3 Estimation

Complementary reading: Chapter 8 (WMS).

3.1 Introduction

REMARK: Up until now (i.e., in STAT 121 and the material so far in STAT 122), we have dealt with **probability models**. These models, as we know, can be generally divided up into two types: discrete and continuous. These models are used to describe populations of individuals.

- In a clinical trial with n patients, let p denote the probability of response to a new drug. A $b(1, p)$ model is assumed for each subject's response (e.g., respond/not).
- In an engineering application, the lifetime of an electrical circuit, Y, is under investigation. An exponential(β) model is assumed.
- In a public-health study, Y, the number of sexual partners in the past year, is recorded for a group of high-risk HIV patients. A Poisson(λ) model is assumed.
- In an ecological study, the amount of dead-weight (measured in g/plot), Y, is recorded. A $\mathcal{N}(\mu, \sigma^2)$ model is assumed.

Each of these situations employs a probabilistic model that is indexed by population parameters. *In real life, these parameters are unknown.* An important statistical problem, thus, involves **estimating** these parameters with a random sample $Y_1, Y_2, ..., Y_n$ (i.e., an iid sample) from the population. We can state this problem generally as follows.

GENERAL PROBLEM: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a population which is described by the model $f_Y(y; \theta)$. Here, $f_Y(y; \theta)$ is a pmf or pdf that describes the population of interest, and θ is a **parameter** that indexes the model. *The statistical problem of interest is to estimate θ with the observed data $Y_1, Y_2, ..., Y_n$.*

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$. A **point estimator** $\widehat{\theta}$ is a function of $Y_1, Y_2, ..., Y_n$ that estimates θ. Since $\widehat{\theta}$ is (in general) a function of $Y_1, Y_2, ..., Y_n$, it is a **statistic**. In practice, θ could be a scalar or vector.

Example 3.1. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson distribution with mean θ. We know that the probability mass function (pmf) for Y is given by

$$f_Y(y;\theta) = \begin{cases} \frac{\theta^y e^{-\theta}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Here, the parameter is $\theta = E(Y)$. What estimator should we use to estimate θ?

Example 3.2. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ distribution. We know that the probability density function (pdf) for Y is given by

$$f_Y(y;\theta) = \begin{cases} \frac{1}{\theta}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

Here, the parameter is θ, the upper limit of the support of Y. What estimator should we use to estimate θ?

Example 3.3. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. We know that the probability density function (pdf) for Y is given by

$$f_Y(y;\boldsymbol{\theta}) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Here, the parameter is $\boldsymbol{\theta} = (\mu, \sigma^2)$, a vector of two parameters (the mean and the variance). What estimator should we use to estimate $\boldsymbol{\theta}$? Or, equivalently, we might ask how to estimate μ and σ^2 separately.

"GOOD" ESTIMATORS: In general, a "good" estimator $\widehat{\theta}$ has the following properties:

(1) $\widehat{\theta}$ is **unbiased** for θ, and

(2) $\widehat{\theta}$ has **small variance**.

3.2 Bias and mean-squared error

TERMINOLOGY: An estimator $\widehat{\theta}$ is said to be **unbiased** for θ if

$$E(\widehat{\theta}) = \theta,$$

for all possible values of θ. If $\widehat{\theta}$ is not an unbiased estimator; i.e., if $E(\widehat{\theta}) \neq \theta$, then we say that $\widehat{\theta}$ is biased. In general, the **bias** of an estimator is

$$B(\widehat{\theta}) \equiv E(\widehat{\theta}) - \theta.$$

If $B(\widehat{\theta}) > 0$, then $\widehat{\theta}$ overestimates θ. If $B(\widehat{\theta}) < 0$, then $\widehat{\theta}$ underestimates θ. If $\widehat{\theta}$ is unbiased, then, of course, $B(\widehat{\theta}) = 0$.

Example 3.1 (revisited). Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson distribution with mean θ. Recall that, in general, the **sample mean** $\overline{Y}$ is an unbiased estimator for a population mean μ. For the Poisson model, the (population) mean is $\mu = E(Y) = \theta$. Thus, we know that

$$\widehat{\theta} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is an unbiased estimator of θ. Recall also that the variance of the sample mean, $V(\overline{Y})$, is, in general, the population variance σ^2 divided by n. For the Poisson model, the (population) variance is $\sigma^2 = \theta$; thus, $V(\widehat{\theta}) = V(\overline{Y}) = \theta/n$. $\square$

Example 3.2 (revisited). Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ distribution, and consider the point estimator $Y_{(n)}$. Intuitively, this seems like a reasonable estimator to use; the largest order statistic should be fairly close to θ, the upper endpoint of the support. To compute $E(Y_{(n)})$, we have to know how $Y_{(n)}$ is distributed, so we find its pdf. For $0 < y < \theta$, the pdf of $Y_{(n)}$ is

$$\begin{aligned} f_{Y_{(n)}}(y) &= n f_Y(y)[F_Y(y)]^{n-1} \\ &= n\left(\frac{1}{\theta}\right)\left(\frac{y}{\theta}\right)^{n-1} = n\theta^{-n}y^{n-1}, \end{aligned}$$

so that

$$E(Y_{(n)}) = \int_0^\theta y \times \underbrace{n\theta^{-n}y^{n-1}}_{= f_{Y_{(n)}}(y)} dy = n\theta^{-n}\left(\frac{1}{n+1}\right) y^{n+1}\Big|_0^\theta = \left(\frac{n}{n+1}\right)\theta.$$

We see that $Y_{(n)}$ is a **biased estimator** of θ (it underestimates θ on average). But,

$$\widehat{\theta} = \left(\frac{n+1}{n}\right) Y_{(n)}$$

is an unbiased estimator because

$$E(\widehat{\theta}) = E\left[\left(\frac{n+1}{n}\right) Y_{(n)}\right] = \left(\frac{n+1}{n}\right) E(Y_{(n)}) = \left(\frac{n+1}{n}\right)\left(\frac{n}{n+1}\right)\theta = \theta. \ \square$$

EXERCISE: Compute $V(\widehat{\theta})$.

Example 3.3 (revisited). Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. To estimate μ, we know that a good estimator is $\overline{Y}$. The sample mean $\overline{Y}$ is unbiased; i.e., $E(\overline{Y}) = \mu$, and, furthermore, $V(\overline{Y}) = \sigma^2/n$ decreases as the sample size n increases. To estimate σ^2, we can use the **sample variance**; i.e.,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2.$$

Assuming the normal model, the sample variance is unbiased. To see this, recall that

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

so that

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1,$$

since the mean of a χ^2 random variable equals its degrees of freedom. Thus,

$$n-1 = E\left[\frac{(n-1)S^2}{\sigma^2}\right] = \left(\frac{n-1}{\sigma^2}\right) E(S^2) \implies E(S^2) = \sigma^2,$$

showing that S^2 is an unbiased estimator of the population variance σ^2. To compute the variance of S^2 as an estimator, recall that

$$V\left[\frac{(n-1)S^2}{\sigma^2}\right] = 2(n-1),$$

since the variance of a χ^2 random variable equals twice its degrees of freedom. Therefore,

$$\begin{aligned} 2(n-1) = V\left[\frac{(n-1)S^2}{\sigma^2}\right] &= \left[\frac{(n-1)^2}{\sigma^4}\right] V(S^2) \\ &\Longrightarrow \quad V(S^2) = \frac{2\sigma^4}{n-1}. \ \square \end{aligned}$$

ESTIMATING FUNCTIONS OF PARAMETERS: In some problems, the goal is to estimate a function of θ, say, $\tau(\theta)$. The following example illustrates how we can find an unbiased estimator of a function of θ.

Example 3.4. Suppose that $Y_1, Y_2, ..., Y_n$ are iid exponential observations with mean θ. Derive an unbiased estimator for $\tau(\theta) = 1/\theta$.

