MA22004 - Statistics and Probability II

Dr Eric Hall

Last updated: 2020-09-10

Contents

Course Documents

Welcome

Welcome to MA22004.

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.

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

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.

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.

> You do not need to purchase these books.

- Devore, *Probability and Statistics for Engineering and the Sciences,* Cengage learning, 2011. [Devore, 2016, §6-10 + 12]
- DeGroot and Schervish, *Probability and Statistics,* Addison-Wesley, 2001. [DeGroot and Schervish, 2001, §7-10]
- Rice, *Mathematical statistics and data analysis,* Cengage Learning, 2006. [Rice, 2006, §6-12]
- Wasserman, *All of Statistics,* Springer-Verlag, 2004. [Wasserman, 2004, Concise general reference]

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.

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

Coursework

Assessed coursework includes:

- six hand-in laboratory reports and
- 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

Chapter 1

Special distributions

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 $X \sim \mathcal{N}(\mu, \sigma^2)$.

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

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

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

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

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

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

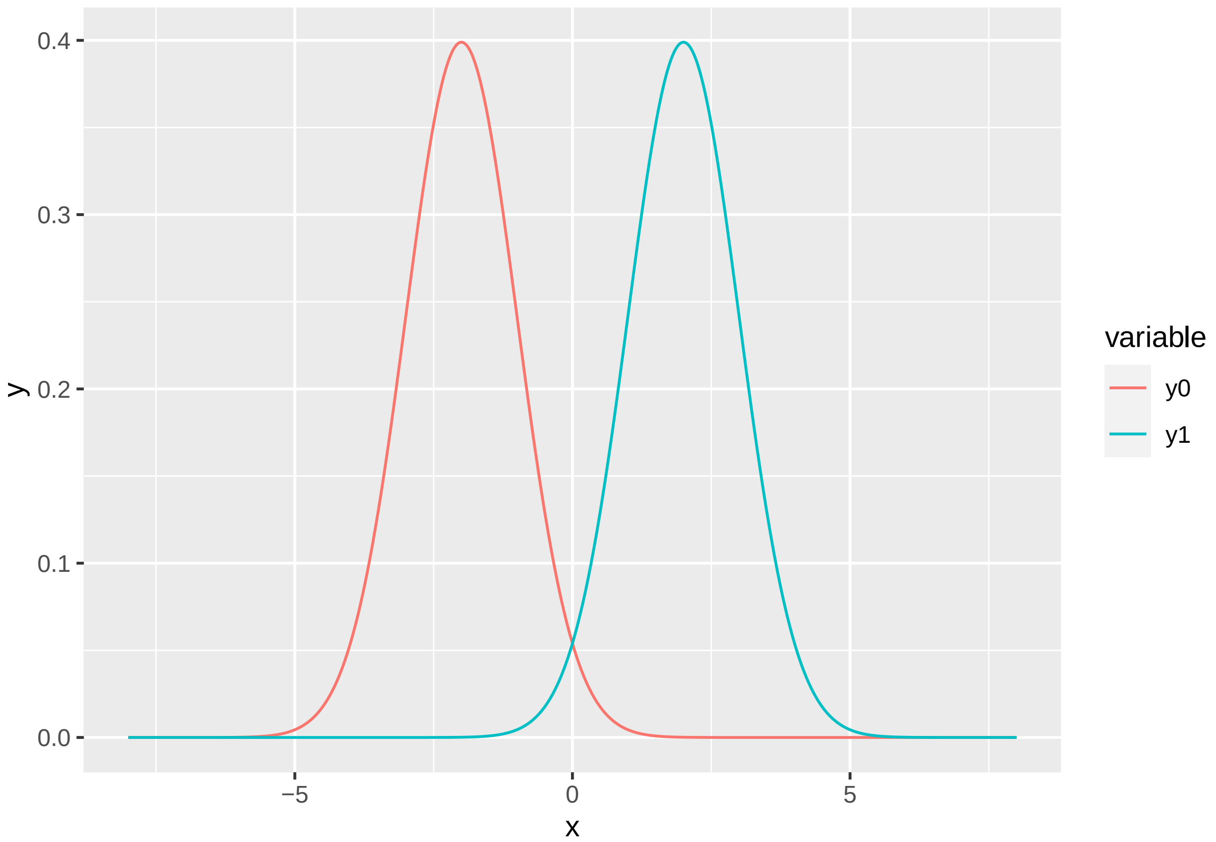

Figure 1.1: The pdfs of two normal rvs with different means and the same standard deviations.

