STAT2201

Analysis of Engineering & Scientific Data

Condensed Course Notes.

Semester 1, 2017.

Last Edited: April 23, 2017
Contains All Units 1 – 10

These condensed notes summarise definitions, procedures, theorems and results relevant for STAT2201. Further material is available in the course book, *Applied Statistics and Probability for Engineers" by D. C. Montgomery and G. C. Runger*, [MonRun2014] and on the course web-site: **https://courses.smp.uq.edu.au/STAT2201/2017a** .
It is recommended to bring printouts of these notes to course lectures and tutorials.

© School of Mathematics and Physics, The University of Queensland, Brisbane QLD 4072, Australia.
Contains materials and contributions developed by Yoni Nazarathy, Melanie Robertson-Dean and Hayden Klok.

About:

These **condensed course notes** are designed as an aid for lecture participation, tutorial participation, assignment preparation and reading of the course book. The course is structured with 10 study units, 1–10. It is recommended to attend lecture and tutorial sessions with printouts of the notes for the relevant unit as well as earlier units. For example, when attending lectures associated with unit 5, bring print outs of at least units 1–5. These will help you follow the lectures and tutorials. See the course website for a detailed lecture and tutorial schedule.

A major goal of this course is to enable students to "talk the language" of probability and statistics. This includes understanding of terms such as probability, events, random variables, distributions, means, moments, estimation, confidence intervals, hypothesis tests, regression and much more. Each of these concepts on its own, entails a variety of associated concepts, results, formulas and properties. Hence the volume of terms and concepts is rather large in comparison to the course duration. This document's goal is to alleviate hardships associated with this, by giving a succinct description of the content.

The course units are as follows:

Unit 1 – Introduction

An overview of the terms: probability, statistics, data-analysis, data science, inference, experimentation, deterministic models, stochastic models, statistical models. Introduction to the Julia language and an overview of alternatives. Simulation examples.

Unit 2 – Probability and Monte Carlo

Probability spaces, outcomes and events. Operations on events as subsets of the sample space. The meaning of probabilistic statements. Independence. Conditional probability and the law of total probability. Pseudorandom number generation and Monte Carlo in a bit more depth.

Unit 3 – Distributions

Random variables. Mathematical description of the distribution of discrete and continuous random variables. Probability mass functions, probability density functions and cumulative distribution functions. Expectation, mean, variance, standard deviation and moments. Quantiles. Discrete uniform and binomial distributions. Uniform (continuous), exponential and normal (Gaussian) distributions. Using the normal distribution table.

Unit 4 – Joint Distributions

Bivariate and multivariate probability distributions. Marginal distributions. Covariance and correlation. Independent random variables. Expectations of functions of several random variables. Means and variance of sums and linear combinations of random variables.

Unit 5 – Descriptive Statistics

Statistics as functions of a sample: sample mean, sample variance and common summarising functions. Box plots. Constructing histograms. Samples with two or more variables. Sample correlation. Scatter plots. Time series. Cumulative plots including probability plots.

Unit 6 – Statistical Inference Ideas

Random samples. Point estimates. The sampling distribution of a statistic. The central limit theorem. Confidence interval idea. Confidence intervals for the mean when the variance is known or for large samples. Prediction intervals. Hypothesis test ideas. Types of error in hypothesis testing. P-value. A general procedure for hypothesis testing.

Unit 7 – Single Sample Inference

The t-distribution. Confidence intervals for the mean of a normal distribution when the variance is not known. Hypothesis tests for the mean of a normal distribution when the variance is not know (the t-test).

Unit 8 – Two Sample Inference

Confidence intervals and hypothesis tests for the difference in means of a normal distribution when the variance is not known (the two sample t-test). Pooled sample variance. Approximations when population variances are assume difference. Understanding model assumptions.

Unit 9 – Linear Regression

Fitting a line through points using least squares. The meaning of linear statistical models. Simple linear regression (slope and intercept). Residual analysis. Properties of the leas square estimators. R squared. Hypothesis tests for regression. A brief discussion of transformations. Logistic regression.

Unit 10 – What more is there

A survey of basic tools from a first statistics course not covered here: Chi-square tests for goodness of fit and independence, non-parameteric tests for fitting distributions, analysis of variance (ANOVA), experimental design, multi-variable regression, principal components analysis, generalized linear models, time-series analysis, survival analysis, reliability modelling, queueing theory and related stochastic models dealing with random processes.

1 Introduction and Julia

➢ **Probability** is a measure of the likelihood of an event occurring and **Probability theory** is a branch of mathematics dealing with probabilities.

➢ **Statistics** is the science of data and involves probability due to: (i) The random nature of sampling data. (ii) Probability theory is useful for devising statistical models and techniques.

➢ **Data Science** is an emerging field, combining statistics, big-data, **machine learning** and computational techniques.

➢ There are thousands of active statistics researchers around the world and tens of thousands **statisticians** and **data scientists**.

➢ Within the **field of statistics**, there are specialisations in **biostatistics**, **mathematical statistics**, **non-parametric statistics**, **machine learning**, **linear models**, **survival analysis**, **Baysian statistics**, **Geo-spatial statistics** and many more.

➢ Within the **field of Probability theory**, there are more theoretical researchers dealing with abstract mathematical models as well as researchers and practitioners dealing with **applied probability**, also related to **stochastic operations research**. Applied probability contains **reliability theory**, **biological population models**, **inventory theory**, **queueing theory** and a few other related fields.

➢ **Data analysis** is the process of curating, organising and analysing data sets to make **inferences**.

➢ **Statistical Inference** is the process of making inferences about **population parameters** (often never fully observed) based on **observations** collected as part of **samples**.

➢ A **statistic** or **summary statistic** is a quantity calculated from a **sample**.

➢ Data can be collected and analysed through a **retrospective study**, **observational study**, **designed experiments** or a combination.

➢ When collecting data, the notion of **time** sometimes plays a key role.

➢ A **deterministic model** of a physical, chemical, biological, financial or related scenario does not involve any randomness.

➢ A **stochastic model** contains built in randomness. Different **realisations** (runs) of the model yield different **outcomes** or **trajectories**.

➢ A **statistical model** is a stochastic model of suited directly for inference.

➢ Some models are **mechanistic** and are based on basic physical (or related) principles. Other models are **empirical** and are of the more "black box" or "grey box" type. Statistical models are often **empirical**, but a statistical model can be incorporated with a **mechanistic** model.

➢ **Simulation** (in the context of mathematical models) is the process of generating observations/-trajectories/outcomes using a computer based on a model. In case of a stochastic or statistical model, this is called **Monte Carlo** (famous casino) simulation and uses random numbers generated on the computer (often **pseudorandom numbers**).

➢ An example of a pseudorandom sequence is the classic **linear congruential generator**:

$$z_{n+1} = (a\, z_n + c) \mod m,$$

where z_0 is some initial **seed** and a, c and m are parameters. For example a known "good" set of parameters is,

$$a = 69069, \qquad c = 1, \qquad m = 2^{32}.$$

➢ In this course we will use the **Julia** programming language, v0.5.0 through uq.juliabox.com.