SOLUTION. Since $E(\overline{Y}) = \theta$, one's intuition might suggest to try $1/\overline{Y}$ as an estimator for $1/\theta$. First, note that

$$E\left(\frac{1}{\overline{Y}}\right) = E\left(\frac{n}{\sum_{i=1}^n Y_i}\right) = nE\left(\frac{1}{T}\right),$$

where $T = \sum_{i=1}^n Y_i$. Recall that $Y_1, Y_2, ..., Y_n$ iid exponential$(\theta) \Longrightarrow T \sim$ gamma(n, θ), so therefore

$$\begin{aligned} E\left(\frac{1}{\overline{Y}}\right) = nE\left(\frac{1}{T}\right) &= n\int_{t=0}^{\infty} \frac{1}{t} \underbrace{\frac{1}{\Gamma(n)\theta^n} t^{n-1}e^{-t/\theta}}_{\text{gamma}(n,\theta) \text{ pdf}} dt \\ &= \frac{n}{\Gamma(n)\theta^n} \underbrace{\int_{t=0}^{\infty} t^{(n-1)-1}e^{-t/\theta} dt}_{=\ \Gamma(n-1)\theta^{n-1}} \\ &= \frac{n\Gamma(n-1)\theta^{n-1}}{\Gamma(n)\theta^n} = \frac{n\Gamma(n-1)}{(n-1)\Gamma(n-1)\theta} = \left(\frac{n}{n-1}\right)\frac{1}{\theta}. \end{aligned}$$

This shows that $1/\overline{Y}$ is a biased estimator of $\tau(\theta) = 1/\theta$. However,

$$E\left(\frac{n-1}{n\overline{Y}}\right) = \left(\frac{n-1}{n}\right) E\left(\frac{1}{\overline{Y}}\right) = \left(\frac{n-1}{n}\right)\left(\frac{n}{n-1}\right)\frac{1}{\theta} = \frac{1}{\theta}.$$

This shows that

$$\widehat{\tau(\theta)} = \frac{n-1}{n\overline{Y}}$$

is an unbiased estimator of $\tau(\theta) = 1/\theta$. $\square$

TERMINOLOGY: The **mean-squared error** (MSE) of a point estimator $\widehat{\theta}$ is given by

$$\text{MSE}(\widehat{\theta}) \equiv E[(\widehat{\theta} - \theta)^2] = V(\widehat{\theta}) + [B(\widehat{\theta})]^2.$$

We see that the MSE combines the

- the precision (variance) of $\widehat{\theta}$ and
- accuracy (bias) of $\widehat{\theta}$.

Of course, if $\widehat{\theta}$ is unbiased for θ, then $\text{MSE}(\widehat{\theta}) = V(\widehat{\theta})$, since $B(\widehat{\theta}) = 0$.

INTUITIVELY: Suppose that we have two unbiased estimators, say, $\widehat{\theta}_1$ and $\widehat{\theta}_2$. Then we would prefer to use the one with the **smaller variance**. That is, if $V(\widehat{\theta}_1) < V(\widehat{\theta}_2)$, then we would prefer $\widehat{\theta}_1$ as an estimator. *Note that it only makes sense to choose an estimator on the basis of its variance when both estimators are unbiased.*

CURIOSITY: Suppose that we have two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ and that both of them are not unbiased (e.g., one could be unbiased and other isn't, or possibly both are biased). On what grounds should we now choose between $\widehat{\theta}_1$ and $\widehat{\theta}_2$? In this situation, a reasonable approach is to choose the estimator with the **smaller mean-squared error**. That is, if $\text{MSE}(\widehat{\theta}_1) < \text{MSE}(\widehat{\theta}_2)$, then we would prefer $\widehat{\theta}_1$ as an estimator.

Example 3.5. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $0 < p < 1$. Define $X = Y_1 + Y_2 + \cdots + Y_n$ and the two estimators

$$\widehat{p}_1 = \frac{X}{n} \quad \text{and} \quad \widehat{p}_2 = \frac{X+2}{n+4}.$$

Which estimator should we use to estimate p?

SOLUTION. First, we should note that $X \sim b(n, p)$, since X is the sum of iid Bernoulli(p) observations. Thus,

$$E(\widehat{p}_1) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$$

(i.e., $\widehat{p}_1$ is unbiased) and

$$E(\widehat{p}_2) = E\left(\frac{X+2}{n+4}\right) = \frac{1}{n+4}E(X+2) = \frac{1}{n+4}[E(X)+2] = \frac{np+2}{n+4}.$$

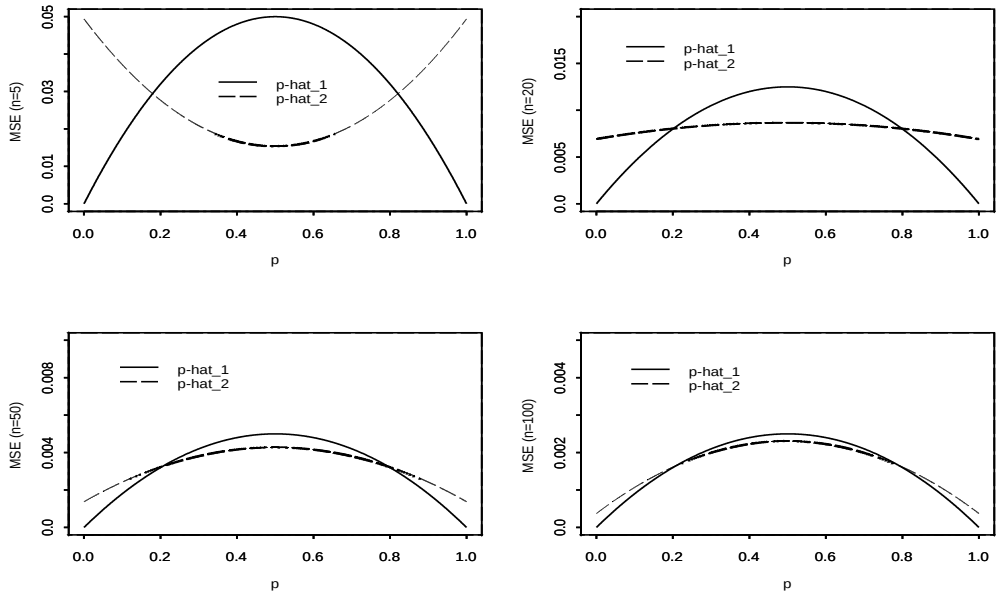

Figure 3.6: *Plots of* MSE($\widehat{p_1}$) *and* MSE($\widehat{p_2}$) *for different sample sizes in Example* 3 .5.

Thus, to compare $\widehat{p}_1$ and $\widehat{p}_2$ as estimators, we should use the estimators' mean-squared errors (since $\widehat{p}_2$ is biased). The variances of $\widehat{p}_1$ and $\widehat{p}_2$ are, respectively,

$$V(\widehat{p}_1) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}[np(1-p)] = \frac{p(1-p)}{n}$$

and

$$V(\widehat{p}_2) = V\left(\frac{X+2}{n+4}\right) = \frac{1}{(n+4)^2}V(X+2) = \frac{1}{(n+4)^2}V(X) = \frac{np(1-p)}{(n+4)^2}.$$

The mean-squared error of $\widehat{p}_1$ is

$$\begin{aligned} \text{MSE}(\widehat{p}_1) &= V(\widehat{p}_1) + [B(\widehat{p}_1)]^2 \\ &= \frac{p(1-p)}{n} + (p-p)^2 = \frac{p(1-p)}{n}, \end{aligned}$$

which is equal to $V(\widehat{p}_1)$ since $\widehat{p}_1$ is unbiased. The mean-squared error of $\widehat{p}_2$ is

$$\begin{aligned} \text{MSE}(\widehat{p}_2) &= V(\widehat{p}_2) + [B(\widehat{p}_2)]^2 \\ &= \frac{np(1-p)}{(n+4)^2} + \left(\frac{np+2}{n+4} - p\right)^2. \end{aligned}$$

ANALYSIS: Figure 8.6 displays values of MSE($\widehat{p}_1$) and MSE($\widehat{p}_2$) graphically for $n = 5$, 20, 50, and 100. We can see that neither estimator is uniformly superior; i.e., neither estimator delivers a smaller MSE for all $0 < p < 1$. However, for smaller sample sizes, $\widehat{p}_2$ often beats $\widehat{p}_1$ (in terms of MSE) when p is in the vicinity of 0.5; otherwise, $\widehat{p}_1$ often provides smaller MSE.

3.3 The standard error of an estimator

TERMINOLOGY: The **standard error** of a point estimator $\widehat{\theta}$ is simply the standard deviation of the estimator. We denote the standard error of $\widehat{\theta}$ by

$$\sigma_{\widehat{\theta}} = \sqrt{V(\widehat{\theta})}.$$

Table 8.1 (WMS, pp 397) summarizes some common point estimators and their standard errors. We now review these.

3.3.1 One population mean

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample with mean μ and variance σ^2 and that interest lies in estimating the **population mean** μ.

POINT ESTIMATOR: To estimate the (population) mean μ, a natural point estimator to use is the **sample mean**; i.e.,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

FACTS: We have shown that, in general,

$$\begin{aligned} E(\overline{Y}) &= \mu \\ V(\overline{Y}) &= \frac{\sigma^2}{n}. \end{aligned}$$

STANDARD ERROR: The **standard error** of $\overline{Y}$ is equal to

$$\sigma_{\overline{Y}} = \sqrt{V(\overline{Y})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

3.3.2 One population proportion

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $0 < p < 1$, and that interest lies in estimating the **population proportion** p. Recall that $X = \sum_{i=1}^{n} Y_i \sim b(n, p)$, since X is the sum of iid Bernoulli(p) observations.

POINT ESTIMATOR: To estimate the (population) proportion p, a natural point estimator to use is the **sample proportion**; i.e.,

$$\widehat{p} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

FACTS: It is easy to show (verify!) that

$$\begin{aligned} E(\widehat{p}) &= p \\ V(\widehat{p}) &= \frac{p(1-p)}{n}. \end{aligned}$$

STANDARD ERROR: The **standard error** of $\widehat{p}$ is equal to

$$\sigma_{\widehat{p}} = \sqrt{V(\widehat{p})} = \sqrt{\frac{p(1-p)}{n}}.$$

3.3.3 Difference of two population means

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$ are iid with mean μ_1 and variance σ_1^2

Sample 2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$ are iid with mean μ_2 and variance σ_2^2

and that interest lies in estimating the **population mean difference** $\theta \equiv \mu_1 - \mu_2$. As noted, we assume that the samples themselves are independent (i.e., observations from one sample are independent from observations in the other sample).

