MA22004 - Statistics and Probability II

Dr Eric Hall

Last updated: 2020-09-16

Contents

Course Documents

Welcome

Welcome to MA22004 at the University of Dundee.

This course covers the basics of statistical inference. The first part of this document contains information of a practical nature regarding the mechanics of the course. The second part of this document contains the content.

These are some trying times.

Please try to stay healthy.

These notes are available at dundeemath.github.io/MA22004/ and also as a PDF (visit the page and click on the PDF icon to download).

Course Guide

Organisation

This module runs for 11 teaching weeks and is worth 20 SCQF credits (equivalently, 10 ECTS points). All organisation and teaching will be carried out by:

Dr Eric Hall
ehall001@dundee.ac.uk
Mathematics Division
Room TBA, Fulton Building
01382 TBA

This course uses Blackboard Ultra (look for course `MA22004_SEM0000_2021`) for communicating all announcements/deadlines and also for running online meetings. This course also uses Gradescope for submission of some of the continuous assessment items and Perusall for collaborative engagement with reading materials.

If you have a problem regarding the course, then you should make an appointment to see Dr Hall. You may also bring matters of concern about the course to the attention of the Mathematics Division Staff/Student Committee, which meets once each semester. A volunteer from Level 2 Mathematics will act as class representative to sit on the Staff/Student Committee (see Ultra for contact details).

Timetable

Due to COVID19, these plans may be subject to change.

The delivery of this module consists of a blend of synchronous and asynchronous content delivered both in-person and online. On an average week, there will be seven planned teaching and learning activities.

The anticipated student effort is 200 hours over the length of the module. You are expected to be "present" for all synchronous timetabled activities except for the online office hours, which are optional. You may engage with the asynchronous material at your own pace, keeping in mind to meet any deadlines for engagement and/or attainment that will be posted to Ultra and discussed.

Activity	Timetabled	Group	Hours	Delivery
Reading	asynchronous	individually & in groups	6	online
Investigation	asynchronous	individually	1	online
Seminar	synchronous	whole class	1	online
Computer Lab	asynchronous	individually	6	online
Workshop Preparation	asynchronous	individually or in groups	2	online
Workshop	synchronous	in groups	1	face-to-face
Office Hours	synchronous	in groups	1	online

Pre-requisites

To take this course, you must have passed module MA12003 or equivalent.

Syllabus

Sampling Distributions Mean and standard deviation of samples, sampling from a single population, sampling from two populations, shape of sampling distributions. Normal distribution, χ^2-square distribution, F-distribution.

Hypothesis tests Null and Alternate hypotheses, inferences, confidence intervals, estimating means, proportions and standard deviations.

Linear Regression Least squares, assessing usefulness of a model, using a model.

Industrial Quality Control Control Charts, acceptance sampling.

R software package Appropriate use of computational software to carry out statistical and probabilistic calculations.

Recommended Books

In addition to the course notes, here are some textbooks you may wish to consult.

- *Probability and Statistics for Engineering and the Sciences,* [Devore, 2016, §6-10, 12, 14, 16]
- *Probability and Statistics,* [DeGroot and Schervish, 2001, §7-10]
- *Mathematical statistics and data analysis,* [Rice, 2006, §6-12]
- *All of Statistics,* [Wasserman, 2004, Concise general reference]

You do not need to purchase these books.

Assessment

The module will be *continuously* assessed using coursework and examinations. Deadlines, as well as test dates, will be posted on Ultra and announcements made in the class hours. The module assessment weighting is as follows.

Coursework

Assessed coursework includes:

- six hand-in laboratory reports and

Assessment	Weight
Assignments	20%
Midterm Exam 1	20%
Midterm Exam 2	20%
Final Exam	40%

- weekly engagement with the reading material using Perusall.

There will also be alternative means of demonstrating your mastery of course material through:

- one (group) lab presentation and
- short seminar quizzes (announced in advance).

Examinations

The **Midterm Exams** will be computer-assessed and will be one (1) hour in scope. These will likely be in weeks 4 and 8.

The **Final Exam** will be a two (2) hour hand-written exam that will be submitted using Gradescope. This process will be thoroughly discussed and trialled with a dummy exam in advance of the real submission. The Final Exam will be in week 11 (i.e., during the last week of the term).

To pass this module, you must:

- obtain an overall grade of at least `D3` in the overall assessment **and**
- obtain a grade of at least `M1` for the exam **and**
- obtain a grade of at least `M1` for the coursework.

For those who fail the module, there may be an opportunity to take a two-hour resit examination paper at the next available exam diet.

Resit marks are based on the resit exam only.

Unless you have mitigating circumstances, if you fail to achieve a module grade of `CF` or above at first attempt, then you may not be permitted to resit the exam. Also, unless you have mitigating circumstances, any pass after a resit will be capped at a grade of `D3` regardless of the weighted average mark obtained.

Your Commitment

You should attend all synchronous timetabled sessions except on medical grounds or with the special permission of Dr Hall. If you are unable to attend the degree examination or complete elements of the coursework on time, then you should inform Dr Hall and submit a medical certificate. Medical certificates should be submitted to your School Office as soon as possible after the absence.

You must also submit a Mitigating Circumstances form to explain which aspects of assessment have been affected by your absence.

A Medical Certificate will only be taken into account if accompanied by a completed Mitigating Circumstances form that refers to the medical certificate.

Approved Calculators

The Casio FX83 and the Casio FX85 are the only calculators approved for use in assessments in the School of Engineering, Physics and Mathematics.

Study Support

If you are having difficulty with the course, you are encouraged to seek help at an early stage by making an appointment with Dr Hall. You may also obtain additional help from the Maths Base (see Ultra for details).

Disability

The University of Dundee is committed to making reasonable, effective and appropriate accommodations to meet the needs of students with disabilities and to create an inclusive and barrier-free campus. If you require accommodation for a documented disability, then you are advised to register with Disability Services. Please communicate any needs you may have directly with Dr Hall and as soon as possible to ensure timely management of any accommodations.

Academic Honesty

Honesty in scholarship and research is integral to the integrity of the academic enterprise of any higher education institution. Therefore, all students at the University of Dundee must practice academic honesty. Academic dishonesty includes cheating, fabrication, plagiarism, and facilitating dishonesty. Cases of academic dishonesty will be subject to appropriate sanctions and ignorance of such standards is not sufficient evidence of lack of intent. Please see the *Code of Practice on Academic Misconduct by Students* for more information about what constitutes academic dishonesty.

End of Module Questionaire

You will have the opportunity to complete a confidential questionnaire regarding the content and presentation of the module periodically. These questionaires form an important element in the University's Academic Standards procedures. Thank you in advance for your cooperation.

Course Notes

Preliminaries

Notation

Uppercase roman letters, e.g., X, will typically denote random variables (rvs); lower case letters, e.g., x, will denote a particular value (observation) of a rv. Rvs have probability distributions. Distributions are typically characterized by *parameters* which are fixed real numbers. Parameters describe population characteristics that are often unknown and must be estimated from data. Statistical inference is a tool that will help us to do this.

Statistical models comprise both rvs and parameters. Be careful not to confuse them!

Abbreviations

Abbreviation	Expanded
pdf	probability density function
cdf	cumulative distribution function
rv	random variable
iid	independent and identically distributed
obs	observations
CI	confidence interval
df	degrees of freedom

Topic 1

Sampling distributions

A **statistic** is a quantity that can be calculated from sample data. Prior to obtaining data, a statistic is an unknown quantity and is therefore a rv. We refer to the probability distribution for a statistic as a **sampling distribution** to emphasize how the distribution will vary across all possible sample data.

1.1 Normal distribution

Normal distributions play an important role in probability and statistics as they describes many natural phenomenon. For instance, the Central Limit Theorem tells us that sums of rvs are approximately normal in distribution.

Definition 1.1. A continuous rv X has a **normal distribution** with parameters μ and σ^2, where $-\infty < \mu < \infty$ and $\sigma > 0$, if X has pdf

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty .$$

We write $X \sim \mathsf{N}(\mu, \sigma^2)$.

For $X \sim \mathsf{N}(\mu, \sigma^2)$, it can be shown that $\mathbf{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, that is, μ is the *mean* and σ^2 is the *variance* of X. The pdf takes the form of a bell-shaped curve that is symmetric about μ. The value σ (*standard deviation*) is the distance from μ to the inflection points of the curve. Thus, the position (location) and spread of the distribution depends on μ and σ.

Definition 1.2. We say that X has a **standard normal distribution** if $\mu = 0$ and $\sigma = 1$ and we will usually denote standard Normal rvs by Z (why? tradition!).

1.1.1 Some useful facts about Normals

Here are some useful facts about how to manipulate Normal rvs.

1. If $X \sim \mathsf{N}(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \quad \sim \mathsf{N}(0, 1)$.
2. If $Z \sim \mathsf{N}(0, 1)$, then $X = \mu + \sigma Z \quad \sim \mathsf{N}(\mu, \sigma^2)$.
3. If $X_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$ for $i = 1, \ldots, n$ are independent rvs, then

$$\sum_{i=1}^{n} X_i \sim \mathsf{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) .$$

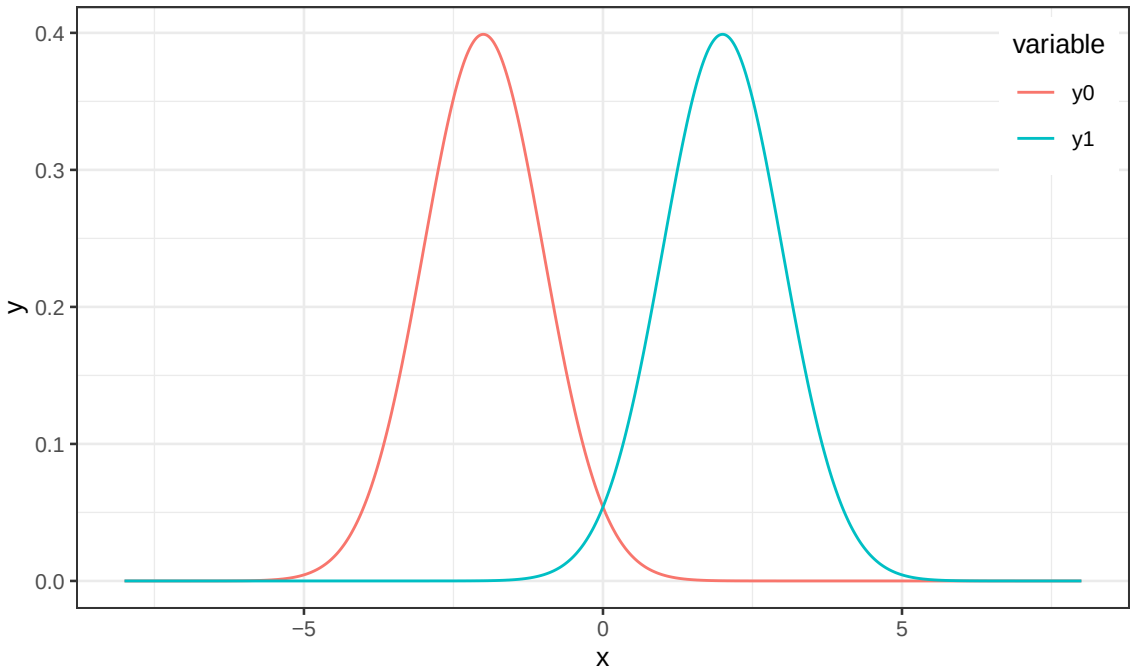

Figure 1.1: The pdfs of two normal rvs with different means and the same standard deviations.

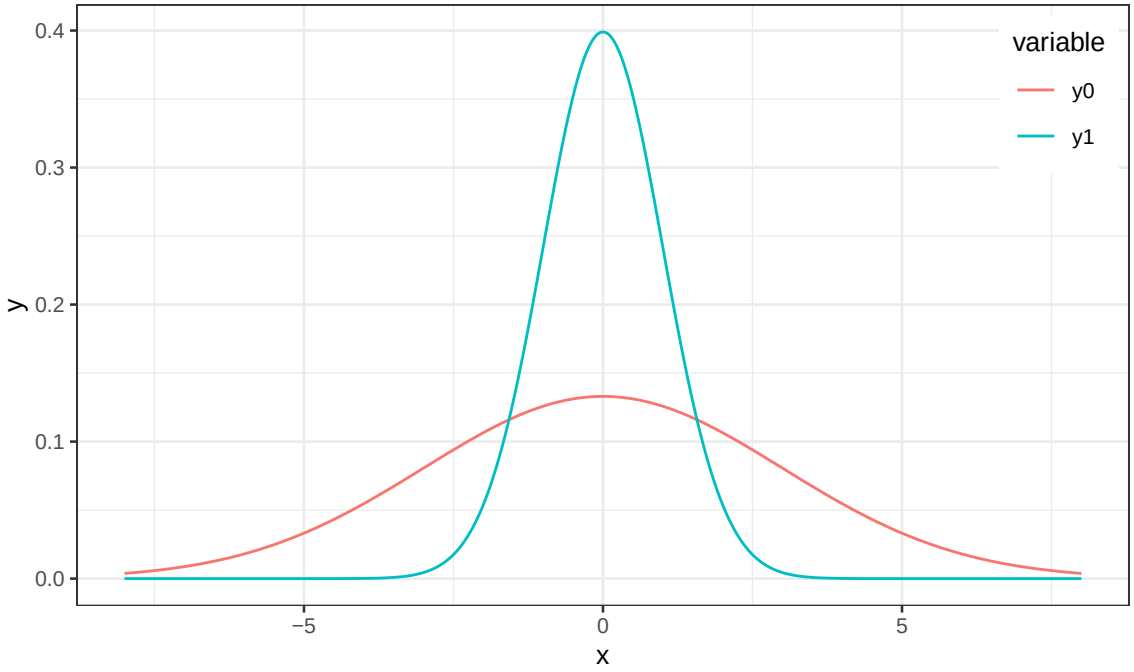

Figure 1.2: The pdfs of two normal rvs with the same means and different standard deviations.

In particular, we note that for differences of independent rvs $X_1 \sim \mathsf{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathsf{N}(\mu_2, \sigma_2^2)$ then the variances also add:

$$X_1 - X_2 \sim \mathsf{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)\,.$$

Probabilities $P(a \le X \le b)$ are found by converting the problem in $X \sim \mathsf{N}(\mu, \sigma^2)$ to the *standard normal* distribution $Z \sim \mathsf{N}(0, 1)$ whose probability values $\Phi(z) = P(Z \le z)$ can then be looked up in a table. From (1.) above,

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\,. \end{aligned}$$

This process is often referred to as *standardizing* (the normal rv).

Example 1.1. Let $X \sim \mathsf{N}(5, 9)$ and find $P(X \ge 5.5)$.

$$\begin{aligned} P(X \ge 5.5) &= P\left(Z \ge \frac{5.5-5}{3}\right) \\ &= P(Z \ge 0.1667) \\ &= 1 - P(Z \le 0.1667) \\ &= 1 - \Phi(0.1667) \\ &= 1 - 0.5662 \\ &= 0.4338\,, \end{aligned}$$

where we look up the value of $\Phi(z) = P(Z \le z)$ in a table of standard normal curve areas.

Alternatively, we can use the `r` code:

```r
pnorm(5.5, mean = 5, sd = 3, lower.tail = FALSE)
```

```
[1] 0.4338162
```

TODO: plot of area under normal curve (right tail) ♢

Example 1.2. Let $X \sim \mathsf{N}(5, 9)$ and find $P(4 \le X \le 5.25)$.

$$\begin{aligned} P(4 \le X \le 5.25) &= P\left(\frac{4-5}{3} \le Z \le \frac{5.25-5}{3}\right) \\ &= P(-0.3333 \le Z \le 0.0833) \\ &= \Phi(0.0833) - \Phi(-0.3333) \\ &= 0.5332 - 0.3694 \\ &= 0.1638\,. \end{aligned}$$

where we look up the value of $\Phi(z) = P(Z \le z)$ in a table of standard normal curve areas.

Alternatively, we can use the `r` code:

```r
pnorm(5.25, mean = 5, sd = 3) - pnorm(4, mean = 5, sd = 3)
```

```
[1] 0.1637654
```

TODO: plot area under normal curve (interior) ♢

1.1.2 Empirical rule ($68-95-99.7$ rule)

For samples from a normal distribution, the percentage of values that lie within one, two, and three standard deviations of the mean are 68.27%, 95.45%, and 99.73%, respectively. That is, for $X \sim \mathsf{N}(\mu, \sigma^2)$,

$$P(\mu - 1\sigma \leq X \leq \mu + 1\sigma) \approx 0.6827\,,$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.9545\,,$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973\,.$$

For a normal population, this means that nearly all the the values lie within "three-sigmas" of the mean.

1.2 t distribution

Student's t distribution gets its peculiar name as it was first published under the pseudonym "Student". This bit of obfuscation was to protect the identity of his employer, and thereby vital trade secrets, in a highly competitive and lucrative industry.

Definition 1.3. A continuous rv X has a t **distribution** with parameter $\nu > 0$, if X has pdf

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}\,, \quad -\infty < x < \infty\,.$$

We write $X \sim \mathsf{t}(\nu)$.

1.2.1 Properties of t distributions

1. The density for $\mathsf{t}(\nu)$ is a bell-shaped curve centered at 0.
2. The density for $\mathsf{t}(\nu)$ is more spread out than the standard normal density (i.e., it has "fatter tails" than the normal).
3. As $\nu \to \infty$, the spread of the corresponding $\mathsf{t}(\nu)$ density converges to the standard normal density (i.e., the spread of the $\mathsf{t}(\nu)$ density decreases relative to the standard normal).

If $X \sim \mathsf{t}(\nu)$, then $\mathbf{E}[X] = 0$ for $\nu > 1$ (otherwise the mean is undefined).

1.3 χ^2 distribution

The χ^2 distribution arises as the distribution of a sum of the squares of ν independent standard normal rvs.

