quantity that is to be predicted in the relationship.

The predicted value of Y will be designated by Y_p to distinguish from the Y coordinate of a potential data point or points with the same X coordinate.

Example:
In the previous example the variable X is the price while Y is the number of breakdowns. This is because the purchaser is in control of how much he wishes to spend on the lawnmower which is the nonrandom input, while the number of breakdowns is that which is uncertain or random, perhaps dependent upon the input price, and what we wish to predict. The ordered pair corresponding to the third lawnmower is $(X, Y) = (\$200, 2 \text{ breakdowns})$. A scatter diagram of the data is



The statistical population in paired data analysis is the collection of all possible ordered pairs that could be observed. The characteristics of this population are described by its **parameters**, just as before we found parameters for describing data belonging to a single variable X such as the mean, μ.

For data with a linear trend, such as the lawnmower data above, it is possible to characterize the population by a best fit straight line which runs through the data. Since the data consists of more than two points, any line can only approximate the points. Note, however, that our goal is not to reconstruct the random scatter of the data with our line, but rather to give an indication of what we expect for a given input value X.

A straight line can be described completely by two parameters; the **slope** and the **vertical axis intercept**.

The vertical axis intercept can be determined by inspection of the graph or by calculating the value of Y when $X = 0$. The intercept parameter is designated by the symbol α in statistics. Recall parameters are designated by Greek letters to show their theoretical nature.

The definition of the slope of a line is the ratio of the rise to the run between any two points along the line. The slope parameter is designated by the symbol β in statistics, and satisfies the relation:

$$\beta = \frac{\Delta Y}{\Delta X}$$

The equation, in terms of its parameters, which governs the relationship between variable X and variable Y along a straight line always takes the form:

$$Y_p = \alpha + \beta X$$

When analyzing single variables we saw how a sample from the population produced an estimate, such as $\bar{X}$ of the true, and usually unknown, population parameter μ. In the same way a sample data set, such as for our lawnmowers, will only produce an estimate of the best-fit line that would theoretically exist through the entire population of lawnmower data. As such a best fit line to the sample data will

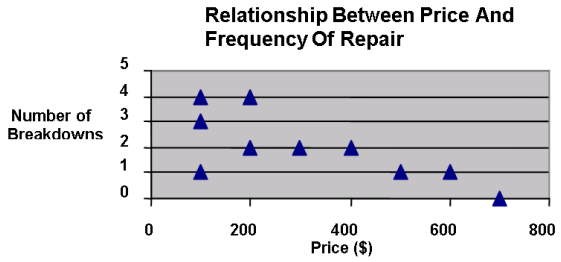
be written

$$\boxed{Y_p = a + bX}$$

where a estimates α, and b estimates β.[13]

Example:
For the lawnmower scatter diagram, use a ruler to draw a best fit straight line through the data. Approximate a by examining the intersection with the vertical axis and b by drawing a triangle of appropriate size and calculating its rise and the run. Fill in your equation for Y_p and use it to predict the number of breakdowns that the municipality would expect if it pays \$400 for a lawnmower.

Solution:



- $a \approx$
- $b \approx \dfrac{\Delta Y}{\Delta X} =$
- $Y_p \approx \left(\quad\right) + \left(\quad\right) X$
- $Y_p \approx \left(\quad\right) + \left(\quad\right) (\$400) =$

Note that your predicted value for the number of breakdowns, Y_p, at $X = \$400$ likely is not the same as the data Y-value, $Y = 2$ breakdowns, found at that same X.

4.2 The Least Squares Criteria

Cases where data points do not line up along a straight line are said to be random in nature. Through the random scatter in our lawnmower example we can discern a downward linear trend. The regression equation

$$\boxed{Y_p = a + bX}$$

provides us with a model for the underlying straight-line trend through the random scatter. We have seen how we could by eye estimate a linear relationship for our data. Statisticians draw this straight line using what is called **the least squares criteria**. The least squares line is that straight line through the data that minimizes the sum of the squared deviations from the observed points to the line.[14] This criteria will be met if we use the statistics a and b to estimate α and β, where a and b are calculated as follows:

$$\boxed{a = \frac{\left(\sum X^2\right) \cdot \left(\sum Y\right) - \left(\sum X\right) \cdot \left(\sum XY\right)}{n \cdot \left(\sum X^2\right) - \left(\sum X\right)^2}}$$

$$\boxed{b = \frac{n \cdot \left(\sum XY\right) - \left(\sum X\right) \cdot \left(\sum Y\right)}{n \cdot \left(\sum X^2\right) - \left(\sum X\right)^2}}$$

Here n is the number of pairs in our sample. These two calculations may be done tabularly.

[13] The student should note that our statistical notation deviates from that commonly used for the slope-intercept form of the line, namely $y = mx + b$. Our b in statistics is the slope not the y-intercept of the line.

[14] This is a mathematical criteria whose solution can be found using mathematics beyond the scope of this course. For our purposes it is enough to know that a unique solution to the problem always exists and is given by the formulae provided.

Example:
For the lawnmower example find the linear regression line. Plot the line on the scatter diagram. Use the regression line to find the predicted number of breakdowns for X=\$400.

Solution:
Noting the required summations of our intercept and slope formulae, the following table will provide us with the information to do the two calculations.