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

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

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

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

Alternatively, we can use the `r` code:

```
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 $X \sim \mathcal{N}(5, 9)$ and find $P(4 \leq X \leq 5.25)$.

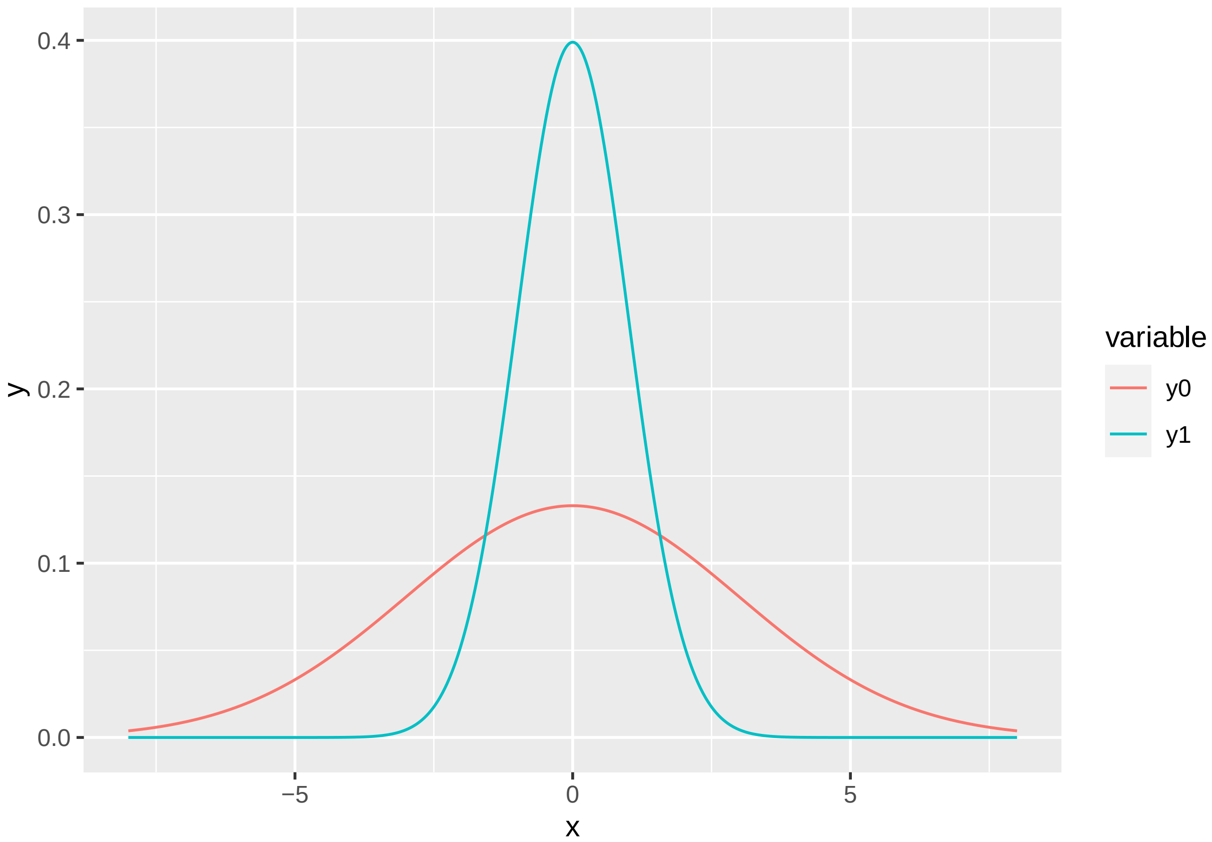

Figure 1.2: The pdfs of two normal rvs with the same means and different standard deviations.

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

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

TODO: plot area under normal curve (interior)

Alternatively, we can use the `r` code:

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