Definition 1.4. A continuous rv X has a χ^2 **distribution** with parameter $\nu \in \mathbf{N}_>$, if X has pdf

$$f(x;\nu) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}x^{(\nu/2)-1}e^{-x/2}\,,$$

with support $x \in (0,\infty)$ if $\nu = 1$, otherwise $x \in [0,\infty)$. We write $X \sim \chi^2(\nu)$.

The pdf $f(x;\nu)$ of the $\chi^2(\nu)$ distribution depends on a positive integer ν referred to as the df. The density $f(x;\nu)$ is positively skewed, i.e., the right tail is longer and hence the mass is concentrated to the left of the figure. The distribution becomes more symmetric as ν increases. We denote critical values of the $\chi^2(\nu)$ distribution by $\chi^2_{\alpha,\nu}$.

Unlike the normal and t distributions, the $\square^2$ distribution is not symmetric. This means that the critical values e.g. $\chi^2_{.99,\nu}$ and $\chi^2_{0.01,\nu}$ are **not** equal. Hence, it will be necessary to look up both values for CIs based on χ^2 critical values.

If $X \sim \chi^2(\nu)$, then $\mathbf{E}[X] = \nu$ and $\mathrm{Var}[X] = 2\nu$.

1.4 F distribution

The F distribution arises as a test statistic when comparing population variances and in ANOVA.

Definition 1.5. A continuous rv X has an F **distribution** with df parameters ν_1 and ν_2, if X has pdf

$$f(x;\nu_1,\nu_2) = \frac{\Gamma\left(\frac{n+m}{2}\right) n^{n/2} m^{m/2}}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{x^{n/2-1}}{(m+nx)^{(n+m)/2}}\,.$$

Theorem 1.1. *If $X_1 \sim \chi^2(\nu_1)$ and $X_2 \sim \chi^2(\nu_2)$ are independent rvs, then the rv*

$$F = \frac{X_1/\nu_1}{X_2/\nu_2} \quad \sim \mathsf{F}(\nu_1,\nu_2)\,,$$

that comprises the ratio of two χ^2 rvs divided by their respective df has an $\mathsf{F}(\nu_1,\nu_2)$ distribution.

Topic 2

Basics of statistical inference

We discuss point estimation, confidence intervals, and hypothesis testing in Sections 2.1, 2.2, and 2.3, respectively. These three tools will form the basis for making inferences about a population.

2.1 Point estimation

Statistical inference seeks to draw conclusions about the characteristics of a population from data. For example, suppose we are botanists interested in taxonomic classification of iris flowers. Let μ denote the true average petal length (in cm) of the *Iris setosa* (AKA the bristle-pointed iris). The parameter μ is a characteristic of the whole population of the *setosa* species. Before we collect data, the petal lengths of m independent *setosa* flowers are denoted by rvs $X_1, X_2, \ldots, X_m$. Any function of the X_i's, such as the sample mean,

$$\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i\,, \tag{2.1}$$

or the sample variance,

$$S^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \overline{X})^2\,, \tag{2.2}$$

is also a rv.

Suppose we actually find and measure the petal length of 50 independent *setosa* flowers resulting in observations $x_1, x_2, \ldots, x_{50}$; the distribution (counts) of 50 such petal length measurements are displayed in Figure 2.1. The sample mean $\overline{x}$ for petal length can then be used to draw a conclusion about the (true) value of the population mean μ. Based on the data in Figure 2.1 and using (2.1), the value of the sample mean is $\overline{x} = 1.462$. The value $\overline{x}$ provides a "best guess" or point estimate for the true value of μ based on the $m = 50$ samples.

The botonist Edgar Anderson's **Iris Data** contains 50 obs. of four features (sepal length [cm], sepal width [cm], petal length [cm], and petal width [cm]) for each of three plant species (*setosa*, *virginica*, *versicolor*) for 150 obs. total. This data set can be accessed in `r` by loading `library(datasets)` and then calling `data(iris)`.

Definition 2.1. A **point estimate** of a parameter θ (recall: a fixed, unknown quantity) is a single number that we regard as a sensible value for θ. Consider iid $X_1, X_2, \ldots, X_m \sim F(\theta)$. A **point estimator** $\hat{\theta}_m$ of θ

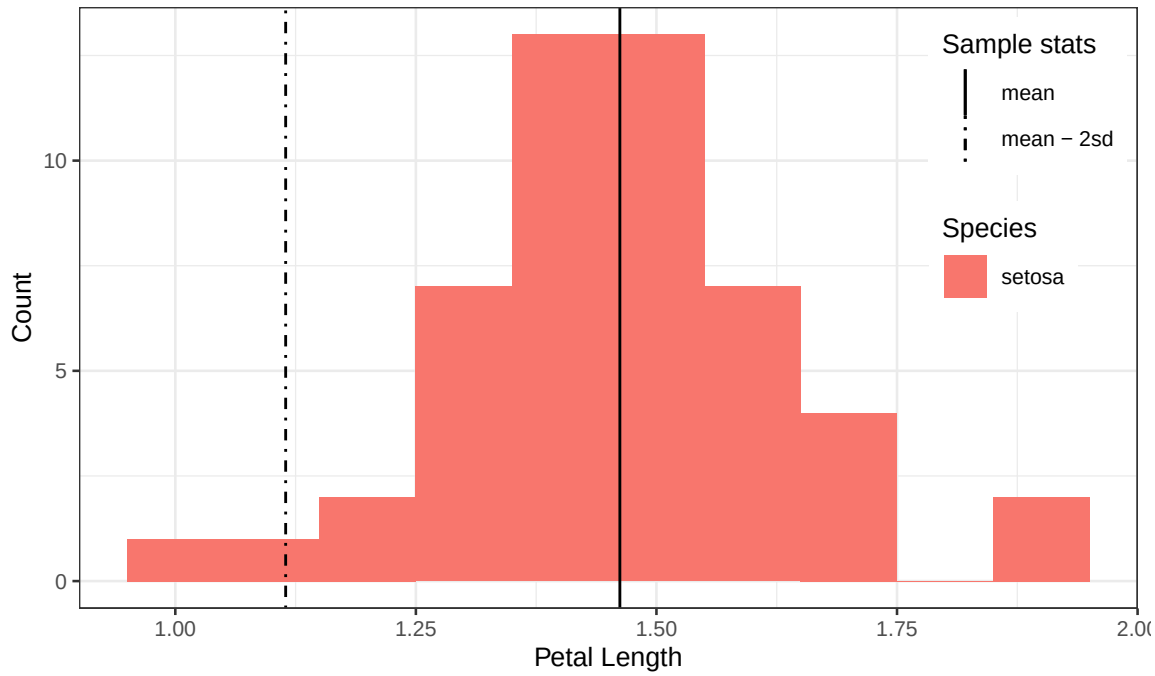

Figure 2.1: The distribution (counts) of $m = 50$ *setosa* petal length measurments.

is obtained by selecting a suitable statistic g,

$$\hat{\theta}_n = g(X_1, \ldots, X_m)\,.$$

A point estimate $\hat{\theta}_m$ can then be computed from the estimator using sample data.

The symbol $\hat{\theta}_m$ (or simply $\hat{\theta}$ when the sample size m is clear from context) is typically used to denote both the estimator and the point estimate resulting from a given sample. Note that writing, e.g., $\hat{\theta} = 42$ does not indicate how the point estimate was obtained. Therefore, it is essential to report both the estimator and the resulting point estimate.

Definition 2.1 does not say how to select an appropriate statistic. For the *setosa* example, the sample mean $\overline{X}$ is suggested as a good estimator of the population mean μ. That is, $\widehat{\mu} = \overline{X}$ or "the point estimator of μ is the sample mean $\overline{X}$". Here, while μ and σ^2 are fixed quantities representing characteristics of the population, $\overline{X}$ and S^2 are rvs with sampling distributions. If the population is *normally distributed* or if the *sample is large* then the sampling distribution for $\overline{X}$ has a known form: $\overline{X}$ is normal with mean $\mu_{\overline{X}} = \mu$ and variance $\sigma^2_{\overline{X}} = \sigma^2/m$, i.e.,

$$\overline{X} \sim \mathsf{N}(\mu, \sigma^2/m)\,,$$

where m is the sample size and μ and σ are the (typically unknown) population parameters.

Example 2.1. Let us consider the heights (measured in inches) of 31 black cherry trees (sorted, for your enjoyment) in Table 2.1.

The **Cherry Tree Data** contains 31 obs. of three features (diameter, height, and volume) and can be accessed in `r` by loading `library(datasets)` and then calling `data(trees)`.

Table 2.1: Observations of $m = 31$ felled black cherry trees.

Height [in]
63, 64, 65, 66, 69, 70, 71, 72, 72, 74, 74, 75, 75, 75, 76, 76, 77, 78, 79, 80, 80, 80, 80, 80, 81, 81, 82, 83, 85, 86, 87

The quantile-quantile plot in Figure 2.2, that compares the quantiles of this data to the quantiles of a normal distribution, is fairly straight.[1] Therefore, we assume that the distribution of black cherry tree heights is (at least approximately) normal with a mean value μ; i.e., that the population of heights is distributed $\mathsf{N}(\mu, \sigma^2)$ where μ is a parameter to be estimated and σ^2 is unknown. The observations $X_1, \ldots, X_{31}$ are then assumed to be a random sample from this normal distribution, i.e., iid

$$X_1, \ldots, X_{31} \sim \mathsf{N}(\mu, \sigma^2)\,.$$

Consider the following three different estimators and the resulting point estimates for μ based on the 31 samples in Table 2.1.

a. Estimator (sample mean) $\overline{X}$ as in (2.1) and estimate $\overline{x} = \sum x_i/n = 2356/31 = 76$.

b. Estimator (average of extreme heights) $\widetilde{X} = [\min(X_i) + \max(X_i)]/2$ and estimate $\tilde{x} = (63 + 87)/2 = 75$.

c. Estimator (10% trimmed mean – i.e., in this instance exclude the smallest and largest three values) $\overline{X}_{\text{tr(10)}}$ and estimate $\overline{x}_{\text{tr(10)}} = (2356 - 63 - 64 - 65 - 87 - 86 - 85)/25 = 76.24$.

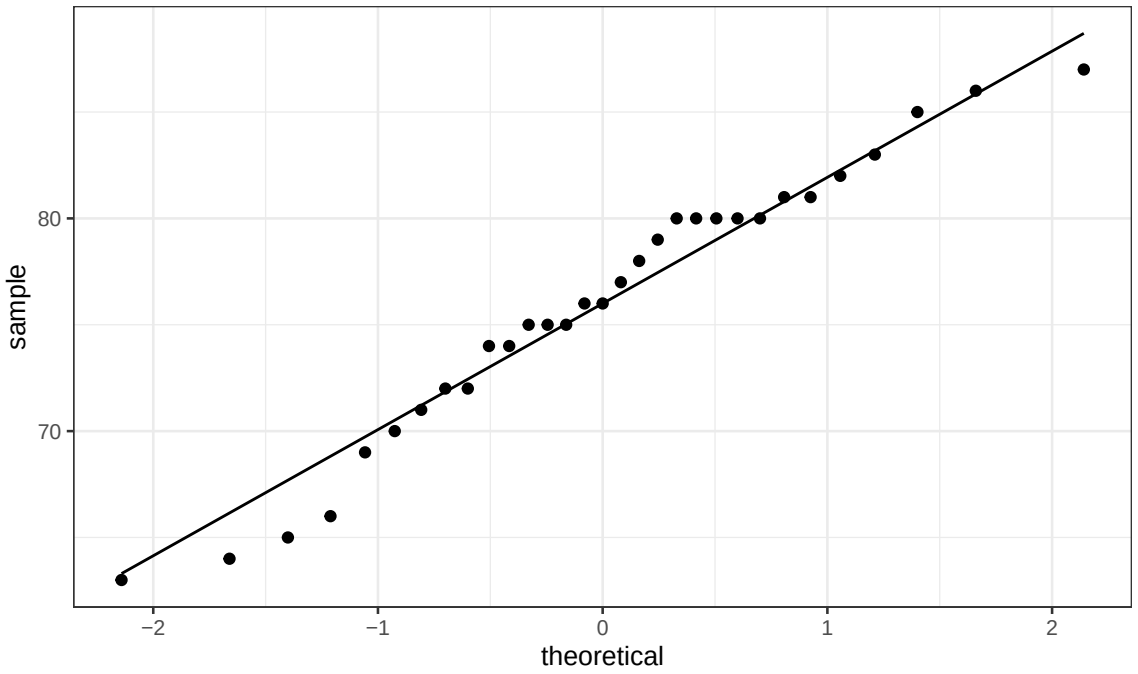

Figure 2.2: Normal quantile-quantile plot for the **Cherry Tree Data**.

[1] How do we tell whether a population is normal? Constructing a normal quantile-quantile plot is one way of assessing whether a normality assumption is reasonable; such a plot compares the quantiles of the sample data x_i against the (theoretical) standard normal quantiles. If the sample data is consistent with a sample from a normal distribution, then the points will lie on a straight line (more or less). Below we display the QQ plot comparing quantiles of cherry tree heights from 2.1 to normal quantiles.

Each estimator above uses a different notion of center for the sample data, i.e., a different statistic. An interesting question to think about is: which estimator will tend to produce estimates closest to the true parameter value? Will the estimators work universally well for all distributions? ♢

In addition to reporting a point estimate together with its estimator, some indication of its precision should be given. One measure of the precision of an estimate is its standard error.

Definition 2.2. The **standard error** of an estimator $\hat{\theta}$ is the standard deviation

$$\sigma_{\hat{\theta}} = \sqrt{\mathrm{Var}(\hat{\theta})}\,.$$

Often, the standard error depends on unknown parameters and must also be estimated. The **estimated standard error** is denoted by $\widehat{\sigma}_{\hat{\theta}}$ or simply $s_{\hat{\theta}}$.[2]

2.2 Confidence intervals

An alternative to reporting a point estimate for a parameter is to report an interval estimate suggesting an entire range of plausible values for the parameter of interest. A confidence interval is an interval estimate that makes a probability statement about the degree of reliability, or the confidence level, of the interval. The first step in computing a confidence interval is to select the confidence level α. A popular choice is a 95% confidence interval which corresponds to level $\alpha = 0.05$.

Definition 2.3. A $100(1-\alpha)\%$ **confidence interval** for a parameter θ is a *random* interval $C_m = (L_m, U_m)$, where $L_m = \ell(X_1, \ldots, X_m)$ and $U_m = u(X_1, \ldots, X_m)$ are functions of the data, such that

$$P_\theta(L_m < \theta < U_m) = 1 - \alpha\,, \tag{2.3}$$

for all $\theta \in \Theta$.

My favorite interpretation of a confidence interval is due to [Wasserman, 2004, p 92]:

> On day 1, you collect data and construct a 95 percent confidence interval for a parameter θ_1. On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_2. On day 3, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_3. You continue this way constructing confidence intervals for a sequence of unrelated parameters θ_1, θ_2, … Then 95 percent of your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over.

This interpretation makes clear that a confidence interval is not a probability statement about the parameter θ. In Definition 2.3, note that θ is fixed (θ is not a rv) and the interval C_m is random. After data has been collected and a point estimator has been calculated, the resulting CIs either contain the true parameter value or they do not, as illustrated in Figure 2.3.

2.3 Hypothesis testing

Sections 2.1 and 2.2 reviewed how to estimate a parameter by a single number (point estimate) or range of plausible values (confidence interval), respectively. Next we discuss methods for determining which of two contradictory claims, or **hypotheses**, about a parameter is correct.

Definition 2.4. The **null hypothesis**, denoted by H_0, is a claim that we intially assume to be true by dafault. The **alternative hypothesis**, denoted by H_a, is an assertion that is contradictory to H_0.

[2] The standard error is sometimes denoted $\mathsf{se} = \mathsf{se}(\hat{\theta})$ and the estimated standard error by $\widehat{\mathsf{se}}$.

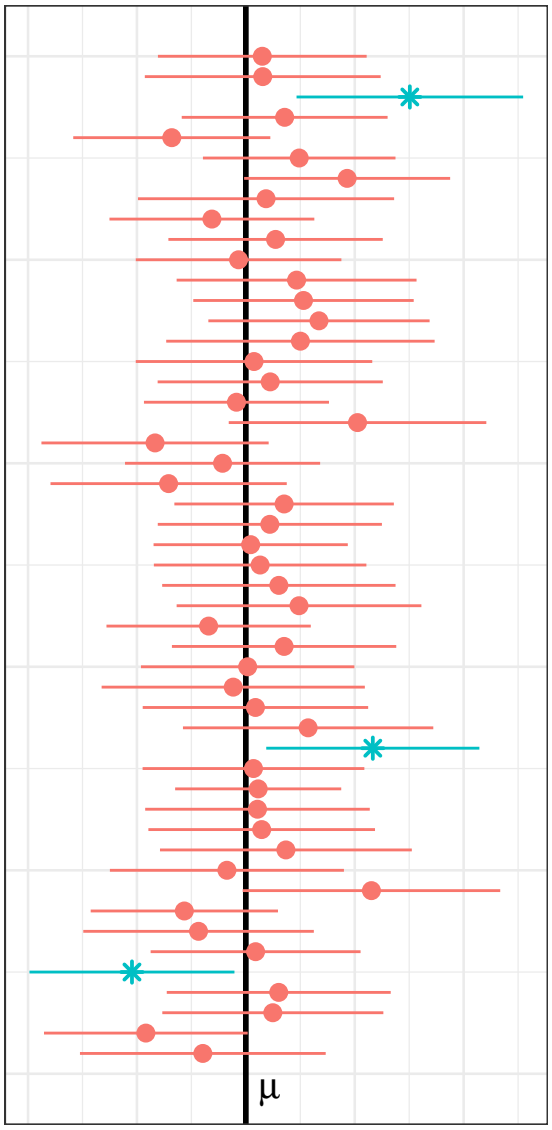

Figure 2.3: Fifty 95% CIs for a population mean μ. After a sample is taken, the computed interval estimate either contains μ or it does not (asterisk identify intervals that do not include μ). When drawing such a large number of 95% CIs, we would anticipate that approximately 5% (ca. 2 or 3) would fail to cover the true parameter μ.

