CHAPTER 9

ONE- AND TWO-SAMPLE ESTIMATION PROBLEMS

9.1 Introduction

The purpose of statistical inference is to draw conclusions from data. We have already examined data and arrived at conclusions many times in the previous chapters. Formal inference emphasizes substantiating our conclusions by probability calculations.

Although there are many specific recipes for inference, there are only a few general types of statistical inference. This chapter and next chapter will introduce the two most common types: *confidence intervals* and *tests of significance*. We usually refer them as the problems of **estimation** and **hypothesis testing**.

9.2 Statistical Inference

Statistical inference draws conclusions about a population or process based on sample data. For instance, one uses the sample mean $\bar{x}$ to make generalization for the population mean μ, and uses the sample standard deviation s for the population standard deviation σ. It also provides a statement, expressed in terms of probability, of how much confidence we can place in our conclusions.

For the problem of estimation of unknown population parameters such as the mean, the proportion, and the variance, the trend is to distinguish between the **classical method**, or **frequentist method**, whereby inferences are based strictly on information obtained from a random sample selected from the population, and the **Bayesian method**, which utilizes prior subjective knowledge about the probability distribution of the unknown parameters in conjunction with the information provided by the sample data. We shall use classical methods to estimate unknown population parameters by computing statistics from random samples and applying the theory of sampling distributions.

Because the methods of formal inference are based on sampling distributions, they require a probability model for the data. Trustworthy probability models can arise in many ways, but the model is most secure and inference is most reliable when the data are produced by a properly randomized design. *When you use statistical inference, you are acting as if the data come from a random sample or a randomized experiment.*

9.3 Classical Methods of Estimation

A **point estimate** of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$. For example, the value $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ of the statistic $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ is a point estimate of the population parameter μ. Similarly, $\hat{p} = x/n$ is a point estimate of the true proportion p for a binomial experiment.

Note that we don't expect an error-free estimation because of the sampling bias and variability. Also, there may be more than one point estimates for a population unknown parameter; we need choose one wisely.

Unbiased Estimator

A statistic Θ is said to be an *unbiased estimator* of the parameter θ if

$$\mu_{\hat{\Theta}} = \mathrm{E}\left(\hat{\Theta}\right) = \theta.$$

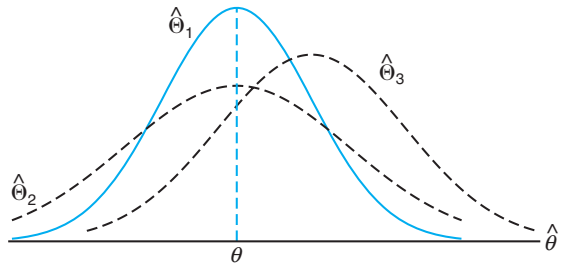
EXAMPLE 9.1. Show that $\overline{X}$ is an unbiased estimator of the parameter μ.

EXAMPLE 9.2. Show that S^2 is an unbiased estimator of the parameter σ^2.

Most Efficient Estimator

If we consider all possible unbiased estimators of some parameter θ, the one with the **smallest variance** is called the *most efficient estimator* of θ.

EXAMPLE 9.3. Which one of the following estimators is the most efficient one?



In many situations, we prefer to determine an interval within which we would expect to find the value of the parameter. Such an interval is called an **interval estimate**.

Interval Estimation

An *interval estimate* of a population parameter θ is an interval of the form

$$\hat{\theta}_L < \theta < \hat{\theta}_U,$$

where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on the value of the statistic $\hat{\Theta}$ for a particular sample and also on the sampling distribution of $\hat{\Theta}$.

9.4 Single Sample: Estimating the Mean

9.4.1 An Introductory Example.

Let us now look at an example.

The heights of the freshmen at UMD are supposed to follow a *normal* distribution with mean μ and standard deviation $\sigma = 10$ (in cm). A random sample of size $n = 36$ is taken, the sample mean $\bar{x} = 160$.