X(\$)	Y(breakdowns)	XY(\$ · breakdowns)	X^2 (\$^{2})
600	1	600	360000
300	2	600	90000
200	2	400	40000
500	1	500	25000
100	4	400	10000
700	0	0	490000
100	3	300	10000
400	2	800	160000
100	1	100	10000
200	4	800	40000
$\sum X = 3200$	$\sum Y = 20$	$\sum XY = 4500$	$\sum X^2 = 1460000$

Substitute the totals into the two statistics formulae to find estimates of the regression line parameters.

$$a = \frac{\left(\sum X^2\right) \cdot \left(\sum Y\right) - \left(\sum X\right) \cdot \left(\sum XY\right)}{n \cdot \left(\sum X^2\right) - \left(\sum X\right)^2} = \frac{(1460000) \cdot (20) - (3200) \cdot (4500)}{10 \cdot (1460000) - (3200)^2} = 3.3945 \text{ breakdowns}$$

$$b = \frac{n \cdot \left(\sum XY\right) - \left(\sum X\right) \cdot \left(\sum Y\right)}{n \cdot \left(\sum X^2\right) - \left(\sum X\right)^2} = \frac{10 \cdot (4500) - (3200) \cdot (20)}{10 \cdot (1460000) - (3200)^2} = -0.0044 \text{ breakdowns/\$}$$

The least squares line is then:

$$Y_p = 3.3945 - 0.0044X$$

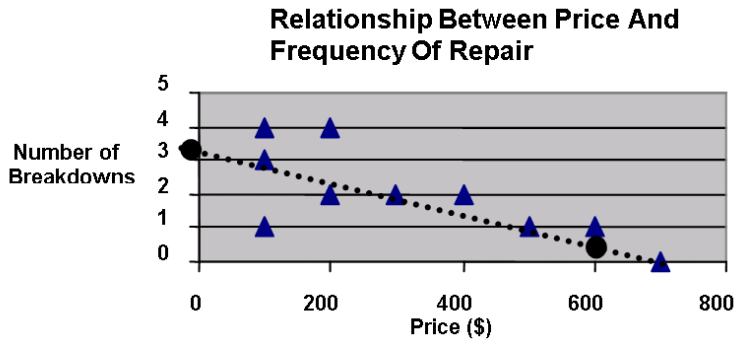
To show the underlying linear trend, one can plot this equation on the scatter diagram by identifying the Y-intercept and calculating a second Y value near the end of the range of X values:

By examining the equation, the Y-intercept= $a = 3.3945$ breakdowns.

An X value at the far end of the range is X = \$600, where

$$Y_p = 3.3945 - 0.0044(600) = 0.75 \text{ breakdowns}$$

Connecting these points on our diagram gives the regression line.



Notice that the points are scattered approximately evenly on either side of this line. If the predictability had been perfect the points would all lie exactly on the line.

We can predict the number of breakdowns at \$400 using our regression line to get:

$$Y_p = 3.3945 - 0.0044(400) = 1.63 \text{ breakdowns .}$$

Note that this value, as well as a and b should all compare favourably to your rough estimates done by eye.

A property of the regression line is that it predicts the mean Y value perfectly on the basis of the mean X value. To see this, note that for the above example:

$$\bar{X} = \frac{\sum X}{n} = \frac{3200}{10} = \$320 \text{ , and}$$

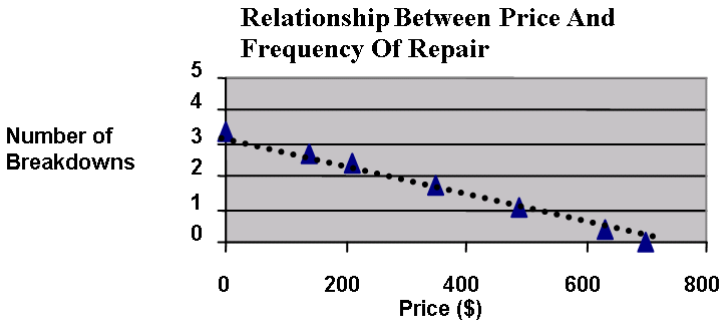
$$\bar{Y} = \frac{\sum Y}{n} = \frac{20}{10} = 2 \text{ breakdowns ,}$$

while placing $\bar{X} = \$320$ in the regression line formula gives:

$$Y_p = 3.3945 - 0.0044(320) = 1.99 \text{ breakdowns .}$$

4.3 Correlation Analysis

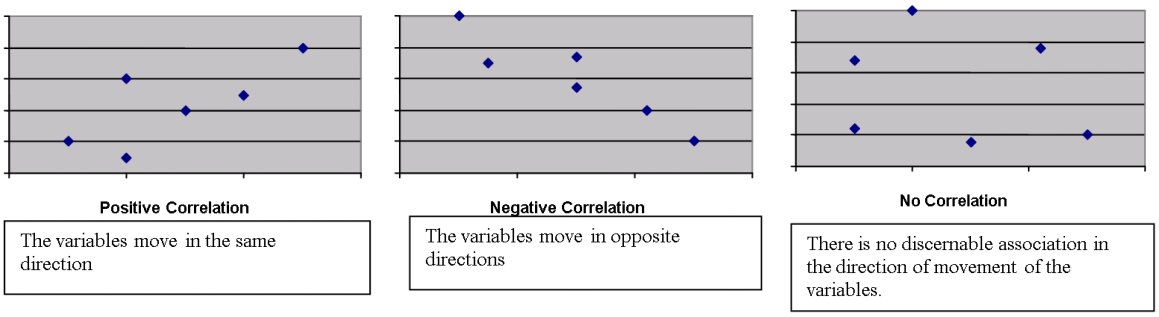
A situation in which there is perfect predictability is said to be a **deterministic model**. In a deterministic model all ordered pairs fall exactly along a straight line. For example, suppose our data for the lawnmowers had looked as follows:



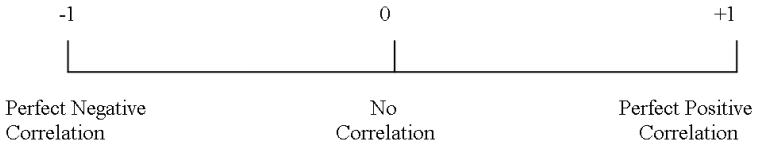
In this case after finding the regression line parameters, predictions about frequency of repair could be made perfectly with the regression line equation.

In this case we would say that all of the variability observed in number of breakdowns is explainable by the variability in price paid. Nothing is left to chance. In practice it would be very rare to see a perfect situation like this as was shown in our actual data.

Correlation analysis measures the strength of the relationship between variables. It is useful for assessing how much of the variation in the Y values is explainable by variation in the X values. If a scatter diagram is analyzed it is possible to analyze strong and weak correlation by the location of the plotted points.



The **Pearson Product Moment Correlation Coefficient** gives a quantifiable measure of the degree to which the points fit a straight line and the direction of movement of the variables. It is designated by the symbol $\boldsymbol{r}$. The correlation coefficient, r, is found in this range of values:



The **sign** of r indicates whether the data are positively or negatively correlated, while the **magnitude** of r indicates the extent of the correlation. If r is either +1 or -1, predictability is perfect and all ordered pairs will lie exactly on a straight line of either positive or negative slope respectively. If r is close to 0 there is no discernible association between the variables and the points will lie very far from a straight line. In general we will interpret the magnitude of $|r|$ as follows:

- 0.00 to 0.40: insignificant (negligible) correlation
- 0.40 to 0.70: substantial correlation
- 0.70 to 0.80: high correlation
- 0.80 to 0.90: very high correlation
- 0.90 to 1.00: extremely high correlation

These guidelines are somewhat subjective and will vary depending on the field of study. In practice a value of $|r|$ that is very high (0.80 to 0.90) is useful for predicting the behaviour of a group of items or individuals, but a larger value is required to predict the behaviour of a single individual.

Look at the three data sets above and qualitatively estimate r on the given number line.

The calculation of r involves finding the square root of the complement of the percentage of the total variability among Y values that is made up of the variability from the least squares line. Without getting into the mathematics of this, the calculation is:

$$r = \frac{n \cdot (\sum XY) - (\sum X) \cdot (\sum Y)}{\sqrt{n \cdot (\sum X^2) - (\sum X)^2} \cdot \sqrt{n \cdot (\sum Y^2) - (\sum Y)^2}}$$

Example:
Using the data for the lawnmower example calculate the correlation coefficient and interpret the result.

Solution:
Looking at the formula, there is one additional column required from our regression calculation, the Y^2 column:

X(\$)	Y(breakdowns)	XY(\$ · breakdowns)	X^2 (\$^{2})	Y^2 (breakdowns^2)
600	1	600	360000	1
300	2	600	90000	4
200	2	400	40000	4
500	1	500	25000	1
100	4	400	10000	16
700	0	0	490000	0
100	3	300	10000	9
400	2	800	160000	4
100	1	100	10000	1
200	4	800	40000	16
$\sum X = 3200$	$\sum Y = 20$	$\sum XY = 4500$	$\sum X^2 = 1460000$	$\sum Y^2 = 56$

Substitute:

$$r = \frac{n \cdot (\sum XY) - (\sum X) \cdot (\sum Y)}{\sqrt{n \cdot (\sum X^2) - (\sum X)^2} \cdot \sqrt{n \cdot (\sum Y^2) - (\sum Y)^2}} = \frac{10 \cdot (4500) - (3200) \cdot (20)}{\sqrt{10 \cdot (1460000) - (3200)^2} \cdot \sqrt{10 \cdot (56) - 20^2}} = -0.72$$

Interpretation:

- The negative sign indicates a negative correlation; as the price of a lawnmower increases the number of breakdowns tends to decrease.
- The magnitude 0.72 means that the correlation is high.

This should not be interpreted as cause and effect. A low price does not cause a high frequency of repair. The two variables simple move together.

The **Coefficient of Determination**, calculated as r^2, measures the proportion of variation in Y that is explainable by variation in X.

The **Coefficient of Nondetermination**, calculated as $(1 - r^2)$, measures the proportion of variation in Ythat is explainable by factors other than the variation in X.

Example:
Find and interpret the coefficient of determination and non-determination for the lawnmower example.

Solution:
Since $r = -0.72$ the coefficient of determination is:

$$r^2 = (-0.72)^2 = 0.52 \ .$$

This mean that 52% of the variation in number of breakdowns is explainable by the variation in the price paid for a lawnmower. The coefficient of nondetermination is:

$$(1 - r^2) = 1 - 0.52 = 0.48 \ .$$

This means that 48% of the variation in the number of breakdowns is explainable by factors other than the variation in the price paid for a lawnmower.

Calculator Keys and Software

The statistical calculator used in this course is capable of processing data as ordered pairs in the statistical mode. Identify this process. Typically it requires going into the statistical mode for linear regression and entering ordered pairs separated by a comma.