Typically, we shall consider hypothesis concerning a parameter $\theta \in \Theta$ taking values in a parameter space. The statistical hypothesis are contradictory in that H_0 and H_a divide Θ into two disjoint sets. For example, for a statistical hypothesis regarding the *equality* of a parameter θ with a fixed quantity θ_0, the null and alternative hypotheses will usually take one of the following forms in Table 2.2.

Table 2.2: Typical null hypothesis and corresponding alternative hypothesis.

Null hypothesis	Alternative hypothesis	Test form
$H_0 : \theta = \theta_0$	$H_a : \theta \neq \theta_0$	two-sided test
$H_0 : \theta \leq \theta_0$	$H_a : \theta > \theta_0$	one-sided test
$H_0 : \theta \geq \theta_0$	$H_a : \theta < \theta_0$	one-sided test

These pairs of hypothesis are associated with either a one-sided or two-sided test; what this means will become clear in the sequel. The value θ_0, called the **null value**, separates the alternative from the null.

Definition 2.5. A **hypothesis test** asks if the available data provides sufficient evidence to reject H_0. If the observations disagree with H_0, then we reject the null hypothesis. If the sample evidence does not strongly contradict H_0, then we continue to believe H_0. The two possible conclustions of a hypothesis test are: *reject* H_0 or *fail to reject* H_0.[3]

A procedure for carrying out a hypothesis test is based on specifying two additional items: a test statistic and a corresponding rejection region. A **test statistic** T is a function of the sample data (like an estimator). The decision to reject or fail to reject H_0 will involve computing the test statistic. The **rejection region** R is the collection of values of the test statistic for which H_0 is to be rejected in favor of the alternative, e.g.,

$$R = \{x : T(x) > c\} ,$$

where c is referred to as a **critical value**. If a given sample falls in the rejection region, then we reject H_0. That is, if $X \in R$ (e.g., the calculated test statistic exceeds some critical value), then we reject H_0. The alternative is that $X \notin R$ and in this case we fail to reject the null.

When carrying out a hypothesis test, two types of errors can be made. The basis for choosing a rejection region involves considering these errors.

Definition 2.6. A **type I** error occurs if H_0 is rejected when H_0 is actually true. A **type II** error is made if we fail to reject H_0 when H_0 is actually false.

If the maximal type I error of a test is fixed at an acceptably small value, then the type II error decreases as the sample size increases. In particular, a conclusion is reached in a hypothesis test by selecting a **significance level** α for the test linked to the maximal type I error rate. Typically, $\alpha = 0.10, 0.05, 0.01$, or 0.001 is selected for the significance level.

Definition 2.7. A **P-value** is the probability, calculated assuming H_0 is true, of obtaining a value of the test statistic at least as contradictory to H_0 as the value calculated from the sample data.

Smaller P-values indicate stronger evidence against H_0 in favor of H_a. If $P \leq \alpha$ then we reject H_0 at significance level α. If $P \geq \alpha$ we fail to reject H_0 at significance level α.

[3] We comment that *fail to reject* H_0 is sometimes phrased *retain* H_0 or (perhaps less accurately) *accept* H_0. Why not just *accept* the null and move on with our lives? Well, if I search the Highlands for the Scottish wildcat (critically endangered) and fail to find any, does that prove they do not exist?

The P-value is a probability calculated assuming that H_0 is true. However, the P-value is **not** the probability that:

1. H_0 is TRUE,
2. H_0 is FALSE, or
3. a wrong conclusion is reached.

Proposition 2.1. *The hypothesis test procedure that*

$$\begin{cases} \textit{rejects } H_0 & \textit{if } P \leq \alpha \\ \textit{fails to reject } H_0 & \textit{otherwise} \end{cases}$$

has $P(\textit{type I error}) = \alpha$.

Example 2.2. Churchill claims that he will receive half the votes for the House of Commons seat for the constituency of Dundee.[4] If we do not believe Churchill's claim and we are doubtful of his popularity, then we would seek to test an alternative hypothesis. How should we write down our research hypotheses?

If we let p be the fraction of the population voting for Churchill, then we have the null hypothesis,

$$H_0 : p = 0.5\,,$$

and the alternative hypothesis (we believe Churchill is less popular than he claims),

$$H_a : p < 0.5\,.$$

Support for the alternative hypothesis is obtain by showing lack of support for it's converse hypothesis (the null hypothesis). $\Diamond$

Example 2.3. Suppose that $m = 15$ voters are selected from Dundee and X, the number favoring Churchill, is recorded. Based on observing X, we construct a rejection region $R = \{x : x \leq k\}$. If k is small compared to m, then the rejection region would provide pretty strong evidence to reject H_0. How should one choose the rejection region?

Assume now that $m = 15$ voters are polled and that we select $k = 2$ to have a rejection region $R = \{x \leq 2\}$. For this choice of k, the rejection region R provides strong support to reject H_0. If we assume the null hypothesis is true, then we would expect that approximately half of the 15 voters (ca. 7) would plump for Churchill. Observing $x = 0$, $x = 1$ or $x = 2$ (the values that would place us in the rejection region) would provide strong evidence *against* H_0.

We can calculate the probability of a type I error. From the definition of type I error,

$$\begin{aligned} \alpha &= P(\text{type I error}) \\ &= P(\text{rejecting } H_0 \text{ when } H_0 \text{ is true}) \\ &= P(X \in R \text{ when } H_0 \text{ is true}) \\ &= P(X \leq 2 \text{ when } p = 0.5)\,. \end{aligned}$$

Since $X \sim \mathsf{Binom}(15, 0.50)$,[5] we calculate that $\alpha = 0.00369$. Thus, for this particular choice of rejection region R, the risk of concluding that Churchill will lose if in fact he is the winner is very small.

[4] Sir Winston Churchill was Member of Parliament for Dundee from 1908–1922. [W]

[5] X is a binomial random variable because it can be modeled as m independent Bernoulli trails each with probability p of success (i.e. votes Churchill) as long as the sample size m is much smaller than the population of Dundee. If we had the means to canvas nearly the whole population, what goes wrong conceptually?

For this rejection region, how good is the test at protecting us from type II errors, i.e., concluding that Churchill is the winner if in fact he will lose? Suppose that Churchill receive 25 of the votes ($p = 0.25$). The probability of a type II error β,

$$\begin{aligned}\beta &= P(\text{type II error}) \\ &= P(\text{fail to reject } H_0 \text{ when } H_0 \text{ false}) \\ &= P(X \notin R \text{ when } H_0 \text{ false}) \\ &= P(X > 2 \text{ when } p = 0.3)\,.\end{aligned}$$

For $X \sim \mathsf{Binom}(15, 0.25)$, we calculate $\beta = 0.764$. If we use $R = \{x \leq 2\}$ then our test will lead us to conclude that Churchill is the winner with probability 0.764 even if p is as low as 0.25!

If we repeat these caclculations for $R^* = \{x \leq 5\}$, we find $\alpha = 0.151$ versus $\beta = 0.148$, even if p is as low as 0.25, which is a much better balance between type I and type II errors. $\Diamond$

To summarize, the elements of a statistical test are:

1. Null hypothesis (H_0)
2. Alternative hypothesis (H_a)
3. Test statistic
4. Rejection region
5. Significance level (α)

Topic 3

Inferences based on a single sample

In a few situations we can derive the sampling distribution for the statistic of interest and use this as the basis for constructing confidence intervals and hypothesis tests. Presently we estimate population means μ in Section 3.1, population proportions p in Section 3.2, and population variances σ^2 in Section 3.3 in some special cases.

3.1 Estimating means

If the parameter of interest is the population mean $\theta = \mu$, then what can be said about the distribution of the sample mean estimator $\hat{\theta} = \overline{X}$ in (2.1)? We will consider three cases,

1. normal population with known σ^2,
2. any population with unknown σ^2, when the sample size m is large,
3. normal population with unknown σ^2, when the sample size m is small,

where the form of the confidence interval and hypothesis test statistic for μ can be derived using the approximate normality of the sample mean.

In general, the confidence intervals for the mean based on normality theory will have the form:

$$\text{point estimate } \mu \pm (\text{critical value of reference dist.}) \cdot (\text{precision of point estimate}), \tag{3.1}$$

where the reference distribution will be the standard normal (for 1. and 2.) and the Student's t distribution (for 3.). The critical value corresponds to the value under the reference distribution that yields the two-sided (symmetric) tail areas summing to $1 - \alpha$.

3.1.1 Mean of a normal population with known variance

When sampling from a normal population with known mean and variance, the estimator for the sample mean is also normal with mean μ and variance σ^2/m where m is the sample size. Standardizing,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{m}} \quad \sim \mathsf{N}(0, 1) \tag{3.2}$$

we see that

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{m}} < z_{\alpha/2}\right) = 1 - \alpha\,.$$

Based on knowing the sampling distribution of the estimator, we state the following CI.

Definition 3.1. A $100(1-\alpha)\%$ **confidence interval** for the mean μ of a normal population when the value of σ^2 is known is given by

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{m}}, \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{m}}\right), \tag{3.3}$$

or $\overline{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{m}$, where m is the sample size.

The CI for the mean (3.3) can be expressed (cf. (3.1)) as

$$\text{point estimate } \mu \pm (z \text{ critical value}) \cdot (\text{standard error of mean}).$$

The z critical value is related to the tail areas under the standard normal curve; we need to find the z-score having a cumulative probability equal to $1-\alpha$ according to Definition 2.3.

Example 3.1. Consider 400 samples from a normal population with a known standard deviation $\sigma = 17000$ with mean $\overline{x} = 20992$ (as depicted in 3.1). How do we construct a 95% confidence interval for μ?

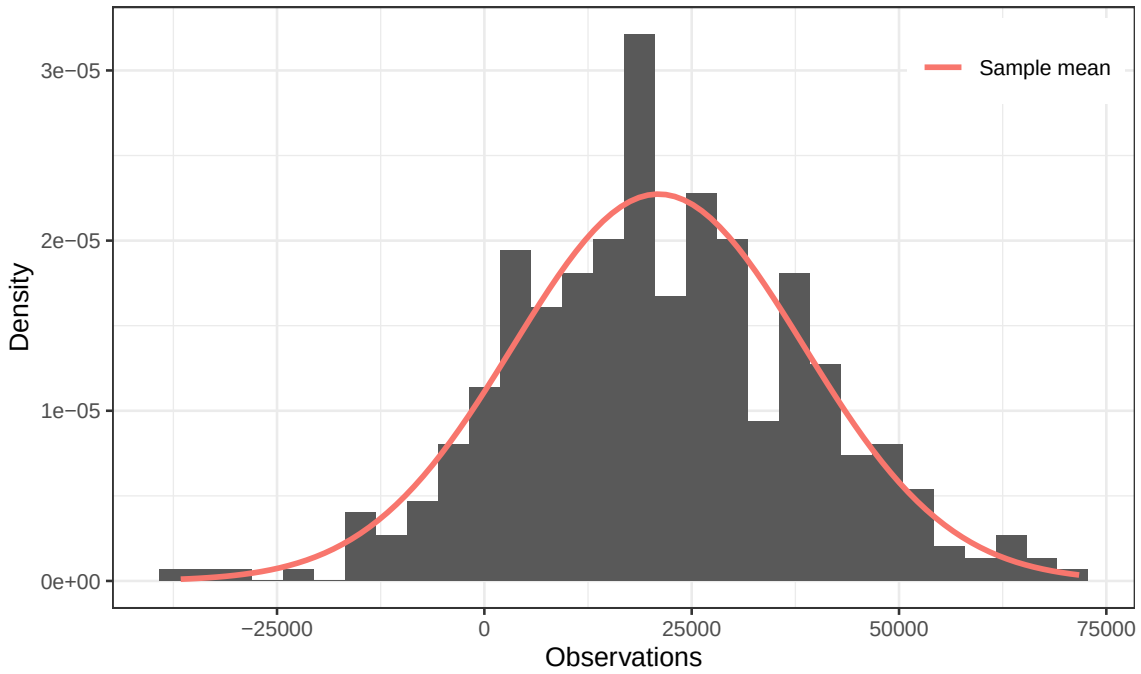

Figure 3.1: 400 samples from a normal population with known variance $\sigma = 17000$ together with the corresponding (normal) sampling distribution for the observed mean.

For $\alpha = 0.05$, the critical value $z_{0.025} = 1.96$; this value can be found by looking in a table of critical z values or using the `r` code `qnorm(1-.05/2)`. From Definition 3.1,

$$\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{m}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{m}}\right) = \left(20992 - 1.96\frac{17000}{\sqrt{400}}, 20992 + 1.96\frac{17000}{\sqrt{400}}\right)$$
$$= (19326, 22658).$$

The data above was generated with a true population parameter $\mu = 21500$, and, incidentally, the CI actually contains the parameter value. $\Diamond$

As noted in (3.1) and (3.3), the width of a CI is related to the estimator's precision. The confidence level (or reliability) is inversely related to this precision. When the population is normal and the variance is known, then an appealing strategy is to determine the sample size necessary to achieve a desired confidence level

and precision. A general formula for the sample size m^* necessary to achieve an interval width w is obtained at confidence level α is obtained by equating w to $2z_{\alpha/2} \cdot \sigma/\sqrt{m^*}$ and then solving for m^*.

Proposition 3.1. *The sample size n required to achieve a CI for μ with width w at level α is given by,*

$$m^* = \left(2z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2 .$$

From Proposition 3.1, we see that the smaller the desired w then the larger m^* must be (and subsequently, the more effort that must be allocated to data collection).

Example 3.2. In Example 3.1 we identified a 95% confidence interval for a normal population with known variance. The range (width) of that interval was $22658 - 19326 = 3332$. By how much would m need to increase to halve the interval width?

Using Proposition 3.1,

$$m = \left(2 \cdot 1.96 \cdot \frac{17000}{1666}\right)^2 = (40)^2 = 1600 .$$

Thus, we find that for the same level $\alpha = 0.05$, we would need to quadruple our original sample size to halve the interval. It is expensive to remove uncertainty! $\diamondsuit$

Suppose now that we would like to consider a hypothesis test for the population mean, such as $H_0 : \mu = \mu_0$. Starting from (3.2) and assuming that the null hypothesis is true, we find

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{m}} .$$

The statistic Z measures the distance (measured in units of $\text{sd}[\overline{X}]$) between $\overline{X}$ and its expected value under the null hypothesis. We will use the statistic Z to determine if there is substantial evidence against H_0 i.e. if the distance is too far in a direction consistent with H_a.

Proposition 3.2. *Assume that we sample $X_1, \ldots, X_m$ from a normal population with mean μ and known variance σ^2.*

Consider $H_0 : \mu = \mu_0$. The test statistic is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{m}} . \tag{3.4}$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \mu > \mu_0$, then $P = 1 - \Phi(z)$, i.e., upper-tail $R = \{z > z_\alpha\}$.

If $H_a : \mu < \mu_0$, then $P = \Phi(z)$, i.e., lower-tail $R = \{z < -z_\alpha\}$.

If $H_a : \mu \neq \mu_0$, then $P = 2(1 - \Phi(|z|))$, i.e., two-tailed $R = \{|z| > z_{\alpha/2}\}$.

We recall that $\Phi(z)$ is the area in the lower-tail of the standard normal density, i.e., to the *left* of the calculated value of z. Thus $1 - \Phi(z)$ is the area in the upper-tail and $2(1 - \Phi(|z|))$ is twice the area capture in the upper-tail by $|z|$ i.e. the sum of the area in the tails corresponding to $\pm z$. If $P < \alpha$ then we reject H_0 at level α as the data provides sufficient evidence at the α level against the null hypothesis.

Example 3.3. Let's return to the data in Example 3.1, where we sample from a normal population with a known standard deviation $\sigma = 17000$. Suppose that someone cliams the true mean is $\mu_0 = 20000$. Does our sample mean $\overline{x} = 20992$ based on $m = 400$ samples provide evidence to contradict this claim at the $\alpha = 0.05$ level?

The first thing to record is our parameter of interest: μ, the true population mean. The null hypothesis, which we assume to be true, is a statement about the value of μ,

$$H_0 : \mu = 20000\,,$$

and the alternative hypothesis is

$$H_a : \mu \neq 20000\,,$$

since we are concerned with a deviation in either direction from $\mu_0 = 20000$.

Since the population is normal with known variance, we compute the test statistic:

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{m}} = \frac{20992 - 20000}{17000/\sqrt{400}} = 1.167\,.$$

That is, the observed sample mean $\overline{x}$ is a little more than 1 standard deviation above what we would expect under H_0. Consulting 3.2, we see that a two-tailed test is indicated for this particular H_a (i.e., containing "$\neq$"). The P-value is the area,[1]

$$P = 2(1 - \Phi(1.167)) = 2(0.1216052) = 0.2432.$$

Thus, since $P = 0.2432 > 0.05 = \alpha$, we fail to reject H_0 at the level 0.05. The data does not support the claim that the true population mean differs from the value 20000 at the 0.05 level. $\diamondsuit$

3.1.2 Mean of a population with unknown variance (large-sample)

Consider samples $X_1, \ldots, X_m$ from a population with mean μ and variance σ^2. Provided that m is large enough, the Central Limit Theorem implies that the estimator for the sample mean $\overline{X}$ in (2.1) has *approximately* a normal distribution. Then

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{m}} < z_{\alpha/2}\right) \approx 1 - \alpha\,, \tag{3.5}$$

since the transformed variable has approximately a standard normal distribution. Thus, computing a point estimate based on a large m of samples yields a CI for the population parameter μ at an *approximate* confidence level α. However, it is often the case that the variance is unknown. When m is large, replacing the population variance σ^2 by the sample variance S^2 in (2.2) will not typically introduce too much additional variability.