- Use $\bar{x} = 160$ to estimate the value of μ. This is a point estimation of μ.
- Is μ equal to 160?
- We would like to convert this point estimate into a statement, like "the value of μ is between 150 cm and 170 cm" and attached to the statement a measure of degree of confidence of it being true.
- From the distribution of $\overline{X}$,

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

- About 95% of the values of $\overline{X}$ are expected to fall within $2(\sigma/\sqrt{n})$ of μ, i.e.,

$$\mathrm{P}\left(\mu - 2\frac{\sigma}{\sqrt{n}} \le \overline{X} \le \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Exchanging the positions of μ and $\overline{X}$,

$$\mathrm{P}\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

- Our sample gives $\bar{x} = 160$, then the interval is from $\left(\bar{x} - 2\frac{\sigma}{\sqrt{n}}\right)$ to $\left(\bar{x} + 2\frac{\sigma}{\sqrt{n}}\right)$, or,

$$\left(160 - 2\frac{10}{\sqrt{36}},\ 160 + 2\frac{10}{\sqrt{36}}\right)$$

or,

$$(157,\ 163)$$

- This is an interval estimate of the unknown parameter of μ.
  - 0.95, confidence level or confidence coefficient
  - $(157,\ 163)$, 95% confidence interval of μ
  - 157, lower confidence limit
  - 163, upper confidence limit
  - $6 = 163 - 157$, interval width

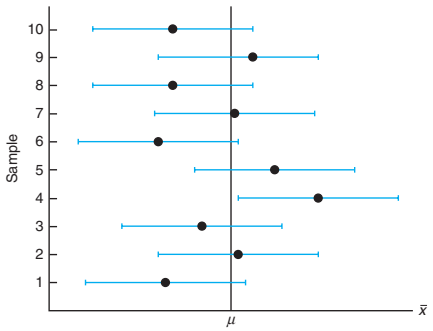
Interpretation of Interval Estimates

From the sampling distribution of $\hat{\Theta}$, we shall be able to determine $\hat{\theta}_L$ and $\hat{\theta}_U$ such that

$$\mathrm{P}\left(\hat{\Theta}_L < \theta < \hat{\Theta}_U\right) = 1 - \alpha,$$

for $0 < \alpha < 1$, then we have a probability of $1 - \alpha$ of selecting a random sample that will produce an interval containing θ.

- The interval $\hat{\theta}_L < \theta < \hat{\theta}_U$, computed from the selected sample, is called a $100(1-\alpha)\%$ **confidence interval**.
- The fraction $1 - \alpha$ is called the **confidence coefficient** or the **degree of confidence**.
- The endpoints, $\hat{\theta}_L$ and $\hat{\theta}_U$, are called the **lower and upper confidence limits**.



Continue on the UMD freshmen height example.

- Interpretation of 95% confidence interval
  - We are 95% confident that the interval from 157 cm and 163 cm will contain the true value of μ.
  - We are 95% confident that the true value of μ lies between 157 cm and 163 cm.
  - If we repeat the sampling processes over and over again, then approximately 95% of the similarly constructed intervals are expected to contain the true value of μ.
- Caution – *Don't say*
  - 95% of all freshmen at UMD are expected to have heights between $157cm$ and $163cm$.
  - We are 95% confident that a randomly selected UMD freshman has a height between $157cm$ and $163cm$.
  - 95% of all simple random samples of 10 UMD freshmen will have mean height between $157cm$ and $163cm$.

In general, let us consider the interval estimate of the unknown population mean μ, in general. If the sample is selected from a normal population or, if n is sufficiently large, we can establish a confidence interval for μ based on the sampling distribution of $\overline{X}$.

9.4.2 The case of known σ

The idea is the same as that in the UMD freshmen height example.