```
[1] 0.1637654
```

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 $X \sim t(\nu)$.

1.2.1 Properties of t distributions

1. The density for $t(\nu)$ is a bell-shaped curve centered at 0.
2. The density for $t(\nu)$ 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 $t(\nu)$ density converges to the standard normal density (i.e., the spread of the $t(\nu)$ density decreases relative to the standard normal).

If $X \sim t(\nu)$, 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 χ^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 rf X has an F **distribution** with parameters ν_1 and ν_2

Chapter 2

Inferences based on a single sample

We discuss the basics of point estimation, confidence intervals, and hypothesis testing for making inferences about a population based on a single sample in Sections 2.1, 2.2, and 2.3, respectively. In particular, we provide details about estimating population means (μ) in Section 2.4, population proportions (p) in Section 2.5, and population variances (σ^2) in Section 2.6.

2.1 Point estimation

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.

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 n independent *setosa* flowers are denoted by rvs $X_1, X_2, \dots, X_n$. Any function of the X_i's, such as the sample mean,

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

or the sample variance,

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

is also a rv.

Suppose we actually find and measure the petal length of 50 independent *setosa* flowers resulting in observations $x_1, x_2, \dots, x_{50}$; 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 $n = 50$ 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. Let $X_1, X_2, \dots, X_n$ be iid samples from a distribution $F(\theta)$. A **point estimator** $\hat{\theta}_n$ of a parameter θ is obtained by selecting a suitable statistic g,

$$\hat{\theta}_n = g(X_1, \dots, X_n)\,.$$

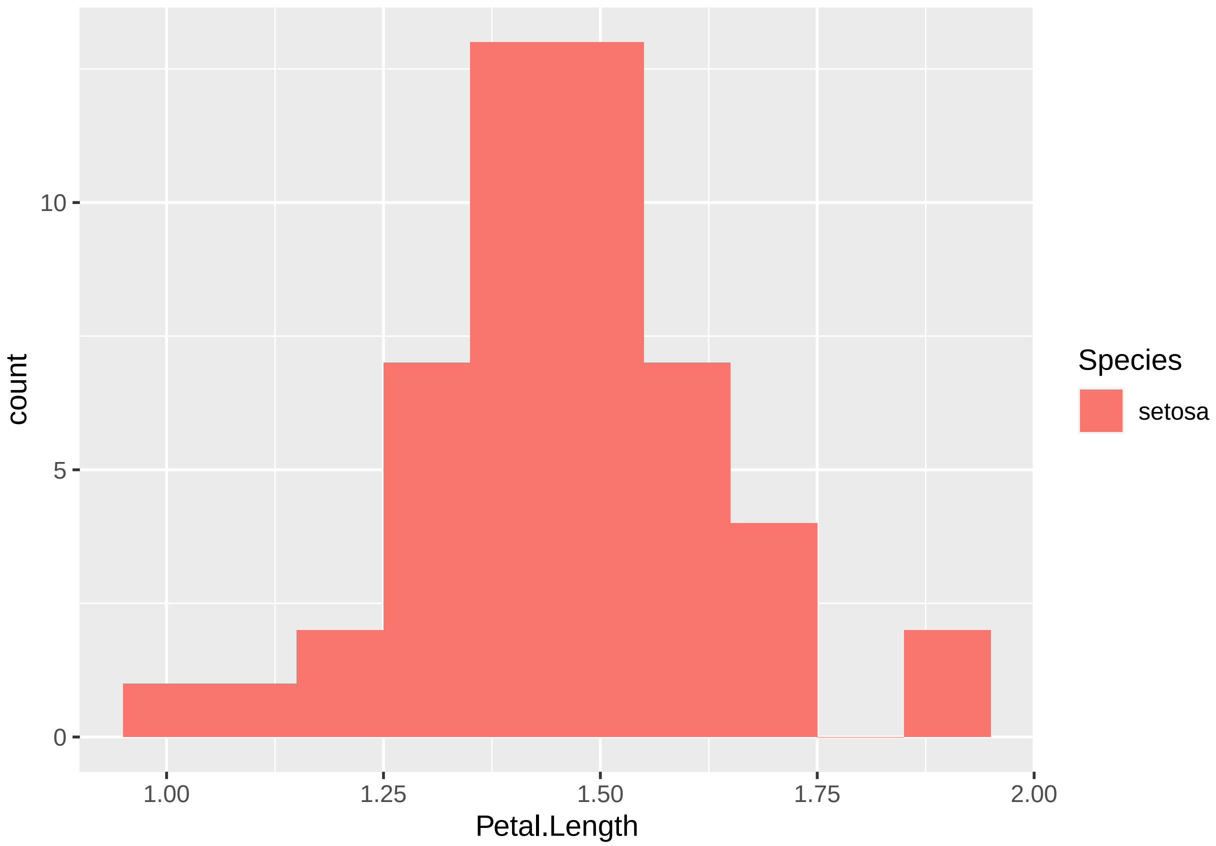

Figure 2.1: The distribution (counts) of 50 *setosa* petal length measurments.

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

> The symbol $\hat{\theta}_n$ (or simply $\hat{\theta}$ when the sample size n 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.

Note that Definition 2.1 does not say how to select the appropriate statistic. For the *setosa* example, the sample mean $\overline{X}$ is suggested as a good estimator of the population mean μ. That is, $\hat{\mu} = \overline{X}$ 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 $\sigma^2_{\overline{X}} = \sigma^2/n$, i.e.,

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

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

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

Example 2.1. Let us consider the heights (measured in inches) of 31 black cherry trees (sorted, for your enjoyment): `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 comparing this data to a normal distribution is fairly straight, so we assume that the distribution of black cherry tree heights is normal with a mean value μ; i.e., that the population of heights is distributed $\mathcal{N}(\mu, \sigma^2)$ where μ is a parameter to be estimated. The observations $X_1, \ldots, X_{31}$ are then assumed to be a random sample from this normal distribution (iid). Consider the following three different stimators and the resulting point estimates for μ based on the 31 samples.

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) $\overline{X}_{\mathrm{tr}(10)}$ and estimate $\overline{x}_{\mathrm{tr}(10)} = (2356 - 63 - 64 - 65 - 87 - 86 - 85)/25 = 76.24$.