Proposition 3.3. *For large sample size m, an approximate* $100(1-\alpha)\%$ *confidence interval for the mean* μ *of any population when the variance is uknown is given by*

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{m}}, \overline{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{m}}\right), \tag{3.6}$$

or $\overline{x} \pm z_{\alpha/2} \cdot s/\sqrt{m}$.

The CI for the mean (3.6) applies regardless of the shape of the population distribution so long as the number of samples is large. A rule of thumb is that $m > 40$ is sufficient.[2] In words, the CI (3.6) can be expressed (cf. (3.1)) as

$$\text{point estimate } \mu \pm (z \text{ critical value}) \cdot (\text{estimated standard error of mean})\,.$$

[1] Note $\Phi(z) = P(Z \leq z)$ is found by calling `pnorm(z)` in `r` or by looking up the value in a Z table.

[2] For $m > 20$, the interval estimate

$$\text{point estimate } \pm 2\,\text{sd}$$

has 95% coverage and is surprisingly robust, i.e. applies to a wide variety of population distributions including the normal. However, this rule of thumb won't apply if you want to consider some different level, say 80% [van Belle, 2008, §1].

Typically, a large-sample CI for a general parameter θ holds that is similar in nature to (3.6) for any estimator $\hat{\theta}$ that satisfies: (1) approximately normal in distribution, (2) approximately unbiased, and (3) an expression for the standard error is available.

To conduct a large-sample hypothesis test regarding the population mean μ, we consider the test statistic

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{m}}$$

under the null hypothesis, i.e., we replace the population standard deviation σ by the sample standard deviation S. When the number of samples m is large (say $m > 40$) then Z will be approximately normal. Substituting this test statistic Z for (3.4), we follow Proposition 3.2 to determine how to calculate the P-value.

3.1.3 Mean of a normal population with unknown variance

In Section 3.1.1, we considered samples $X_1, \ldots, X_m$ from a normal population with a known μ and σ^2. In contrast, here we consider samples from a normal population and assume the population parameters μ and σ^2 are unknown. If the number of samples is large, the discussion in Section 3.1.2 indicates that the rv $Z = (\overline{X} - \mu)\sqrt{m}/S$ has approximately a standard normal distribution. However, if m is not sufficiently large[3] then the transformed variable will be more spread out than a standard normal distribution.

Theorem 3.1. *For the sample mean $\overline{X}$ based on m samples from a normal distribution with mean μ, the rv*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{m}} \quad \sim \mathsf{t}(m-1)\,, \tag{3.7}$$

that is, T has Student's t *distribution with $\nu = m - 1$ df.*

This leads us to consider a CI for the population parameter μ that is based on critical values of the t distribution.

Proposition 3.4. *A $100(1-\alpha)\%$ **confidence interval** for the mean μ of a normal population, when σ^2 is unknown, is given by*

$$\left(\overline{x} - t_{\alpha/2,m-1} \cdot \frac{s}{\sqrt{m}}\,, \overline{x} + t_{\alpha/2,m-1} \cdot \frac{s}{\sqrt{m}}\right), \tag{3.8}$$

or $\overline{x} \pm t_{\alpha/2,m-1} \cdot s/\sqrt{m}$. Here $\overline{x}$ and s are the sample mean and sample standard deviation, respectively.

Example 3.4. Let us return to the the height of 31 felled black cherry trees from the **Cherry Tree Data** in Table 2.1. Give a 99% CI for the population mean μ.

For $m = 31$, the critical value of the reference distribution is $t_{0.005,30} \approx 2.7499$, which can looked up in a table of critical values for $\mathsf{t}(\nu = m-1)$ or found using the `r` command `qt(1-0.01/2, df = 31-1)`. The sample mean $\overline{x} = 76$ (computed in Example 2.1) is combined with the sample standard deviation,

$$\begin{aligned} s &= \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2} \\ &= \sqrt{\frac{1}{30}\left((63-76)^2 + \cdots + (87-76)^2\right)} \\ &= 6.372\,, \end{aligned}$$

[3] Recall that we would consider $m > 40$ to be large.

to form the interval estimate

$$\left(\overline{x} - t_{\alpha/2,m-1} \cdot \frac{s}{\sqrt{m}}, \overline{x} + t_{\alpha/2,m-1} \cdot \frac{s}{\sqrt{m}}\right)$$
$$= \left(76 - 2.750 \cdot \frac{6.372}{\sqrt{31}}, 76 + 2.750 \cdot \frac{6.372}{\sqrt{31}}\right)$$
$$= (72.85, 79.15) .$$

For comparison, the critical value $t_{.01/2,\nu}$ for $\nu = 13, \ldots, 30$

```
qt(1-0.01/2, df = seq(12:39))
```

```
 [1] 63.656741  9.924843  5.840909  4.604095  4.032143  3.707428  3.499483  3.355387
 [9]  3.249836  3.169273  3.105807  3.054540  3.012276  2.976843  2.946713  2.920782
[17]  2.898231  2.878440  2.860935  2.845340  2.831360  2.818756  2.807336  2.796940
[25]  2.787436  2.778715  2.770683  2.763262
```

can deviate significantly from the corresponding $z_{0.01/2} = 2.575829$. In particular, if we had erroneously used the large sample estimate (3.6) then we would have obtained 99% CI (73.05, 78.95) which might give us a false sense of security as it is narrower. ◊

In contrast to Proposition 3.1, it is difficult to select the sample size m to control the width of the t-based CI as the width involves the unknown (before the sample is acquired) s and because m also enters through $t_{\alpha/2,m-1}$. A one-sample t test based on (3.7) can be used to test hypothesis about the population mean when the population is normal and σ^2 is unknown.

Proposition 3.5. *Assume that we sample $X_1, \ldots, X_m$ from a normal population with mean μ and unknown variance σ^2.*

Consider $H_0 : \mu = \mu_0$. The test statistic is

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{m}} .$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \mu > \mu_0$, then P-value is the area under $\mathsf{t}(m-1)$ to the right of t.

If $H_a : \mu < \mu_0$, then P-value is the area under $\mathsf{t}(m-1)$ to the left of t.

If $H_a : \mu \neq \mu_0$, then P-value is twice the area under $\mathsf{t}(m-1)$ to the right of $|t|$.

Example 3.5. From the **Cherry Tree Data**, let's look at the average volume of timber, given in Table 3.1. The distribution for this data is approximately normal.[4] We might ask if the data provides compelling evidence, say at level 0.05, for concluding that the true average volume of timber exceeds 21.3 cubic feet.[5]

Let's carry out a significance test for the true average volume of timber μ at level $\alpha = 0.05$. We assume the null hypothesis

$$H_0 : \mu = 21.3 .$$

[4] After looking at the normal quantile-quantile plot, I decided to test a hypothesis. For level 0.01, I ran a Kolmogorov–Smirnov test for the null hypothesis that the data is consistent with $\mathsf{N}(\overline{x}, s^2)$ vs the alternative that the data is not consistent with the specified reference distribution. The P-value attained was $P = 0.2532 > 0.10$, and therefore I fail to reject the null hypothesis. The data is consistent with being drawn from a normal population.

[5] How much wood is that? About a sixth of a cord. A full cord of chopped fire wood in the US is 124 cu ft; about enough to keep you warm through a New England winter (according to my inlaws).

Table 3.1: Observations of $m = 31$ felled black cherry trees.

Volume [cu ft]
10.2, 10.3, 10.3, 15.6, 16.4, 18.2, 18.8, 19.1, 19.7, 19.9, 21.0, 21.3, 21.4, 22.2, 22.6, 24.2, 24.9, 25.7, 27.4, 31.7, 33.8, 34.5, 36.3, 38.3, 42.6, 51.0, 51.5, 55.4, 55.7, 58.3, 77.0

An appropriate null hypothesis is

$$H_a : \mu > 21.3\,,$$

that is, we will adopt the stance that the true average exceeds $\mu_0 = 21.3$ only if the null is rejected.

From our $m = 31$ samples, we find that $\overline{x} = 30.17$ and that $s = 16.44$. The computed value of the one-sample t-statistic is given by

$$\begin{aligned} t &= \frac{\overline{x} - \mu_0}{s/\sqrt{m}} \\ &= \frac{30.17 - 21.3}{16.44/\sqrt{31}} \\ &= 3\,. \end{aligned} \tag{3.9}$$

The test is based on $\nu = 31 - 1$ df, and $P = 0.04096$. This is the upper-tail area, i.e. the area to the right of t (see Figure 3.2).

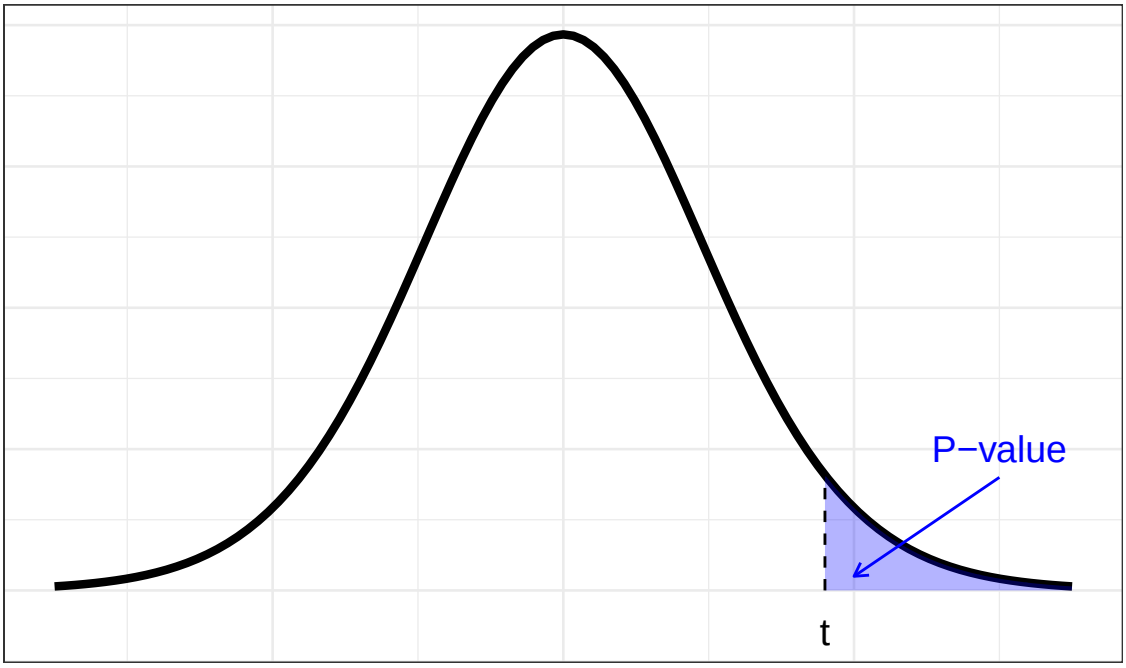

Figure 3.2: For this test, the P-value is the upper-tail area, i.e., to the right of the computed statistic t.

Since $P \ll \alpha$, we reject the null hypothesis that the population mean is 21.3. The data provides sufficient evidence that the population mean is different from 21.3. $\Diamond$

3.2 Estimating proportions

Consider a population of size M in which a proportion p of the population satisfies a given property. The $p \in (0, 1)$ is a parameter characterizing the population, with distribution $F(p)$, that we might be interested

in estimating.[6] A sample, $X_1, \ldots, X_m \sim F(p)$, from the population contains a proportion,

$$\hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i\,, \tag{3.10}$$

satisfying the given property. The estimator $\hat{p}$ varies with the sample and for large m it's sampling distribution has the following properties:

$$\mu_{\hat{p}} = \mathbf{E}[X_i] = p$$

and

$$\sigma^2_{\hat{p}} = \frac{\mathrm{Var}[X_i]}{m} = \frac{p(1-p)}{m}\,, \tag{3.11}$$

provided that m is small relative to M (a rule of thumb is $m \leq 0.05M$).[7] Moreover, by invoking the Central Limit Theorem we have the distribution of $\hat{p}$ is approximately normal for sufficiently large m as (3.10) is a sample mean. Indeed, this normal approximation works well for moderately large m as long as p is not too close to zero or one; a rule of thumb is that $mp > 5$ and $m(1-p) > 5$.

Proposition 3.6. *For large samples n, a* $100(1-\alpha)\%$ *confidence interval for the parameter p is given by*

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{m}}\,. \tag{3.12}$$

This follows from Proposition 3.3 by observing that (3.10) is a sample mean and replacing the standard error $\sigma_{\hat{p}}$ from (3.11) by the estimated standard error,

$$\widehat{\mathsf{se}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{m}}\,;$$

recall the s in (3.6) is the sample variance for the *population* and $s/\sqrt{m} = \mathsf{se}$ is the standard error of the point estimator.

Proposition 3.7. *Let X be the count of members with a given property based on a sample of size m from a population where a proportion p share the property. Then* $\hat{p} = X/m$ *is an estimator of p. Assume* $mp_0 \geq 10$ *and* $m(1-p_0) \geq 10$.

Consider $H_0 : p = p_0$. *The test statistic is*

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/m}}\,.$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : p > p_0$, *then P-value is the area under* $\mathsf{N}(0,1)$ *to the right of z.*

If $H_a : p < p_0$, *then P-value is the area under* $\mathsf{N}(0,1)$ *to the left of z.*

If $H_a : p \neq p_0$, *then P-value is twice the area under* $\mathsf{N}(0,1)$ *to the right of* $|t|$.

[6] Here we write F for a general distribution, but what special distribution might this be?

[7] Note that if m is large relative to M ($m > 0.05M$) then the variance (3.11) must be adjusted by a factor (related to the hypergeometric distribution):

$$\sigma^2_{\hat{p}} = \frac{p(1-p)}{m}\,\frac{M-m}{M-1}\,,$$

where for fixed m the factor converges to 1 as $M \to \infty$.

Example 3.6. Let us revisit Example 2.2, where we were considering Churchill's claim that he will receive half the votes for the House of Commons seat for the constituency of Dundee. We are skeptical that he is as popular as he says. Suppose 116 out of 263 Dundonians polled claimed that they intended to vote for Churchill. Can it be concluded at significance level 0.10 that more than half of all eligible Dundonains will vote for Churchill?

The parameter of interest is p, the proportion of votes for Churchill. The null hypothesis is $H_0 : p = 0.5$. The alternative hypothesis is $H_a : p < 0.5$, since we . Since $263(0.5) = 131.5 > 10$, we satisfy the assumptions stated in Proposition 3.7.

Based on the sample, $\hat{p} = 116/263 = 0.4411$. The test statistic value is

$$\begin{aligned} z &= \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/m}} \\ &= \frac{0.4411 - 0.5}{\sqrt{0.5(1-0.5)/263}} \\ &= -1.91\,. \end{aligned}$$

The P-value for this lower-tailed z test is $P = \Phi(-1.91) = 0.028$. Since $P < 0.10 = \alpha$, we reject the null hypothesis at the 0.05 level. The evidence for concluding that the true proportion is different from $p_0 = 0.5$ at the 0.10 level is compelling.[8] $\Diamond$

3.3 Estimating variances

Next we consider estimates of the population variance (and standard deviation) when the population is assumed to have a normal distribution. In this case, the sample variance S^2 in (2.2) provides the basis for inferences. Consider iid samples $X_1, \ldots, X_m \sim \mathsf{N}(\mu, \sigma^2)$. We provide the following theorem without proof.

Theorem 3.2. *For the sample variance S^2 based on m samples from a normal distribution with variance σ^2, the rv*

$$V = \frac{(m-1)S^2}{\sigma^2} = \frac{\sum_i (X_i - \overline{X})^2}{\sigma^2} \qquad \sim \chi^2_{m-1}\,,$$

that is, V has a χ^2 distribution with $\nu = m - 1$ df.

Based on Theorem 3.2,

$$P\left(\chi^2_{1-\alpha/2,m-1} < \frac{(m-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,m-1}\right) = 1 - \alpha\,,$$

i.e., the area captured between the right and left tail critical χ^2 values is $1 - \alpha$. The expression above can be further manipulated to obtain an interval for the unknown parameter σ^2:

$$P\left(\frac{(m-1)s^2}{\chi^2_{\alpha/2,m-1}} < \sigma^2 < \frac{(m-1)s^2}{\chi^2_{1-\alpha/2,m-1}}\right) = 1 - \alpha\,,$$

where we substitute the computed value of the point estimate s^2 for the estimator into the limits to give a CI for σ^2. If we take square roots of the inequality above, we obtain a CI for the population standard deviation σ.

[8] Churchill took ca. 44% of the vote in the 1908 by-election to become MP for Dundee. [W]

Proposition 3.8. *A* $100(1-\alpha)\%$ *confidence interval for the variance of a normal population is*

$$\left((m-1)s^2/\chi^2_{\alpha/2,m-1}\,,(m-1)s^2/\chi^2_{1-\alpha/2,m-1}\right).$$

A $100(1-\alpha)\%$ *confidence interval for the standard deviation* σ *of a normal population is given by taking the square roots of the lower and upper limits in* (3.8).

Example 3.7. For the **Cherry Tree Data** in Table 3.1 concerning the timber volume of 31 felled black cherry trees, give a 95 CI for the variance.

We are interested in estimating the true variance σ^2 of the volume of timber based on $m = 31$ samples. Recall that the mean of our data is $\overline{x} = 30.17$ cu ft and that the sample variance is $s^2 = 270.2$ using the estimator (2.2). The critical values for the $\chi^2_{.975,30} = 16.7908$ and $\chi^2.025, 30 = 46.9792$ can be found by checking a table of critical values of the $\chi^2(\nu = 30)$ distribution or by using the `r` code `qchisq(1-0.05/2, df=30, lower.tail = FALSE)` and `qchisq(0.05/2, df=df, lower.tail = FALSE)`, respectively (see 3.3).