➢ There are dozens of possible software packages and languages for statistics, data-analysis, data-science and scientific computing. Here is a partial list:

1. **R** - Has become the language of choice for the statistics community.
2. **Matlab** - Often the language of choice for engineering, especially, IEEE systems and control.
3. **Octave** - An open source alternative to Matlab.
4. **Scilab** - Another open source alternative to Matlab.
5. **Mathematica** - A general purpose mathematics, data-science, numerical computing platform. Made it's name with superb symbolic computation abilities. Very powerful in many other domains.
6. **Maple** - Similar to Mathematica with strong symbolic capabilities.
7. **Wolfram Alpha Pro** - A more user friendly version of Mathematica, uses more natural language and less programming.
8. **Python** (including **NumPy**) - Often the language of choice for Data Science.
9. **General programming languages** – C/C++, Java, C#, Go, Swift, Fortran (to name a few)– Not specifically tailored for scientific computing and data-analysis, although they often have many libraries.
10. **Languages (mostly) from the past** – Lisp, Smalltalk, Pascal – Influenced current languages.
11. **Javscript** – Runs on web-browser (together with **html** and **css**). Not really tailored for scientific computation.
12. **PHP, NodesJS** – Server side for supporting web-pages. Not really tailored for scientific computation.
13. **Excel** (and similar such as Apple's **Numbers**) - General spreadsheet software. Often powerful enough for quick simple computations or much more. May require plug-in language such as **Visual Basic** in Excel for specific macros.
14. **Excel like Statistics packages - SPSS** and **Minitab**. More geared towards observations (rows) and variables (columns) than Excel and often yield very easy to use output of common statistical procedures.
15. **Specific Statistics and Machine Learning software and packages** – E.g. **WinBUGS**, **H20**, **Google Tensor Flow** and much more.
16. **Specific civil, mechanical and aerospace modelling languages and software packages** – You will use these in your career.
17. **Latex** - The typesetting language we use for these notes. Not a scientific computing language.
18. **SQL** - A Database Query Language for relational databases. Not a scientific computing language.
19. **Assembly, LLVM** – Low level (not for us).
20. **Julia** - A relatively new (still in version 0.5.0) open sourced general purpose scientific computing language. Similar in nature to Python and Matlab, but strongly typed and Just In Time compiled hence very fast in execution.

➢ We are using **iJulia**, a **Jupyter** style notebook running on a web browser through **JuliaBox**. Current alternatives include a local install of iJulia, the Julia REPL (command line), a plug-in into the Atom Editor - we won't use these.

Here is an example of statistical analysis of the linear congruential generator in Julia:

```
using PyPlot

#Define the parameters of the generator
a = 69069
c = 1
m = 2^32;

#This is a function in Julia, it takes z and returns the next value
# according to the LCG.
#Note that '%' is modulo
function nextZ(z)
    return (a*z + c) % m
end

#Let's set an arbitrary number for the initial seed
z = 2017

#Run for NN steps.
NN=10^6

#Will store all the values of the LCG in an array of integers.
intData=Int64[]

#Loop and iterate
for i in 1:NN
    push!(intData,z)
    z = nextZ(z)
end

#Now create also real values (floating point) on the interval [0,1].
#Julia is strongly typed and to divide z by m we need to cast
#Note that [value(z) for z in data] is called a comprehension in Julia
realData = [z/m for z in intData]

#Plot a segment of the datapoints
subplot(211)
PyPlot.plot(1:1000,intData[1:1000],".")

#Plot a histogram. Does it look uniform?
subplot(212)
PyPlot.plt[:hist](realData[120:242],10)

#Look at the sample mean, sample variance and 1/12.
#The latter is the theoretical
#variance of uniform(0,1) random variables.
mean(realData), var(realData), 1/12.
```

➢ Continuing with this example, we see the Central Limit Theorem in Action:

```
#sampleSize
kk = 2 #try this for k=1, k=2, k=4, k=10 and k=100

#chop the daata into samples, each of size kk
samples = [realData[kk*i+1:kk*(i+1)]
                for i in 0:Int64(length(realData)/kk)-2]

#map the mean function (this is what "." does) on each of the samples.
means=mean.(samples)
PyPlot.plt[:hist](means,100);
```

2 Probability and Monte Carlo

➢ An experiment that can result in different outcomes, even though it is repeated in the same manner every time, is called a **random experiment**.

➢ The set of all possible outcomes of a random experiment is called the **sample space** of the experiment, and is denoted as Ω.

- A sample space is **discrete** if it consists of a finite or countable infinite set of outcomes.
- A sample space is **continuous** if it contains an interval (either finite or infinite) of real numbers, vectors or similar objects.

➢ An **event** is a subset of the sample space of a random experiment.

- The **union** of two events is the event that consists of all outcomes that are contained in either of the two events or both. We denote the union as $E_1 \cup E_2$.
- The **intersection** of two events is the event that consists of all outcomes that are contained in both of the two events. We denote the intersection as $E_1 \cap E_2$.
- The **complement** of an event in a sample space is the set of outcomes in the sample space that are not in the event. We denote the complement of the event E as $\overline{E}$. The notation E^C is also used in other literature to denote the complement. Note that $E \cup \overline{E} = \Omega$.

➢ Two events, denoted E_1 and E_2 are **mutually exclusive** if: $E_1 \cap E_2 = \varnothing$ where $\varnothing$ is called the **empty set** or **null event**.

➢ A collection of events, $E_1, E_2, \ldots, E_k$ is said to be **mutually exclusive** if for all pairs,

$$E_i \cap E_j = \varnothing.$$

➢ The definition of the complement of an event implies that: $(E^c)^c = E$.

➢ The distributive law for set operations implies that

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C) \qquad \text{and} \qquad (A \cap B) \cup C = (A \cup C) \cap (B \cup C).$$

➢ DeMorgan's laws imply that

$$(A \cup B)^c = A^c \cap B^c \qquad \text{and} \qquad (A \cap B)^c = A^c \cup B^c.$$

➢ Union and intersection are commutative operations: $A \cap B = B \cap A \qquad$ and $\qquad A \cup B = B \cup A$.

➢ **Probability** is used to quantify the likelihood, or chance, that an outcome of a random experiment will occur.

➢ Whenever a sample space consists of a finite number N of possible outcomes, each **equally likely**, the probability of each outcome is $1/N$.

➢ For a discrete sample space, the **probability of an event** E, denoted as $P(E)$, equals the sum of the probabilities of the outcomes in E.

➢ If Ω is the sample space and E is any event in a random experiment,

(1) $P(\Omega) = 1$.

(2) $0 \leq P(E) \leq 1$.

(3) For two events E_1 and E_2 with $E_1 \cap E_2 = \varnothing$ (disjoint),

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

(4) $P(E^c) = 1 - P(E)$.
(5) $P(\emptyset) = 0$.

➢ The probability of event A or event B occurring is,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

➢ If A and B are mutually exclusive events,

$$P(A \cup B) = P(A) + P(B)$$

➢ For a collection of **mutually exclusive events**,

$$P(E_1 \cup E_2 \cup \cdots \cup E_k) = P(E_1) + P(E_2) + \ldots P(E_k)$$

➢ The probability of an event B under the knowledge that the outcome will be in event A is denoted $P(B \mid A)$ and is called the **conditional probability** of B given A.

➢ The **conditional probability** of an event B given an event A, denoted as $P(B \mid A)$, is

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \quad \text{for} \quad P(A) > 0.$$

➢ The **multiplication rule** for probabilities is: $P(A \cap B) = P(B \mid A)P(A) = P(A \mid B)P(B)$.

➢ For an event B and a collection of mutual exclusive events, $E_1, E_2, \ldots, E_k$ where their union is Ω. The **law of total probability** yields,

$$\begin{aligned} P(B) &= P(B \cap E_1) + P(B \cap E_2) + \cdots + P(B \cap E_k) \\ &= P(B \mid E_1)P(E_1) + P(B \mid E_2)P(E_2) + \cdots + P(B \mid E_k)P(E_k). \end{aligned}$$

➢ Two events are **independent** if any one of the following equivalent statements is true:

(1) $P(A \mid B) = P(A)$.
(2) $P(B \mid A) = P(B)$.
(3) $P(A \cap B) = P(A)P(B)$.

Observe that **independent** events and **mutually exclusive** events, are completely different concepts. Don't confuse these concepts.

➢ For **multiple events** $E_1, E_2, \ldots, E_n$ are independent if and only if for any subset of these events

$$P(E_{i_1} \cap E_{i_2} \cap \cdots \cap E_{i_k}) = P(E_{i_1})\ P(E_{i_2})\ \ldots\ P(E_{i_k}).$$

➢ A **pseudorandom sequence** is a sequence of numbers $U_1, U_2, \ldots$ with each number, U_k depending on the previous numbers $U_{k-1}, U_{k-2}, \ldots, U_1$ through a well defined functional relationship and similarly U_1 depending on the **seed** $\tilde{U}_0$. Hence for any seed, $\tilde{U}_0$, the resulting sequence $U_1, U_2, \ldots$ is fully defined and repeatable. A pseudorandom sequence often lives within a discrete domain as $\{0, 1, \ldots, 2^{64} - 1\}$. It can then be **normalised** to floating point numbers with,

$$R_k = \frac{U_k}{2^{64} - 1}.$$

➢ A good pseudorandom sequence has the following attributes among others:

1. It is quick and easy to compute the next element in the sequence.
2. The sequence of numbers $R_1, R_2, \ldots$ resembles properties as an i.i.d. sequence of uniform(0,1) random variables (i.i.d. is defined in Unit 4).