NEW NOTATION: Because we have two samples, we need to adjust our notation accordingly. Here, we use the conventional notation Y_{ij} to denote the jth observation from

sample i, for $i = 1, 2$ and $j = 1, 2, ..., n_i$. The symbol n_i denotes the sample size from sample i. It is not necessary that the sample sizes n_1 and n_2 are equal.

POINT ESTIMATOR: To estimate the population mean difference $\theta = \mu_1 - \mu_2$, a natural point estimator to use is the **difference of the sample means**; i.e.,

$$\widehat{\theta} \equiv \overline{Y}_{1+} - \overline{Y}_{2+},$$

where

$$\overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad \overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j}.$$

This notation is also standard; the "+" symbol is understood to mean that the subscript it replaces has been "summed over."

FACTS: It is easy to show (verify!) that

$$\begin{aligned} E(\overline{Y}_{1+} - \overline{Y}_{2+}) &= \mu_1 - \mu_2 \\ V(\overline{Y}_{1+} - \overline{Y}_{2+}) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned}$$

STANDARD ERROR: The **standard error** of $\widehat{\theta} = \overline{Y}_{1+} - \overline{Y}_{2+}$ is equal to

$$\sigma_{\overline{Y}_{1+} - \overline{Y}_{2+}} = \sqrt{V(\overline{Y}_{1+} - \overline{Y}_{2+})} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

3.3.4 Difference of two population proportions

SITUATION: Suppose that we have two **independent** samples; i.e.,

$$\begin{aligned} &\text{Sample 1:} \quad Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ are iid Bernoulli}(p_1) \\ &\text{Sample 2:} \quad Y_{21}, Y_{22}, ..., Y_{2n_2} \text{ are iid Bernoulli}(p_2) \end{aligned}$$

and that interest lies in estimating the **population proportion difference** $\theta \equiv p_1 - p_2$. Again, it is not necessary that the sample sizes n_1 and n_2 are equal. As noted, we assume that the samples themselves are independent (i.e., observations from one sample are independent from observations in the other sample). Define

$$X_1 = \sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad X_2 = \sum_{j=1}^{n_2} Y_{2j}.$$

We know that $X_1 \sim b(n_1, p_1)$, $X_2 \sim b(n_2, p_2)$, and that X_1 and X_2 are independent (since the samples are). The **sample proportions** are

$$\widehat{p}_1 = \frac{X_1}{n_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad \widehat{p}_2 = \frac{X_2}{n_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j}.$$

POINT ESTIMATOR: To estimate the population proportion difference $\theta = p_1 - p_2$, a natural point estimator to use is the **difference of the sample proportions**; i.e.,

$$\widehat{\theta} \equiv \widehat{p}_1 - \widehat{p}_2.$$

FACTS: It is easy to show (verify!) that

$$\begin{aligned} E(\widehat{p}_1 - \widehat{p}_2) &= p_1 - p_2 \\ V(\widehat{p}_1 - \widehat{p}_2) &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \end{aligned}$$

STANDARD ERROR: The **standard error** of $\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2$ is equal to

$$\sigma_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{V(\widehat{p}_1 - \widehat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

3.4 Estimating the population variance

RECALL: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample with mean μ and variance σ^2. The **sample variance** is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

In Example 3.3 (notes), we showed that if $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, then the sample variance S^2 is an **unbiased estimator** of the population variance σ^2.

NEW RESULT: That S^2 is an unbiased estimator of σ^2 holds in general; that is, as long as $Y_1, Y_2, ..., Y_n$ is an iid sample with mean μ and variance σ^2,

$$E(S^2) = \sigma^2;$$

that is, S^2 is an unbiased estimator of σ^2, *regardless of the population distribution*, as long as $\sigma^2 < \infty$. The proof of this result is given on pp 398-399 in WMS.

3.5 Error bounds and the Empirical Rule

TERMINOLOGY: We are often interested in understanding how close our estimator $\widehat{\theta}$ is to a population parameter θ. Of course, in real life, θ is unknown, so we can never know for sure. However, we can make probabilistic statements regarding the closeness of $\widehat{\theta}$ and θ. We call $\epsilon = |\widehat{\theta} - \theta|$ the **error in estimation**.

THE EMPIRICAL RULE: Suppose the estimator $\widehat{\theta}$ has an approximate normal **sampling distribution** with mean θ and variance $\sigma^2_{\widehat{\theta}}$. It follows then that

- about 68 percent of the values of $\widehat{\theta}$ will fall between $\theta \pm \sigma_{\widehat{\theta}}$
- about 95 percent of the values of $\widehat{\theta}$ will fall between $\theta \pm 2\sigma_{\widehat{\theta}}$
- about 99.7 percent (or nearly all) of the values of $\widehat{\theta}$ will fall between $\theta \pm 3\sigma_{\widehat{\theta}}$.

These facts follow directly from the normal distribution. For example, with $Z \sim \mathcal{N}(0,1)$, we compute

$$\begin{aligned} P\left(\theta - \sigma_{\widehat{\theta}} < \widehat{\theta} < \theta + \sigma_{\widehat{\theta}}\right) &= P\left(\frac{\theta - \sigma_{\widehat{\theta}} - \theta}{\sigma_{\widehat{\theta}}} < \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} < \frac{\theta + \sigma_{\widehat{\theta}} - \theta}{\sigma_{\widehat{\theta}}}\right) \\ &\approx P(-1 < Z < 1) \\ &= F_Z(1) - F_Z(-1) \\ &= 0.8413 - 0.1587 = 0.6826, \end{aligned}$$

where $F_Z(\cdot)$ denotes the cdf of the standard normal distribution.

REMARK: Most estimators $\widehat{\theta}$, with probability "in the vicinity of" 0.95, will fall within two standard deviations (standard errors) of its mean. Thus, if $\widehat{\theta}$ is an unbiased estimator of θ, or is approximately unbiased, then $b = 2\sigma_{\widehat{\theta}}$ serves as a good approximate **upper bound** for the error in estimation; that is, $\epsilon = |\widehat{\theta} - \theta| \leq 2\sigma_{\widehat{\theta}}$ with "high" probability.

Example 3.6. In an agricultural experiment, we observe an iid sample of n yields, say, $Y_1, Y_2, ..., Y_n$, measured in kg/area per plot. We can estimate the (population) mean yield

μ with $\overline{Y}$, the sample mean; from the Central Limit Theorem, we know that

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right),$$

for large n. Thus, $b = 2\sigma/\sqrt{n}$ serves as an approximate 95 percent bound on the error in estimation $\epsilon = |\overline{Y} - \mu|$. $\square$

Example 3.7. In a public-health study involving intravenous drug users, subjects are tested for HIV. Denote the HIV statuses by $Y_1, Y_2, ..., Y_n$ and assume these statuses are iid Bernoulli(p) random variables (e.g., 1, if positive; 0, otherwise). The sample proportion of HIV infecteds, then, is given by

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Recall that for n large,

$$\widehat{p} \sim \mathcal{AN}\left[p, \frac{p(1-p)}{n}\right].$$

Thus, $b = 2\sqrt{p(1-p)/n}$ serves as an approximate 95 percent bound on the error in estimation $\epsilon = |\widehat{p} - p|$. $\square$

REMARK: To use the Empirical Rule, we need the sampling distribution of $\widehat{\theta}$ to be normally distributed, or, at least, approximately normally distributed. Otherwise, the Empirical Rule may provide incorrect results. If we have an estimator $\widehat{\theta}$ that does not follow a normal distribution, we could use Chebyshev's Inequality to put a bound on the error in estimation ϵ. Recall that **Chebyshev's Inequality** says

$$P(|\widehat{\theta} - \theta| < k\sigma_{\widehat{\theta}}) \geq 1 - \frac{1}{k^2},$$

for any value $k > 0$. For example, if $k = 2$, then $b = 2\sigma_{\widehat{\theta}}$ is an **at least** 75 percent bound on the error in estimation $\epsilon = |\widehat{\theta} - \theta|$.

3.6 Confidence intervals and pivotal quantities

REMARK: A **point estimator** $\widehat{\theta}$ provides a "one-shot guess" of the value of an unknown parameter θ. On the other hand, an interval estimator, or **confidence interval**, provides a range of values that is likely to contain θ.

Table 3.1: *Manufacturing part length data. These observations are modeled as* $n = 10$ *realizations from a* $\mathcal{N}(\mu, \sigma^2)$ *distribution.*

12.2	12.0	12.2	11.9	12.4	12.6	12.1	12.2	12.9	12.4

Example 3.8. The length of a critical part, measured in mm, in a manufacturing process varies according to a $\mathcal{N}(\mu, \sigma^2)$ distribution (this is a model assumption). Engineers plan to observe an iid sample of $n = 10$ parts and record $Y_1, Y_2, ..., Y_{10}$. The observed data from the experiment are given in Table 3.1.