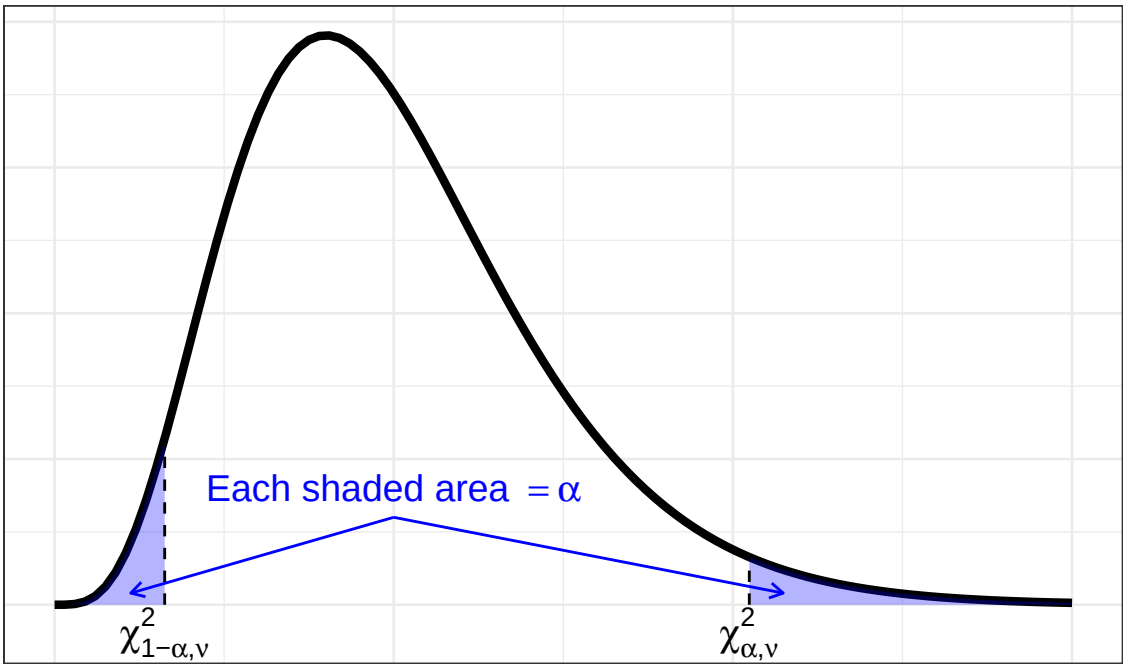

Figure 3.3: As the χ^2 distribution is not symmetric, the upper and lower critical values will not be the same (the shaded areas are equal).

Pulling everything together, a 95% CI for the population variance is given by

$$\begin{aligned}&\left((m-1)s^2/\chi^2_{\alpha/2,m-1}\,,(m-1)s^2/\chi^2_{1-\alpha/2,m-1}\right)\\ &\quad=((30)270.2/46.9792\,,(30)270.2/16.7908)\\ &\quad=(172.5\,,482.8)\,.\end{aligned}$$

Note the position of the critical values—don't swap them round. ◊

Topic 4

Inferences based on two samples

We consider inferences—estimators, confidence intervals, and hypothesis testing—for comparing means, proportions, and variances based on two independent samples from different populations, respectively, in Sections 4.1, 4.3, 4.4. We also consider inferences when the samples are not independent, so-called paired samples, in Section 4.2.

4.1 Comparing means

Let us assume that we have two normal populations with iid samples

$$X_1, \dots, X_m \sim \mathsf{N}(\mu_X, \sigma_X^2)$$

and

$$Y_1, \dots, Y_n \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$$

and, moreover, that the X and Y samples are independent of one another. When comparing the means of two populations, the quantity of interest is the difference: $\mu_X - \mu_Y$.

Proposition 4.1. *If we consider the sample means $\overline{X}$ and $\overline{Y}$, then the mean of the variable $\overline{X} - \overline{Y}$ is,*

$$\mu_{\overline{X}-\overline{Y}} = \mathbf{E}\left[\overline{X} - \overline{Y}\right] = \mu_X - \mu_Y\,,$$

and the variance is,

$$\sigma^2_{\overline{X}-\overline{Y}} = \mathrm{Var}\left[\overline{X} - \overline{Y}\right] = \frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\,.$$

Proposition 4.1 follows directly from the definition of the sample mean in (2.1) and properties of expectation and variance. If our parameter of interest is

$$\theta = \mu_1 - \mu_2\,,$$

then its estimator,

$$\hat{\theta} = \overline{X} - \overline{Y}\,,$$

is normally distributed with mean and variance given by Proposition 4.1. If the samples sizes m and n are large, then the estimator is approximately normally distributed by the Central Limit Theorem regardless of the population. We now discuss CIs and hypothesis tests for comparing population means $\theta = \mu_X - \mu_Y$. We consider three cases when comparing means:

1. normal populations when the variances σ_X^2 and σ_Y^2 are known,

2. any populations with unknown variances σ_X^2 and σ_Y^2, when the sample sizes m and n are large,
3. normal populations when the variances σ_X^2 and σ_Y^2 are unknown, when the sample sizes m and n are small,

noting that the development largely reflects that of Section 3.1.

4.1.1 Comparing means of normal populations when variances are known

When σ_X^2 and σ_Y^2 are known, standardizing $\overline{X} - \overline{Y}$ yields the standard normal variable:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \quad \sim \mathsf{N}(0,1)\,. \tag{4.1}$$

Inferences proceed by treating the parameter of interest θ as in the single sample case using the test statistic (4.1).

Proposition 4.2. *A* $100(1-\alpha)\%$ *CI for the parameter* $\theta = \mu_X - \mu_Y$ *based on samples of size* m *from a normal population* $\mathsf{N}(\mu_X, \sigma_X^2)$ *and of size* n *from* $\mathsf{N}(\mu_Y, \sigma_Y^2)$ *with known variances, is given by*

$$(\overline{x} + \overline{y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}\,.$$

Proposition 4.3. *Assume that we sample iid* $X_1, \ldots, X_m \sim \mathsf{N}(\mu_X, \sigma_X^2)$ *and iid* $Y_1, \ldots, Y_n \sim \mathsf{N}(\mu_Y, \sigma_Y^2)$ *and that the* X *and* Y *samples are independent.*

Consider $H_0 : \mu_X - \mu_Y = \theta_0$*. The test statistic is*

$$Z = \frac{\overline{X} - \overline{Y} - \theta_0}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}\,. \tag{4.2}$$

For a hypothesis test at level α*, we use the following procedure:*

If $H_a : \mu_X - \mu_Y > \theta_0$*, then* $P = 1 - \Phi(z)$*, i.e., upper-tail* $R = \{z > z_\alpha\}$.

If $H_a : \mu_X - \mu_Y < \theta_0$*, then* $P = \Phi(z)$*, i.e., lower-tail* $R = \{z < -z_\alpha\}$.

If $H_a : \mu_X - \mu_Y \neq \theta_0$*, then* $P = 2(1 - \Phi(|z|))$*, i.e., two-tailed* $R = \{|z| > z_{\alpha/2}\}$.

4.1.2 Comparing means when the sample sizes are large

When the samples are large, then the assumptions about normality of the populations and knowledge of the variances σ_X^2 and σ_Y^2 can be relaxed. For sufficiently large m and n, the difference of the sample means, $\overline{X} - \overline{Y}$, has approximately a normal distribution for any underlying population distributions by the Central Limit Theorem. Moreover, if m and n are large enough, then replacing the population variances with the sample variances S_X^2 and S_Y^2 will not increase the variability of the estimator or the test statistic too much.

Proposition 4.4. *For* m *and* n *sufficiently large, an approximate* $100(1-\alpha)\%$ *CI for* $\mu_X - \mu_Y$ *for two samples from populations with any underlying distribution is given by*

$$(\overline{x} + \overline{y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}$$

Proposition 4.5. *Under the same assumptions and procedures as in Proposition 4.3, a large-sample, i.e., $m > 40$ and $n > 40$, test statistic,*

$$Z = \frac{\overline{X} - \overline{Y} - \theta_0}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}},$$

can be used in place of (4.2) *for hypothesis testing.*

4.1.3 Comparing means of normal populations when variances are unknown and the sample size is small

If σ_X and σ_Y are unknown and either sample is small (e.g., $m < 30$ or $n < 30$), but both populations are normally distributed, then we can use Student's t distribution to make inferences. We provide the following theorem without proof.

Theorem 4.1. *When both population distributions are normal, the standardized variable*

$$T = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \quad \sim \mathsf{t}(\nu)$$

where the df ν is estimated from the data. Namely, ν is given by (round ν down to the nearest integer):

$$\nu = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{(s_X^2/m)^2}{m-1} + \frac{(s_Y^2/n)^2}{n-1}} = \frac{\left(s_{\overline{X}}^2 + s_{\overline{Y}}^2\right)^2}{\frac{s_{\overline{X}}^4}{m-1} + \frac{s_{\overline{Y}}^4}{n-1}} \tag{4.3}$$

where s_X^2 and s_Y^2 are point estimators of the sample variances; alternatively, we see that the formula (4.3) *can also be written in terms of the standard error of the sample means:*

$$s_{\overline{X}} = \frac{s_X}{\sqrt{m}} \quad \textit{and} \qquad s_{\overline{Y}} = \frac{s_Y}{\sqrt{n}}.$$

The formula (4.3) for the data-driven choice of ν calls for the computation of the standard error of the sample means.

Proposition 4.6. *A $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$ for two samples of size m and n from normal populations where the variances are unknown is given by*

$$(\overline{x} - \overline{y}) \pm t_{\alpha/2,\nu} \sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}},$$

where we recall that $t_{\alpha/2,\nu}$ is the $\alpha/2$ critical value of $\mathsf{t}(\nu)$ with ν given by (4.3).

Proposition 4.7. *Assume that we sample iid $X_1, \ldots, X_m$ and iid $Y_1, \ldots, Y_n$ from normal populations with with unknown variances and means μ_X and μ_Y, respecitvely, and that the X and Y samples are independent.*

Consider $H_0 : \mu_X - \mu_Y = \theta_0$. The test statistic is

$$T = \frac{\overline{X} - \overline{Y} - \theta_0}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}. \tag{4.4}$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \mu_X - \mu_Y > \theta_0$, then P-value is the area under $\mathsf{t}(\nu)$ *to the right of t, i.e., upper-tail $R = \{t > t_{\alpha,\nu}\}$.*

If $H_a : \mu_X - \mu_Y < \theta_0$, then P-value is the area under $\mathsf{t}(\nu)$ *to the left of t, i.e., lower-tail $R = \{t < -t_{\alpha,\nu}\}$.*

If $H_a : \mu_X - \mu_Y \neq \theta_0$, then P-value is twice the area under $\mathsf{t}(\nu)$ *to the right of $|t|$, i.e., two-tailed $R = \{|t| > t_{\alpha/2,\nu}\}$.*

Here ν is given by (4.3).

If the variances of the normal populations are unknown but are the same, $\sigma_X^2 = \sigma_Y^2$, then deriving CIs and test statistics for comparing the means can be simplified by considering a combined or pooled estimator for the single parameter σ^2. If we have two samples from populations with variance σ^2, then each sample provides an estimate for σ^2. That is, S_X^2, based on the m observations of the first sample, is one estimator for σ^2 and another is given by S_Y^2, based on n observations of the second sample. The correct way to combine these two estimators into a single estimator for the sample variance is to consider the **pooled estimator** of σ^2,

$$S_{\mathsf{p}}^2 = \frac{m-1}{m+n-2} S_X^2 + \frac{n-1}{m+n-2} S_Y^2 \,. \tag{4.5}$$

The pooled estimator is a weighted average that adjusts for differences between the sample sizes m and n.[1]

Proposition 4.8. *A $100(1-\alpha)\%$ CI for $\mu_X - \mu_Y$ for two samples of size m and n from normal populations where the variance σ^2 is unknown is given by*

$$(\overline{x} - \overline{y}) \pm t_{\alpha/2,m+n-2} \cdot \sqrt{s_{\mathsf{p}}^2 \left(\frac{1}{m} + \frac{1}{n}\right)},$$

where we recall that $t_{\alpha/2,m+n-2}$ is the $\alpha/2$ critical value of the $\mathsf{t}(\nu)$ *with $\nu = m + n - 2$ df.*

Similarly, one can consider a pooled t test, i.e., a hypothesis test based on the pooled estimator for the variance as opposed to the two-sample t test in Proposition 4.7. In the case of a pooled t test, the test statistic

$$T = \frac{\overline{X} - \overline{Y} - \theta_0}{\sqrt{S_{\mathsf{p}}^2 \left(\frac{1}{m} + \frac{1}{n}\right)}},$$

with the pooled estimator of the variance, replaces (4.4) in Proposition 4.7 and the same procedures are followed for determining the P-value with $\nu = m + n - 2$ in place of (4.3). If you have reasons to believe that $\sigma_X^2 = \sigma_Y^2$, these pooled t procedures are appealing because ν is very easy to compute.

Pooled *t* procedures are not robust if the assumption of equalized variance is violated. Theoretically, you could first carry out a statistical test $H_0 : \sigma_X^2 = \sigma_Y^2$ on the equality of variances and then use a pooled t procedure if the null hypothesis is not rejected. However, there is no free lunch: the typical F test for equal variances (see Section 4.4) is sensitive to normality assumptions. The two sample t procedures, with the data-driven choice of ν in (4.3), are therefore recommended unless, of course, you have a very compelling reason to believe $\sigma_X^2 = \sigma_Y^2$.

[1] If $m \neq n$, then the estimator with *more* samples will contain *more* information about the parameter σ^2. Thus, the simple average $(S_X^2 + S_Y^2)/2$ wouldn't really be fair, would it?

4.2 Comparing paired samples

The preceding analysis for comparing population means was based on the assumption that a random sample $X_1, \dots, X_m$ is drawn from a distribution with mean μ_X and that a completely independent random sample $Y_1, \dots, Y_n$ is drawn from a distribution with mean μ_Y. Some situations, e.g., comparing observations before and after a treatment or exposure, necessitate the consideration of paired values.

Consider a random sample of iid pairs

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

with $\mathbf{E}[X_i] = \mu_X$ and $\mathbf{E}[Y_i] = \mu_Y$. If we are interested in making inferences about the difference $\mu_X - \mu_Y$ then the paired differences

$$D_i = X_i - Y_i\,, \quad i = 1, \dots, n\,,$$

constitute a sample with mean $\mu_D = \mu_X - \mu_Y$ that can be treated using single-sample CIs and tests, e.g., see Section 3.1.3.

4.3 Comparing proportions

Consider a population containing a proportion p_X of individuals satisfying a given property. For a sample of size m from this population, we denote the sample proportion by $\hat{p}_X$. Likewise, we consider a population containing a proportion p_Y of individuals satisfying the same given property. For a sample of size n from this population, we denote the sample proportion by $\hat{p}_Y$. We assume the samples from the X and Y populations are independent. The natural estimator for the difference in population proportions $p_X - p_Y$ is the difference in the sample proportions $\hat{p}_X - \hat{p}_Y$.

Provided the samples are much smaller than the population sizes (i.e., the populations are about 20 times larger than the samples),

$$\mu_{(\hat{p}_X - \hat{p}_Y)} = \mathbf{E}[\hat{p}_X - \hat{p}_Y] = p_X - p_Y\,,$$

and

$$\sigma^2_{(\hat{p}_X - \hat{p}_Y)} = \mathrm{Var}[\hat{p}_X - \hat{p}_Y] = \frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}\,,$$

by considering the fact that the count of individuals satisfying the given property in each population will be independent draws from $\mathsf{Binom}(m, p_X)$ and $\mathsf{Binom}(n, p_Y)$, respectively. Further, if m and n are large (e.g., $m \geq 30$ and $n \geq 30$), then $\hat{p}_X$ and $\hat{p}_Y$ are (approximately) normally distributed. Standardizing $\hat{p}_X - \hat{p}_Y$,

$$Z = \frac{\hat{p}_X - \hat{p}_Y - (p_X - p_Y)}{\sqrt{\frac{p_X(1-p_X)}{m} + \frac{p_Y(1-p_Y)}{n}}} \quad \sim \mathsf{N}(0,1)\,.$$

A CI for $\hat{p}_X - \hat{p}_Y$ then follows from the large-sample CI considered in Section 3.1.2.

Proposition 4.9. *An approximate* $100(1-\alpha)\%$ *CI for* $p_X - p_Y$ *is given by*

$$\hat{p}_X - \hat{p}_Y \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{m} + \frac{\hat{p}_Y(1-\hat{p}_Y)}{n}}\,,$$

and, as a rule of thumb, can be reliably used if $m\hat{p}_X$, $m(1-\hat{p}_X)$, $n\hat{p}_Y$, *and* $n(1-\hat{p}_Y)$ *are greater than or equal to* 10.

Proposition 4.9 does not pool the estimators for the population proportions. However, if we are considering a hypothesis test concerning the equality of the population proportions with the null hypothesis

$$H_0 : p_X - p_Y = 0\,,$$

then we assume $p_X = p_Y$ as our default position. Therefore, as a matter of consistency, we should replace the standard error in (4.9) with a pooled estimator for the standard error of the population proportion,

$$\hat{p} = \frac{m}{m+n}\hat{p}_X + \frac{n}{m+n}\hat{p}_Y \,.$$

Proposition 4.10. *Assume that* $m\hat{p}_X$, $m(1-\hat{p}_X)$, $n\hat{p}_Y$, $n(1-\hat{p}_Y)$ *are all greater than* 10.

Consider $H_0 : p_X - p_Y = 0$. *The test statistic is*

$$Z = \frac{\hat{p}_X - \hat{p}_Y}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{m}+\frac{1}{n}\right)}} \,.$$

For a hypothesis test at level α, *we use the following procedure:*

If $H_a : p_X - p_Y > 0$, *then* $P = 1 - \Phi(z)$, *i.e., upper-tail* $R = \{z > z_\alpha\}$.

If $H_a : p_X - p_Y < 0$, *then* $P = \Phi(z)$, *i.e., lower-tail* $R = \{z < -z_\alpha\}$.

If $H_a : p_X - p_Y \neq 0$, *then* $P = 2(1 - \Phi(|z|))$, *i.e., two-tailed* $R = \{|z| > z_{\alpha/2}\}$.