➢ Computer simulation of random experiments is called **Monte Carlo** and is typically carried out by setting the seed to either a reproducible value or an arbitrary value such as system time.

➢ Random experiments may be replicated on a computer using Monte Carlo simulation.

3 Distributions

➢ A **random variable** X is a numerical (integer, real, complex, vector etc.) summary of the outcome of the random experiment. The **range** or **support** of the random variable is the set of possible values that it may take. Random variables are usually denoted by capital letters.

➢ A **discrete random variable** is an integer/real-valued random variable with a finite (or countably infinite) range.

➢ A **continuous random variable** is a real valued random variable with an interval (either finite or infinite) of real numbers for its range.

➢ The **probability distribution** of a random variable X is a description of the probabilities associated with the possible values of X. There are several common alternative ways to describe the probability distribution, with some differences between discrete and continuous random variables.

➢ While not the most popular in practice, a unified way to describe the distribution of any scalar valued random variable X (real or integer) is the **cumulative distribution function**,

$$F(x) = P(X \leq x).$$

➢ It holds that

(1) $0 \leq F(x) \leq 1$.
(2) $\lim_{x \to -\infty} F(x) = 0$.
(3) $\lim_{x \to \infty} F(x) = 1$.
(4) If $x \leq y$, then $F(x) \leq F(y)$. That is, $F(\cdot)$ is non-decreasing.

➢ Distributions are often summarised by numbers such as the **mean**, μ, **variance**, σ^2, or **moments**. These numbers, in general do not identify the distribution, but hint at the general location, spread and shape.

➢ The **standard deviation** of X is $\sigma = \sqrt{\sigma^2}$ and is particularly useful when working with the Normal distribution.

➢ Given a discrete random variable X with possible values $x_1, x_2, \ldots, x_n$, the **probability mass function** of X is,

$$p(x) = P(X = x).$$

Note: In [MonRun2014] and many other sources, the notation used is $f(x)$ (as a pdf of a continuous random variable).

➢ A probability mass function, $p(x)$ satisfies:

(1) $p(x_i) \geq 0$.
(2) $\sum_{i=1}^{n} p(x_i) = 1$.

➢ The **cumulative distribution function** of a discrete random variable X, denoted as $F(x)$, is

$$F(x) = \sum_{x_i \leq x} p(x_i).$$

➢ $P(X = x_i)$ can be determined from the *jump* at the value of x. More specifically

$$p(x_i) = P(X = x_i) = F(x_i) - \lim_{x \uparrow x_i} F(x_i).$$

➢ The **mean** or **expected value** of a discrete random variable X, is

$$\mu = E(X) = \sum_x x\, p(x).$$

➢ The **expected value** of $h(X)$ for some function $h(\cdot)$ is:

$$E\Big[h(X)\Big] = \sum_x h(x)\, p(x).$$

➢ The k'th **moment** of X is,

$$E(X^k) = \sum_x x^k\, p(x).$$

➢ The **variance** of X, is

$$\sigma^2 = V(X) = E(X-\mu)^2 = \sum_x (x-\mu)^2\, p(x) = \sum_x x^2\, p(x) - \mu^2.$$

➢ A random variable X has a **discrete uniform distribution** if each of the n values in its range, $x_1, x_2, \ldots, x_n$, has equal probability. I.e.

$$p(x_i) = 1/n.$$

➢ Suppose that X is a discrete uniform random variable on the consecutive integers $a, a+1, a+2, \ldots, b$, for $a \le b$. The **mean** and **variance** of X are

$$E(X) = \frac{b+a}{2} \qquad \text{and} \qquad V(X) = \frac{(b-a+1)^2 - 1}{12}.$$

➢ The setting of n **independent and identical Bernoulli trials** is as follows:

(1) There are n trials.

(1) The trials are independent.

(2) Each trial results in only two possible outcomes, labelled as "success" and "failure".

(3) The probability of a success in each trial denoted as p is the same for all trials.

➢ The random variable X that equals the number of trials that result in a success is a **binomial random variable** with parameters $0 \le p \le 1$ and $n = 1, 2, \ldots$. The probability mass function of X is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

➢ Useful to remember from algebra: the binomial expansion for constants a and b is

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

➢ If X is a binomial random variable with parameters p and n, then,

$$E(X) = n\,p \qquad \text{and} \qquad V(X) = n\,p\,(1-p).$$

➢ Given a continuous random variable X, the **probability density function** (pdf) is a function, $f(x)$ such that,

(1) $f(x) \geq 0$.

(2) $f(x) = 0$ for x not in the range.

(3) $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

(4) For small Δx, $f(x)\,\Delta x \approx P(X \in [x, x + \Delta\, x))$.

(5) $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx =$ area under $f(x)$ from a to b.

➢ Given the PDF, $f(x)$ we can get the CDF as follows:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du \qquad \text{for} \qquad -\infty < x < \infty.$$

➢ Given the CDF, $F(x)$ we can get the PDF:

$$f(x) = \frac{d}{dx}\,F(x).$$

➢ The **mean** or **expected value** of a continous random variable X, is

$$\mu = E(X) = \int_{-\infty}^{\infty} x\,f(x)dx.$$

➢ The **expected value** of $h(X)$ for some function $h(\cdot)$ is:

$$E\Big[h(X)\Big] = \int_{-\infty}^{\infty} h(x)f(x)\,dx.$$

➢ The k'th **moment** of X is,

$$E(X^k) = \int_{-\infty}^{\infty} x^k\,f(x)\,dx.$$

➢ The **variance** of X, is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2.$$

➢ A continuous random variable X with probability density function

$$f(x) = \frac{1}{b-a}, \qquad a \leq x \leq b.$$

is a **continuous uniform random variable** or "uniform random variable" for short.

➢ If X is a continuous uniform random variable over $a \leq x \leq b$, the **mean** and **variance** are:

$$\mu = E(X) = \frac{a+b}{2} \quad \text{and} \quad \sigma^2 = V(X) = \frac{(b-a)^2}{12}.$$

➢ A random variable X with probability density function

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

is a **normal random variable** with parameters μ where $-\infty < \mu < \infty$, and $\sigma > 0$. For this distribution, the parameters map directly to the mean and variance,

$$E(X) = \mu \quad \text{and} \quad V(X) = \sigma^2.$$

The notation $N(\mu, \sigma^2)$ is used to denote the distribution. Note that some authors and software packages use σ for the second parameter and not σ^2.

➢ A normal random variable with a mean and variance of:

$$\mu = 0 \quad \text{and} \quad \sigma^2 = 1$$

is called a **standard normal random variable** and is denoted as Z. The cumulative distribution function of a standard normal random variable is denoted as

$$\Phi(z) = F_Z(z) = P(Z \leq z),$$

and is tabulated in a table.

➢ It is very common to compute $P(a < X < b)$ for $X \sim N(\mu, \sigma^2)$. This is the typical way:

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

We get:

$$F_X(b) - F_X(a) = F_Z\Big(\frac{b-\mu}{\sigma}\Big) - F_Z\Big(\frac{a-\mu}{\sigma}\Big).$$

➢ The **exponential distribution** with parameter $\lambda > 0$ is given by the **survival function**,

$$\overline{F}(x) = 1 - F(x) = P(X > x) = e^{-\lambda x}.$$

➢ The random variable X that equals the distance between successive events from a Poisson process with mean number of events per unit interval $\lambda > 0$.

➢ The probability density function of X is

$$f(x) = \lambda e^{-\lambda x} \quad \text{for} \quad 0 \leq x < \infty.$$

Note that sometimes a different parameterisation, $\theta = 1/\lambda$ is used (e.g. in the Julia Distributions package).