Each estimator above uses a different notion of center for the sample data. 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?

> 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 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).

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 $\sigma_{\hat{\theta}} = \sqrt{\mathrm{Var}(\hat{\theta})}$ (sometimes denoted $\mathsf{se} = \mathsf{se}(\hat{\theta})$). Often, the standard error depends on unknown parameters and must also be estimated. The **estimated standard error** is denoted by $\hat{\sigma}_{\hat{\theta}}$ or $s_{\hat{\theta}}$ or $\widehat{\mathsf{se}}$.

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. 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 $C_n = (L_n, U_n)$ where $L_n = \ell(X_1, \dots, X_n)$ and $U_n = u(X_1, \dots, X_n)$ are functions of the data such that

$$P_\theta(L_n < \theta < U_n) = 1 - \alpha\,, \tag{2.2}$$

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_n 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 (see).

TODO: fix above plot EG devore fig 7.3 that illustrates ca. 50 95% CIs (with asterisks identifying intervals that do not include μ).

2.3 Hypothesis testing

In Sections 2.1 and 2.2 we 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.

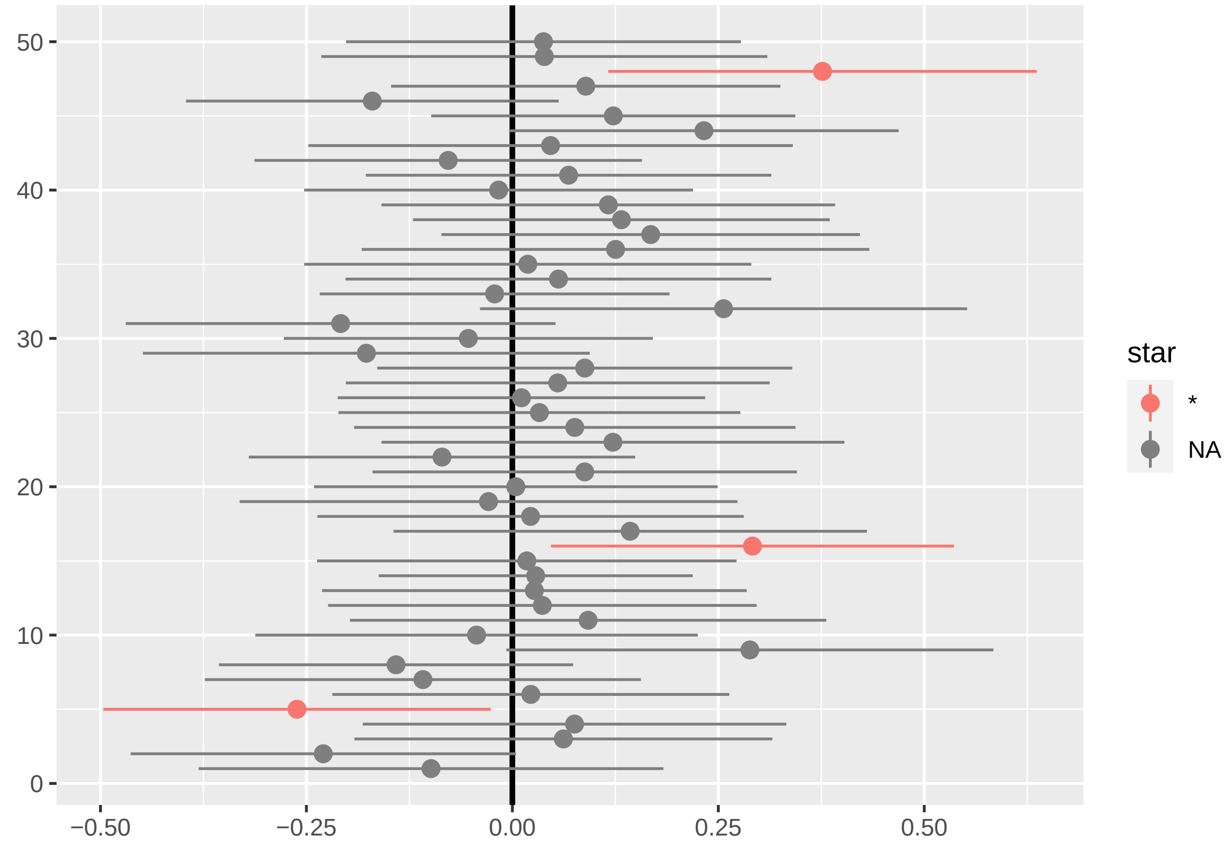

Figure 2.2: Fifty 95% CIs for a population mean μ. After a sample is taken, the computed interval estimate either contains μ or it does not (asteriks identify intervals that do not include μ). When drawing such a large number of 95% CIs, we would anticipate that approximately 5% (ca. 2.5) would fail to cover the true parameter μ.

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.

For a statistical hypothesis regarding the *equality* of a parameter θ with a fixed quantity θ_0, the null and alternative hypotheses will take one of the following forms.

Null hypothesis	Alternative hypothesis
$H_0 : \theta \leq \theta_0$	$H_a : \theta > \theta_0$
$H_0 : \theta \geq \theta_0$	$H_a : \theta < \theta_0$
$H_0 : \theta = \theta_0$	$H_a : \theta \neq \theta_0$

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.[1]

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** is a function of the sample data (like an estimator). The statistical decision to reject or fail to reject the null hypothesis will involve computing the test statistic. The **rejection region** are the values of the test statistic for which the null hypothesis is to be rejected in favor of the alternative. That is, we compute the test statistic based on a given sample; the test statistic either falls in the rejection region—in which case we reject the null H_0—or it does not fall in the rejection region—in which case we fail to reject the null H_0.

Example 2.2. **TODO**: example hypothesis test

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

Example 2.3. **TODO**: example hypothesis error types

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

2.4 Estimating means

If the parameter of interest is the population mean $\theta = \mu$, then the sample mean estimator $\hat{\theta} = \overline{X}$ in (2.1) has (at least approximately) a normal distribution for sufficiently large n. We will consider three cases where the form of the confidence interval can be derived using the approximate normality of the sample mean:

1. CI for μ of a normal population with known σ^2,
2. CI for μ of any population with unknown σ^2, when the sample size n is large,