POINT ESTIMATES: The sample mean computed with the observed data is $\bar{y} = 12.3$ and sample variance is $s^2 = 0.09$ (verify!). The sample mean $\bar{y} = 12.3$ is a **point estimate** for the population mean μ. Similarly, the sample variance $s^2 = 0.09$ is a **point estimate** for the population variance σ^2. However, neither of these estimates has a measure of variability associated with it; that is, both estimates are just single "one-number" values.

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution (probability model) described by $f_Y(y; \theta)$. *Informally, a confidence interval is an interval of plausible values for a parameter* θ. More specifically, if θ is our parameter of interest, then we call $(\widehat{\theta}_L, \widehat{\theta}_U)$ a $100(1-\alpha)$ **percent confidence interval** for θ if

$$P(\widehat{\theta}_L \leq \theta \leq \widehat{\theta}_U) = 1 - \alpha,$$

where $0 < \alpha < 1$. We call $1 - \alpha$ the **confidence level**. In practice, we would like the confidence level $1 - \alpha$ to be large (e.g., 0.90, 0.95, 0.99, etc.).

IMPORTANT: Before we observe $Y_1, Y_2, ..., Y_n$, the interval $(\widehat{\theta}_L, \widehat{\theta}_U)$ is a **random** interval. This is true because $\widehat{\theta}_L$ and $\widehat{\theta}_U$ are random quantities as they will be functions of $Y_1, Y_2, ..., Y_n$. On the other hand, θ is a **fixed** parameter; it's value does not change. After we see the data $Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n$, like those data in Table 3.1, the numerical interval $(\widehat{\theta}_L, \widehat{\theta}_U)$ based on the realizations $y_1, y_2, ..., y_n$ is no longer random.

Example 3.9. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, {\sigma_0}^2)$ sample, where the mean μ is unknown and variance σ_0^2 is **known**. In this example, we focus on the population mean μ. From past results, we know that $\overline{Y} \sim \mathcal{N}(\mu, \sigma_0^2/n)$. Thus,

$$Z = \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}(0,1);$$

i.e., Z has a standard normal distribution. We know there exists a value $z_{\alpha/2}$ such that

$$\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= P\left(-z_{\alpha/2} < \frac{\overline{Y}-\mu}{\sigma_0/\sqrt{n}} < z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} < \overline{Y}-\mu < z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\
&= P\left(z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} > \mu - \overline{Y} > -z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\
&= P\left(\overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} > \mu > \overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\
&= P\Bigg(\underbrace{\overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}}_{\widehat{\theta}_L} < \mu < \underbrace{\overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}}_{\widehat{\theta}_U}\Bigg).
\end{aligned}$$

These calculations show that

$$\overline{Y} \pm z_{\alpha/2}\left(\frac{\sigma_0}{\sqrt{n}}\right)$$

is a $100(1-\alpha)$ percent confidence interval for the population mean μ. The probability that the **random** interval $(\widehat{\theta}_L, \widehat{\theta}_U)$ includes the mean μ is $1-\alpha$. $\square$

Example 3.8 (revisited). In Example 8.8, suppose that the population variance for the distribution of part lengths is ${\sigma_0}^2 = 0.1$, so that $\sigma_0 \approx 0.32$ (we did not make this assumption before) and that we would like to construct a 95 percent confidence interval for μ, the mean length. From the data in Table 3.1, we have $n = 10$, $\bar{y} = 12.3$, $\alpha = 0.05$, and $z_{0.025} = 1.96$ (z-table). A 95 percent confidence interval for μ is

$$12.3 \pm 1.96\left(\frac{0.32}{\sqrt{10}}\right) \Longrightarrow (12.1, 12.5).$$

INTERPRETATION: We are 95 percent confident that the population mean length μ is between 12.1 and 12.5 mm. $\square$

NOTE: The interval $(12.1, 12.5)$ is no longer random! Thus, it is not theoretically appropriate to say that "the mean length μ is between 12.1 and 12.5 with probability 0.95." A confidence interval, after it has been computed with actual data (like above), no longer possesses any randomness. We only attach probabilities to events involving random quantities.

INTERPRETATION: Instead of attaching the concept of probability to the interpretation of a confidence interval, here is how one must think about them. *In repeated sampling, approximately* $100(1-\alpha)$ *percent of the confidence intervals will contain the true parameter* θ*. Our calculated interval is just one of these.*

TERMINOLOGY: We call the quantity Q a **pivotal quantity**, or a **pivot**, if its sampling distribution does not depend on any unknown parameters. Note that Q can depend on unknown parameters, but its sampling distribution can not. *Pivots help us derive confidence intervals.* Illustrative examples now follow.

Example 3.10. In Example 3.9, the quantity

$$Z = \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Since the standard normal distribution does not depend on any unknown parameters, Z is a pivot. We used this fact to derive a $100(1-\alpha)$ confidence interval for the population mean μ, when $\sigma^2 = \sigma_0^2$ was known. $\square$

Example 3.11. The time (in seconds) for a certain chemical reaction to take place is assumed to follow a $\mathcal{U}(0,\theta)$ distribution. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of such times and that we would like to derive a $100(1-\alpha)$ percent confidence interval for θ, the maximum possible time. Intuitively, the largest order statistic $Y_{(n)}$ should be "close" to θ, so let's use $Y_{(n)}$ as an estimator. From Example 8.2, the pdf of $Y_{(n)}$ is given by

$$f_{Y_{(n)}}(y) = \begin{cases} n\theta^{-n}y^{n-1}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

As we will now show,

$$Q = \frac{Y_{(n)}}{\theta}$$

is a pivot. We can show this using a transformation argument. With $q = y_{(n)}/\theta$, the inverse transformation is given by $y_{(n)} = q\theta$ and the Jacobian is $dy_{(n)}/dq = \theta$. Thus, the pdf of Q, for values of $0 < q < 1$ (why?), is given by

$$\begin{aligned} f_Q(q) &= f_{Y_{(n)}}(q\theta) \times |\theta| \\ &= n\theta^{-n}(q\theta)^{n-1} \times \theta \\ &= nq^{n-1}. \end{aligned}$$

You should recognize that $Q \sim \text{beta}(n, 1)$. Since Q has a distribution free of unknown parameters, Q is a pivot, as claimed.

USING THE PIVOT: Define b as the value that satisfies $P(Q > b) = 1 - \alpha$. That is, b solves

$$1 - \alpha = P(Q > b) = \int_b^1 nq^{n-1}dq = 1 - b^n,$$

so that $b = \alpha^{1/n}$. Recognizing that $P(Q > b) = P(b < Q < 1)$, it follows that

$$\begin{aligned} 1 - \alpha = P(\alpha^{1/n} < Q < 1) &= P\left(\alpha^{1/n} < \frac{Y_{(n)}}{\theta} < 1\right) \\ &= P\left(\alpha^{-1/n} > \frac{\theta}{Y_{(n)}} > 1\right) \\ &= P\left(Y_{(n)} < \theta < \alpha^{-1/n}Y_{(n)}\right). \end{aligned}$$

This argument shows that

$$(Y_{(n)}, \alpha^{-1/n}Y_{(n)})$$

is a $100(1 - \alpha)$ percent confidence interval for the unknown parameter θ. $\square$

Example 3.11 (revisited). Table 3.2 contains $n = 36$ chemical reaction times, modeled as iid $\mathcal{U}(0, \theta)$ realizations. The largest order statistic is $y_{(36)} = 9.962$. With $\alpha = 0.05$, a 95 percent confidence interval for θ is

$$(9.962,\ (0.05)^{-1/36} \times 9.962) \Longrightarrow (9.962, 10.826).$$

Thus, we are 95 percent confident that the maximum reaction time θ is between 9.962 and 10.826 seconds. $\square$

Table 3.2: *Chemical reaction data. These observations are modeled as $n = 36$ realizations from $\mathcal{U}(0, \theta)$ distribution.*

0.478	0.787	1.102	0.851	8.522	5.272	4.113	7.921	3.457
3.457	9.159	6.344	6.481	4.448	5.756	0.076	3.462	**9.962**
2.938	3.281	5.481	1.232	5.175	5.864	8.176	2.031	1.633
4.803	8.249	8.991	7.358	2.777	5.905	7.762	8.563	7.619