➢ The **mean** and **variance** are:

$$\mu = E(X) = \frac{1}{\lambda} \quad \text{and} \quad \sigma^2 = V(X) = \frac{1}{\lambda^2}$$

➢ The exponential distribution is the only continuous distribution with range $[0, \infty)$ exhibiting the **lack of memory property**. For an exponential random variable X,

$$P(X > t + s \,|\, X > t) = P(X > s).$$

➢ Monte Carlo simulation makes use of methods to transform a uniform random variable in a manner where it follows an arbitrary given given distribution. One example of this is if $U \sim \text{Uniform}(0, 1)$ then $X = -\frac{1}{\lambda}\log(U)$ is exponentially distributed with parameter λ.

4 Joint Probability Distributions

➢ A joint probability distribution of two random variables is also referred to as **bivariate probability distribution**.

➢ A **joint probability mass function** for discrete random variables X and Y, denoted as $p_{XY}(x, y)$, satisfies the following properties:

(1) $p_{XY}(x, y) \geq 0$ for all x, y.

(2) $p_{XY}(x, y) = 0$ for (x, y) not in the range.

(3) $\sum\sum p_{XY}(x, y) = 1$, where the summation is over all (x, y) in the range.

(4) $p_{XY}(x, y) = P(X = x,\ Y = y)$.

➢ A **joint probability density function** for continuous random variables X and Y, denoted as $f_{XY}(x, y)$, satisfies the following properties:

(1) $f_{XY}(x, y) \geq 0$ for all x, y.

(2) $f_{XY}(x, y) = 0$ for (x, y) not in the range.

(3) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x, y)\ dx\ dy = 1$.

(4) For small $\Delta x,\ \Delta y$: $\quad f_{XY}(x, y)\,\Delta x\,\Delta y \approx P\Big((X, Y) \in [x, x + \Delta\, x) \times [y, y + \Delta\, y)\Big)$.

(5) For any region R of two-dimensional space,

$$P\Big((X, Y) \in R\Big) = \iint_R f_{XY}(x, y)\ dx\ dy.$$

➢ A **joint probability density function** can also be defined for $n > 2$ random variables (as can be a **joint probability mass function**). The following needs to hold:

(1) $f_{X_1X_2\ldots X_p}(x_1, x_2, \ldots, x_n) \geq 0$.

(2) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{X_1X_2\ldots X_p}(x_1, x_2, \ldots, x_n)dx_1\ dx_2 \ldots dx_n = 1$.

➢ Most of the concepts in this section, carry over from bivariate to general multivariate distributions ($n > 2$).

➢ The **marginal distributions** of X and Y as well as **conditional distributions** of X given a specific value $Y = y$ and vice versa can be obtained from the joint distribution.

➢ If the random variables X and Y are independent, then $f_{XY}(x, y) = f_X(x)\, f_Y(y)$ and similarly in the discrete case.

➢ The **expected value of a function of two random variables** is:

$$E\Big[h(X, Y)\Big] = \iint h(x, y) f_{XY}(x, y)\ dx\ dy \qquad \text{for } X, Y \text{ continuous.}$$

➢ The **covariance** is a common measure of the relationship between two random variables (say X and Y). It is denoted as $\text{cov}(X, Y)$ or σ_{XY}, and is given by:

$$\sigma_{XY} = E\Big[(X - \mu_X)(Y - \mu_Y)\Big] = E(XY) - \mu_X\mu_Y.$$

➢ The covariance of a random variable with itself is its variance.

➢ The **correlation** between the random variables X and Y, denoted as ρ_{XY}, is

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

➢ For any two random variables X and Y, $-1 \leq \rho_{XY} \leq 1$.

➢ If X and Y are independent random variables, $\sigma_{XY} = 0$ and $\rho_{XY} = 0$. The opposite case does not always hold: In general $\rho_{XY} = 0$ does not imply independence. But for jointly Normal random variables it does. In any case, if $\rho_{XY} = 0$ then the random variables are called uncorrelated.

➢ When considering several random variables, it is common to consider the (symmetric) **Covariance Matrix**, Σ with $\Sigma_{i,j} = \text{cov}(X_i, X_j)$.

➢ The **probability density function** of a **bivariate normal distribution** is

$$f_{XY}(x,y;\sigma_X,\sigma_Y,\mu_X,\mu_Y,\rho) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} \right] \right\}$$

for $-\infty < x < \infty$ and $-\infty < y < \infty$,
with parameters $\sigma_X > 0$, $\sigma_Y > 0$, $-\infty < \mu_X < \infty$, $-\infty < \mu_Y < \infty$, and $-1 < \rho < 1$.

➢ Given random variables $X_1, X_2, \ldots, X_n$ and constants $c_1, c_2, \ldots, c_n$, the (scalar) **linear combination**

$$Y = c_1X_1 + c_2X_2 + \cdots + c_nX_n$$

is often a random variable of interest.

➢ The mean of the linear combination is the linear combination of the means,

$$E(Y) = c_1E(X_1) + c_2E(X_2) + \cdots + c_nE(X_n).$$

This holds even if the random variables are not independent.

➢ The variance of the linear combination is as follows:

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \cdots + c_n^2V(X_n) + 2\sum_{i<j}\sum c_ic_j\text{cov}(X_i,X_j)$$

➢ If $X_1, X_2, \ldots, X_n$ are **independent** (or even if they are just uncorrelated).

$$V(Y) = c_1^2V(X_1) + c_2^2V(X_2) + \cdots + c_n^2V(X_n).$$

➢ In case the random variables $X_1, \ldots, X_n$ were jointly Normal then, $Y \sim \text{Normal}\Big(E(Y), V(Y)\Big)$. That is, **linear combinations of Normal random variables remain Normally distributed**.

➢ A collection of random variables, $X_1, \ldots, X_n$ is said to be **i.i.d.**, or **independent and identically distributed** if they are mutually independent and identically distributed. This means that the (n - dimensional) joint probability density is a product of the individual densities.

➢ In the context of statistics, a **random sample** is often modelled as an i.i.d. vector of random variables. $X_1, \ldots, X_n$.

➢ An important linear combination associated with a random sample is the **sample mean**:

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \ldots + \frac{1}{n}X_n.$$

➢ If X_i has mean μ and variance σ^2 then sample mean (of an i.i.d. sample) has,

$$E(\overline{X}) = \mu, \qquad V(\overline{X}) = \frac{\sigma^2}{n}.$$

5 Descriptive Statistics

➢ **Descriptive statistics** deals with summarizing **data** using numbers, qualitative summaries, tables and graphs.

➢ Here are some types of **data configurations**:

1. Single sample: $x_1, x_2, \ldots, x_n$.
2. Single sample over time (time series): $x_{t_1}, x_{t_2}, \ldots, x_{t_n}$ with $t_1 < t_2 < \ldots < t_n$.
3. Two samples: $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$.
4. Generalizations from two samples to k samples (each of potentially different sample size, $n_1, \ldots, n_k$).
5. Observations in tuples: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
6. Generalizations from tuples to vector observations (each vector of length ℓ),

$$(x_1^1, \ldots, x_1^\ell), \ldots, (x_n^1, \ldots, x_n^\ell).$$

➢ Individual **variables** may be **categorical** or **numerical**. Categorical variables (taking values in one of several categories) may be **ordinal** meaning that they be sorted (e.g. "a", "b", "c", "d"), or not (e.g. "cat", "dog", "fish").