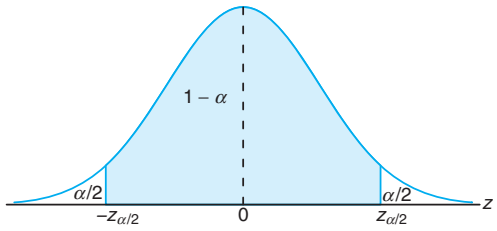
$$\begin{aligned}
\mathrm{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) &= 1-\alpha \\
\mathrm{P}\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) &= 1-\alpha \\
\mathrm{P}\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha \\
\mathrm{P}\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha
\end{aligned}$$

Confidence Interval on μ, when σ Known

If $\bar{x}$ is the mean of a random sample of size n from a population with known standard deviation σ, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the z-value such that $\mathrm{P}\left(Z > z_{\alpha/2}\right) = \alpha/2$.



NOTE. Sometimes, it is easier to use the t table to find the z-values.

NOTE. The C.I. is exact if the population is normal; it is approximate if the population is non-normal and n is sufficiently large.

EXAMPLE 9.4. High school students who take the SAT mathematics exam a second time generally score higher than on their first try. The change in score has a normal distribution with variance $\sigma^2 = 2500$. A random sample of 1000 students gains an average of $\bar{x} = 22$ points on their second try.

(a) Construct a 90% confidence interval for the mean score gain μ in the population of all students.

(b) Interpret the C.I. in part (a).

(c) Repeat part (a) for levels of confidence of 95% and 99%.

(d) How does increasing the confidence level affect the width of a confidence interval?

A wise user of statistics never plans data collection without at the same time planning the inference. We could arrange to have both high confidence and a small error.

If e is a pre-fixed (perhaps, desired and specified) amount that the error $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ can not exceed, we set

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < e$$

Solve for n, we have

Sample Size Determination

If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error will not exceed a specified amount e when the sample size is

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

NOTE. When solving for the sample size, n, we round all fractional values up to the **next whole number**. This way, we can be sure that our degree of confidence never falls below $100(1-\alpha)\%$.

EXAMPLE 9.5. A community health nutritionist wishes to conduct a survey among a population of teenage girls to determine their average daily protein intake (measured in grams). Assume that the population of protein intakes is normally distributed with a standard deviation of 20 grams. If she wants a 95% confidence interval with an error of no more than 5 grams, how many teenage girls should be interviewed?

One-Sided Confidence Interval on μ, σ Known

If $\bar{x}$ is the mean of a random sample of size n from a population with standard deviation σ, the one-sided $100(1-\alpha)\%$ confidence intervals for μ are given by

$$\text{upper one-sided C.I.:} \quad -\infty < \mu < \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}$$

$$\text{lower one-sided C.I.:} \quad \bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \infty.$$

The derivation can be similarly done. For the lower one-sided C.I.,

$$1-\alpha = \mathrm{P}(Z < z_\alpha) = \mathrm{P}\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_\alpha\right)$$
$$= \mathrm{P}\left(\overline{X} < z_\alpha \frac{\sigma}{\sqrt{n}} + \mu\right) = \mathrm{P}\left(\mu > \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

EXAMPLE 9.6. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 40 hours. If a sample of 30 bulbs has an average life of 780 hours, find a 98% lower one-sided confidence interval for the population mean of all bulbs produced by this firm.

9.4.3 The case of unknown σ

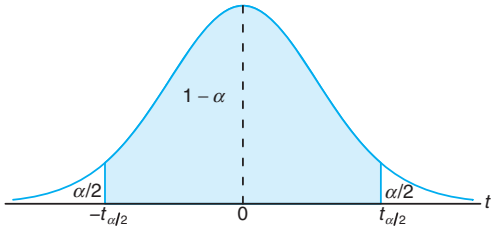
It is more practical and important that the population standard deviation σ is assumed unknown.

Confidence Interval on μ, when σ Unknown

If $\bar{x}$ and s are the mean and the standard deviation of a random sample of size n from a population with unknown standard deviation σ, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the t-value with $(n-1)$ degrees of freedom such that $\mathrm{P}\left(T(n-1) > t_{\alpha/2}\right) = \alpha/2$.