Once the data is keyed in, the three statistics a, b, and r can be retrieved from the storage registers with the appropriate label. The coefficient of determination can be found by squaring the r value once it is in the display. Also note that the intermediate sums required in the formula can also be retrieved by appropriate keystrokes.

On the computer, spreadsheet programs can plot scatter diagrams, produce the least squares equation and also calculate the correlation coefficient. For more sophisticated statistical work there are statistical programs available, many of them free. Google "statistics open source".

Assignment:

1. The table below presents data for a random sample of eight students showing the amount of outside class study time spent and the grade earned at the end of a one month statistics course.

Study Time (hr)	Grade (%)			
20	64			
16	61			
34	84			
23	70			
27	88			
32	92			
18	72			
22	77			

   Answer the following questions by completing the table and check your work on your calculator.

   (a) Determine the regression equation for estimating the examination grade given the hours of study. Carry your calculations to one place beyond the decimal place.

   (b) Use the regression equation to estimate the examination grade for a student who devotes 30 hours to study.

   (c) Compute and interpret the values of the coefficients of correlation, determination, and nondetermination.

2. The following gives the peak power load in megawatts for a power plant and the daily high temperature for a random sample of 10 days:

Day	1	2	3	4	5	6	7	8	9	10
Daily High Temp (°F)	95	82	90	81	99	100	93	95	93	87
Peak Load (MW)	214	152	156	129	254	266	210	204	213	150

   Do the following in a tabular format.

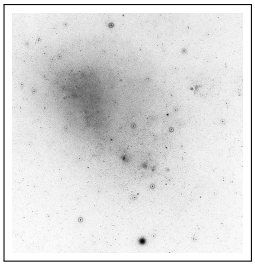
   (a) Determine the best-fitting straight line (linear regression equation) to predict peak load, Y, based on the daily high temperature, X.

   (b) Using the prediction equation determined in (a), find the peak load for a daily high temperature of 85°F.

   (c) What is the coefficient of correlation? What two things does the coefficient of correlation tell us about the relationship between daily high temperature, X, and the peak load, Y?

   (d) Calculate the coefficient of determination and interpret your result.

3. Cepheid variable stars have the interesting property that their brightness changes in a periodic (regular) way over time. Measuring the time between when a star is at its brightest gives its period. In 1908, Henrietta S. Leavitt published the following data concerning sixteen Cepheid variable stars in a nearby dwarf galaxy called the Small Magellanic Cloud (SMC). The following data gives the (natural logarithm of) the period in days, X, of each variable star along with its apparent magnitude, Y, when it is at its maximum brightness.



The SMC (photo courtesy ESA/Hubble)

X	Y	X	Y
2.34	13.6	1.51	14.3
4.84	11.2	0.51	14.8
3.46	12.2	0.57	14.8
4.19	11.4	0.23	14.8
2.60	13.4	0.63	15.1
1.46	14.6	1.67	14.4
2.13	13.9	1.67	14.3
1.89	14.1	1.61	14.3

(a) Using your calculator, determine the correlation coefficient for the data and interpret your result.

(b) Using your calculator, determine the linear regression line and use it to determine the expected maximum apparent magnitude of a variable star in the SMC of period 60.2 days. (Remember X is the natural logarithm of this time.)

Answers

page 13:

1. (a) $P(\$9550 < \bar{X} < \$9650)$
$= P(-2.50 < Z < 2.50) = 0.9876$
   (b) $P(\$9660 < \bar{X})$
$= P(3.00 < Z) = 0.0013$
3. (a) $P(2300 \text{ kg} < \bar{X} < 2700 \text{ kg})$
$= P(-1.80 < Z < 1.80) = 0.9282$
   (b) $P(\bar{X} < 2000 \text{ kg or } 3000 \text{ kg} < \bar{X})$
$= P(Z < -4.50) + P(4.50 < Z) \approx 0$
5. (a) $P(76\% < \bar{X}) = P(1.12 < Z) = .1314$
   (b) 95% of all sample means would fall between 73.25% and 76.75%
6. (a) $\sigma_{\bar{X}} = \sigma$. The mean of a sample of size 1 is the same as measuring X.
   (b) $\sigma_{\bar{X}} = 0$ since $F.C.F. = 0$. No uncertainty as the sample mean is the population mean.
7. (a) {2,4}, {2,6}, {2,8}, {4,6}, {4,8}, {6,8}
Note that when sampling without replacement we use selections (order does not matter), while when sampling with replacement the samples are permutations. (Hence our use of braces here rather than parentheses.)
   (b) $\mu_{\bar{X}} = 5$ occupants $= \mu$
   (c) $\sigma_{\bar{X}} = \sqrt{\frac{5}{3}}$ occupants $= \sqrt{\frac{4-2}{4-1}} \cdot \frac{\sqrt{5}}{\sqrt{2}}$
$= (F.C.F.) \cdot \frac{\sqrt{\sigma}}{\sqrt{n}}$
   (d) The distribution $P(\bar{X})$ is not normal. (Graph it!)

page 16:

1. $P(p < .43 \text{ or } .47 < p)$
$= P(Z < -1.29) + P(1.29 < Z) = .1970$
2. (a) $P(.1 < p < .2)$
$= P(-1.37 < Z < 0.00) = .4147$
   (b) $P(3 \le X \le 6)$
$= P(3) + P(4) + P(5) + P(6)$
$= .0785 + .1325 + .1723 + .1795 = .5628$
   (c) $P(.0833 < p < .2166)$
$= P(-1.60 < Z < 0.23) = .5362$
Comparing to (b) shows a smaller difference than between (a) and (b). Note that $n\pi = 6$ which shows we are close to being outside of the region of validity for p to be approximately normal so some difference is expected.

page 23:

1. $P(.300 - .016 \text{ ppm} < \mu < .300 + .016 \text{ ppm})$
$= P(.284 \text{ ppm} < \mu < .316 \text{ ppm}) = 98\%$
2. $P(\$512 - \$9.54 < \mu < \$512 + \$9.54)$
$= P(\$502.46 < \mu < \$521.54) = 90\%$
3. (a) $P(\$85 - \$3.02 < \mu < \$85 + \$3.02)$
$= P(\$81.98 < \mu < \$88.02) = 95\%$
   (b) $P(\$85 - \$3.94 < \mu < \$85 + \$3.94)$
$= P(\$81.06 < \mu < \$88.94) = 99\%$
4. n must be 30 or greater to use the C.L.T. .
5. (a) $P(82 - 1.15\% < \mu < 82 + 1.15\%)$
$= P(80.85\% < \mu < 83.15\%) = 99\%$
   (b) $P(82 - 1.04\% < \mu < 82 + 1.04\%)$
$= P(80.96\% < \mu < 83.04\%) = 99\%$

page 26:

1. (a) Yes. $\pi \approx p = .005794$, so $n\pi \approx 197 > 5$ and $n(1-\pi) \approx 33803 > 5$.
   (b) $P(.0058 - .0007 < \pi < .0058 + .0007)$
$= P(.0051 < \pi < .0065) = 90\%$
   (c) Yes, since the selection process is not random. Sick animals are selected and infected areas have increased hunting quotas. This would mean our estimate will be too high. (Alternatively, CWD tests are likely difficult with infected animals not having observable symptoms resulting in false negatives.)
2. (a) $\pi \approx p = .520$
   (b) $P(.520 - .175 < \pi < .520 + .175)$
$= P(.345 < \pi < .695) = 99\%$
3. (a) $P(.56 - .03 < \pi < .56 + .03)$
$= P(.53 < \pi < .59) = 95\%$
Note a pollster would say "Support of the PP is 56%. This result is accurate to within ±3%, 19 times out of 20."
   (b) Yes, since .50 is not in the interval.
   (c) Yes, $P(.56 - .04 < \pi < .56 + .04)$
$= P(.52 < \pi < .60) = 99\%$
which still does not contain .50 .

page 29:

1. 97 measurements
2. 16,588 Quebec citizens (if the 1995 result is used as an estimate)
3. 75 vehicles
4. (a) 601 burial plots
   (b) 505 burial plots
   (c) Lower the confidence and/or increase max sampling error. For example, with $P = 90\%$ and $E = 10\% = .10$, we need only $n = 57$ burial plots.

page 33:

1. (a) $t = 2.624$
   (b) $t = 1.703$
   (c) $t = 4.032$

(d) $Z = 3.291$ (This is effectively normal as n=560¿30. Use $df = \infty$ on t table.)

2. (a) $P(156.3 - 78.0\ \frac{\text{m}^3}{\text{s}} < \mu < 156.3 + 78.0\ \frac{\text{m}^3}{\text{s}})$
$= P(78.3\ \frac{\text{m}^3}{\text{s}} < \mu < 234.3\ \frac{\text{m}^3}{\text{s}}) = 95\%$

   (b) Use 30 or more years of data so the C.L.T. applies and the underlying distribution does not matter.

3. (a) $P(\$2.03 - \$.66 < \mu < \$2.03 + \$.66)$
$= P(\$1.37 < \mu < \$2.69) = 90\%$

   (b) \$1096 < total < \$2152

page 34:

1. (a) $P(18 \text{ wks} < X) = P(1.50 < Z) = 0.0668$

   (b) $P(18 \text{ wks} < \bar{X}) = P(15.00 < Z) \approx 0$

   (c) i. $E = 0.31$ wks

   ii. $P(11.80 - .31 \text{ wks} < \mu < 11.80 + .31 \text{ wks})$
$= P(11.49 \text{ wks} < \mu < 12.11 \text{ wks}) = 99\%$

   (d) i. $E = 2.39$ wks

   ii. $P(11.80 - 2.39_{\text{wks}} < \mu < 11.80 + 2.39_{\text{wks}})$
$= P(9.41 \text{ wks} < \mu < 14.19 \text{ wks}) = 99\%$

   (e) i. $E = 0.02$ wks

   ii. $P(11.80 - .02 \text{ wks} < \mu < 11.80 + .02 \text{ wks})$
$= P(11.78 \text{ wks} < \mu < 11.82 \text{ wks}) = 99\%$

   (f) $n = 217$ unemployed people

2. $P(35,000 - 412 \text{ km} < \mu < 35,000 + 412 \text{ km})$
$= P(34,588 \text{ km} < \mu < 35,412 \text{ km}) = 98\%$

3. $P\left(7.2 - .7\ \frac{l}{100 \text{ km}} < \mu < 7.2 + .7\ \frac{l}{100 \text{ km}}\right)$
$= P\left(6.5\ \frac{l}{100 \text{ km}} < \mu < 7.9\ \frac{l}{100 \text{ km}}\right) = 90\%$

4. (a) $E = 2.7\%$ for π in favour. The maximum sampling error depends upon the proportion asked. For the proportion against changing to DST for instance $E = 2.9\%$ while for the proportion with no opinion it is $E = 1.5\%$. The largest sampling error for any possible proportion will occur when $p = .5$ is used and this results in $E = 3.1\%$ which is larger than all of them and quoted.