➢ A **statistic** is a quantity computed from a sample (assume here a single sample $x_1, \ldots, x_n$). Here are very common and useful statistics:

1. The **sample mean**: $\bar{x} = \dfrac{x_1 + \cdots + x_n}{n} = \dfrac{\sum_{i=1}^{n} x_i}{n}$.
2. The **sample variance**: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \dfrac{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}{n-1}$.
3. The **sample standard deviation**: $s = \sqrt{s^2}$.
4. **Order statistics** work as follows: Sort the sample to obtain the sequence of sorted observations, denoted $x_{(1)}, \ldots, x_{(n)}$ where, $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. Some common order statistics:
   (a) The **minimum** $\min(x_1, \ldots, x_n) = x_{(1)}$.
   (b) The **maximum** $\max(x_1, \ldots, x_n) = x_{(n)}$.
   (c) The **median**
   $$\text{median} = \begin{cases} x_{(\frac{n+1}{2})} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\big(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\big) & \text{if } n \text{ is even.} \end{cases}$$
   Note that the median is the 50'th percentile and the 2nd quartile (see below).
   (d) The q th **quantile** ($q \in [0,1]$) or alternatively the $p = 100q$ **percentile** (measured in percents instead of a decimal), is the observation such that p percent of the observations are less than it and $(1-p)$ percent of the observations are greater than it. In cases (as is typical) that there is not such a precise observation, it is a linear interpolation between two neighbouring observations (as is done for the median when n is even). In terms of order statistics, the q th quantile is approximately (not taking linear interpolations into account) $x_{([q*n])}$. Here $[z]$ denotes the nearest integer in $\{1, \ldots, n\}$ to z.
   (e) The first **quartile**, denoted $Q1$ is the 25th percentile. The second quartile ($Q2$) is the median. The **third quartile**, denoted $Q3$ is the 75th percentile. Thus half of the observations lie between $Q1$ and $Q3$. In other words, the quartiles break the sample into 4 quarters. The difference $Q3 - Q1$ is the **interquartile range**.
   (f) The **sample range** is $x_{(n)} - x_{(1)}$.

➢ **Constructing a Histogram (Equal Bin Widths)**

(1) Label the bin (class interval) boundaries on a horizontal scale.

(2) Mark and label the vertical scale with **frequencies** or **counts**.

(3) Above each bin, draw a rectangle where height is equal to the frequency (or count).

➢ A **Kernel Density Estimate** (KDE) is a way to construct a **Smoothed Histogram**. While construction is not as straightforward as steps (1)–(3) above, automated tools can be used.

➢ Both the histogram and the KDE are not unique in the way they summarize data. With these methods, different settings (e.g. number of bins in histograms or **bandwidth** in a KDE) may yield different representations of the same data set. Nevertheless, they are both very common, sensible and useful visualisations of data.

➢ The **box plot** is a graphical display that simultaneously describes several important features of a data set, such as centre, spread, **departure from symmetry**, and identification of unusual observations or **outliers**. It is often common to plot several box plots next to each other for comparison.

➢ An anachronistic, but useful way for summarising small data-sets is the **stem and leaf diagram**.

➢ In a **cumulative frequency plot** the height of each bar is the total number of observations that are less than or equal to the upper limit of the bin.

➢ The **Empirical Cumulative Distribution Function** (ECDF) is,

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \leq x\}.$$

Here $\mathbf{1}\{\cdot\}$ is the **indicator function**. The ECDF is a function of the data, defined for all x.

➢ Given a **candidate distribution** with CDF $F(x)$, a **probability plot** is a plot of the ECDF (or sometimes just it's jump points) with the y-axis stretched by the inverse of the CDF $F^{-1}(\cdot)$. The monotonic transformation of the y-axis is such that if the data comes from the candidate $F(x)$, the points would appear to lie on a straight line. Names of variations of probability plots are the **P-P plot** and **Q-Q plot** (these plots are similar to the probability plot). A very common probability plot is the **Normal probability plot** where the candidate distribution is taken to be Normal$(\overline{x}, s^2)$.

➢ The Normal probability plot can be useful in identifying distributions that are symmetric but that have tails that are "heavier" or "lighter" than the Normal.

➢ A **time series plot** is a graph in which the vertical axis denotes the observed value of the variable and the horizontal axis denotes time.

➢ A **scatter diagram** is constructed by plotting each pair of observations with one measurement in the pair on the vertical axis of the graph and the other measurement in the pair on the horizontal axis.

➢ The **sample correlation coefficient** r_{xy} is an estimate for the correlation coefficient, ρ, presented in the previous unit:

$$r_{xy} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

6 Statistical Inference Ideas

➢ **Statistical Inference** is the process of forming judgements about the **parameters of a population**, typically on the basis of **random sampling**.

➢ The random variables $X_1, X_2, \ldots, X_n$ are an (i.i.d.) **random sample** of size n if

(a) the X_i's are independent random variables and

(b) every X_i has the same probability distribution.

➢ A **statistic** is any function of the observations in a random sample, and the probability distribution of a statistic is called the **sampling distribution**.

➢ Any function of the observation, or any **statistic**, is also a random variable. We call the probability distribution of a statistic a **sampling distribution**. A **point estimate** of some population parameter θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$. The statistic $\hat{\Theta}$ is called the **point estimator**.

➢ The most common statistic we consider is the **sample mean**, $\overline{X}$, with a given value denoted by $\overline{x}$. As an estimator, the sample mean is an estimator of the population mean, μ.

➢ **Central Limit Theorem** (for sample means):
If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population with mean μ and finite variance σ^2 and if $\overline{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$, is the standard normal distribution.

➢ This implies that $\overline{X}$ is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.

➢ The **standard error** of $\overline{X}$ is given by $\sigma/\sqrt{n}$. In most practical situations σ is not known but rather estimated in this case, the **estimated standard error**, (denoted in typical computer output as "SE"), is $s/\sqrt{n}$ where s is the point estimator,

$$s = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\,\overline{x}^2}{n-1}}.$$

➢ **Central Limit Theorem** (for sums):
Manipulate the central limit theorem (for sample means and use $\sum_{i=1}^{n} X_i = n\overline{X}$. This yields,

$$Z = \frac{\sum_{i=1}^{n} X_i - n\,\mu}{\sqrt{n\sigma^2}},$$

which follows a standard normal distribution as $n \to \infty$.

➢ This implies that $\sum_{i=1}^{n} X_i$ is approximately normally distributed with mean $n\,\mu$ and variance $n\,\sigma^2$.

➢ Knowing the sampling distribution (or the approximate sampling distribution) of a statistic is the key for the two main tools of statistical inference that we study:

(a) **Confidence intervals** – a method for yielding error bounds on **point estimates**.

(b) **Hypothesis testing** – a methodology for making conclusions about population parameters.

➢ The formulas for most of the statistical procedures use **quantiles of the sampling distribution**. When the distribution is $N(0,1)$ (standard normal), the α's quantile is denoted z_α and satisfies:

$$\alpha = \int_{-\infty}^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}\, dx.$$

A common value to use for α is 0.05 and in procedures the expressions $z_{1-\alpha}$ or $z_{1-\alpha/2}$ appear. Note that in this case $z_{1-\alpha/2} = 1.96 \approx 2$.

➢ A **confidence interval** estimate for μ is an interval of the form $l \leq \mu \leq u$, where the end-points l and u are computed from the sample data. Because different samples will produce different values of l and u, these end points are values of random variables L and U, respectively. Suppose that

$$P\big(L \leq \mu \leq U\big) = 1 - \alpha.$$

The resulting **confidence interval** for μ is

$$l \leq \mu \leq u.$$

The end-points or bounds l and u are called the **lower-** and **upper-confidence limits** (bounds), respectively, and $1 - \alpha$ is called the **confidence level**.

➢ If $\bar{x}$ is the sample mean of a random sample of size n from a normal population with known variance σ^2, a $100(1-\alpha)\%$ **confidence interval** on μ is given by

$$\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

➢ Note that it is roughly of the form, $\overline{x} - 2 \text{ SE} \leq \ \mu \ \leq \overline{x} + 2 \text{ SE}$.

➢ Confidence interval formulas give insight into the **required sample size**: If $\bar{x}$ is used as an estimate of μ, we can be $100(1-\alpha)\%$ confident that the error $|\bar{x} - \mu|$ will not exceed a specified amount Δ when the sample size is not smaller than

$$n = \left(\frac{z_{1-\alpha/2}\ \sigma}{\Delta}\right)^2.$$

➢ A **statistical hypothesis** is a statement about the parameters of one or more populations. The **null hypothesis**, denoted H_0 is the claim that is initially assumed to be true based on previous knowledge. The **alternative hypothesis**, denoted H_1 is a claim that contradicts the null hypothesis.