The derivation can be done in the similar fashion.

$$1-\alpha = \mathrm{P}\left(-t_{\alpha/2} < T < t_{\alpha/2}\right)$$
$$= \mathrm{P}\left(-t_{\alpha/2} < \frac{\overline{X}-\mu}{s/\sqrt{n}} < t_{\alpha/2}\right)$$
$$= \mathrm{P}\left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

EXAMPLE 9.7. In an experiment on the metabolism of insects, American cockroaches were fed measured amounts of a sugar solution after being deprived of food for a week and of water for 3 days. After 2, 5, and 10 hours, the researchers dissected some of the cockroaches and measured the amount of sugar in various tissues. Five cockroaches fed the sugar D-glucose and dissected after 10 hours had the following amounts (in micrograms) of D-glucose in their hindguts:

55.95 68.24 52.73 21.50 23.78

(a) List the conditions that are required for this interval estimation.

(b) Find a 99% confidence interval for the mean amount of D-glucose in cockroach hindguts under these conditions.

EXAMPLE 9.8. *How much do users pay for Internet service?* Here are the monthly fees (in dollars) paid by a random sample of 50 users of commercial Internet service providers in August 2000: (Data from the August 2000 supplement to the Current Population Survey, from the Census Bureau Web site, www.census.gov.)

20	40	22	22	21	21	20	10	20	20
20	13	18	50	20	18	15	8	22	25
22	10	20	22	22	21	15	23	30	12
9	20	40	22	29	19	15	20	20	20
20	15	19	21	14	22	21	35	20	22

(a) Is it appropriate to use t confidence interval to analyze the data? Briefly explain.

(b) Give a 95% confidence interval for the mean monthly cost of Internet access in August 2000.

One-Sided Confidence Interval on μ, σ Unknown

If $\bar{x}$ and s are the mean and the standard deviation of a random sample of size n from a population with standard deviation σ, the one-sided $100(1-\alpha)\%$ confidence intervals for μ are given by

upper one-sided C.I.: $$-\infty < \mu < \bar{x} + t_\alpha \frac{s}{\sqrt{n}}$$

lower one-sided C.I.: $$\bar{x} - t_\alpha \frac{s}{\sqrt{n}} < \mu < \infty.$$

EXAMPLE 9.9. A meat inspector has randomly selected 30 packs of 95% lean beef. The sample resulted in a mean of 96.2% with a sample standard deviation of 0.8%. Find a 90% upper one-sided confidence interval for the leanness of all packs. Assume normality.

9.4.4 Large-Sample Confidence Interval

Assume that the sample size n is greater than 30 and the population distribution is not too skewed. We may utilize both the z-values and the sample standard deviation s for estimating the population mean μ

$$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

This is often referred to as a *large-sample confidence interval*.

NOTE. This can be regarded as a normal approximation (t-value becomes z-value when n is sufficiently large); the quality of the approximation becomes better as the sample size gets larger.

EXAMPLE 9.10. Due to the decrease in interest rates, the First Citizens Bank received a lot of mortgage applications. A recent sample of 100 mortgage loans resulted in an average loan amount of \$255,500 with a standard deviation of \$25,000. Construct a 95% confidence interval for the loan amount. for all customers who fill out mortgage applications.

9.4.5 Summary

For estimating population means based on a single sample, it is essential to require that (i) the statistics (i.e., $\bar{x}$ and s) must be from a **random sample**, and (ii) the population is **normal**, or, if failing, $\mathbf{n \geq 30}$.

In general, when constructing the 2-sided C.I., we

- use $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, if σ is known.
- use $\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$, if σ is unknown and n small.
- use $\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$, if σ is unknown and n large.

9.5 Standard Error of a Point Estimate

The **standard error of** $\overline{X}$ is the standard deviation of $\overline{X}$.

$$\text{s.e.}(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

The **estimated standard error of** $\overline{X}$ is defined by the estimator of $\sigma/\sqrt{n}$.

$$\widehat{\text{s.e.}}(\bar{x}) = \frac{s}{\sqrt{n}}$$

It is also called as the standard error of $\overline{X}$ in many statistical computing packages.