3. CI for μ of a normal population with unknown σ^2, when the sample size n is small.

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

$$\text{point estimate } \mu \pm (\text{critical value of reference dist.}) \cdot (\text{precision of point estimate}) , \tag{2.3}$$

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 two-sided (symmetric) tail areas under the reference distribution.

[1] We comment that *fail to reject* H_0 is sometimes phrased *retain* H_0 or (perhaps less accurately) *accept* H_0.

2.4.1 CI for mean of a Normal population with known variance

Definition 2.7. 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

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) , \tag{2.4}$$

or $\overline{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$.

The CI for the mean (2.9) can be expressed 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. Below we provide a table containing commonly used normal critical values (note: indexed by $\alpha/2$).

\$\alpha/2 =\$ single tail area	central area \$= 1 − \alpha\$	\$z_{\alpha/2}\$
\$0.10\$	\$0.80\$	\$z_{.10} = 1.28\$
\$0.05\$	\$0.90\$	\$z_{.05} = 1.645\$
\$0.025\$	\$0.95\$	\$z_{.025} = 1.96\$
\$0.01\$	\$0.98\$	\$z_{.01} = 2.33\$
\$0.005\$	\$0.99\$	\$z_{.005} = 2.58\$

TODO: add example

As noted in (2.3) and (2.9), 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 n necessary to achieve an interval width w is obtained at confidence level α is obtained by equating w to $2z_{\alpha/2} \cdot \sigma/\sqrt{n}$ and then solving for n.

Proposition 2.1. *The sample size n required to achieve a CI for μ with width w at level α is given by,*

$$n = \left(2z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2 . \tag{2.5}$$

From Proposition 2.1, we see that the smaller the desired w then the larger n must be (and subsequently, the more effort that must be allocated to data collection).

TODO: add example of sample size calculation

2.4.2 Large-sample CI for mean of a population with unknown variance

Consider samples $X_1, \ldots, X_n$ from a population with mean μ and variance σ^2. Provided that n 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

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha , \tag{2.6}$$

since the transformed variable has approximately a standard normal distribution. Thus, computing a point estimate based on a large n 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 n is large, replacing the population variance σ^2 by the sample variance S^2 in (2.1) will not typically introduce too much additional variability.

Proposition 2.2. *For large samples n, an approximate* $100(1-\alpha)\%$ *confidence interval for the mean μ of any population when the variance is uknown is given by*

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}, \overline{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}\right) , \tag{2.7}$$

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

The CI for the mean (2.7) applies regardless of the shape of the population distribution so long as the number of samples is large. A rule of thumb is that $n > 40$ is sufficient. In words, the CI (2.7) can be expressed as

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

Typically, a large-sample CI for a general parameter θ similar to (2.7) holds for any estimator $\widehat{\theta}$ that satisfies: (1) approximately normal in distribution, (2) approximately unbiased, and (3) an expression for the standard error is available.

TODO: add example

2.4.3 CI for mean of a normal population with unknown variance