➢ For some arbitrary value μ_0, a **two-sided alternative hypothesis** would be expressed as follows:

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

whereas a **one-sided alternative hypothesis** would be expressed as:

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu < \mu_0 \qquad \text{or} \qquad H_0 : \mu = \mu_0 \qquad H_1 : \mu > \mu_0.$$

➢ The standard scientific research use of hypothesis is to "hope to reject" H_0 so as to have statistical evidence for the validity of H_1.

➢ An hypothesis test is based on a **decision rule** that is a function of the **test statistic**. For example: Reject H_0 if the test statistic is below a specified threshold, otherwise don't reject.

➢ Rejecting the null hypothesis H_0 when it is true is defined as a **type I error**. Failing to reject the null hypothesis H_0 when it is false is defined as a **type II error**.

	H_0 **Is True**	H_0 **Is False**
Fail to reject H_0:	No error	Type II error
Reject H_0:	Type I error	No error

$\alpha = P(\text{type I error}) = P(\text{reject } H_0 \mid H_0 \text{ is true})$.

$\beta = P(\text{type II error}) = P(\text{fail to reject } H_0 \mid H_0 \text{ is false and } H_1 \text{ is true})$.

➢ The **power** of a statistical test is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis is true.

➢ A typical example of a **simple hypothesis test** has $H_0 : \mu = \mu_0 \qquad H_1 : \mu = \mu_1$, where μ_0 and μ_1 are some specified values for the population mean. This test isn't typically practical but is useful for understanding the concepts at hand.

➢ Assuming that $\mu_0 < \mu_1$ and setting a threshold, τ, reject H_0 if the $\overline{x} > \tau$, otherwise don't reject.

➢ Explicit calculation of the relationships of τ, α, β, n, σ, μ_0 and μ_1 is possible in this case.

➢ In most hypothesis tests used in practice (and in this course), a specified level of type I error, α is predetermined (e.g. $\alpha = 0.05$) and the type II error is not directly specified.

➢ The probability of making a type II error β increases (power decreases) rapidly as the true value of μ approaches the hypothesized value.

➢ The probability of making a type II error also depends on the sample size n - increasing the sample size results in a decrease in the probability of a type II error.

➢ The population (or natural) variability (e.g. described by σ) also affects the power.

➢ The **P-value** is the smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data. That is, the P-value is based on the data. It is computed by considering the location of the test statistic under the sampling distribution based on H_0.

➢ It is customary to consider the test statistic (and the data) significant when the null hypothesis H_0 is rejected; therefore, we may think of the P-value as the smallest α at which the data are significant. In other words, the P-value is the **observed significance level**.

➢ Clearly, the P-value provides a measure of the credibility of the null hypothesis. Computing the exact P-value for a statistical test is not always doable by hand.

➢ It is typical to report the P-value in studies where H_0 was rejected (and new scientific claims were made). Typical ("convincing") values can be of the order 0.001.

➢ **A General Procedure for Hypothesis Tests is**

(1) **Parameter of interest:** From the problem context, identify the parameter of interest.

(2) **Null hypothesis,** H_0**:** State the null hypothesis, H_0.

(3) **Alternative hypothesis,** H_1**:** Specify an appropriate alternative hypothesis, H_1.

(4) **Test statistic:** Determine an appropriate test statistic.

(5) **Reject** H_0 **if:** State the rejection criteria for the null hypothesis.

(6) **Computations:** Compute any necessary sample quantities, substitute these into the equation for the test statistic, and compute the value.

(7) **Draw conclusions:** Decide whether or not H_0 should be rejected and report that in the problem context.

7 Single Sample Inference

➢ The setup is a sample $x_1, \ldots, x_n$ (collected values) modelled by an i.i.d. sequence of random variables, $X_1, \ldots, X_n$.

➢ The parameter at question in this unit is the population mean, $\mu = E[X_i]$. A point estimate is $\overline{x}$ (described by the random variable $\overline{X}$).

➢ We devise hypothesis tests and confidence intervals for μ, distinguishing between the (unrealistic but simpler) case where the population variance, σ^2, is known, and the more realistic case where it is not known and estimated by the sample variance, s^2.

➢ For very small samples, the results we present are valid only if the population is normally distributed. But for non-small samples (e.g. $n > 20$, although there isn't a clear rule), the central limit theorem provides a good approximation and the results are approximately correct.

➢ **Testing Hypotheses on the Mean, Variance Known (Z-Tests)**

Model: $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with μ unknown but σ^2 known.

Null hypothesis: $H_0 : \mu = \mu_0$.

Test statistic: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}, \qquad Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu \neq \mu_0$	$P = 2\big[1 - \Phi\big(\lvert z\rvert\big)\big]$	$z > z_{1-\alpha/2}$ or $z < z_{\alpha/2}$
$H_1 : \mu > \mu_0$	$P = 1 - \Phi\big(z\big)$	$z > z_{1-\alpha}$
$H_1 : \mu < \mu_0$	$P = \Phi\big(z\big)$	$z < z_{\alpha}$

➢ Note: For $H_1 : \mu \neq \mu_0$, a procedure identical to the preceding fixed significance level test is:

$$\text{Reject } H_0 : \mu = \mu_0 \quad \text{if either} \quad \overline{x} < a \text{ or } \overline{x} > b$$
$$\text{where}$$
$$a = \mu_0 - z_{1-\alpha/2}\tfrac{\sigma}{\sqrt{n}} \quad \text{and} \quad b = \mu_0 + z_{1-\alpha/2}\tfrac{\sigma}{\sqrt{n}}$$

Compare these results with the confidence interval formula (presented in previous unit):

$$\overline{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

➢ In this case, if H_0 is not true and H_1 holds with a specific value of $\mu = \mu_1$, then it is possible to compute the probability of type II error, β.

➢ In the (very realistic) case where σ^2 is not known, but rather estimated by S^2, we would like to replace the test statistic, Z, above with,

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

but in general, T no longer follows a Normal distribution.

➢ Under $H_0 : \mu = \mu_0$, and for moderate or large samples (e.g. $n > 100$) this statistic is approximately Normally distributed just like above. In this case, the procedures above work well.

➢ But for smaller samples, the distribution of T is no longer Normally distributed. Nevertheless, it follows a well known and very famous distribution of classical statistics: **The Student-t Distribution**.

➢ The probability density function of a Student-t Distribution with a parameter k, referred to as **degrees of freedom**, is,

$$f(x) = \frac{\Gamma\big[(k+1)/2\big]}{\sqrt{\pi k}\Gamma(k/2)} \cdot \frac{1}{\Big[\big(x^2/k\big)+1\Big]^{(k+1)/2}} \qquad -\infty < x < \infty,$$

where $\Gamma(\cdot)$ is the Gamma-function. It is a symmetric distribution about 0 and as $k \to \infty$ it approaches a standard Normal distribution.

➢ The following mathematical result makes the t-distribution useful: Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a Normal distribution with mean μ and variance σ^2. The random variable, T has a t distribution with $n-1$ degrees of freedom.

➢ Now, knowing the distribution of T (and noticing it depends on the sample size, n), allows us to construct hypothesis tests and confidence intervals when σ^2 is not known, analogous to the (Z-tests and confidence intervals) presented above.

➢ If $\bar{x}$ and s are the mean and standard deviation of a random sample from a normal distribution with unknown variance σ^2, a $100(1-\alpha)\%$ **confidence interval** on μ is given by

$$\bar{x} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

where $t_{1-\alpha/2,n-1}$ is the $1-\alpha/2$ quantile of the t distribution with $n-1$ degrees of freedom.

➢ A related concept is a $100(1-\alpha)\%$ **prediction interval** (PI) on a single future observation from a normal distribution is given by

$$\bar{x} - t_{1-\alpha/2,n-1}s\sqrt{1+\frac{1}{n}} \le X_{n+1} \le \bar{x} + t_{1-\alpha/2,n-1}s\sqrt{1+\frac{1}{n}}.$$

This is the range where we expect the $n+1$ observation to be, after observing n observations and computing $\overline{x}$ and s.

