MA22004 - Statistics and Probability II

Dr Eric Hall

Last updated: 2020-09-03

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. [§6-10, 12]
- DeGroot and Schervish, *Probability and Statistics,* Addison-Wesley, 2001. [§7-10]
- Rice, *Mathematical statistics and data analysis,* Cengage Learning, 2006. [§6-12]

- Wasserman, *All of Statistics,* Springer-Verlag, 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

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 0.1. 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

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

The statement that X is normally distributed with parameters μ and σ^2 is abbreviated $X \sim \mathcal{N}(\mu, \sigma^2)$. For $X \sim \mathcal{N}(\mu, \sigma^2)$, it can be shown that $\mathrm{E}(X) = \mu$ 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.

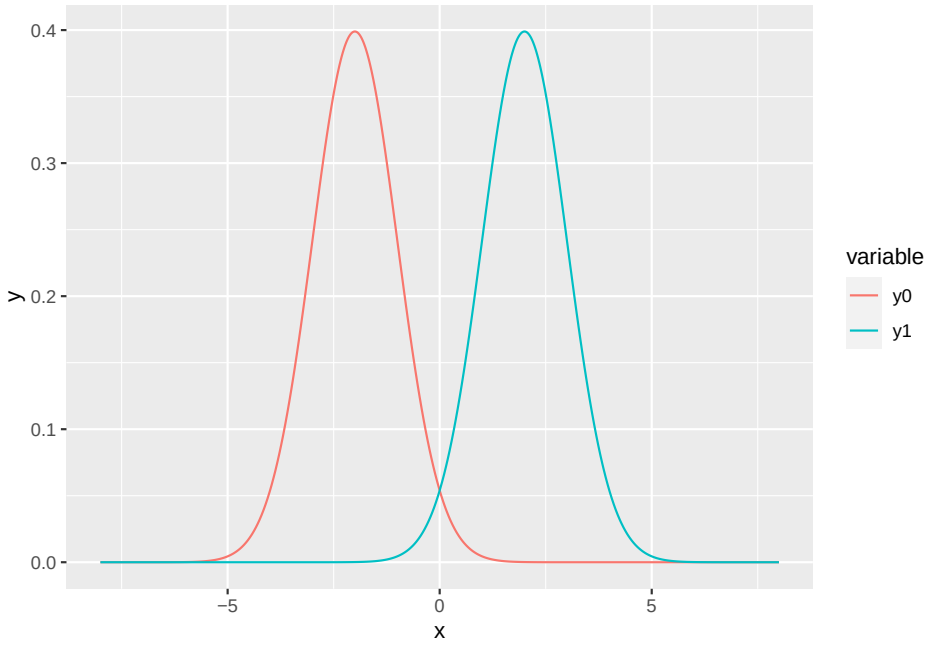

Figure 1: The pdfs of two normal rvs with different means and the same standard deviations.

Definition 0.2. 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!).

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 $i = 1, \ldots, n$ are independent rvs, then

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

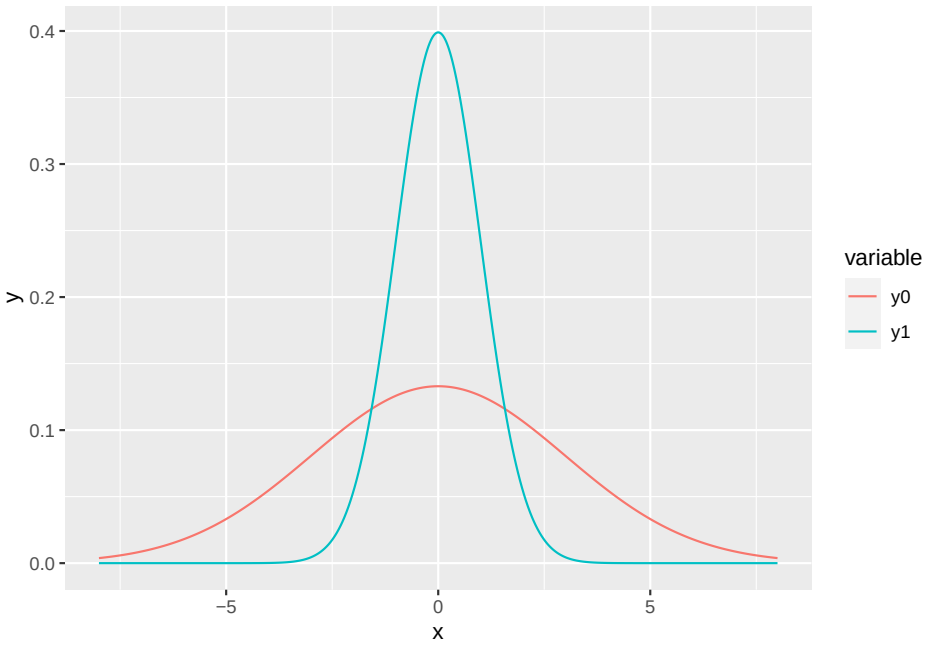

Figure 2: The pdfs of two normal rvs with the same means and different standard deviations.