Example 3.12. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential distribution with mean θ and that we would like to estimate θ with a $100(1-\alpha)$ percent confidence interval. Recall that

$$T = \sum_{i=1}^{n} Y_i \sim \text{gamma}(n, \theta)$$

and that

$$Q = \frac{2T}{\theta} \sim \chi^2(2n).$$

Thus, since Q has a distribution free of unknown parameters, Q is a pivot. Because $Q \sim \chi^2(2n)$, we can trap Q between two quantiles from the $\chi^2(2n)$ distribution with probability $1-\alpha$. In particular, let $\chi^2_{2n,1-\alpha/2}$ and $\chi^2_{2n,\alpha/2}$ denote the lower and upper $\alpha/2$ quantiles of a $\chi^2(2n)$ distribution; that is, $\chi^2_{2n,1-\alpha/2}$ solves

$$P(Q < \chi^2_{2n,1-\alpha/2}) = \alpha/2$$

and $\chi^2_{2n,\alpha/2}$ solves

$$P(Q > \chi^2_{2n,\alpha/2}) = \alpha/2.$$

Recall that the χ^2 distribution is tabled in Table 6 (WMS); the quantiles $\chi^2_{2n,1-\alpha/2}$ and $\chi^2_{2n,\alpha/2}$ can be found in this table (or by using R). We have that

$$\begin{aligned} 1-\alpha = P(\chi^2_{2n,1-\alpha/2} < Q < \chi^2_{2n,\alpha/2}) &= P\left(\chi^2_{2n,1-\alpha/2} < \frac{2T}{\theta} < \chi^2_{2n,\alpha/2}\right) \\ &= P\left(\frac{1}{\chi^2_{2n,1-\alpha/2}} > \frac{\theta}{2T} > \frac{1}{\chi^2_{2n,\alpha/2}}\right) \\ &= P\left(\frac{2T}{\chi^2_{2n,\alpha/2}} < \theta < \frac{2T}{\chi^2_{2n,1-\alpha/2}}\right). \end{aligned}$$

Table 3.3: *Observed explosion data. These observations are modeled as $n = 8$ realizations from an exponential distribution with mean θ.*

3.690	14.091	1.989	0.047	8.114	4.996	20.734	6.975

This argument shows that

$$\left(\frac{2T}{\chi^2_{2n,\alpha/2}}, \frac{2T}{\chi^2_{2n,1-\alpha/2}} \right)$$

is a $100(1-\alpha)$ percent confidence interval for θ. $\square$

Example 3.12 (revisited). Explosive devices used in mining operations produce nearly circular craters when detonated. The radii of these craters, measured in feet, follow an exponential distribution with mean θ. An iid sample of $n = 8$ explosions is observed and the radii observed in the explosions are catalogued in Table 3.3. With these data, we would like to write a 90 percent confidence interval for θ. The sum of the radii is $t = \sum_{i=1}^{8} y_i = 60.636$. With $n = 8$ and $\alpha = 0.10$, we find (from WMS, Table 6),

$$\begin{aligned} \chi^2_{16,0.95} &= 7.96164 \\ \chi^2_{16,0.05} &= 26.2962. \end{aligned}$$

A 90 percent confidence interval for θ based on these data is

$$\left(\frac{2 \times 60.636}{26.2962}, \frac{2 \times 60.636}{7.96164} \right) \Longrightarrow (4.612, 15.232).$$

Thus, we are 90 percent confident that the mean crater radius θ is between 4.612 and 15.232 feet. $\square$

3.7 Large-sample confidence intervals

TERMINOLOGY: The terms "large-sample" and/or "asymptotic" are used to describe confidence intervals that are constructed from asymptotic theory. Of course, the main asymptotic result we have seen so far is the **Central Limit Theorem**. This theorem provides the basis for the large-sample intervals studied in this subsection.

GOALS: In particular, we will present **large-sample confidence intervals** for

1. one population mean μ
2. one population proportion p
3. the difference of two population means $\mu_1 - \mu_2$
4. the difference of two population proportions $p_1 - p_2$.

Because these are "large-sample" confidence intervals, this means that the intervals are approximate, so their true confidence levels are "close" to $1 - \alpha$ for large sample sizes.

LARGE-SAMPLE APPROACH: In each of the situations listed above, we will use a point estimator, say, $\widehat{\theta}$, which satisfies

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{AN}(0, 1),$$

for large sample sizes. In this situation, we say that Z is an **asymptotic pivot** because its large-sample distribution is free of all unknown parameters. Because Z follows an approximate standard normal distribution, we can find a value $z_{\alpha/2}$ that satisfies

$$P\left(-z_{\alpha/2} < \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} < z_{\alpha/2}\right) \approx 1 - \alpha,$$

which, after straightforward algebra (verify!), can be restated as

$$P\left(\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}} < \theta < \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}\right) \approx 1 - \alpha.$$

This shows that

$$\widehat{\theta} \pm z_{\alpha/2}\sigma_{\widehat{\theta}}$$

is an **approximate** $100(1 - \alpha)$ percent confidence interval for the parameter θ.

PROBLEM: As we will see shortly, the standard error $\sigma_{\widehat{\theta}}$ will often depend on unknown parameters (either θ itself or other unknown parameters). This is a problem, because we are trying to compute a confidence interval for θ, and the standard error $\sigma_{\widehat{\theta}}$ depends on population parameters which are not known.

SOLUTION: If we can substitute a "good" estimator for $\sigma_{\widehat{\theta}}$, say, $\widehat{\sigma}_{\widehat{\theta}}$, then the interval

$$\widehat{\theta} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}$$

should remain valid in large samples. The theoretical justification as to why this approach is, in fact, reasonable will be seen in the next chapter.

"GOOD" ESTIMATOR: In the preceding paragraph, the term "good" is used to describe an estimator $\widehat{\sigma}_{\widehat{\theta}}$ that "approaches" the true standard error $\sigma_{\widehat{\theta}}$ (in some sense) as the sample size(s) become(s) large. Thus, we have two approximations at play:

- the Central Limit Theorem that approximates the true sampling distribution of
$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}}$$
- the approximation arising from using $\widehat{\sigma}_{\widehat{\theta}}$ as an estimate of $\sigma_{\widehat{\theta}}$.

TERMINOLOGY: I like to call $\widehat{\sigma}_{\widehat{\theta}}$ the **estimated standard error**. It is simply a point estimate of the true standard error $\sigma_{\widehat{\theta}}$.

APPROXIMATE CONFIDENCE INTERVALS: We will use

$$\widehat{\theta} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}$$

as an **approximate** $100(1-\alpha)$ percent confidence interval for θ. We now present this interval in the context of our four scenarios described earlier. *Each of the following intervals is valid for large sample sizes.* These intervals may not be valid for small sample sizes.

3.7.1 One population mean

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample with mean μ and variance σ^2 and that interest lies in estimating the population mean μ. In this situation, the Central Limit Theorem says that

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{AN}(0,1)$$

for large n. Here,

$$\begin{aligned} \theta &= \mu \\ \widehat{\theta} &= \overline{Y} \\ \sigma_{\widehat{\theta}} &= \frac{\sigma}{\sqrt{n}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \frac{S}{\sqrt{n}}, \end{aligned}$$

where S denotes the sample standard deviation. Thus,

$$\overline{Y} \pm z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

is an approximate $100(1-\alpha)$ percent confidence interval for the population mean μ.

Example 3.13. The administrators for a hospital would like to estimate the mean number of days required for in-patient treatment of patients between the ages of 25 and 34 years . A random sample of $n = 500$ hospital patients between these ages produced the following sample statistics:

$$\begin{aligned} \bar{y} &= 5.4 \text{ days} \\ s &= 3.1 \text{ days.} \end{aligned}$$

Construct a 90 percent confidence interval for μ, the (population) mean length of stay for this cohort of patients.

SOLUTION. Here, $n = 500$ and $z_{0.10/2} = z_{0.05} = 1.65$. Thus, a 90 percent confidence interval for μ is

$$5.4 \pm 1.65\left(\frac{3.1}{\sqrt{500}}\right) \implies (5.2, 5.6) \text{ days.}$$

We are 90 percent confident that the true mean length of stay, for patients aged 25-34, is between 5.2 and 5.6 days. □

3.7.2 One population proportion

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $0 < p < 1$, and that interest lies in estimating the population proportion p. In this situation, the

Central Limit Theorem says that

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{AN}(0,1)$$

for large n, where $\widehat{p}$ denotes the sample proportion. Here,

$$\begin{aligned} \theta &= p \\ \widehat{\theta} &= \widehat{p} \\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{p(1-p)}{n}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}. \end{aligned}$$

Thus,

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for p.

Example 3.14. The Women's Interagency HIV Study (WIHS) is a large observational study funded by the National Institutes of Health to investigate the effects of HIV infection in women. The WIHS study reports that a total of 1288 HIV-infected women were recruited to examine the prevalence of childhood abuse. Of the 1288 HIV positive women, a total of 399 reported that, in fact, they had been a victim of childhood abuse. Find a 95 percent confidence interval for p, the true proportion of HIV infected women who are victims of childhood abuse.