NOTE. All the 2-sided confidence intervals that we have constructed in preceding section can be written as

$$\bar{x} \pm (z_{\alpha/2} \text{ or } t_{\alpha/2}) \cdot \text{s.e.}(\bar{x})$$

More generally, a 2-sided $100(1-\alpha)\%$ C.I. for θ is expressible of

$$\hat{\theta} \pm \text{ (critical value)} \cdot \text{s.e.}(\hat{\theta})$$

9.6 Prediction Intervals

Take STAT 3612 for "Prediction Intervals"

9.7 Tolerance Limits

Take STAT 3612 for "Tolerance Limits"

9.8 Two Samples: Estimating the Difference between Two Means

We will now conduct statistical inference procedures for estimating $\mu_1 - \mu_2$, the difference between two population means, based on *independent* samples.

As in Subsection 8.4.3, suppose that we have two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. We take two *independent* random samples, one from each population, of sizes n_1 and n_2. Then it is quite nature to have the difference between two sample means $\overline{X}_1 - \overline{X}_2$ as a nature **point estimator** of the difference between two population means $\mu_1 - \mu_2$.

For an interval estimate of $\mu_1 - \mu_2$, we must consider the sampling distribution of $\overline{X}_1 - \overline{X}_2$.

9.8.1 Variances Known

Assume that both σ_1^2 and σ_2^2 are known, we have from Subsection 8.4.3, that

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim \mathrm{N}(0,1)$$

Now,

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\ &= \mathrm{P}\left(-z_{\alpha/2} < \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < z_{\alpha/2}\right) \end{aligned}$$

follows

C.I. for $\mu_1 - \mu_2$, when both σ_1^2 and σ_2^2 known

If $\bar{x}_1$ and $\bar{x}_2$ are means of independent random samples of sizes n_1 and n_2 from populations with **known variances** σ_1^2 and σ_2^2, respectively, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &< \mu_1 - \mu_2 \\ &< (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \end{aligned}$$

where $z_{\alpha/2}$ is the z-value defined previously.

NOTE. The confidence interval is exact when two independent samples are taken from normal populations. For non-normal populations, the Central Limit Theorem allows for a pretty good approximation for reasonable size samples.

EXAMPLE 9.11. We would like to compare the mean tar content in regular cigarettes and light cigarettes. We take simple random samples of regular and light cigarettes of a particular brand and measure the tar content (in mg) of each cigarette. The data are as follows:

Regular:	11.3	12.1	12.6	11.5	12.2	12.8
Light:	9.5	9.8	9.3	8.9	10.0	

It is known that tar content for regular cigarettes of this brand follows a normal distribution with standard deviation 0.4 mg and tar content for light cigarettes of this brand follows a normal distribution with standard deviation 0.3 mg. Find a 95% confidence interval for the difference in mean tar content for all regular cigarettes and all light cigarettes of this brand.

9.8.2 Variances Unknown but Equal

Assume that the population variances σ_1^2 and σ_2^2 are unknown but equal, i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We can show that the statistic

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}} \Bigg/ \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2\,(n_1+n_2-2)}}$$

follows the Student t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom.

Define the **pooled estimate of variance**, or, the pooled sample variance, as

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

NOTE. The pooled variance is just a weighted average of the variances of X_1 and X_2, where the weights are the respective degrees of freedom.

Then the above statistic becomes

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T(n_1+n_2-2)$$

Now,

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) \\ &= \mathrm{P}\left(-t_{\alpha/2} < \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{\alpha/2}\right) \end{aligned}$$

follows

C.I. for $\mu_1 - \mu_2$, when $\sigma_1^2 = \sigma_2^2 = \sigma^2$, but Unknown

If $\bar{x}_1$ and $\bar{x}_2$ are means of independent random samples of sizes n_1 and n_2 from populations with **unknown but equal variances**, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\begin{aligned} (\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &< \mu_1 - \mu_2 \\ &< (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \end{aligned}$$

where $t_{\alpha/2}$ is the t-value defined previously and

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

EXAMPLE 9.12. An insurance company would like to know if men drive faster on average than women. The company took a random sample of 52 cars driven by men on a highway and found the mean speed to be 114 km/h with a standard deviation of 10 km/h. Another sample of 30 cars driven by women on the same highway gave a mean speed of 108 km/h with a standard deviation of 7 km/h. Construct a 98% confidence interval for the true difference between the mean speeds of cars driven by men and women on this highway.