4.4 Comparing variances

For a random sample

$$X_1, \ldots, X_m \sim \mathsf{N}(\mu_X, \sigma_X^2)$$

and an independent random sample

$$Y_1, \ldots, Y_n \sim \mathsf{N}(\mu_Y, \sigma_Y^2) \,,$$

the rv

$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \quad \sim \mathsf{F}(m-1, n-1) \,, \tag{4.6}$$

that is, F has an F distribution with df $\nu_1 = m-1$ and $\nu_2 = n-1$. The statistic F in (4.6) comprises the *ratio* of variances σ_X^2/σ_Y^2 and not the difference; therefore, the plausibility of $\sigma_X^2 = \sigma_Y^2$ will be based on how much the ratio differs from 1.

Proposition 4.11. *For the null hypothesis* $H_0 : \sigma_X^2 = \sigma_Y^2$, *the test statistic to consider is:*

$$f = \frac{s_X^2}{s_Y^2}$$

and the P-values are determined by the $\mathsf{F}(m-1, n-1)$ *curve where m and n are the respective sample sizes.*

A $100(1-\alpha)\%$ CI for the ratio σ_X^2/σ_Y^2 is based on forming the probability,

$$P(F_{1-\alpha/2,\nu_1,\nu_2} < F < F_{\alpha/2,\nu_1,\nu_2}) = 1 - \alpha \,,$$

where $F_{\alpha/2,\nu_1,\nu_2}$ is the $\alpha/2$ critical value from the $\mathsf{F}(\nu_1 = m-1, \nu_2 = n-1)$ distribution. Substituting (4.6) with point estimates for F and manipulating the inequalities it is possible to isolate the ratio σ_X^2/σ_Y^2,

$$P\left(\frac{1}{F_{\alpha/2,\nu_1,\nu_2}}\frac{s_X^2}{s_Y^2} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{1}{F_{1-\alpha/2,\nu_1,\nu_2}}\frac{s_X^2}{s_Y^2}\right) = 1 - \alpha \,.$$

Proposition 4.12. *A* $100(1-\alpha)\%$ *CI for the ratio of population variances* σ_X^2/σ_Y^2 *is given by*

$$\left(F_{\alpha/2,m-1,n-1}^{-1} s_X^2/s_Y^2 \, , F_{1-\alpha/2,m-1,n-1}^{-1} s_X^2/s_Y^2 \right) .$$

Proposition 4.13. *Assume the population distributions are normal and the random samples are both independent of one another.*

Consider $H_0 : \sigma_X^2 = \sigma_Y^2$. *The test statistic is*

$$F = S_X^2/S_Y^2 \, .$$

For a hypothesis test at level α, *we use the following procedure:*

If $H_a : \sigma_X^2 > \sigma_Y^2$, *then P-value is* $A_R =$ *area under the* $\mathsf{F}(m-1, n-1)$ *curve to the right of* f.

If $H_a : \sigma_X^2 < \sigma_Y^2$, *then P-value is* $A_L =$ *area under the* $\mathsf{F}(m-1, n-1)$ *curve to the left of* f.

If $H_a : \sigma_X^2 \neq \sigma_Y^2$, *then P-value is* $2 \cdot \min(A_R, A_L)$.

Topic 5

Analysis of variance (ANOVA)

Analysis of variance, shortened as ANOVA, is a collection of statistical models and estimation procedures for analyzing the variation among different groups. In particular, a single-factor ANOVA provides a hypothesis test regarding the equality of two or more population means, thereby generalizing the one-sample and two-sample t tests considered in Sections 3.1.3 and 4.1.3.

5.1 Single factor ANOVA test

Suppose that we have k normally distributed populations[1] with different means $\mu_1, \dots, \mu_k$ and equal variances σ^2. We denote the rv for the jth measurement taken from the ith population by X_{ij} and the corresponding sample observation by x_{ij}. For samples of size $m_1, \dots, m_k$, we denote the sample means

$$\overline{X}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} X_{ij} \,,$$

and sample variances

$$S_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} (X_{ij} - \overline{X}_i)^2 \,,$$

for each $i = 1, \dots, k$; likewise, we denote the associated point estimates for the sample means $\overline{x}_1, \dots, \overline{x}_k$ and the sample variances $s_1^2, \dots, s_k^2$. The average over all observations $m = \sum m_i$, called the **grand mean**, is denoted by

$$\overline{X} = \frac{1}{m} \sum_{i=1}^{k} \sum_{j=1}^{m_i} X_{ij} \,.$$

The sample variances s_i^2, and hence the sample standard deviations, will generally vary even when the k populations share the same variance; a rule of thumb is that the equality of variances is reasonable if the largest s_i is not much more than two times the smallest.

We wish to test the equality of the population means, given by the null hypothesis,

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \,,$$

versus the alternative hypothesis,

$$H_a : \text{at least two } \mu_i \text{ differ} \,.$$

[1] In the context of ANOVA, these k populations are often referred to as **treatment distributions**.

Note that if $k = 3$ then H_0 is true only if all three means are the same, i.e., $\mu_1 = \mu_1 = \mu_3$, but there are a number of ways which the alternative might hold: $\mu_1 \neq \mu_2 = \mu_3$ or $\mu_1 = \mu_2 \neq \mu_3$ or $\mu_1 = \mu_3 \neq \mu_2$ or $\mu_1 \neq \mu_2 \neq \mu_3$.

The test procedure is based on comparing a measure of difference in variation among the sample means, i.e., the variation between x_i's, to a measure of variation within each sample.

Definition 5.1. The **mean square for treatments** is

$$\mathsf{MSTr} = \frac{1}{k-1}\sum_{i=1}^{k} m_i(\overline{X}_i - \overline{X})^2\,,$$

and the **mean square error** is

$$\mathsf{MSE} = \frac{1}{m-k}\sum_{i=1}^{k}(m_i - 1)S_i^2\,.$$

The MSTr and MSE are statistics that measure, respectively, the variation among sample means and the variation within samples. We will also use MSTr and MSE to denote calculated values of these statistics.

Proposition 5.1. *The test statistic*

$$F = \frac{\mathsf{MSTr}}{\mathsf{MSE}}$$

is the appropriate test statistic for the single-factor ANOVA problem involving k populations (or treatments) with a random sample of size $m_1, \ldots, m_k$ from each. When H_0 is true,

$$F \sim \mathsf{F}(\nu_1 = k-1, \nu_2 = m-k)\,.$$

In the present context a large test statistic value is more constradictory to H_0 than a smaller value, therefore the test is upper-tailed, i.e., consider the area F_α to the right of the critical value F_{α,ν_1,ν_2}. We reject H_0 if the value of the test statistic $F > F_\alpha$.

Example 5.1. Consider the average salary data from lcocal councils in Table 5.1. Is the expected average salary in each nation the same at the 5% level?

Table 5.1: **Average Salary Data** reported from 20 local councils.

Nation	Average salaries ('000 £)	Size (m_i)	Sample Mean $(\overline{x}_i)$	Sample SD (s_i)
England	17, 12, 18, 13, 15, 12	6	14.5	2.588
N Ireland	11, 7, 9, 13	4	10.0	2.582
Scotland	15, 10, 13, 14, 13	5	13.0	1.871
Wales	10, 12, 8, 7, 9	5	9.2	1.924

Table 5.1 presents the **Average Salary Data** (in thousands of pounds) reported from 20 local councils classified by nation (England, N Ireland, Scotland, and Wales). The sample means together with the sample standard deviations are summarized in the table as well as presented using using box plots in Figure 5.1.

For $\alpha = 0.05$, we compute the upper-tail area $F_{0.05}$ i.e. to the right of the critical value $F_{0.05,3,16}$ by consulting a statistical table or by using r to find $F_{0.05} = 3.2388715$.

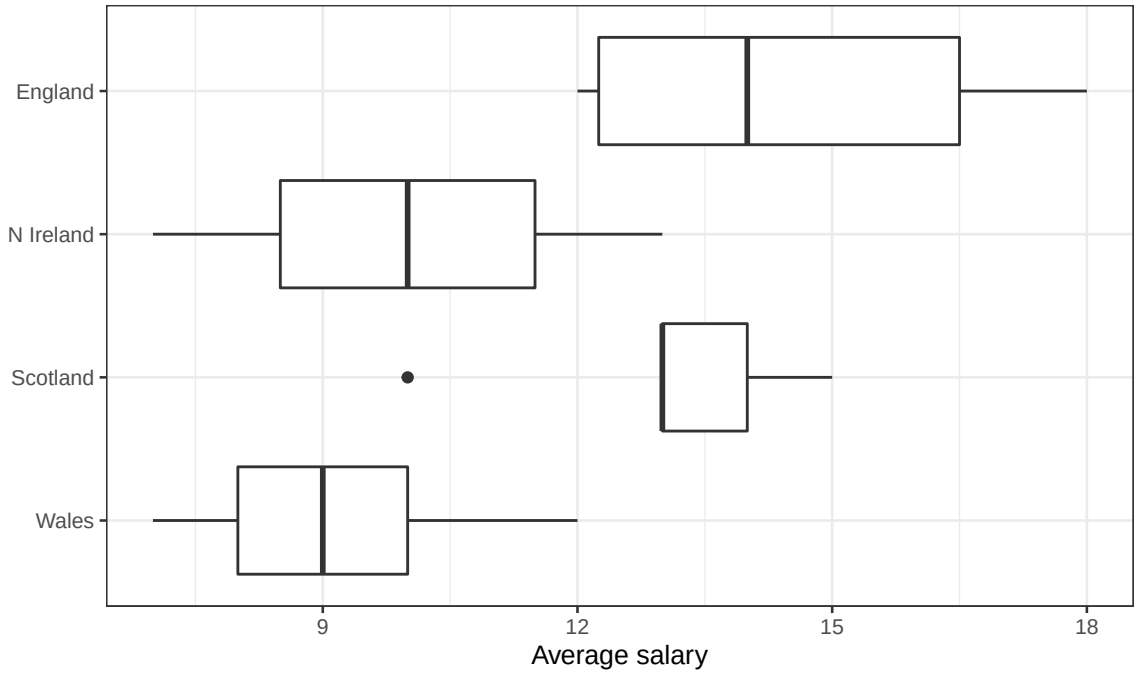

Figure 5.1: Box plots of the average mean salary data in Table 5.1 indicate five summary statistics: the median, two hinges (first and third quartiles) and two whiskers (extending from the hinge to the most extreme data point within $1.5 \cdot \mathsf{IQR}$).

```
# alt: qf(.05, df1 = 3, df2 = 16, lower.tail = FALSE)
qf(1-.05, df1 = 4-1, df2 = 20-4)
```

```
[1] 3.238872
```

The grand mean is

$$\overline{x} = \frac{17 + 12 + 18 + \cdots + 8 + 7 + 9}{20} = 11.9\,,$$

and hence the variation among sample means is given by,

$$\begin{aligned}\mathsf{MSTr} &= \frac{1}{4-1}\left(m_1(\overline{x}_1 - \overline{x})^2 + \cdots + m_4(\overline{x}_4 - \overline{x})^2\right) \\ &= \left(6(14.5 - 11.9)^2 + 4(10.0 - 11.9)^2 + 5(13.0 - 11.9)^2 + 5(9.2 - 11.9)^2\right)/3 \\ &= 32.5\,.\end{aligned}$$

The mean square error is

$$\begin{aligned}\mathsf{MSE} &= \frac{1}{20-4}\left((m_1 - 1)s_1^2 + \cdots (m_4 - 1)s_4^2\right) \\ &= \frac{5(2.588)^2 + 3(2.582)^2 + 4(1.871)^2 + 4(1.924)^2}{16} \\ &= 5.14366\end{aligned}$$

yielding the test statistic value

$$F = \frac{\mathsf{MSTr}}{\mathsf{MSE}} = \frac{32.5}{5.14366} = 6.3184581\,.$$

Since $F > F_\alpha$ we reject H_0. The data does not support they hypothesis that the mean salaries in each nation are identical at the 5% level. $\Diamond$

5.2 Confidence intervals

In Section 4.1 we gave a CI for comparing population means that involved the difference $\mu_X - \mu_Y$. In some settings, we would like to give CIs for more complicated functions of population means μ_i. Let

$$\theta = \sum_{i=1}^{k} c_i \mu_i \,,$$

for constants c_i. As we assume the $X_i j$ are normally distributed with $\mathbf{E}[X_{ij}] = \mu_i$ and $\text{Var}[X_{ij}] = \sigma^2$, the estimator

$$\hat{\theta} = \sum_{i=1}^{k} c_i \overline{X}_i \,,$$

is normally distributed with

$$\text{Var}[\hat{\theta}] - \sum_{i=1}^{k} c_i^2 \,\text{Var}[\overline{X}_i] = \sigma^2 \sum_{i=1}^{k} \frac{c_i}{m_i} \,.$$

Estimating σ^2 by the MSE and forming $\widehat{\sigma}_{\hat{\theta}}$ results in a t variable

$$\frac{\hat{\theta} - \theta}{\widehat{\sigma}_{\hat{\theta}}} \,.$$

Proposition 5.2. *A* $100(1-\alpha)\%$ *CI for* $\sum c_i \mu_i$ *is given by*

$$\sum_{i=1}^{k} c_i \overline{x}_i \pm t_{\alpha/2, m-k} \sqrt{\text{MSE} \sum_{i=1}^{k} \frac{c_i}{m_i}} \,.$$

Example 5.2. Determine a 90% CI for the difference in mean average salary for councils in Scotland and England, based on the data available in Table 5.1

For $\alpha = 0.10$, the critical value $t_{0.05,16} = 1.7458837$ is found by looking in a table of t critical values or by using `r`:

```
# alt: qt(0.1/2, 16, lower.tail = FALSE)
qt(1-0.1/2, df = 20 - 4)
```

```
[1] 1.745884
```

Then for the function $\overline{x}_2 - \overline{x_1}$,

$$\begin{aligned}
(\overline{x}_{Eng} - \overline{x}_{Sco}) \pm t_{0.05,16} \sqrt{\text{MSE}} \sqrt{\frac{1}{m_{Eng}} + \frac{1}{m_{Sco}}} \\
&= (14.5 - 13.0) \pm 1.7458837 \sqrt{5.14366} \sqrt{\frac{1}{6} + \frac{1}{5}} \\
&= 1.5 \pm 2.3976575 \,.
\end{aligned}$$

Thus a 90% confidence interval for $\mu_{Eng} - \mu_{Sco}$ is $(-0.8977\,, 3.898)$. $\Diamond$

How does the result in Example 5.2 compare to the t method in Section 4.1.3?

Topic 6

Linear regression

Regression analysis allows us to study the relationship among two or more rvs. Typically, we are interested in the relationship between a **response** or **dependent** rv Y and a **covariate** X.[1] The relationship between X and Y will be explained through a **regression function**,

$$r(x) = \mathbf{E}[Y \mid X = x] = \int y f(y \mid x) dy \,.$$

In particular, we shall assume that r is linear,

$$r(x) = \beta_0 + \beta_1 x \,, \tag{6.1}$$

and estimate the intercept β_0 and slope β_1 of this linear model from sample data

$$(Y_1, X_1), \ldots, (Y_m, X_m) \sim F_{Y,X} \,.$$

6.1 Simple linear regression models

The simplest regression is when X_i is one-dimensional and $r(x)$ is linear as in (6.1). A linear regression posits the expected value of Y_i is a linear function of the data X_i, but that Y deviates from its expected value by a random amount for fixed x_i.

Definition 6.1. The **simple linear regression model** relates a random response Y_i to a set of independent variables X_i,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \,, \tag{6.2}$$

where the intercept β_0 and slope β_1 are unknown parameters and the **random deviation** or **random error** ϵ_i is a rv assumed to satisfy:

1. $\mathbf{E}[\epsilon_i \mid X_i = x_i] = 0$,
2. $\mathrm{Var}[\epsilon_i \mid X_i = x_i] = \sigma^2$ does not depend on x_i,
3. ϵ_i and ϵ_j are independent for $i, j = 1, \ldots, m$.

From the assumptions on ϵ_i, the linear model (6.2) implies

$$\mathbf{E}[Y_i \mid X_i = x_i] = \beta_0 + \beta_1 x_i \,.$$

Thus, if $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators of β_0 and β_1, then the **fitted line** is

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

[1] The covariates X are also called **predictor variables**, **explanatory variables**, **independent variables**, and/or **features** depending on who you are talking to.

and the **predicted** or **fitted value** $\widehat{Y}_i = \hat{r}(X_i)$ is an estimator for $\mathbf{E}[Y_i \mid X_i = x_i]$. The **residuals** are defined to be

$$\hat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) . \tag{6.3}$$

The **residual sums of squares**,[2]

$$\mathsf{RSS} = \sum_{i=1}^{m} \hat{\epsilon}_i^2 , \tag{6.4}$$

measures how well the regression line $\hat{r}$ fits the data $(Y_1, X_1), \ldots, (Y_m, X_m)$. The **least squares estimates** of $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values that minimize the RSS in (6.4).

Theorem 6.1. *The **least squares estimates** for $\hat{\beta}_1$ and $\hat{\beta}_0$ are given by, respectively,*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{m}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{m}(X_i - \overline{X})^2} = \frac{S_{xy}}{S_{xx}} , \tag{6.5}$$

and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} . \tag{6.6}$$

Equation (6.4) is a function of $\hat{\beta}_0$ and $\hat{\beta}_1$ from the definition of the residuals (6.3). Then (6.5) and (6.6) follow by equating the partial derivatives of (6.4) to zero. The $\hat{\beta}_0$ and $\hat{\beta}_1$ are the unique solution to this linear system.

Example 6.1. In Figures 6.1 and 6.2, we consider the **Cherry Tree Data** (see Table 2.1 and discussion). We fit a least squares regression of the volume of timber (response variable) to the diameter of the tree (independent variable). As you would expect, the timber yield increases with diameter.