   (b) $P(.270 - .052 < \pi < .270 + .052)$
$= P(.218 < \pi < .322) = 99.9\%$
"The percentage in favour of switching to DST was 27%. The result is accurate to within 5.2%, 999 times out of 1000." In some sense this is more useful because the 50% needed for a successful plebiscite is still not in the interval thereby demonstrating the unlikelihood of it succeeding. There is only a .1% chance of the population value lying outside of this interval versus 5% in the one quoted.

5. $P(.20 - .04 < \pi < .20 + .04)$
$= P(0.16 < \pi < 0.24) = 99\%$

6. $P(140,840 - 16,080 < \mu < 140,840 + 16,080)$
$= P(124,760 \text{ veh.} < \mu < 156,920 \text{ veh.}) = 95\%$

7. $n = 2401$ voters

8. $n = 2653$ entries

page 39:

1. $H_0 : \mu = \$52.00$/wk, $H_a : \mu < \$52.00$/wk

2. $H_0 : \mu = \$52.00$/wk, $H_a : \mu > \$52.00$/wk

3. $H_0 : \pi = 0.18$, $H_a : \pi > 0.18$

4. $H_0 : \mu = 10$ cm, $H_a : \mu \neq 10$ cm

page 42:

1. $Z_{\text{critical}} = \pm 1.645$, $Z_{\text{calculated}} = 3.54$, therefore reject H_0 at $\alpha = .10$. Evidence suggests the population mean differs from \$560 .

2. $t_{\text{critical}} = \pm 1.729$, $t_{\text{calculated}} = 1.57$, therefore fail to reject H_0 at $\alpha = .10$ Evidence does not suggest the population mean differs from \$560 . Assume population normality.

3. $Z_{\text{critical}} = -1.645$, $Z_{\text{calculated}} = -5.42$, therefore reject H_0 at $\alpha = .05$. Evidence suggests the population mean is less than 454 g.

4. $\alpha = 0.05$

5. The butter weighs less than 454 grams on average and the sample average we randomly chose led us to believe that the it does weigh 454 grams on average. That is, we accepted the null hypothesis when it was false.

6. $t_{\text{critical}} = \pm 2.947$, $t_{\text{calculated}} = 8.639$, therefore reject H_0 at $\alpha = .01$. Evidence suggests the cluster mean metallicity differs from the general surrounding stellar population. The -0.50 star is an outlier (its Z value in the cluster data is -3.19) which may not be a member of the cluster but happens to lie along the line of sight or perhaps migrated into the region of the cluster at some time.

7. $t_{\text{critical}} = 1.833$, $t_{\text{calculated}} = 0.94$, therefore fail to reject H_0 at $\alpha = .05$. The evidence does not support the conclusion that the Regina rent is greater than \$881 .

page 45:

1. $Z_{\text{critical}} = -1.28$, $Z_{\text{calculated}} = -2.04$, therefore reject H_0 at $\alpha = .1$. Evidence suggests that less than 40% of households watched the program.

2. $Z_{\text{critical}} = -2.05$, $Z_{\text{calculated}} = -698.30$, therefore reject H_0 at $\alpha = .02$. Evidence suggests a (very) significant decrease in the percent of women unaware occurred since 2008.

3. $Z_{\text{critical}} = -1.645$, $Z_{\text{calculated}} = -1.89$, therefore reject H_0 at $\alpha = .05$. The counselor is incorrect; evidence suggests that less than 30% of students hold a part-time job.

4. $Z_{\text{critical}} = \pm 2.326$, $Z_{\text{calculated}} = 0.554$, therefore fail to reject H_0 at $\alpha = .02$. The evidence does not support a change in the failure rate.
5. No, this is not a proportion hypothesis test since the underlying variable is not a binomial (yes/no) qualitative variable. Here the underlying variable X itself is the fraction of a litre of gas that is ethanol which is quantitative. The regulator could take a random sample of n measurements (buying one litre of gasoline at n different outlets and/or times) and measure the fraction of ethanol in each one and take its average $\bar{X}$ and then test the alternative hypothesis regarding the population mean that $\mu < .06$.

page 46:

1. $Z_{\text{critical}} = 1.645$, $Z_{\text{calculated}} = 6.60$, therefore reject H_0 at $\alpha = .05$. The evidence suggests an increase.
2. $Z_{\text{critical}} = 2.33$, $Z_{\text{calculated}} = 10.00$, therefore reject H_0 at $\alpha = .01$. Evidence suggests the trucks weigh more than 25,000 kg.
3. $Z_{\text{critical}} = -1.28$, $Z_{\text{calculated}} = -2.00$, therefore reject H_0 at $\alpha = .10$. Evidence suggests less than 50% are male.
4. $t_{\text{critical}} = \pm 3.499$, $t_{\text{calculated}} = 0.007$, therefore fail to reject H_0 at $\alpha = .01$. Evidence does not support an adjustment be made.
5. $Z_{\text{critical}} = 1.645$, $Z_{\text{calculated}} = 1.267$, therefore fail to reject H_0 at $\alpha = .05$. Evidence does not support an increase.