SOLUTION. Here, $n = 1288$, $z_{0.05/2} = z_{0.025} = 1.96$, and the sample proportion of HIV childhood abuse victims is

$$\widehat{p} = \frac{399}{1288} \approx 0.31.$$

Thus, a 95 percent confidence interval for p is

$$0.31 \pm 1.96\sqrt{\frac{0.31(1-0.31)}{1288}} \implies (0.28, 0.34).$$

We are 95 percent confident that the true proportion of HIV infected women who are victims of childhood abuse is between 0.28 and 0.34. □

3.7.3 Difference of two population means

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$ are iid with mean μ_1 and variance σ_1^2

Sample 2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$ are iid with mean μ_2 and variance σ_2^2

and that interest lies in estimating the population mean difference $\mu_1 - \mu_2$. In this situation, the Central Limit Theorem says that

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{AN}(0,1)$$

for large n_1 and n_2. Here,

$$\begin{aligned} \theta &= \mu_1 - \mu_2 \\ \widehat{\theta} &= \overline{Y}_{1+} - \overline{Y}_{2+} \\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \end{aligned}$$

where S_1^2 and S_2^2 are the respective sample variances. Thus,

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for the mean difference $\mu_1 - \mu_2$.

Example 3.15. A botanist is interested in comparing the growth response of dwarf pea stems to different levels of the hormone indoleacetic acid (IAA). Using 16-day-old pea plants, the botanist obtains 5-millimeter sections and floats these sections on solutions with different hormone concentrations to observe the effect of the hormone on the growth of the pea stem. Let Y_1 and Y_2 denote, respectively, the independent growths that can be attributed to the hormone during the first 26 hours after sectioning for $\frac{1}{2}10^{-4}$ and 10^{-4} levels of concentration of IAA (measured in mm). Summary statistics from the study are given in Table 3.4.

Table 3.4: *Botany data. Summary statistics for pea stem growth by hormone treatment.*

Treatment	Sample size	Sample mean	Sample standard deviation
$\frac{1}{2}10^{-4}$ mm IAA	$n_1 = 53$	$\overline{y}_{1+} = 1.03$	$s_1 = 0.49$
10^{-4} mm IAA	$n_2 = 51$	$\overline{y}_{2+} = 1.66$	$s_2 = 0.59$

The researcher would like to construct a 99 percent confidence interval for $\mu_1 - \mu_2$, the mean difference in growths for the two IAA levels. This confidence interval is

$$(1.03 - 1.66) \pm 2.58\sqrt{\frac{(0.49)^2}{53} + \frac{(0.59)^2}{51}} \implies (-0.90, -0.36).$$

That is, we are 99 percent confident that the mean difference $\mu_1 - \mu_2$ is between -0.90 and -0.36. Note that, because this interval does not conclude 0, this analysis suggests that the two (population) means are, in fact, truly different. □

3.7.4 Difference of two population proportions

SITUATION: Suppose that we have two **independent** samples; i.e.,

$$\begin{aligned} &\text{Sample 1:} \quad Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ are iid Bernoulli}(p_1) \\ &\text{Sample 2:} \quad Y_{21}, Y_{22}, ..., Y_{2n_2} \text{ are iid Bernoulli}(p_2) \end{aligned}$$

and that interest lies in estimating the population proportion difference $p_1 - p_2$. In this situation, the Central Limit Theorem says that

$$Z = \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathcal{AN}(0, 1)$$

for large n_1 and n_2, where $\widehat{p}_1$ and $\widehat{p}_2$ are the sample proportions. Here,

$$\begin{aligned} \theta &= p_1 - p_2 \\ \widehat{\theta} &= \widehat{p}_1 - \widehat{p}_2 \\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}. \end{aligned}$$

Thus,

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for the population proportion difference $p_1 - p_2$.

Example 3.16. An experimental type of chicken feed, Ration 1, contains a large amount of an ingredient that enables farmers to raise heavier chickens. However, this feed may be too strong and the mortality rate may be higher than that with the usual feed. One researcher wished to compare the mortality rate of chickens fed Ration 1 with the mortality rate of chickens fed the current best-selling feed, Ration 2. Denote by p_1 and p_2 the population mortality rates (proportions) for Ration 1 and Ration 2, respectively. She would like to get a 95 percent confidence interval for $p_1 - p_2$. Two hundred chickens were randomly assigned to each ration; of those fed Ration 1, 24 died within one week; of those fed Ration 2, 16 died within one week.

- Sample 1: 200 chickens fed Ration 1 $\Longrightarrow \widehat{p}_1 = 24/200 = 0.12$
- Sample 2: 200 chickens fed Ration 2 $\Longrightarrow \widehat{p}_2 = 16/200 = 0.08$.

An approximate 95 percent confidence interval for the true difference $p_1 - p_2$ is

$$(0.12 - 0.08) \pm 1.96\sqrt{\frac{0.12(1-0.12)}{200} + \frac{0.08(1-0.08)}{200}} \Longrightarrow (-0.02, 0.10).$$

Thus, we are 95 percent confident that the true difference in mortality rates is between -0.02 and 0.10. Note that this interval does include 0, so we do not have strong (statistical) evidence that the mortality rates (p_1 and p_2) are truly different. $\square$

3.8 Sample size determinations

MOTIVATION: In many research investigations, it is of interest to determine how many observations are needed to write a $100(1-\alpha)$ percent confidence interval with a given precision. For example, we might want to construct a 95 percent confidence interval for a

population mean in a way so that the confidence interval length is no more than 5 units (e.g., days, inches, dollars, etc.). Sample-size determinations ubiquitously surface in agricultural experiments, clinical trials, engineering investigations, epidemiological studies, etc., and, in most real problems, there is no "free lunch." Collecting more data costs money! Thus, one must be cognizant not only of the statistical issues associated with sample-size determination, but also of the practical issues like cost, time spent in data collection, personnel training, etc.

3.8.1 One population mean

SIMPLE SETTING: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma_0^2)$ population, where σ_0^2 is known. In this situation, an exact $100(1-\alpha)$ percent confidence interval for μ is given by

$$\overline{Y} \pm \underbrace{z_{\alpha/2}\left(\frac{\sigma_0}{\sqrt{n}}\right)}_{=B, \text{ say}},$$

where B denotes the bound on the error in estimation; this bound is called the **margin of error**.

SAMPLE SIZE FORMULA: In the setting described above, it is possible to determine the sample size n necessary once we specify these two pieces of information:

- the confidence level, $100(1-\alpha)$
- the margin of error, B.

This is true because

$$B = z_{\alpha/2}\left(\frac{\sigma_0}{\sqrt{n}}\right) \iff n = \left(\frac{z_{\alpha/2}\sigma_0}{B}\right)^2.$$

Example 3.17. In a biomedical experiment, we would like to estimate the mean remaining life of healthy rats that are given a high dose of a toxic substance. This may be done in an early phase clinical trial by researchers trying to find a maximum tolerable dose for

humans. Suppose that we would like to write a 99 percent confidence interval for μ with a margin of error equal to $B = 2$ days. From past studies, remaining rat lifetimes are well-approximated by a normal distribution with standard deviation $\sigma_0 = 8$ days. How many rats should we use for the experiment?

SOLUTION. Here, $z_{0.01/2} = z_{0.005} = 2.58$, $B = 2$, and $\sigma_0 = 8$. Thus,

$$n = \left(\frac{z_{\alpha/2}\sigma_0}{B}\right)^2 = \left(\frac{2.58 \times 8}{2}\right)^2 \approx 106.5.$$

Thus, we would need $n = 107$ rats to achieve these goals. □

3.8.2 One population proportion

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $0 < p < 1$, and that interest lies in writing a confidence interval for p with a prescribed length. In this situation, we know that

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for p.

SAMPLE SIZE: To determine the sample size for estimating p with a $100(1-\alpha)$ percent confidence interval, we need to specify the **margin of error** that we desire; i.e.,

$$B = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

We would like to solve this equation for n. However, note that B depends on $\widehat{p}$, which, in turn, depends on n! This is a small problem, but we can overcome it by replacing $\widehat{p}$ with p^*, a **guess** for the value of p. Doing this, the last expression becomes

$$B = z_{\alpha/2}\sqrt{\frac{p^*(1-p^*)}{n}},$$

and solving this equation for n, we get

$$n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p^*(1-p^*).$$

This is the desired sample size to find a $100(1-\alpha)$ percent confidence interval for p with a prescribed margin of error equal to B.

Example 3.18. In a Phase II clinical trial, it is posited that the proportion of patients responding to a certain drug is $p^* = 0.4$. To engage in a larger Phase III trial, the researchers would like to know how many patients they should recruit into the study. Their resulting 95 percent confidence interval for p, the true population proportion of patients responding to the drug, should have a margin of error no greater than $B = 0.03$. What sample size do they need for the Phase III trial?

SOLUTION. Here, we have $B = 0.03$, $p^* = 0.4$, and $z_{0.05/2} = z_{0.025} = 1.96$. The desired sample size is

$$n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p^*(1-p^*) = \left(\frac{1.96}{0.03}\right)^2 (0.4)(1-0.4) \approx 1024.43.$$

Thus, their Phase III trial should recruit around 1025 patients. □

CONSERVATIVE APPROACH: If there is no sensible guess for p available, use $p^* = 0.5$. In this situation, the resulting value for n will be as large as possible. Put another way, using $p^* = 0.5$ gives the most **conservative** solution (i.e., the largest sample size, n). This is true because

$$n = n(p^*) = \left(\frac{z_{\alpha/2}}{B}\right)^2 p^*(1-p^*),$$

when viewed as a function of p^*, is maximized when $p^* = 0.5$.