In Section 2.4.1, we considered samples $X_1, \dots, X_n$ 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 2.4.2 indicates that the rv $Z = (\overline{X} - \mu)\sqrt{n}/S$ has approximately a standard normal distribution. However, if n is not sufficiently large then the transformed variable will be more spread out than a standard normal distribution.

Theorem 2.1. *For the sample mean $\overline{X}$ based on n samples from a normal distribution with mean μ, the rv*

$$T = \frac{\overline{X} - \mu}{S}\sqrt{n} \quad \sim t_{n-1} \,, \tag{2.8}$$

that is, T has Student's t distribution with $\nu = n - 1$ degrees of freedom (df).

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

Proposition 2.3. *A $100(1-\alpha)\%$ **confidence interval** for the mean μ of a normal population when the value of σ^2 is unknown is given by*

$$\left(\overline{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \,, \overline{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right) \,, \tag{2.9}$$

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

In contrast to Proposition 2.1, it is difficult to select the sample size n to control the width of the t-based CI as the width involves the unknown (before the sample is acquired) s and because n also enters through $t_{\alpha/2,n-1}$.

TODO: add example

Proposition 2.4. ***TODO**: rule of thumb normal CI*

TODO: add example using rule of thumb

2.5 Estimating proportions

Consider a population of size N 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 $F(p)$,[2] that we might be interested in estimating. A sample, $X_1, \dots, X_n \sim F(p)$, of size n from the population contains a proportion,

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \,, \tag{2.10}$$

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

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

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

and

$$\sigma_{\widehat{p}}^2 = \frac{\mathrm{Var}[X_i]}{n} = \frac{p(1-p)}{n}\,, \tag{2.11}$$

provided that n is small relative to N (a rule of thumb is $n \leq 0.05N$).[3] Moreover, by invoking the Central Limit Theorem we have the distribution of $\widehat{p}$ is approximately normal for sufficiently large n as (2.10) is a sample mean. Indeed, this normal approximation works well for moderately large n as long as p is not too close to zero or one; a rule of thumb is that $np > 5$ and $n(1-p) > 5$.

Proposition 2.5. *For large samples n, a $100(1-\alpha)\%$ confidence interval for the parameter p is given by*

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\,. \tag{2.12}$$

This follows from Proposition 2.2 by observing that (2.10) is a sample mean and replacing the standard error $\sigma_{\widehat{p}}$ from (2.11) by the estimated standard error,

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

recall the s in (2.7) is the sample variance for the *population* and $s/\sqrt{n} = \mathsf{se}$ is the standard error of the point estimator.

Example 2.4. **TODO**: Examples of sampling distribution for p

Example 2.5. **TODO**: Example confidence interval for p

Example 2.6. **TODO**: Example hypothesis test

2.6 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 (2.1) provides the basis for inferences. Consider iid samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. We provide the following theorem without proof.

Theorem 2.2. *For the sample variance S^2 based on n samples from a normal distribution with variance σ, the rv*

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

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