The `r` code below can be used to calculate the least squares regression and residuals.

```
data(trees)
y <- trees$Volume
x <- trees$Girth # NB: this is diameter; data mislabeled!
fit <- lm(y ~ x)
e <- resid(fit)
yhat <- predict(fit)
```

The `fit` data frame contains the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$:

```
fit$coefficients
```

```
(Intercept)           x
 -36.943459    5.065856
```

Both Figures 6.1 and 6.2 are scatter plots of the observed values y. In Figure 6.1, the regression line $\hat{y}$ is plotted along with the residuals $\hat{\epsilon}$. In Figure 6.2, the sample mean $\overline{y}$ is plotted together with the deviations $y - \overline{y}$. $\diamondsuit$

6.2 Estimating σ^2 for linear regressions

The parameter σ^2 (the variance of the random deviation) determines the variability in the regression model.

[2]The RSS is sometimes referred to as the **error sum of squares** and abbreviated SSE (no, the order is not a typo a typo).

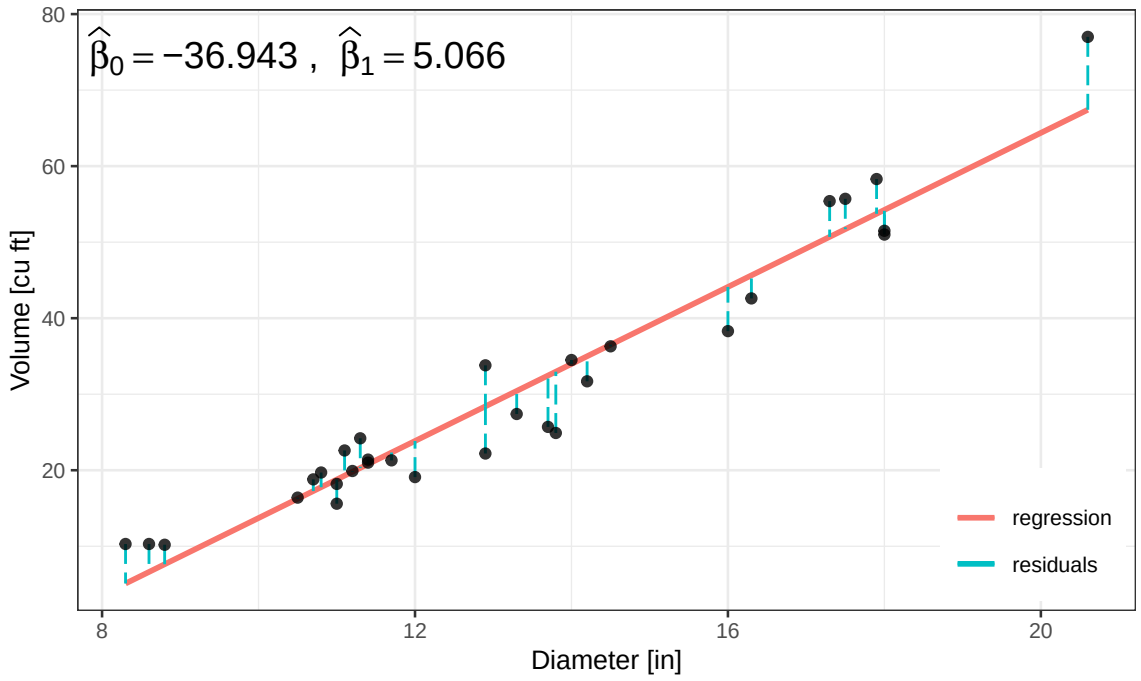

Figure 6.1: Linear regression (or least squares fit) of Volume to Diameter from the **Cherry Tree Data**. The vertical bars between the observed data point and the regression line indicate the error in the fit (the least squares residual). The residuals are squared and summed to yield the RSS (alt: SSE).

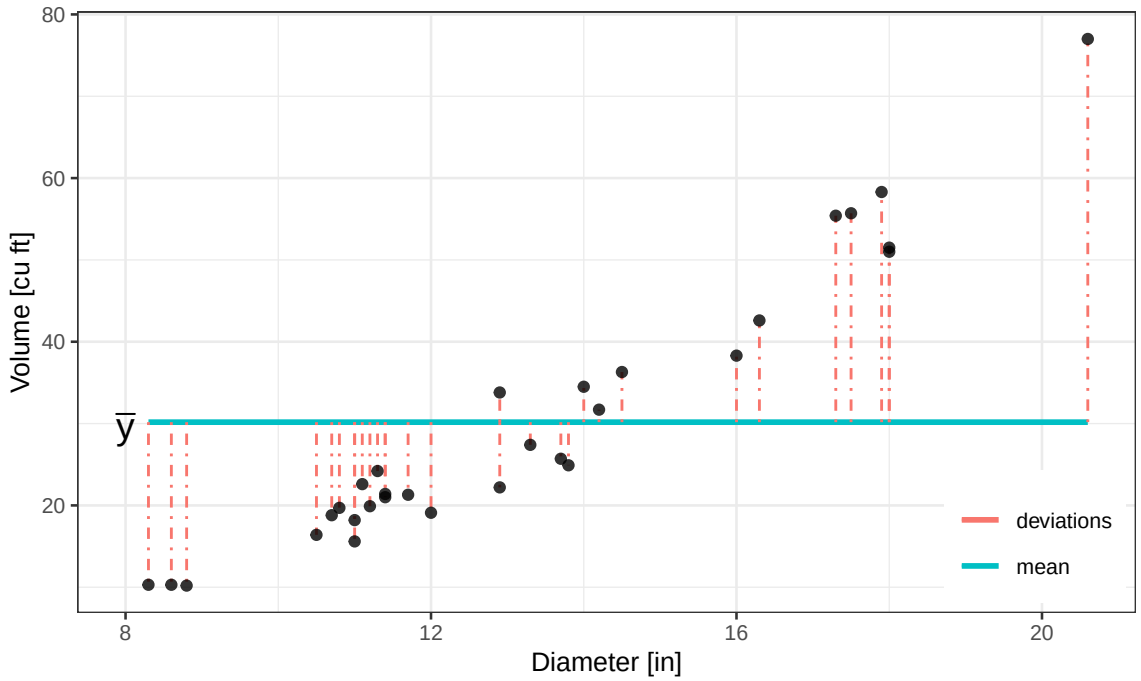

Figure 6.2: The deviations about the sample mean $\bar{y}$. The sum of the squared deviations or SST (total sum of squares) is a measure of the total variation in the observations.

Theorem 6.2. *An unbiased estimate of σ^2 is given by*

$$\widehat{\sigma}^2 = s^2 = \frac{\mathsf{RSS}}{m-2} = \frac{1}{m-2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2\,. \tag{6.7}$$

In Figure 6.3, we present a least squares regression of timber volume on both tree diameter and height (for the **Cherry Tree Data**). As expected, the regressions indicate the volume increases with both covariates. Estimates for the variance of the random deviation (6.7) in both regression models, σ_D^2 and σ_H^2, respectively, are computed to be $s_D^2 = 18.08$ and $s_H^2 = 179.48$. Thus, we see that small variances lead to observations of (x_i, y_i) that sit tightly around the regression line, in contrast to large variances that lead to a large cloud of points.

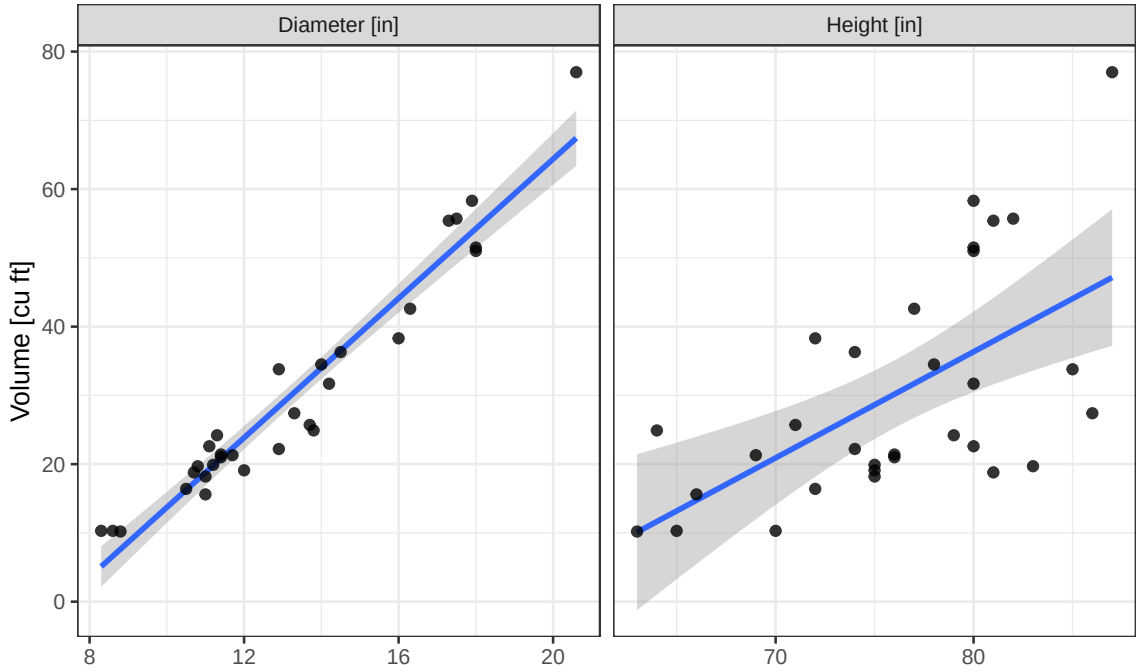

Figure 6.3: For the **Cherry Tree Data**, we estimate the variance to be $s_D^2 = 18.08$ (for Diameter) and $s_H^2 = 179.48$ (for Height); small variances lead to observations of (x_i, y_i) that sit tightly around the regression line, in contrast to large variances that lead to a large cloud of points.

In Theorem 6.2, the number in the denominator is the df associated with the RSS and s^2. To calculate RSS you must estimate two parameters β_0 and β_1 which results in the loss of two df. Hence the $m-2$.

We note for the purposes of making inferences, the statistic

$$S^2 = \frac{\mathsf{RSS}}{m-2}$$

is an unbiased estimator or σ^2 and the random variable

$$\frac{(m-2)S^2}{\sigma^2} \sim \chi^2(m-2)\,.$$

Moreover, the statistic S^2 is independent of both $\hat{\beta}_0$ and $\hat{\beta}_1$.

6.3 Inferences for least-squares parameters

If ϵ_i in (6.2) is assumed to be normally distributed, then we can derive the sampling distributions of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Hence, we can use these sampling distributions to make inferences about the parameters β_0 and β_1.

Provided iid $\epsilon_i \mid X_i \sim \mathsf{N}(0, \sigma^2)$, the least-squares estimators possess the following properties.

1. Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed.
2. Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, i.e., $\mathbf{E}[\hat{\beta}_i] = \beta_i$ for $i = 0, 1$.
3. $\mathrm{Var}[\hat{\beta}_0] = c_{00}\sigma^2$ where $c_{00} = \sum_{i=1}^{m} x_i^2/(mS_{xx})$.
4. $\mathrm{Var}[\hat{\beta}_1] = c_{11}\sigma^2$ where $c_{11} = 1/S_{xx}$.
5. $\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1] = c_{01}\sigma^2$ where $c_{01} = -\overline{x}/S_{xx}$.

These properties can be determined by working directly from (6.5) and (6.6).

Proposition 6.1. *Consider $H_0 : \beta_i = \beta_{i0}$. The test statistic is*

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}.$$

For a hypothesis test at level α, we use the following procedure:

If $H_a : \beta_i > \beta_{i0}$, then P-value is the area under $\mathsf{t}(m-2)$ to the right of t.

If $H_a : \beta_i < \beta_{i0}$, tthen P-value is the area under $\mathsf{t}(m-2)$ to the left of t.

If $H_a : \beta_i \neq \beta_{i0}$, then P-value is twice the area under $\mathsf{t}(m-2)$ to the right of $|t|$.

A confidence interval for β_i, based on the statistic (6.1), can be given following the procedures in 3.

Proposition 6.2. *A $100(1-\alpha)\%$ CI for β_i is given by*

$$\hat{\beta}_i \pm t_{\alpha/2,m-2} S\sqrt{c_{ii}}.$$

6.4 Correlation

Let $(X_1, Y_1), \ldots, (X_m, Y_m)$ denote a random sample from a bivariate normal distribution with $\mathbf{E}[X_i] = \mu_X$, $\mathbf{E}[Y_i] = \mu_Y$, $\mathrm{Var}[X_i] = \sigma_X^2$, $\mathrm{Var}[Y_i] = \sigma_Y^2$, and correlation coefficient ρ. The sample correlation coefficient is given by,

$$r = \frac{\sum_{i=1}^{m}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{m}(X_i - \overline{X})^2 \sum_{i=1}^{m}(Y_i - \overline{Y})^2}}, \tag{6.8}$$

which can be rewritten in terms of S_{xx}, S_{xy}, and S_{yy}:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}},$$

using (6.5) and we see that r and $\hat{\beta}_1$ have the same sign. A $|r|$ close to 1 means that the regression line is a good fit to the data and, similarly, an $|r|$ close to 0 means a poor fit to the data. Note that the correlation coefficient (and the least squares regression) are only suitable for describing *linear* relationships; a nonlinear relationship can also yield r near zero (see Figure 6.4).

Once a model is fit, it can be used to predict a value of y for a given x. However, the model only gives the most likely value of y; a corresponding **prediction interval** is usually more appropriate.

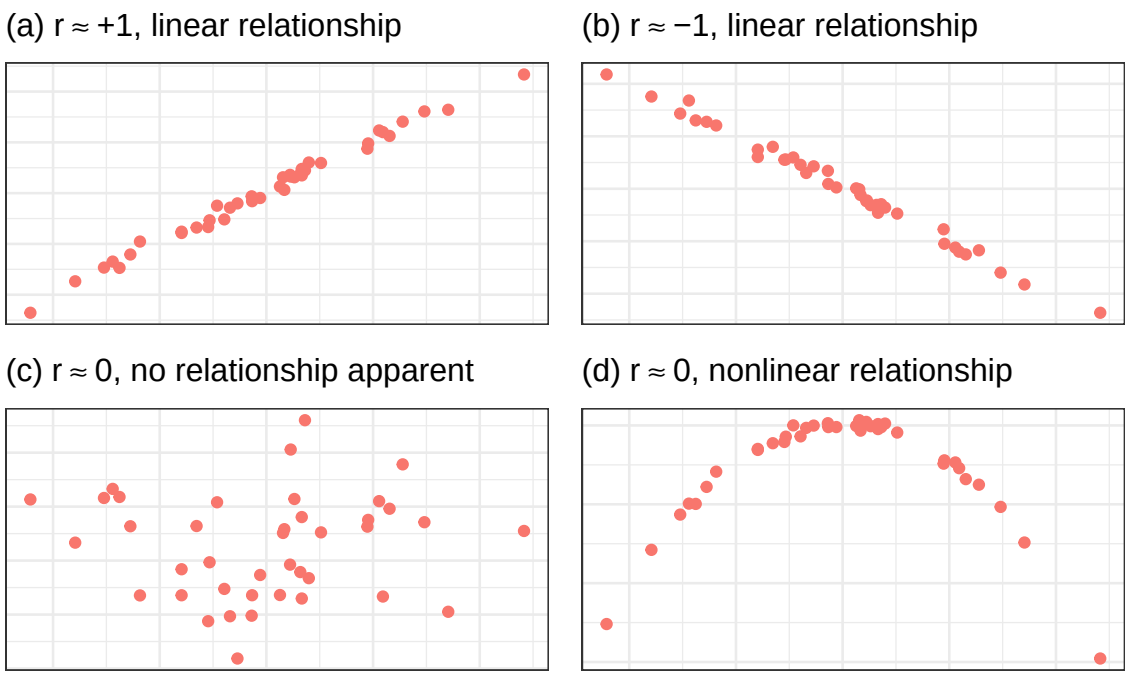

Figure 6.4: Correlations range from -1 to 1 with $|r| = 1$ indicating a strong linear relationship and r near zero indicating the absence of a linear relationship.

Proposition 6.3. *A* $100(1-\alpha)\%$ ***prediction interval*** *for an actual value of* Y *when* $x = x^*$ *is given by*

$$(\hat{\beta}_0 + \hat{\beta}_1 x^*) \pm t_{\alpha/2,m-2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}\,.$$

The prediction interval is not the same as the confidence interval for expected Y. Note that the length of the *confidence interval* for $\mathbf{E}[Y]$ when $x = x^*$ is given by

$$2 \cdot t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

whereas the length for the *prediction interval* of Y is

$$2 \cdot t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}\,.$$

Thus the prediction intervals for an actual value of Y are longer than the confidence intervals for $\mathbf{E}[Y]$ if both are determined for the same value x^*.

The linear model

$$\mathbf{E}[Y \mid X = x] = \beta_0 + \beta_1 x\,,$$

assumes that the conditional expectation of Y for a fixed value of X is a linear function of the x value. If we assume that (X, Y) has a bivariate normal distribution, then

$$\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho\,,$$

and thus, for the simple hypothesis tests we have considered (Table 2.2), statistical tests for β and ρ are equivalent.

Topic 7

Categorical data

7.1 Multinomial experiments

Suppose we have a population that is divided into $k > 2$ distinct categories. We consider an experiment where we select m individuals (or objects) from the population and categorize each one. We denote the proportion of the population in the ith category by p_i. If the sample size m is much smaller than the population size M (so that the m trials are independent), this experiment will be approximately *multinomial* with success probability p_i for each category, $i = 1, \dots, k$.

Before the experiment is performed, we denote the number (or count) of the trials resulting in category i by the rv N_i. The expected number of trails that result in category i is given by

$$\mathbf{E}[N_i] = mp_i\,, \quad i = 1, \dots, k\,. \tag{7.1}$$

After the experiment is performed, we denote the corresponding observed value by n_i. Since the trials result in distinct categories,

$$\sum_{i=1}^{k} N_i = \sum_{i=1}^{k} n_i = m\,,$$

which indicates that, for a given m, we only need to observe $k - 1$ of the variables to be able to work out what the kth variable should be.