page 53:

1. $Z_{\text{critical}} = \pm 1.96$, $Z_{\text{calculated}} = 2.61$, therefore reject H_0 at $\alpha = .05$. The evidence suggests that there is a difference between the average monthly household incomes in the two communities.
2. $t_{\text{critical}} = 2.681$, $t_{\text{calculated}} = 0.45$, therefore fail to reject H_0 at $\alpha = .01$. The evidence does not support an increase in sales..
3. No. $Z_{\text{critical}} = -1.28$, $Z_{\text{calculated}} = -1.06$, therefore fail to reject H_0 at $\alpha = .10$. The evidence does not support the hypothesis that a person with enhanced sleep performed the task in less time with statistical significance.

page 57:

1. $Z_{\text{critical}} = 1.645$, $Z_{\text{calculated}} = 1.05$, therefore fail to reject H_0 at $\alpha = .05$. There is no evidence to suggest that a higher proportion of business students have part-time jobs than science students.
2. (a) $Z_{\text{critical}} = \pm 1.28$, $Z_{\text{calculated}} = 1.82$, therefore reject H_0 at $\alpha = .20$. The evidence supports a difference in the ability of the two swallow types to carry coconuts.
   (b) It becomes a right-tailed test ($H_a : \pi_A - \pi_E > 0$) with $Z_{\text{critical}} = 0.84$. Still reject H_0 and accept H_a but now evidence supports that African swallows are better carriers than European ones.

page 58:

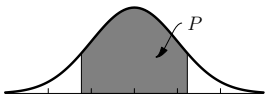
1. $Z_{\text{critical}} = \pm 2.575$, $Z_{\text{calculated}} = 2.89$, therefore reject H_0 at $\alpha = .01$. Evidence suggests that the proportions preferring an accountant differ.
2. $t_{\text{critical}} = -2.457$, $t_{\text{calculated}} = -1.88$, therefore fail to reject H_0 at $\alpha = .01$. The evidence does not support that the firm's professionals earn less.
3. $Z_{\text{critical}} = 1.645$, $Z_{\text{calculated}} = 3.06$, therefore reject H_0 at $\alpha = .05$. Evidence suggests foreign cars are more expensive to repair on average.

page 67:

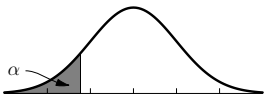
1. (a) $a = 40.08$ %, $b = 1.497$ %/hr, $Y_p = a + bX = 40.1 + 1.5X$
   (b) $Y_p = 85\%$
   (c) $r = 0.8621$, very high positive correlation; $r^2 = .7432 = 74.32\%$ of the variation in the exam grade is due to the hours of study; $(1 - r^2) = .2568 = 25.68\%$ of the variation in the exam grade is due to other factors than study (such as ability, interest, experience, etc.)
2. (a) $Y_p = -419.8 + 6.7X$ (MW)
   (b) $Y_p = 151.1$ MW
   (c) $r = .9441$; magnitude $|r| = .9441$ implies extremely high correlation; sign of r positive implies positive correlation (As temperature (X) increases, peak load (Y) increases.
   (d) $r^2 = 89.13\%$; 89.13% of the variation in peak load (Y) is related to the variation in temperature (X).
3. (a) $r = -0.9722$, an extremely high negative correlation.
   (b) $Y_p = 15.58 - 0.8966X$; For $X = \ln(60.2) = 4.10$, $Y_p = 11.9$.

   Note: Cepheid variable stars are of critical importance to astronomers precisely because one can measure their period to determine their intrinsic brightness (absolute magnitude). Then measuring their apparent brightness allows an exact measurement of their distance from us. They are one of the *standard candles* for measuring distances far beyond our galaxy.

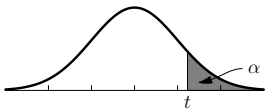
Student's t Distribution



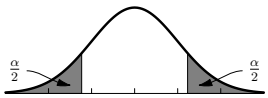
Confidence Interval



Left-tailed Test



Right-tailed Test



Two-tailed Test

df	Confidence Intervals, P					
	80%	90%	95%	98%	99%	99.9%
	Level of Significance for One-Tailed Test, α					
	0.100	0.050	0.025	0.010	0.005	0.0005
	Level of Significance for Two-Tailed Test, α					
	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646