NOTE. It is practically important to determine whether the population variances can be assumed to be equal or not. A rule of thumb is to look at the ratio of the sample standard deviations. If $\frac{1}{2} < \frac{s_1}{s_2} < 2$, the equal variance can be assumed; otherwise unequal.

9.8.3 Variances Unknown and Unequal

Assume that the population variances σ_1^2 and σ_2^2 are unknown but unequal, i.e., $\sigma_1^2 \neq \sigma_2^2$. The statistic

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \;\dot{\sim}\; \mathrm{T}(\nu)$$

where $\nu = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{[(s_1^2/n_1)^2/(n_1-1)] + [(s_2^2/n_2)^2/(n_2-1)]}$.

NOTE. The expression for ν above is an estimate of the degrees of freedom. In applications, it is rarely a whole number, and we should **round it down** to the nearest integer to achieve the desired confidence.

C.I. for $\mu_1 - \mu_2$, when $\sigma_1^2 \neq \sigma_2^2$, but Unknown

If $\bar{x}_1$ and s_1^2 and $\bar{x}_2$ and s_2^2 are the means and variances of independent random samples of sizes n_1 and n_2, respectively, from approximately normal populations with **unknown and unequal variances**, an approximate $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $t_{\alpha/2}$ is the t-value defined previously with ν as above.

EXAMPLE 9.13. The gasoline prices (in cents/litre) for a random sample of 8 Winnipeg gas stations and 5 Calgary gas stations are recorded one day and are shown below:

Winnipeg:	119.9	122.4	121.7	120.9	
	121.0	122.9	119.9	121.7	
Calgary:	117.9	120.4	118.4	122.9	117.0

Find a 95% confidence interval for the difference in mean gas prices for the two cities.

9.9 Paired Observations

Take STAT 3612 for "Paired Observations"

9.10 Single Sample: Estimating a Proportion

Suppose that we draw a random sample of size n from a large population having population proportion p of successes. Let X be the count of successes in the sample that follows the Binomial distribution with parameters n and p.

Define the sample proportion of successes

$$\widehat{P} = \frac{X}{n}.$$

When n is large, the sampling distribution of $\widehat{P}$ is *approximately* normal with mean

$$\mu_{\widehat{P}} = \mathrm{E}\left(\widehat{P}\right) = \mathrm{E}\left(\frac{X}{n}\right) = \frac{np}{n} = p$$

and variance

$$\sigma_{\widehat{P}}^2 = \mathrm{Var}\left(\widehat{P}\right) = \mathrm{Var}\left(\frac{X}{n}\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

That is,

$$\widehat{P} \,\dot{\sim}\, \mathrm{N}\left(p,\ \sqrt{\frac{p(1-p)}{n}}\right), \quad n \to \infty.$$

or,

$$Z = \frac{\widehat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \;\to\; \mathrm{N}(0,1)$$

NOTE. As a thumb rule, this approximation requires $np \geq 5$ and $n(1-p) \geq 5$.

Now,

$$\begin{aligned}
1-\alpha &= \mathrm{P}\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) \\
&= \mathrm{P}\left(-z_{\alpha/2} < \frac{\widehat{P}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) \\
&\approx \mathrm{P}\left(-z_{\alpha/2} < \frac{\widehat{P}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < z_{\alpha/2}\right) \\
&= \mathrm{P}\left(\widehat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \widehat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)
\end{aligned}$$

where we used the point estimate $\hat{p} = x/n$ to replace p under the radical sign.