Based on Theorem 2.2,

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

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:

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

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.

[3] Note that if n is large relative to N ($n > 0.05N$) then the variance (2.11) must be adjusted by a factor (related to the hypergeometric distribution):

$$\sigma_{\widehat{p}}^2 = \frac{p(1-p)}{n}\frac{N-n}{N-1}\,,$$

where for fixed n the factor converges to 1 as $N \to \infty$.

Proposition 2.6. *A* $100(1-\alpha)\%$ *confidence interval for the variance of a normal population is*

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

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* (2.6)*.*

Example 2.7. TODO: Example CI for variance (using the tree data?)

Chapter 3

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 3.1, 3.3, 3.4. We also consider inferences when the samples are not independent, so-called paired samples, in Section 3.2.

3.1 Comparing means

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

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

and

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

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 3.1. *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 3.1 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 3.1. 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 $\theta = \mu_X - \mu_Y$. The development largely reflects that of Section 2.4:

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

3.1.1 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:

$$Z = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \quad \sim \mathcal{N}(0,1)\,. \tag{3.1}$$

Inferences proceed by treating the parameter of interest θ as a single sample using (3.1).

Proposition 3.2. *A $100(1-\alpha)\%$ CI for the parameter $\theta = \mu_X - \mu_Y$ based on samples of size m from a normal population $\mathcal{N}(\mu_X, \sigma_X^2)$ and of size n from $\mathcal{N}(\mu_Y, \sigma_Y^2)$ 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 3.3. *For null hypothesis $H_0 : \mu_X - \mu_Y = \theta_0$, the test statistic to consider is:*

$$z = \frac{\overline{x} - \overline{y} - \theta_0}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}$$

3.1.2 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 3.4. *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 3.5. *For null hypothesis $H_0 : \mu_X - \mu_Y = \theta_0$, the test statistic to consider is:*

$$z = \frac{\overline{x} - \overline{y} - \theta_0}{\sqrt{\frac{s_X^2}{m} + \frac{s_Y^2}{n}}}$$

3.1.3 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 3.1. *When both population distributions are normal, the standardized variable*

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

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

$$\nu = \frac{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right)^2}{\frac{(s_X^2/m)^2}{m-1} + \frac{(s_Y^2/n)^2}{n-1}} = \frac{\left(s_{\overline{X}}^2 + s_{\overline{Y}}^2\right)^2}{\frac{s_{\overline{X}}^4}{m-1} + \frac{s_{\overline{Y}}^4}{n-1}} \tag{3.2}$$

where s_X^2 and s_Y^2 are point estimators of the sample variances; alternatively, we see that the formula (3.2) *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 (3.2) for the data-driven choice of ν calls for the computation of the standard error of the sample means.

Proposition 3.6. *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 the* t *distribution with* ν *given by* (3.2)*.*