7.2 Goodness-of-fit for a single factor

We are interested in making inferences about the proportion parameters p_i. Specifically, we will consider the null hypothesis,

$$H_0 : p_1 = p_{10}\,, p_2 = p_{20}\,, \cdots\,, p_k = p_{k0}\,, \tag{7.2}$$

that completely specifies a value p_{i0} for each p_i.[1] The alternative hypothesis H_a will state that H_0 is not true, i.e., that at least one p_i is different from the value p_{i0} claimed under the null H_0.

Provided the null hypothesis in (7.2) is true, the expected values (7.1) can be written in terms of the expected frequencies,

$$\mathbf{E}[N_i] = mp_{i0}\,, \quad i = 1, \dots, k\,.$$

Often the n_i, referred to as the **observed cell counts**, and the corresponding mp_{i0}, referred to as the **expected cell counts**, are tabulated, for example, as in Table 7.1.

[1] Here for $i = 1, \dots, k$ we use the notation p_{i0} to denote the value of p_i claimed under the null hypothesis.

Table 7.1: Observed and expected cell counts.

Category	$i=1$	$i=2$	$\cdots$	$i=k$	Row total
Observed	n_1	n_2	$\cdots$	n_k	m
Expected	mp_{10}	mp_{20}	$\cdots$	mp_{k0}	m

The test procedure assesses the discrepancy between the value of the observed and expected cell counts. This discrepancy, or **goodness of fit**, is measured by the squared deviations divided by the expected count.[2]

Theorem 7.1. *For $mp_i \geq 5$ for $i = 1, \ldots, k$, the rv*

$$V = \sum_{i=1}^{k} \frac{(N_i - mp_i)^2}{mp_i} \quad \sim \chi^2(k-1)\,,$$

that is, V has approximately a χ^2 distribution with $\nu = k - 1$ df.

Proposition 7.1. *Consider the null*

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \cdots, p_k = p_{k0}\,,$$

and the alternative

$$H_a : p_i \neq p_{i0} \text{ for at least one } i\,.$$

The test statistic is

$$V = \sum_{i=1}^{k} \frac{(N_i - mp_{i0})^2}{mp_{i0}}\,.$$

As a rule of thumb, provided $mp_{i0} \geq 5$ for all $i = 1, \ldots, k$, then the P-value is the area under $\chi^2(k-1)$ to the right of v.

If $mp_{i0} < 5$ for some i then it may be possible to combine the categories such that the new categorizations satisfy the assumptions of Proposition 7.1.

Things are much more complicated if the category probabilities are not completely specified.

7.3 Test for independence of factors

In Section 7.2 we considered the categorization of a population into a single factor. We now consider a single population where each individual is categorized into two factors with I distinct categories for the first factor and J distinct categories for the second factor. That is, each individual from the population belongs to exactly one of the I categories of the first factor and exactly one of the J categories of the second factor. We want to determine whether or not there is any dependency between the two factors.

For a sample of m individuals, we denote by n_{ij} the count of the m samples that fall both in category i of the first factor and category j of the second factor, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. A **contingency (data) table** with I rows and J columns (i.e., IJ cells) will be used to record the n_{ij} counts (in the obvious way).[3] Let p_{ij} be the proportion of individuals in the population who belong in category i of factor 1

[2]The division by the expected cell counts is to account for possible differences in the relative magnitude of the observed/expected counts.

[3]Contingency is another word for dependency in this context.

and category j of factor 2. Then, the probability that a randomly selected individual falls in category i of factor 1 is found by summing over all j:

$$p_i = \sum_{j=1}^{J} p_{ij} \,,$$

and likewise, the probability that a randomly selected individual falls in category j of factor 2 is found by summing over all i:

$$p_j = \sum_{i=1}^{I} p_{ij} \,.$$

The null hypothesis that we will be interested in adopting is

$$H_0 \,:\, p_{ij} = p_i \cdot p_j \;\forall (i,j) \,,$$

that is, that an individual's category in factor 1 is independent of the category in factor 2.

Following the same program as for the single category goodness-of-fit test, we note that, assuming the null hypothesis (7.3) is true, then the expected count in cell i, j is

$$\mathbf{E}[N_{ij}] = mp_{ij} = mp_i p_j \,;$$

and we estimate p_i and p_j by the appropriate sample proportion:

$$\hat{p}_i = \frac{n_i}{m} \,, \qquad n_i = \sum_j n_{ij} \quad \text{(row totals)} \,,$$

and

$$\hat{p}_j = \frac{n_j}{m} \,, \qquad n_j = \sum_i n_{ij} \quad \text{(column totals)} \,.$$

Thus, the expected cell count is given by

$$\hat{e}_{ij} = m\hat{p}_i\hat{p}_j = \frac{n_i n_j}{m} \,,$$

and we assess the goodness of fit between the observed cell count $n_i j$ and the expected cell count $\hat{e}_i j$.

Proposition 7.2. *Assume the null hypothesis*

$$H_0 \,:\, p_{ij} = p_i p_j \text{ for all } i = 1, \ldots, I \,, j = 1, \ldots, J \,,$$

against the alternative hypothesis

$$H_a \,:\, H_0 \text{ is not true} \,.$$

The test statistic is

$$V = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(N_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \,.$$

As a rule of thumb, provided $\hat{e}_{ij} \geq 5$ for all i, j and when H_0 is true, then the test statistic has approximately a $\chi^2(\nu)$ distribution with $\nu = (I-1)(J-1)$ df. For a hypothesis test at level α, the procedure is upper-tailed and the P-value is the area under $\chi^2(\nu)$ to the right of v.

Topic 8

Quality Control

Quality control is an area of applies statistics that seeks to make interventions to maintain or improve the outcome of industrial processes. Random variations in output processes might negatively impact the quality of a product. We would like to identify the sources of random variations in output processes that might have *assignable causes*. **Control charts** are a tool that help us to recognize when industrial processes are no longer controlled so that one might then seek to identify assignable causes.

The basic elements of control charting involve specifying a control region and then analyzing time-series data. We will specify a base-line value along with an upper and lower limit of control and assume that a process is under control unless a test statistic suggests otherwise.[1] To construct a control chart, one collects data about a process at fixed points of time and calculates the running value of quality statistic. If the quality statistic exceeds the upper or lower control limits, the process is deemed to out of control and the quality of the product is assumed to be negatively impacted.

The process of creating a control chart is best illustrated through an extended example, like Example 8.1 provided below.

Example 8.1. Here we consider the typical 3σ control charting for a process mean $\overline{X}$ based on estimated parameters. That is, we assume the generating process X is normally distributed with unknown parameters μ and σ^2. We seek to estimate the mean $\overline{X}$. Our control region is specified to be three standard deviations; that is, the process is in-control if it remains within three standard devaitions of a baseline value.

The **Beer Data** contains measurements of the features OG, ABV, pH, and IBU for 50 batches of each of three types of product (Premium Lager, IPA, and Light Lager). We are interested the IPA's pH value, which influences saccharification. We assume that 3 batches of IPA are produced per day are we prepare the data as follows.

```
ipa <- beer %>%
 select(Batch_Id, pH, Beer) %>%
 filter(Beer == "IPA") %>%
 rename(Day = Batch_Id)
ipa$Day[1:48] <- rep(1:16, each = 3)
ipa <- ipa[1:48,]
m <- 3    # three batches per day
k <- 16   # number of days
```

We first observe that the pH measurements are (at least approximately) normal, as can be seen from the

[1] The default position here will be reminiscent of hypothesis testing.

quantile-quantile plot in Figure 8.1.

```
ggplot(ipa, aes(sample = pH)) + stat_qq() + stat_qq_line()
```

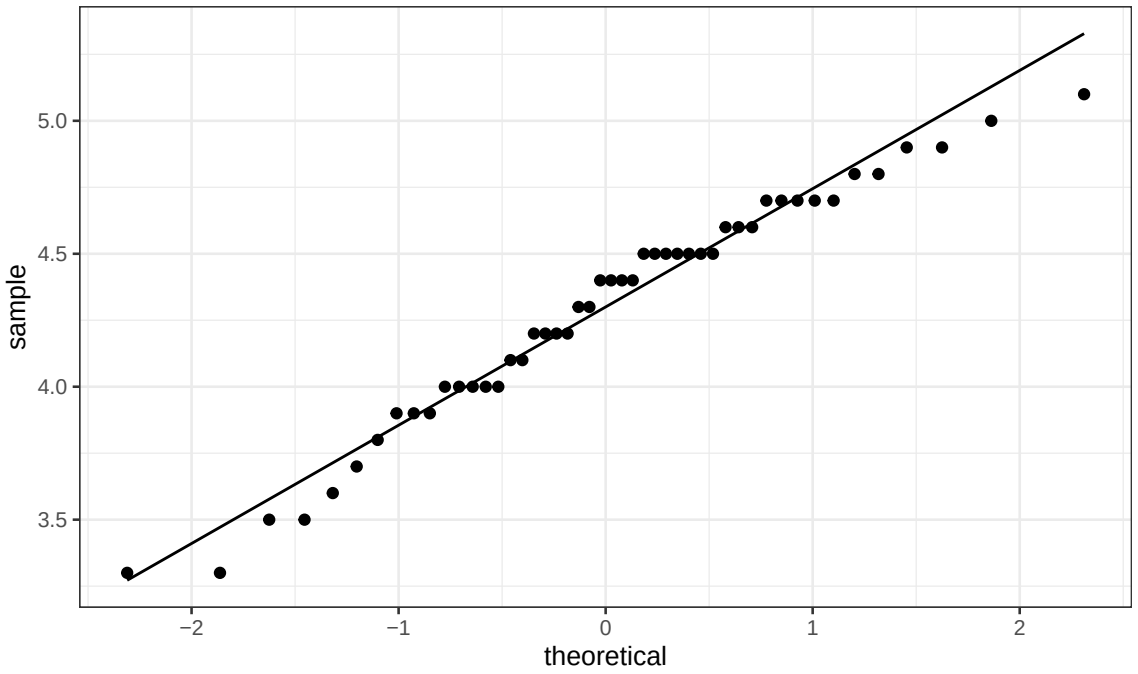

Figure 8.1: Normal quantile-quantile plot of observed pH measurements of the IPA batches in Table 8.1.

We consider the data for pH readings from three batches of IPA taken over sixteen days ($k = 16$) presented in Table 8.1. The Table includes the sample mean per day $\overline{x}$, the sample standard deviation s, and the range of values $\max x_i - \min x_i$ per day (each based on $m = 3$ batches).

```
ipa_stat <- ipa %>%
 group_by(Day) %>%
 summarise(observations = list(pH), mean = signif(mean(pH), digits = 4),
           sd = signif(sd(pH), digits = 4), range = max(pH) - min(pH))
kbl(ipa_stat, align = "llccc",
      caption = "Observations and summary statistics for the **Beer Data**.",
      col.names = c('Day', "pH Observations", '$\\overline{x}$', '$s$', 'Range'),
    booktabs = T) %>%
   kable_styling(latex_options = c("striped", "hold_position"))
```

We estimate of the mean

$$\widehat{\mu} = \frac{1}{k}\sum_{i=1}^{k} \overline{x}_i\,,$$

by averaging the means found for the k days and, similarly, estimate the mean of the sample standard deviation,

$$\overline{s} = \frac{1}{k}\sum_{i=1}^{k} s_i\,,$$

by averaging the sample standard deviations for the k days. It can be shown that

$$\widehat{\sigma} = \frac{\overline{S}}{a_m}$$

Table 8.1: Observations and summary statistics for the **Beer Data**.

Day	pH Observations	$\overline{x}$	s	Range
1	4.7, 4.5, 4.9	4.700	0.20000	0.4
2	4.0, 4.6, 4.5	4.367	0.32150	0.6
3	4.7, 3.3, 4.6	4.200	0.78100	1.4
4	3.9, 3.5, 4.2	3.867	0.35120	0.7
5	4.0, 4.7, 3.6	4.100	0.55680	1.1
6	4.4, 4.5, 4.1	4.333	0.20820	0.4
7	4.5, 3.9, 4.8	4.400	0.45830	0.9
8	4.0, 4.9, 4.7	4.533	0.47260	0.9
9	4.3, 4.4, 4.8	4.500	0.26460	0.5
10	5.0, 4.5, 3.5	4.333	0.76380	1.5
11	3.8, 3.7, 3.9	3.800	0.10000	0.2
12	5.1, 4.5, 4.5	4.700	0.34640	0.6
13	4.7, 4.4, 4.1	4.400	0.30000	0.6
14	4.0, 4.4, 4.6	4.333	0.30550	0.6
15	4.0, 3.3, 4.2	3.833	0.47260	0.9
16	4.2, 4.2, 4.3	4.233	0.05774	0.1

is an unbiased estimator of σ where

$$a_m = \frac{\sqrt{2}\Gamma(m/2)}{\sqrt{m-1}\Gamma\left((n-1)/2\right)}.$$

Thus, we compute the 3σ upper and lower control limits, respectively,

$$\text{UCL} = \hat{\mu} + 3\frac{\bar{s}}{a_m\sqrt{m}}$$

and

$$\text{LCL} = \hat{\mu} - 3\frac{\bar{s}}{a_m\sqrt{m}}.$$

The computations in `r` follow, along with the resulting **control chart** in Figure 8.2.

```
a <- function(m){ sqrt(2) * gamma(m/2) / (sqrt(m-1) * gamma((m-1)/2)) }
muhat = sum(ipa_stat$mean) / k
sbar = sum(ipa_stat$sd) / k
lcl = muhat - 3*sbar / (a(m) * sqrt(m))
ucl = muhat + 3*sbar / (a(m) * sqrt(m))

ggplot(ipa_stat, aes(x = Day)) + geom_point(aes(y = mean)) +
 geom_hline(aes(yintercept = muhat, color = "Mean"), size = 1) +
 geom_hline(aes(yintercept = lcl, color = "LCL"), size = 1.5) +
 geom_hline(aes(yintercept = ucl, color = "UCL"), size = 2) + ylab("pH") +
   theme(legend.justification = c(1,1), legend.position = c(1,1),
         legend.title = element_blank(),
         legend.box.margin = margin(c(4, 4, 4, 4), unit = "pt"))
```

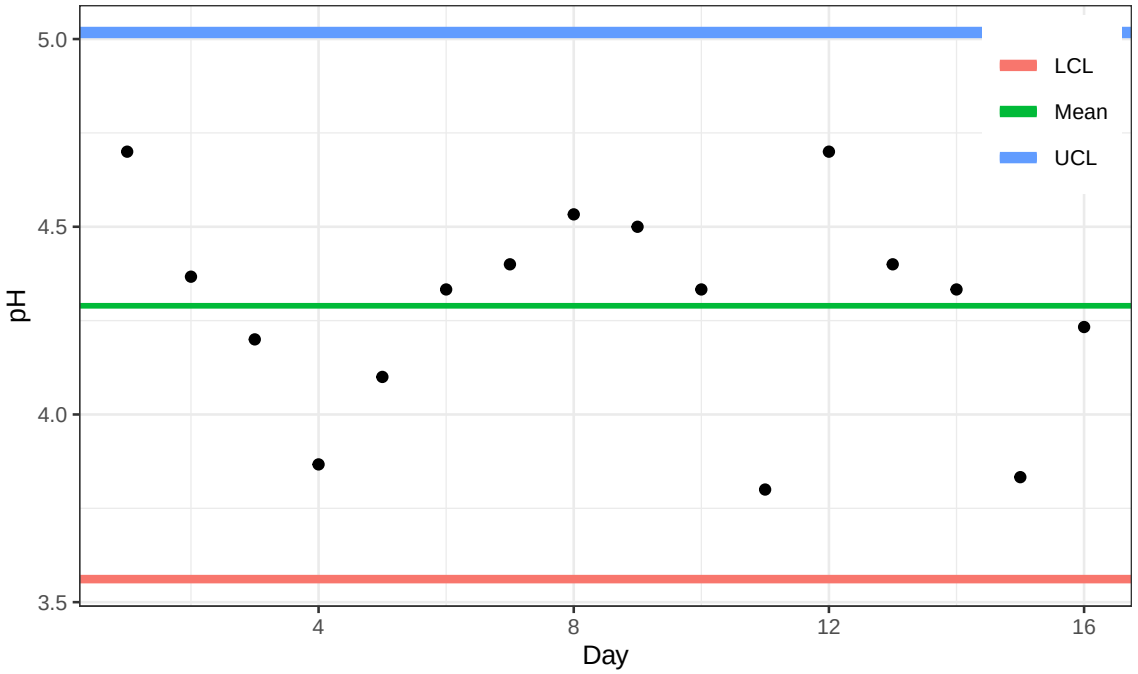

Figure 8.2: The 3σ **control chart** illustrates that with respect to pH the brewing process is in-control over the selected timeframe as the observations fall within the $(\mathsf{LCL}, \mathsf{UCL})$ control interval.

From Figure 8.2, we observe for each day the process is in-control as the observed mean pH values fall within the control limits $(\mathsf{LCL}, \mathsf{UCL})$. If this were not the case, then our initial assumption, that process is in-control, would be violated. The violation of the assumption would require that we seek to identify an assignable cause for the variation. If a cause could be identified, then the we would need to recompute our control limits with the observations that were out of control removed. $\Diamond$

References

M DeGroot and M Schervish. *Probability and Statistics*. Addison-Wesley, 3rd edition, 2001.

Jay L Devore. *Probability and Statistics for Engineering and the Sciences*. Cengage Learning, 9th edition, 2016.

John A Rice. *Mathematical statistics and data analysis*. Cengage Learning, 2006.

Gerald van Belle. *Statistical rules of thumb*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2008. ISBN 978-0-470-14448-0. doi: 10.1002/9780470377963. URL https://mathscinet.ams.org/mathscinet-getitem?mr=2437790.

Larry Wasserman. *All of Statistics*. Springer-Verlag, New York, 2004. doi: 10.1007/978-0-387-21736-9.