In particular, we note that for differences of independent rvs $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ 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 $P(a \leq X \leq b)$ 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 $\Phi(z) = P(Z \leq z)$ can then be looked up in a table. From (1.) above,

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

Example 0.1. 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 0.2. Let $X \sim \mathcal{N}(5, 9)$ and find $P(4 \leq X \leq 5.25)$.

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

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
```

Chapter 1

Inferences based on a single sample

1.1 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 for the *setosa* species. Before we collect data, the petal lengths of n independent *setosa* flowers are denoted by rvs $X_1, X_2, \ldots, X_n$. Any function of the X_i's, such as the sample mean,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \,, \tag{1.1}$$

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 $x_1, x_2, \ldots, x_{50}$; the distribution (counts) of 50 such petal length measurements are displayed in Figure 1.1. The sample mean

$\overline{x}$ for petal length could then be used to draw a conclusion about the value of the population mean μ. Similarly, if σ^2 is the variance of the *setosa* petal length then s^2 can be used to infer something about the value of σ^2. Based on the data in Figure 1.1 and using (1.1), 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.

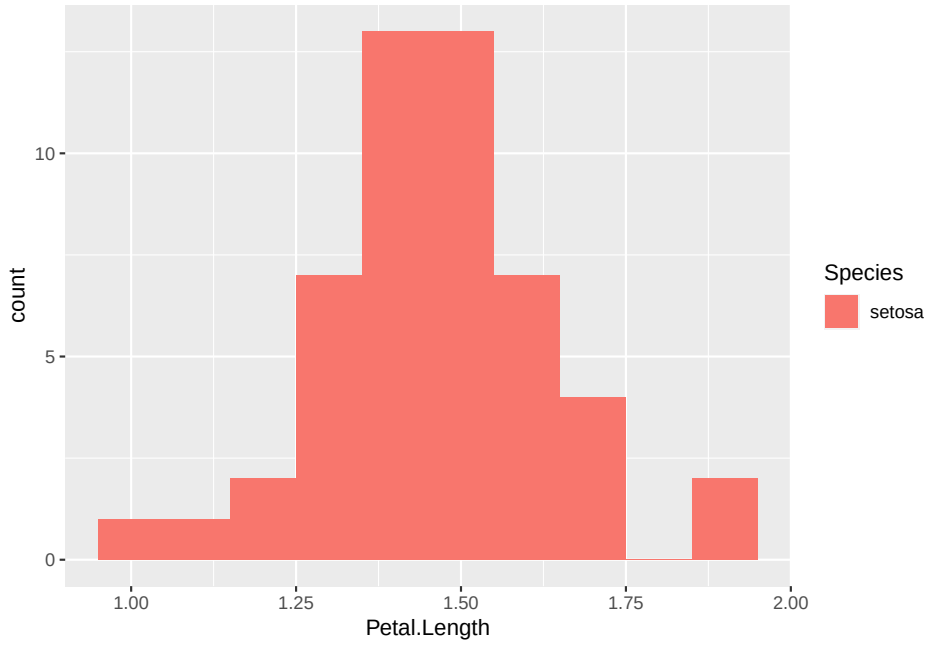

Figure 1.1: The distribution (counts) of 50 *setosa* petal length measurments.

> 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 1.1. 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 $X_1, X_2, \ldots, X_n$ 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,

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

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 1.1 above 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.

Example 1.1. 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 normal probability plot of this data is fairly straight, so we assume that the distribution of black cherry tree heights is normal with a mean value μ. The observations $X_1, \ldots, X_{31}$ are then assumed to be a random sample from this normal distribution. Consider the following estimators and resulting point estimates for μ.

a. Estimator (sample mean) $\overline{X}$ as in (1.1) 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?

> 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)`.

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

Definition 1.2. 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 $\widehat{\mathsf{se}}$.

1.2 Confidence intervals

An alternative to reporting a point estimate for a parameter is to report an interval estimate that suggests 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 always to select the confidence level. A popular choice is a 95% confidence interval which corresponds to level $\alpha = 0.05$.

Definition 1.3. A $(1-\alpha)$ **confidence interval** for a parameter θ is a *random* interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and $b = b(X_1, \ldots, X_n)$ are functions of the data such that

$$P_\theta(\theta \in C_n) = 1 - \alpha\,, \tag{1.2}$$

for all $\theta \in \Theta$.

In Definition 1.3, note that the interval C_n is random and the parameter θ is fixed; a confidence interval is not a probability statement about θ (θ is not a rv).

My favorite interpretation of a confidence interval is due to Wasserman in [?]:

> 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.

The confidence level is not

TODO: CI for mean normal population with known variance [Devore p272]

TODO: sample size necessary for CI to have predetermined width [Devore p273]

TODO: large-sample CI for mean of any population with unknown variance [Devore p277]

TODO: small-sample CI for mean of normal population with unknown variance [Devore p288]

1.3 Hypothesis testing