Large-Sample Confidence Intervals for p

If $\hat{p}$ is the proportion of successes in a random sample of size n an approximate $100(1-\alpha)\%$ confidence interval, for the binomial parameter p is given by

$$\widehat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \widehat{P} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

NOTE. The (estimated) standard error of $\widehat{P}$ is defined by

$$\text{s.e.}(\widehat{P}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

EXAMPLE 9.14. A question in a Christmas tree market survey was "Did you have a Christmas tree last year?" Of the 500 respondents, 421 answered "Yes."

(a) Find the sample proportion and its standard error.

(b) Give a 90% confidence interval for the proportion of Indiana households who had a Christmas tree this year.

EXAMPLE 9.15. When trying to hire managers and executives, companies sometimes verify the academic credentials described by the applicants. One company that performs these checks summarized its findings for a six-month period. Of the 84 applicants whose credentials were checked, 15 lied about having a degree. (*Data provided by Jude M. Werra & Associates, Brookfield, Wisconsin.*)

(a) Find the proportion of applicants who lied about having a degree and its standard error.

(b) Consider these data to be a random sample of credentials from a large collection of similar applicants. Give a 95% confidence interval for the true proportion of applicants who lie about having a degree.

In a similar fashion, if e is a pre-fixed amount that the error can not exceed, we set $z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < e$. and solve for n to determine the sample size.

Sample Size Determination

If $\hat{p}$ is used as an estimate of p, we can be $100(1-\alpha)\%$ confident that the error will be less than a specified amount e when the sample size is approximately

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

NOTE. In order to ensure the the confidence degree is no less than $100(1-\alpha)\%$, we round all fractional values up to the **next whole number**.

EXAMPLE 9.16. An automobile manufacturer would like to know what proportion of its customers are dissatisfied with the service received from their local dealer. The customer relations department will survey a random sample of customers and compute a 95% confidence interval for the proportion that are dissatisfied. From past studies, they believe that this proportion will be about 0.25. Find the sample size needed if the error of the confidence interval is to be no more than 0.02.

NOTE. If we have no idea what the value of p might be, we can use $\hat{p} = 0.5$ in the sample size formula, i.e.,

$$n = \frac{(z_{\alpha/2})^2}{4e^2}.$$

This is the most conservative estimate of the sample size.

EXAMPLE 9.17. The use of email is growing rapidly and is having a dramatic effect on the way we communicate. Suppose that we want to determine the current proportion of Canadian households using email. How many households must be surveyed to estimate the proportion with a 90% confidence and an error of no more than 3%?

9.11 Two Samples: Estimating the Difference between Two Proportions

We will now turn our attention to the case where we wish to compare two population proportions and would like to estimate the difference in population proportions $p_1 - p_2$, where p_1 and p_2 are the true proportions of all individuals in Population 1 and Population 2 who have some attribute, respectively.

To do this, we will take a random sample of size n_1 from Population 1 and a random sample of size n_2 from Population 2, and then calculate $\hat{p}_1$ and $\hat{p}_2$, the sample proportions from the first and second samples, respectively.

Hence it is quite nature that our point estimate of $p_1 - p_2$ is $\hat{p}_1 - \hat{p}_2$. The mean of $p_1 - p_2$ is

$$\begin{aligned}\mu_{\hat{p}_1-\hat{p}_2} &= \mathsf{E}(\hat{p}_1 - \hat{p}_2) \\ &= \mathsf{E}(\hat{p}_1) - \mathsf{E}(\hat{p}_2) = p_1 - p_2\end{aligned}$$

and, since the sample proportions are independent, the variance of $\hat{p}_1 - \hat{p}_2$ is

$$\begin{aligned}\sigma^2_{\hat{p}_1-\hat{p}_2} &= \mathsf{Var}(\hat{p}_1 - \hat{p}_2) \\ &= \mathsf{Var}(\hat{p}_1) + \mathsf{Var}(\hat{p}_2) \\ &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\end{aligned}$$