➢ **Testing Hypotheses on the Mean, Variance Unknown (T-Tests)**

Model: $X_i \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ with both μ and σ^2 unknown.

Null hypothesis: $H_0 : \mu = \mu_0$.

Test statistic: $t = \dfrac{\bar{x}-\mu_0}{s/\sqrt{n}}, \qquad T = \dfrac{\overline{X}-\mu_0}{S/\sqrt{n}}.$

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu \neq \mu_0$	$P = 2\big[1 - F_{n-1}\big(\lvert t\rvert\big)\big]$	$t > t_{1-\alpha/2,n-1}$ or $t < t_{\alpha/2,n-1}$
$H_1 : \mu > \mu_0$	$P = 1 - F_{n-1}\big(t\big)$	$t > t_{1-\alpha,n-1}$
$H_1 : \mu < \mu_0$	$P = F_{n-1}\big(t\big)$	$t < t_{\alpha,n-1}$

Note that here, $F_{n-1}(\cdot)$ denotes the CDF of the t-distribution with $n-1$ degrees of freedom. As opposed to $\Phi(\cdot)$, it is not tabulated in standard tables and like $\Phi(\cdot)$ it cannot be explicitly evaluated. So to calculate P-values, we use software.

8 Two Sample Inference

➢ The setup is a sample $x_1, \ldots, x_{n_1}$ modelled by an i.i.d. sequence of random variables, $X_1, \ldots, X_{n_1}$ and another sample $y_1, \ldots, y_{n_2}$ modelled by an i.i.d. sequence of random variables, $Y_1, \ldots, Y_{n_1}$. Observations, x_i and y_i (for same i) are not paired. In fact, it is possible that $n_1 \neq n_2$ (unequal sample sizes).

➢ The model assumed is, $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2)$, $\quad Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2)$.
Variations are: (i) **equal variances:** $\sigma_1^2 = \sigma_2^2 := \sigma^2$. (ii) **unequal variances:** $\sigma_2^2 \neq \sigma_2^2$.

➢ We could carry single sample inference for each population separately. Specifically, for $\mu_1 = E[X_i]$ and $\mu_2 = E[X_2]$. However we focus on,

$$\Delta_\mu := \mu_1 - \mu_2 = E[X_i] - E[Y_i].$$

For this **difference in means** we can carry out inference jointly.

➢ It is very common to ask if Δ_μ $(=, <, >)$ 0, i.e. if μ_1 $(=, <, >)$ μ_2. But we can also replace the "0" with other values, e.g. $\mu_1 - \mu_2 = \Delta_0$ for some Δ_0.

➢ A point estimator for Δ_μ is $\overline{X} - \overline{Y}$ (difference in sample means). The estimate from the data is denoted by $\overline{x} - \overline{y}$ (the difference in the individual sample means), with,

$$\overline{x} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i, \qquad \overline{y} = \frac{1}{n_2}\sum_{i=1}^{n_2} y_i.$$

➢ In the case (ii) of **unequal variances**: Point estimates for σ_1^2 and σ_2^2 are the individual sample variances,

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_i - \overline{x})^2, \qquad s_2^2 = \frac{1}{n_2 - 2}\sum_{i=1}^{n_2}(y_i - \overline{y})^2.$$

➢ In case (i) of **equal variances**, both S_1^2 and S_2^2 estimate σ^2. In this case, a more reliable estimate can be obtained via the **pooled variance estimator**

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

➢ In case (i), under H_0:

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \quad \sim \quad t\big(n_1 + n_2 - 2\big).$$

That is, the T test statistic follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

➢ In case (ii), under H_0, there is only the approximate distribution,

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}} \quad \sim^{\text{approx}} \quad t\big(v\big).$$

where the degrees of freedom are

$$v = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(s_s^2/n_s\right)^2}{n_s - 1}}.$$

If v is not an integer, may round down to the nearest integer (for using a table).

➢ **Case (i):**
Testing Hypotheses on Differences of Mean, Variance Unknown and Assumed Equal (two sample T-Tests with equal variance)

Model: $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2), \qquad Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2).$

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0.$

Test statistic: $t = \dfrac{\bar{x} - \bar{y} - \Delta_0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}, \qquad T = \dfrac{\overline{X} - \overline{Y} - \Delta_0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}.$

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$P = 2\big[1 - F_{n_1+n_2-2}\big(\lvert t\rvert\big)\big]$	$t > t_{1-\alpha/2,n_1+n_2-2}$ or $t < t_{\alpha/2,n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$P = 1 - F_{n_1+n_2-2}\big(t\big)$	$t > t_{1-\alpha,n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$P = F_{n_1+n_2-2}\big(t\big)$	$t < t_{\alpha,n_1+n_2-2}$

➢ **Case (ii):**
Testing Hypotheses on Differences of Mean, Variance Unknown and NOT Equal (two sample T-Tests with unequal variance)

Model: $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2), \qquad Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2).$

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0.$

Test statistic: $t = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}, \qquad T = \dfrac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}.$

Alternative Hypotheses	***P*-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$P = 2\big[1 - F_v\big(\lvert t\rvert\big)\big]$	$t > t_{1-\alpha/2,v}$ or $t < t_{\alpha/2,v}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$P = 1 - F_v\big(t\big)$	$t > t_{1-\alpha,v}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$P = F_v\big(t\big)$	$t < t_{\alpha,v}$

➢ Case (i) (Equal variances) - confidence interval:

$$\bar{x} - \bar{y} - t_{1-\alpha/2,n_1+n_2-2}\; s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \;\leq\; \mu_1 - \mu_2 \;\leq\; \bar{x} - \bar{y} + t_{1-\alpha/2,n_1+n_2-2}\; s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

➢ Case (ii) (NOT Equal variances) - confidence interval:

$$\bar{x} - \bar{y} - t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \;\leq\; \mu_1 - \mu_2 \;\leq\; \bar{x} - \bar{y} + t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

9 Linear Regression

➢ The collection of statistical tools that are used to model and explore relationships between variables that are related in a nondeterministic manner is called **regression analysis**. Of key importance is the conditional expectation,

$$E(Y \mid x) = \mu_{Y \mid x} = \beta_0 + \beta_1 x \qquad \text{with} \qquad Y = \beta_0 + \beta_1 x + \epsilon,$$

where x is not random and ϵ is a Normal random variable with $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$.

➢ **Simple Linear Regression** is the case where both x and y are scalars, in which case the data is,

$$(x_1, y_1), \ldots, (x_n, y_n).$$

Then given estimates of β_0 and β_1 denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ we have

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad i = 1, 2, \ldots, n,$$

where e_i, are the **residuals** and we can also define the **predicted observation**,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Ideally it would hold that $y_i = \hat{y}_i$ ($e_i = 0$) and thus **total mean squared error**

$$L := SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

would be zero. But in practice, unless $\sigma^2 = 0$ (and all points lie on the same line), we have that $L > 0$.

➢ The standard (classic) way of determining the statistics ($\hat{\beta}_0$, $\hat{\beta}_1$) is by minimisation of L. The solution, called the **least squares estimators** must satisfy

$$\frac{\partial L}{\partial \beta_0}\Big|_{\hat{\beta}_0 \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1}\Big|_{\hat{\beta}_0 \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Simplifying these two equations yields

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

These are called the **least squares normal equations**. The solution to the normal equations results in the **least squares estimators** $\hat{\beta}_0$ and $\hat{\beta}_1$. Using the sample means, $\overline{x}$ and $\overline{y}$ the estimators are,

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}.$$

➢ The following quantities are also of common use:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x^i\right)\left(\sum_{i=1}^{n} y^i\right)}{n}$$

Hence,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

Further,

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \qquad SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

➢ The **Analysis of Variance Identity** is

$$\sum_{i=1}^{n}\Big(y_i - \bar{y}\Big)^2 = \sum_{i=1}^{n}\Big(\hat{y}_i - \bar{y}\Big)^2 + \sum_{i=1}^{n}\Big(y_i - \hat{y}_i\Big)^2$$

or,

$$SS_T = SS_R + SS_E.$$

Also, $SS_R = \hat{\beta}_1 S_{xy}$.