3.9 Small-sample confidence intervals for normal means

RECALL: We have already discussed how one can use large-sample arguments to justify the use of **large-sample confidence intervals** like

$$\overline{Y} \pm z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

for estimating a single population mean, μ, and

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

for estimating the difference of two population means, $\mu_1 - \mu_2$.

CURIOSITY: What happens if the sample size n (or the sample sizes n_1 and n_2 in the two-sample case) is/are not large? How appropriate are these intervals? Unfortunately, neither of these confidence intervals is preferred when dealing with small sample sizes. Thus, we need to treat small-sample problems differently. In doing so, we will assume (at least initially) that we are dealing with normally distributed data.

3.9.1 One population mean

SETTING: Suppose that $Y_1, Y_2, ...Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population. If $\sigma^2 = \sigma_0^2$ is **known**, we have already seen that

$$\overline{Y} \pm z_{\alpha/2}\left(\frac{\sigma_0}{\sqrt{n}}\right)$$

is an exact $100(1-\alpha)$ percent confidence interval for μ.

PROBLEM: In most real problems, rarely will anyone tell us the value of σ^2. That is, it is almost always the case that the population variance σ^2 is an unknown parameter. One might think to try using S as a point estimator for σ and substituting it into the confidence interval formula above. This is certainly not illogical, but, the sample standard deviation S is not an unbiased estimator for σ! Thus, if the sample size is small, there is no guarantee that sample standard deviation S will be "close" to the population standard deviation σ (it likely will be "close" if the sample size n is large). Furthermore, when the sample size n is small, the bias and variability associated with S (as an estimator of σ) could be large. To obviate this difficulty, we recall the following result.

RECALL: Suppose that $Y_1, Y_2, ...Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population. From past results, we know that

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

Note that since the sampling distribution of T is free of all unknown parameters, T a **pivotal quantity**. So, just like before, we can use this fact to derive an exact $100(1-\alpha)$ percent confidence interval for μ.

DERIVATION: Let $t_{n-1,\alpha/2}$ denote the upper $\alpha/2$ quantile of the $t(n-1)$ distribution. Then, because $T \sim t(n-1)$, we can write

$$\begin{aligned}
1-\alpha &= P(-t_{n-1,\alpha/2} < T < t_{n-1,\alpha/2}) \\
&= P\left(-t_{n-1,\alpha/2} < \frac{\overline{Y}-\mu}{S/\sqrt{n}} < t_{n-1,\alpha/2}\right) \\
&= P\left(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \overline{Y}-\mu < t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\
&= P\left(t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} > \mu-\overline{Y} > -t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\
&= P\Bigg(\underbrace{\overline{Y}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}}_{\widehat{\theta}_L} < \mu < \underbrace{\overline{Y}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}}_{\widehat{\theta}_U}\Bigg).
\end{aligned}$$

This argument shows that

$$\overline{Y} \pm t_{n-1,\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

is an exact $100(1-\alpha)$ percent confidence interval for the population mean μ. This interval is "exact" only if the underlying probability distribution is normal. □

Example 3.19. In an agricultural experiment, a random sample of $n = 10$ plots produces the yields below (measured in kg per plot). From past studies, it has been observed that plot yields vary according to a normal distribution. The goal is to write a 95 percent confidence interval for μ, the population mean yield. Here are the sample yields:

```
23.2   20.1   18.8   19.3   24.6   27.1   33.7   24.7   32.4   17.3
```

From these data, we compute $\overline{y} = 24.1$ and $s = 5.6$. Also, with $n = 10$, the degrees of freedom is $n - 1 = 9$, and $t_{n-1,\alpha/2} = t_{9,0.025} = 2.262$ (WMS Table 5). The 95 percent confidence interval is

$$24.1 \pm 2.262\left(\frac{5.6}{\sqrt{10}}\right) \implies (20.1, 28.1).$$

Thus, based on these data, we are 95 percent confident that the population mean yield μ is between 20.1 and 28.1 kg/plot. □

3.9.2 Difference of two population means

TWO-SAMPLE SETTING: Suppose that we have two **independent** samples:

$$\begin{aligned} \text{Sample 1}: &\quad Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \text{iid } \mathcal{N}(\mu_1, \sigma_1^2) \\ \text{Sample 2}: &\quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \text{iid } \mathcal{N}(\mu_2, \sigma_2^2) \end{aligned}$$

and that we would like to construct a $100(1-\alpha)$ percent confidence interval for the difference of population means $\mu_1 - \mu_2$. As before, we define the statistics

$$\begin{aligned} \overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} &= \text{sample mean for sample 1} \\ \overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j} &= \text{sample mean for sample 2} \\ S_1^2 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{1j} - \overline{Y}_{1+})^2 &= \text{sample variance for sample 1} \\ S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_{2j} - \overline{Y}_{2+})^2 &= \text{sample variance for sample 2.} \end{aligned}$$

We know that

$$\overline{Y}_{1+} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \overline{Y}_{2+} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Furthermore, since $\overline{Y}_{1+}$ and $\overline{Y}_{2+}$ are both normally distributed, the difference $\overline{Y}_{1+} - \overline{Y}_{2+}$ is too since it is just a linear combination of $\overline{Y}_{1+}$ and $\overline{Y}_{2+}$. By straightforward calculation, it follows that

$$\overline{Y}_{1+} - \overline{Y}_{2+} \sim \mathcal{N}\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Standardizing, we get

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

Also recall that $(n_1-1)S_1^2/\sigma_1^2 \sim \chi^2(n_1-1)$ and that $(n_2-1)S_2^2/\sigma_2^2 \sim \chi^2(n_2-1)$. It follows that

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_1+n_2-2).$$

REMARK: The population variances are **nuisance parameters** in the sense that they are not the parameters of interest here. Still, they have to be estimated. We want to write a confidence interval for $\mu_1 - \mu_2$, but exactly how this interval is constructed depends on the true values of σ_1^2 and σ_2^2. In particular, we consider two cases:

- $\sigma_1^2 = \sigma_2^2 = \sigma^2$; that is, the two population variances are **equal**
- $\sigma_1^2 \neq \sigma_2^2$; that is, the two population variances are **not equal**.

EQUAL-VARIANCE ASSUMPTION: When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we have

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1)$$

and

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{\sigma_2^2} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_1+n_2-2).$$

Thus,

$$\frac{\frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2}/(n_1+n_2-2)}} = \frac{\text{“}\mathcal{N}(0,1)\text{”}}{\text{“}\chi^2(n_1+n_2-2)\text{”}/(n_1+n_2-2)} \sim t(n_1+n_2-2).$$

The last distribution results because the numerator and denominator are independent (why?). But, algebraically, the last expression reduces to

$$T = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2),$$

where

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

is the **pooled sample variance estimator** of the common population variance σ^2.

PIVOTAL QUANTITY: Since T has a sampling distribution that is free of all unknown parameters, it is a pivotal quantity. We can use this fact to construct a $100(1-\alpha)$ percent

confidence interval for mean difference $\mu_1 - \mu_2$. In particular, because $T \sim t(n_1 + n_2 - 2)$, we can find the value $t_{n_1+n_2-2,\alpha/2}$ that satisfies

$$P(-t_{n_1+n_2-2,\alpha/2} < T < t_{n_1+n_2-2,\alpha/2}) = 1 - \alpha.$$

Substituting T into the last expression and performing the usual algebraic manipulations (verify!), we can conclude that

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is an exact $100(1-\alpha)$ percent confidence interval for the mean difference $\mu_1 - \mu_2$. $\square$

Example 3.20. In the vicinity of a nuclear power plant, marine biologists at the EPA would like to determine whether there is a difference between the mean weight in two species of a certain fish. To do this, they will construct a 90 percent confidence interval for the mean difference $\mu_1 - \mu_2$. Two independent random samples were taken, and here are the recorded weights (in ounces):

- Species 1: `29.9, 11.4, 25.3, 16.5, 21.1`
- Species 2: `26.6, 23.7, 28.5, 14.2, 17.9, 24.3`

Out of necessity, the scientists assume that each sample arises from a normal distribution with $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (i.e., they assume a common population variance). Here, we have $n_1 = 5$, $n_2 = 6$, $n_1 + n_2 - 2 = 9$, and $t_{9,0.05} = 1.833$. Straightforward computations show that $\overline{y}_{1+} = 20.84$, $s_1^2 = 52.50$, $\overline{y}_{2+} = 22.53$, $s_2^2 = 29.51$, and that

$$s_p^2 = \frac{4(52.50) + 5(29.51)}{9} = 39.73.$$

Thus, the 90 percent confidence interval for $\mu_1 - \mu_2$, based on these data, is given by

$$(20.84 - 22.53) \pm 1.833\sqrt{39.73}\sqrt{\frac{1}{5} + \frac{1}{6}} \implies (-8.69, 5.31).$$

We are 90 percent confident that the mean difference $\mu_1 - \mu_2$ is between -8.69 and 5.31 ounces. Since this interval includes 0, this analysis does not suggest that the mean species weights, μ_1 and μ_2, are truly different. $\square$

UNEQUAL-VARIANCE ASSUMPTION: When $\sigma_1^2 \neq \sigma_2^2$, the problem of constructing a $100(1-\alpha)$ percent confidence interval for $\mu_1 - \mu_2$ becomes markedly more difficult. The reason why this is true stems from the fact that there is no "obvious" pivotal quantity to construct (go back to the equal-variance case and see how this assumption simplified the derivation). However, in this situation, we can still write an **approximate** confidence interval for $\mu_1 - \mu_2$; this interval is given by

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where the degree of freedom parameter ν is approximated by

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}.$$

This formula for ν is called **Satterwaite's formula**. The derivation of this interval is left to another day.