df	Confidence Intervals, P					
	80%	90%	95%	98%	99%	99.9%
	Level of Significance for One-Tailed Test, α					
	0.100	0.050	0.025	0.010	0.005	0.0005
	Level of Significance for Two-Tailed Test, α					
	0.20	0.10	0.05	0.02	0.01	0.001
31	1.309	1.696	2.040	2.453	2.744	3.633
32	1.309	1.694	2.037	2.449	2.738	3.622
33	1.308	1.692	2.035	2.445	2.733	3.611
34	1.307	1.691	2.032	2.441	2.728	3.601
35	1.306	1.690	2.030	2.438	2.724	3.591
36	1.306	1.688	2.028	2.434	2.719	3.582
37	1.305	1.687	2.026	2.431	2.715	3.574
38	1.304	1.686	2.024	2.429	2.712	3.566
39	1.304	1.685	2.023	2.426	2.708	3.558
40	1.303	1.684	2.021	2.423	2.704	3.551
41	1.303	1.683	2.020	2.421	2.701	3.544
42	1.302	1.682	2.018	2.418	2.698	3.538
43	1.302	1.681	2.017	2.416	2.695	3.532
44	1.301	1.680	2.015	2.414	2.692	3.526
45	1.301	1.679	2.014	2.412	2.690	3.520
46	1.300	1.679	2.013	2.410	2.687	3.515
47	1.300	1.678	2.012	2.408	2.685	3.510
48	1.299	1.677	2.011	2.407	2.682	3.505
49	1.299	1.677	2.010	2.405	2.680	3.500
50	1.299	1.676	2.009	2.403	2.678	3.496
51	1.298	1.675	2.008	2.402	2.676	3.492
52	1.298	1.675	2.007	2.400	2.674	3.488
53	1.298	1.674	2.006	2.399	2.672	3.484
54	1.297	1.674	2.005	2.397	2.670	3.480
55	1.297	1.673	2.004	2.396	2.668	3.476
60	1.296	1.671	2.000	2.390	2.660	3.460
80	1.292	1.664	1.990	2.374	2.639	3.416
100	1.290	1.660	1.984	2.364	2.626	3.390
200	1.286	1.653	1.972	2.345	2.601	3.340
∞	1.282	1.645	1.960	2.326	2.576	3.291

Sampling Distribution Formulae

Standard Error Formulae for Single Means and Proportions

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad s_{\bar{X}} = \frac{s}{\sqrt{n}} \qquad \sigma_{p} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}} \qquad s_{p} = \sqrt{\frac{p \cdot (1-p)}{n}}$$

$$\textbf{Finite Correction Factor} \quad \Rightarrow \quad F.C.F. = \sqrt{\frac{N-n}{N-1}}$$

Sampling Error (Precision) Formulae for Single Means And Proportions

$$E = Z \cdot \sigma_{\bar{X}} \qquad E = Z \cdot \sigma_{p}$$

$$E = Z \cdot s_{\bar{X}} \qquad E = t \cdot s_{\bar{X}} \qquad E = Z \cdot s_{p}$$

Confidence Intervals for Population Means and Proportions

$$P\left(\left[\bar{X} - E\right] < \mu < \left[\bar{X} + E\right]\right) = P\% \qquad P\left(\left[p - E\right] < \pi < \left[p + E\right]\right) = P\%$$

Sample Sizes for Estimating Means And Proportions

$$n = \left[\frac{Z \cdot \sigma}{E}\right]^2 \qquad n = \pi \cdot (1-\pi) \cdot \left[\frac{Z}{E}\right]^2$$

Standard Scores for Single Means and Proportions

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \qquad Z = \frac{p - \pi}{\sigma_{p}}$$

$$Z = \frac{\bar{X} - \mu}{s_{\bar{X}}} \qquad t = \frac{\bar{X} - \mu}{s_{\bar{X}}} \qquad Z = \frac{p - \pi}{s_{p}}$$

Other Formulae

$$_{N}C_{n} \quad \text{or} \quad N^{n} \qquad df = n - 1$$

Standard Error Formulae for Differences of Means and Proportions

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad s_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{(n_1-1)\cdot s_1^2 + (n_2-1)\cdot s_2^2}{n_1+n_2-2} \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

$$p_{\text{Pool}} = \frac{X_1 + X_2}{n_1 + n_2} \qquad s_{p_1-p_2} = \sqrt{p_{\text{Pool}} \cdot (1 - p_{\text{Pool}}) \cdot \left[\frac{1}{n_1} + \frac{1}{n_2}\right]}$$

Standard Scores For Differences Between Means And Proportions

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - 0}{s_{\bar{X}_1-\bar{X}_2}} \qquad t = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - 0}{s_{\bar{X}_1-\bar{X}_2}} \qquad Z = \frac{(p_1 - p_2) - 0}{s_{p_1-p_2}}$$

Other Formulae

$$df = n_1 + n_2 - 2$$

Linear Regression/Correlation Analysis

$$Y_p = a + bX$$

$$a = \frac{\left(\sum X^2\right)\cdot\left(\sum Y\right) - \left(\sum X\right)\cdot\left(\sum XY\right)}{n\cdot\left(\sum X^2\right) - \left(\sum X\right)^2}$$

$$b = \frac{n\cdot\left(\sum XY\right) - \left(\sum X\right)\cdot\left(\sum Y\right)}{n\cdot\left(\sum X^2\right) - \left(\sum X\right)^2}$$

$$r = \frac{n\cdot\left(\sum XY\right) - \left(\sum X\right)\cdot\left(\sum Y\right)}{\sqrt{n\cdot\left(\sum X^2\right) - \left(\sum X\right)^2}\cdot\sqrt{n\cdot\left(\sum Y^2\right) - \left(\sum Y\right)^2}}$$

GNU Free Documentation License

Preamble

The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "**Document**", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "**you**". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "**Modified Version**" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "**Secondary Section**" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "**Invariant Sections**" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "**Cover Texts**" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "**Transparent**" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "**Opaque**".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "**Title Page**" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "**publisher**" means any person or entity that distributes copies of the Document to the public.

A section "**Entitled XYZ**" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "**Acknowledgements**", "**Dedications**", "**Endorsements**", or "**History**".) To "**Preserve the Title**" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
D. Preserve all the copyright notices of the Document.
E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
H. Include an unaltered copy of this License.
I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
O. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION

You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING

"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with ... Texts." line with this:

> with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.