If both sample sizes are large, we have the approximate distribution of $\hat{p}_1 - \hat{p}_2$:

$$\hat{p}_1 - \hat{p}_2 \sim \mathrm{N}\left(p_1 - p_2,\ \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

and so

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \dot{\sim} \mathrm{N}(0,\ 1)$$

NOTE. The (estimated) standard error of $\hat{p}_1 - \hat{p}_2$ is given by the estimate of the standard deviation

$$\text{s.e.}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Large-Sample Confidence Interval for $p_1 - p_2$

If $\hat{p}_1$ and $\hat{p}_2$ are the proportions of successes in random samples of sizes n_1 and n_2, respectively, an approximate $100(1-\alpha)\%$ confidence interval for the difference of two binomial parameters, $p_1 - p_2$, is given by

$$\begin{aligned}(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\,\text{s.e.}(\hat{p}_1 - \hat{p}_2) < p_1 - p_2 \\ < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\,\text{s.e.}(\hat{p}_1 - \hat{p}_2)\end{aligned}$$

or

$$\begin{aligned}(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < p_1 - p_2 \\ < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\end{aligned}$$

EXAMPLE 9.18. Do older adults and young adults have different views on Canada's involvement in the war in Afghanistan? A sample of 150 older adults (aged 40 - 65) and a sample of 120 young adults (aged 18 - 30) were selected. Respondents were asked if they approved of Canada's continued involvement in Afghanistan. Of the older adults, 87 said they agree with the decision, while 54 of the young adults said they agree. Let p_1 be the true population proportion of all older adults who agree with the war and let p_2 be the true population proportion of all young adults who agree with the war. Calculate a 95% confidence interval for the difference in population proportions $p_1 - p_2$.

9.12 Single Sample: Estimating the Variance

We have shown that the sample variance S^2 is an unbiased estimator of the population variance σ^2. Thus, S^2 is a point estimate of σ^2.

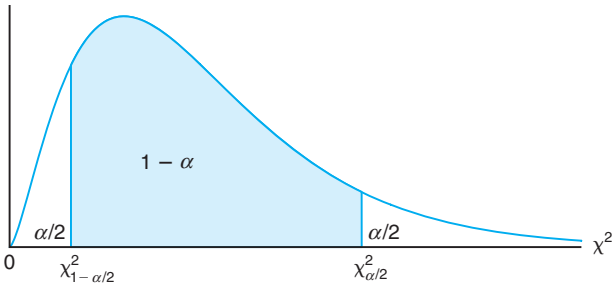
We have also shown that the statistic

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

if random samples of size n are selected from a normal population.

Based on this,

$$\text{P}\left(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$$



And,

$$\begin{aligned}1-\alpha &= \text{P}\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right) \\ &= \text{P}\left(\frac{1}{\chi^2_{\alpha/2}} < \frac{\sigma^2}{(n-1)S^2} < \frac{1}{\chi^2_{1-\alpha/2}}\right) \\ &= \text{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right)\end{aligned}$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are as we defined previously.

Confidence Intervals for σ^2

If s^2 is the variance of a random sample of size n from a normal population, a $100(1-\alpha)\%$ confidence interval for σ^2 is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are χ^2-values with $\nu = n-1$ degrees of freedom, leaving areas of $\alpha/2$ and $1-\alpha/2$, respectively, to the right.

EXAMPLE 9.19. The bottlers of a new soft drink are experiencing problems with the filling mechanism for their 16 floz bottles. To estimate the standard deviation of the fill volume, the filled volume for 20 bottles was measured, yielding a sample standard deviation of 0.1 floz. Compute a 95% confidence interval for the population variance.

NOTE. Under the same setups, a $100(1-\alpha)\%$ confidence interval for σ is given by

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}}$$

EXAMPLE 9.20. Suppose that the data collected from a random sample of 20 observations from a normal population and the sample variance is 100. Construct a 90% confidence interval for the population standard deviation σ.

This page is intentionally blank.