3.9.3 Robustness of the t procedures

REMARK: In the derivation of the one and two-sample confidence intervals for normal means (based on the t distribution), we have explicitly assumed that the underlying population distribution(s) was/were normal. Under the normality assumption,

$$\overline{Y} \pm t_{n-1,\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

is an **exact** $100(1-\alpha)$ percent confidence interval for the population mean μ. Under the normal, independent sample, and constant variance assumptions,

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is an **exact** $100(1-\alpha)$ percent confidence interval for the mean difference $\mu_1 - \mu_2$. Of course, the natural question arises:

*"What if the data are **not** normally distributed?"*

ROBUSTNESS: A statistical inference procedure (like constructing a confidence interval) is said to be **robust** if the quality of the procedure is not affected by a departure from the assumptions made.

IMPORTANT: The t confidence interval procedures are based on the population distribution being normal. However, these procedures are fairly robust to departures from normality; i.e., even if the population distribution(s) is/are nonnormal, we can still use the t procedures and get approximate results. The following guidelines are common:

- $n < 15$: Use t procedures only if the population distribution appears normal and there are no outliers.
- $15 \leq n \leq 40$: Be careful about using t procedures if there is strong skewness and/or outliers present.
- $n > 40$: t procedures should be fine regardless of the population distribution shape.

REMARK: These are just guidelines and should not be taken as "truth." Of course, if we know the distribution of $Y_1, Y_2, ..., Y_n$ (e.g., Poisson, exponential, etc.), then we might be able to derive an exact $100(1-\alpha)$ percent confidence interval for the mean directly by finding a suitable pivotal quantity. In such cases, it may be better to avoid the t procedures altogether.

3.10 Confidence intervals for variances

MOTIVATION: In many experimental settings, the researcher is concerned not with the mean of the underlying population, but with the population variance σ^2 instead. For example, in a laboratory setting, chemists might wish to estimate the variability associated with a measurement system (e.g., scale, caliper, etc.) or to estimate the unit-to-unit variation of vitamin tablets. In large-scale field trials, agronomists are often likely to compare variability levels for different cultivars or genetically-altered varieties. In clinical trials, the FDA is often concerned whether or not there is significant variation among various clinic sites. We examine the one and two-sample problems here.

3.10.1 One population variance

RECALL: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. In this case, we know

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

Because Q has a distribution that is free of all unknown parameters, Q is a pivot. We will use this pivot to derive an exact $100(1-\alpha)$ percent confidence interval for σ^2.

DERIVATION: Let $\chi^2_{n-1,\alpha/2}$ denote the upper $\alpha/2$ quantile and let $\chi^2_{n-1,1-\alpha/2}$ denote the lower $\alpha/2$ quantile of the $\chi^2(n-1)$ distribution; i.e., $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ satisfy

$$P[\chi^2(n-1) > \chi^2_{n-1,\alpha/2}] = \alpha/2 \quad \text{and} \quad P[\chi^2(n-1) < \chi^2_{n-1,1-\alpha/2}] = \alpha/2,$$

respectively. Then, because $Q \sim \chi^2(n-1)$,

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

This argument shows that

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$

is an exact $100(1-\alpha)$ percent confidence interval for the population variance σ^2. □

NOTE: Taking the square root of both endpoints in the $100(1-\alpha)$ percent confidence interval for σ^2 gives a $100(1-\alpha)$ percent confidence interval for σ.

Example 3.21. Entomologists studying the bee species *Euglossa mandibularis* Friese measure the wing-stroke frequency for $n = 4$ bees for a fixed time. The data are

235 225 190 188

Assuming that these data are an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, find a 90 percent confidence interval for σ^2.

SOLUTION. Here, $n = 4$ and $\alpha = 0.10$, so we need $\chi^2_{3,0.95} = 0.351846$ and $\chi^2_{3,0.05} = 7.81473$ (Table 6, WMS). I used R to compute $s^2 = 577.6667$. The 90 percent confidence interval is thus

$$\left[\frac{3(577.6667)}{7.81473}, \frac{3(577.6667)}{0.351846}\right] \Longrightarrow (221.76, 4925.45).$$

That is, we are 90 percent confident that the true population variance σ^2 is between 221.76 and 4925.45; i.e., that the true population standard deviation σ is between 14.89 and 70.18. Of course, both of these intervals are quite wide, but remember that $n = 4$, so we shouldn't expect notably precise intervals. □

3.10.2 Ratio of two variances

TWO-SAMPLE SETTING: Suppose that we have two **independent** samples:

$$\begin{aligned} \text{Sample 1}: &\quad Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \text{iid } \mathcal{N}(\mu_1, \sigma_1^2) \\ \text{Sample 2}: &\quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \text{iid } \mathcal{N}(\mu_2, \sigma_2^2) \end{aligned}$$

and that we would like to construct a $100(1-\alpha)$ percent confidence interval for $\theta = \sigma_2^2/\sigma_1^2$, the **ratio** of the population variances. Under these model assumptions, we know that $(n_1-1)S_1^2/\sigma_1^2 \sim \chi^2(n_1-1)$, that $(n_2-1)S_2^2/\sigma_2^2 \sim \chi^2(n_2-1)$, and that these two quantities are independent. It follows that

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} \sim \frac{\text{“}\chi^2(n_1-1)\text{”}/(n_1-1)}{\text{“}\chi^2(n_2-1)\text{”}/(n_2-1)} \sim F(n_1-1, n_2-1).$$

Because F has a distribution that is free of all unknown parameters, F is a pivot, and we can use it to derive $100(1-\alpha)$ percent confidence interval for $\theta = \sigma_2^2/\sigma_1^2$. Let $F_{n_1-1,n_2-1,\alpha/2}$ denote the upper $\alpha/2$ quantile and let $F_{n_1-1,n_2-1,1-\alpha/2}$ denote the lower $\alpha/2$ quantile of the $F(n_1-1, n_2-1)$ distribution. Because $F \sim F(n_1-1, n_2-1)$, we can write

$$\begin{aligned} 1-\alpha &= P\left(F_{n_1-1,n_2-1,1-\alpha/2} < F < F_{n_1-1,n_2-1,\alpha/2}\right) \\ &= P\left(F_{n_1-1,n_2-1,1-\alpha/2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{n_1-1,n_2-1,\alpha/2}\right) \\ &= P\left(\frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,1-\alpha/2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,\alpha/2}\right). \end{aligned}$$

This argument shows that

$$\left(\frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,1-\alpha/2}, \frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,\alpha/2} \right)$$

is an exact $100(1-\alpha)$ percent confidence interval for the ratio $\theta = \sigma_2^2/\sigma_1^2$. □

Example 3.22. Snout beetles cause millions of dollars worth of damage each year to cotton crops. Two different chemical treatments are used to control this beetle population using 13 randomly selected plots. Below are the percentages of cotton plants with beetle damage (after treatment) for the plots:

- Treatment 1: `22.3, 19.5, 18.6, 24.3, 19.9, 20.4`
- Treatment 2: `9.8, 12.3, 16.2, 14.1, 15.3, 10.8, 18.3`

Under normality, and assuming that these two samples are independent, find a 95 percent confidence interval for $\theta = \sigma_2^2/\sigma_1^2$, the ratio of the two treatment variances.

SOLUTION. Here, $n_1 = 6$, $n_2 = 7$, and $\alpha = 0.05$, so that $F_{5,6,0.025} = 5.99$ (WMS, Table 7). To find $F_{5,6,0.975}$, we can use the fact that

$$F_{5,6,0.975} = \frac{1}{F_{6,5,0.025}} = \frac{1}{6.98} \approx 0.14$$

(WMS, Table 7). Again, I used R to compute $s_1^2 = 4.40$ and $s_2^2 = 9.27$. Thus, a 95 percent confidence interval for $\theta = \sigma_2^2/\sigma_1^2$ is given by

$$\left(\frac{9.27}{4.40} \times 0.14, \frac{9.27}{4.40} \times 5.99 \right) \implies (0.29, 12.62).$$

We are 95 percent confident that the ratio of variances $\theta = \sigma_2^2/\sigma_1^2$ is between 0.29 and 12.62. Since this interval includes 1, we can not conclude that the two treatment variances are significantly different. □

NOTE: Unlike the t confidence intervals for means, the confidence interval procedures for one and two population variances are **not robust** to departures from normality. Thus, one who uses these confidence intervals is placing strong faith in the underlying normality assumption.