Proposition 3.7. *For the null hypothesis* $H_0 = \mu_X - \mu_Y = \theta_0$*, the test statistic to consider is:*

$$t = \frac{(\overline{x}-\overline{y})-\theta_0}{\sqrt{\frac{s_X^2}{m}+\frac{s_Y^2}{n}}}$$

If the normal populations have the same variance (i.e., the populations differ only in *location* not in *spread*), then matters simplify. If the variances of the normal populations are unknown but are the same (i.e., $\sigma_X^2 = \sigma_Y^2$), 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,

$$S_{\mathsf{p}}^2 = \frac{m-1}{m+n-2}S_X^2 + \frac{n-1}{m+n-2}S_Y^2\,. \tag{3.3}$$

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

Proposition 3.8. *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* t *distribution 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 3.7. 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 3.4) is sensitive to normality assumptions. The two sample t procedures, with the data-driven choice of ν in (3.2), are therefore recommended unless, of course, you have a very compelling reason to believe $\sigma_X^2 = \sigma_Y^2$.

3.2 Comparing paired samples

The preceding analysis for comparing population means was based on the assumption that a random sample $X_1, \ldots, X_m$ is drawn from a distribution with mean μ_X and that a completely independent random sample $Y_1, \ldots, Y_n$ 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

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

[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?

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

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

constitute a sample with mean $\mu_D = \mu_X - \mu_Y$ that can be treated using one-sample CIs and tests, e.g., see Section 2.4.3.

3.3 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 $\widehat{p}_X$. 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 $\widehat{p}_Y$. 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 $\widehat{p}_X - \widehat{p}_Y$.

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

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

and

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

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 $\widehat{p}_X$ and $\widehat{p}_Y$ are (approximately) normally distributed. Standardizing $\widehat{p}_X - \widehat{p}_Y$,

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

A CI for $\widehat{p}_X - \widehat{p}_Y$ then follows from the large-sample CI considered in Section 2.4.2.

Proposition 3.9. *An approximate* $100(1-\alpha)\%$ *CI for* $p_X - p_Y$ *is given by*

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

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

Proposition 3.9 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 are assuming that $p_X = p_Y$ (unless the data suggests otherwise). Therefore as a matter of consistency we should replace the standard error in (3.9) with a pooled estimator for the standard error of the population proportion,

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

Proposition 3.10. *For the null hypothesis* $H_0 = p_X - p_Y = 0$*, the test statistic to consider is:*

$$z = \frac{\widehat{p}_X - \widehat{p}_Y}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{m} + \frac{1}{m}\right)}}$$

3.4 Comparing variances

Chapter 4

Categorical data and Goodness-of-Fit tests

Chapter 5

Linear models

5.1 Regression

5.2 ANOVA

References

Bibliography

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.

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