➢ An **Estimator of the Variance**, σ^2 is

$$\hat{\sigma}^2 := MS_E = \frac{SS_E}{n-2}$$

➢ A widely used measure for a regression model is the following ratio of sum of squares, which is often used to judge the adequacy of a regression model:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

$$E\Big(\hat{\beta}_0\Big) = \beta_0, \qquad V\Big(\hat{\beta}_0\Big) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right]$$

$$E\Big(\hat{\beta}_1\Big) = \beta_1, \qquad V\Big(\hat{\beta}_1\Big) = \frac{\sigma^2}{S_{XX}}.$$

➢ In simple linear regression, the **estimated standard error of the slope** and the **estimated standard error of the intercept** are

$$se\Big(\hat{\beta}_1\Big) = \sqrt{\frac{\hat{\sigma}^2}{S_{XX}}} \qquad \text{and} \qquad se\Big(\hat{\beta}_0\Big) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{XX}}\right]}$$

➢ The **Test Statistic for the Slope** is

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{XX}}}$$

$$H_0 : \beta_1 = \beta_{1,0} \qquad H_1 : \beta_1 \neq \beta_{1,0}$$

Under H_0 the test statistic T follows a **t - distribution** with "$n-2$ degree of freedom".

➢ An alternative is to use the F statistic as is common in **ANOVA** (Analysis of Variance) – not covered fully in the course.

$$F = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E}.$$

Under H_0 the test statistic F follows an **F - distribution** with "1 degree of freedom in the numerator and $n-2$ degrees of freedom in the denominator".

Analysis of Variance Table for Testing Significance of Regression

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_E	
Total	SS_T	$n-1$		

➢ There are also confidence intervals for β_0 and β_1 as well as prediction intervals for observations. We don't cover these formulas.

➢ To check the regression model assumptions we plot the residuals e_i and check for (i) Normality. (ii) Constant variance. (iii) Independence.

Logistic Regression:

➢ Take the response variable, Y_i as a Bernoulli random variable. In this case notice that $E(Y) = P(Y = 1)$.

➢ The **logit response function** has the form

$$E\Big(Y\Big) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp\Big(\beta_0 + \beta_1 x\Big)}.$$

➢ Fitting a logistic regression model to data yields estimates of β_0 and β_1.

➢ The following formula is called the **odds**

$$\frac{E\Big(Y\Big)}{1 - E\Big(Y\Big)} = \exp\Big(\beta_0 + \beta_1 x\Big).$$

10 Further Stats Overview

The course covered details associated with most of the key ideas appearing in chapters 1–11 of [MonRun2014], our suggested introductory textbook for statistics and probability, suitable to a general engineering audience. However, we also skipped quite a few sections from these chapters. We also completely omitted chapters 12–15.

Based on the material that we did cover in detail (the bulk of chapters 1–11) and the exercises carried out, the student should be equipped with a basic understanding of the terms and concepts appearing in probability and statistics. In future academic and/or professional tasks that the student will carry out, she may often need to expand her knowledge beyond the basic material covered in detail and perhaps beyond the basic material covered in [MonRun2014].

In this respect, we now overview further aspects of probability and statistics from the book, not covered in detail in the course. The aim is to give general feel about those additional subjects.

Material from Chapters 1–11

- 2–7: **Bayes' Theorem** – This is an elementary addition to the subject of conditional probability that has far reaching consequences. Conditional probability is all about the probability of A given B. With Bayes's theorem, the question becomes, "given that A happened, what do we know about B". That is, the roles of A and B are reversed.

- 3–7: **Geometric and Negative Binomial Distributions** – These discrete probability distributions are friends of the Binomial distribution and appear in the context of repeated Bernoulli trials. The geometric distribution describes the number of attempts/failures until success and it's generalization, the Negative Binomial distribution describes the number of attempts/failures until a given number of successes.

- 3–9: **Poisson Distribution** – This discrete distribution occurs naturally when considering the number of events occurring over a time interval and/or a region in space.

- 3–8: **Hypergeometric Distribution** – This discrete distribution describes the situation of random sampling from a population without replacement. In this respect it is related to the Binomial distribution (can be used to describe sampling from a population with replacement).

- 4–7: **Normal Approximation to the Binomial and Poisson Distributions** – The central limit theorem acts on the binomial and poisson distribution. These two discrete distributions can be approximated by the Normal distribution.

- 4–9: **Erlang and Gamma Distributions** – This continuous asymmetric distribution of a non-negative random variable is useful for modelling non-negative random variables. Erlang is a special case of Gamma appearing as a sum of a finite number of i.i.d. exponential random variables. It generalizes the exponential distribution.

- 4–10: **Weibull Distribution** – This continuous asymmetric distribution of a non-negative random variable is often used in reliability analysis. It generalises the exponential distribution.

- 4–11: **Lognormal Distribution** – This continuous asymmetric distribution of a non-negative random variable appears when transforming a normal random variable, X, by e^X.

- 4–12: **Beta Distribution** – This continuous asymmetric distribution generalizes the uniform distribution allowing the distribution to have other shapes on finite support.

- 5–3.1: **Multinomial Distribution** – This discrete multi-variable distribution generalizes the Binomial distribution to the case of several possible outcomes per Bernoulli trial.

- 5–6: **Moment-Generating Functions** – A moment generating function is another analytical way to view the distribution of a random variable (in addition to the CDF, PDF/PMF).
- 7–3.1: **Unbiased Estimators** – An estimator is unbiased if in expectation it gets the correct value. The study of this subject explains the $n-1$ in the denominator of the sample variance (as opposed to having the more intuitive value n).
- 7–3: **Other general concepts of point estimation** – Point estimation is a key area of statistics and is studied much further in mathematical statistics courses. There are more ways to quantify and understand the behaviour of an estimator.
- 7–4.1: **Method of moments for point estimation** – This method of point estimation works by solving equations specified by moment estimators in the data.
- 7–4.2: **Method of maximum likelihood for point estimation** – This general method of point estimation has much theory associated with it. It is used in Logistic regression in the course. It is a central part of the study of Mathematical statistics.
- 7–4.3: **Bayesian Estimation of Parameters** – Bayesian statistics is a large branch of statistics which allows to incorporate prior beliefs and knowledge in the estimation process. This is often done by employing a **prior distribution** on the unknown parameter and then after carrying out inference obtaining a **posterior distribution** of the unknown parameter. In Bayesian statistics, as opposed to classic frequentist statistics (taught in most of this course), the parameter is always considered a random variable.
- 8–4: **Large Sample Confidence Intervals for a Population Proportion** – This is about estimating a population proportion. For e.g. "proportion of people who support a given candidate". In case of large samples, confidence intervals for the population proportion can be obtained using the Normal approximation to the Binomial distribution.
- 8–6: **Bootstrap Confidence Interval** – A general way of creating approximate confidence intervals based on Monte Carlo simulation sampling of the estimators.
- 9–4: **Tests on the Variance and Standard Deviation of a Normal Distribution** – Methods to check hypothesis about variance parameters of a population. These methods use the **Chi-squared** (χ^2) **distribution**.
- 9–5: **Tests on a Population Proportion** – Methods to make hypothesis about a population proportion as the unknown parameter.
- 9–7: **Testing for Goodness of Fit** – Methods for checking if a postulated distribution agrees with the empirical distribution. Here one common method uses the fact that sums of squared deviations of expected and observed proportions are approximately χ^2 distributed.
- 9–8: **Contingency Table Tests** – A suite of methods that check if different categorial quantities are independent or not. Here again, the χ^2 distribution is used.
- 9–9: **Nonparametric Procedures** – Nonparametric statistics is a large branch of statistics where there is no assumed family of distributions in the model. One method already studied in Unit 5 is the **Kernel Density Estimation**. Some more classic nonparametric methods that can be used as alternatives to the T-test and Z-test are the **signed rank test**. A plus of this method is that it does not assume any specific distribution. A drawback comes with less statistical power.
- 10–3: **Wilcoxon Rank-Sum Nonparametric test for the difference of two means** – This is a another classic nonparametric method that can replace Z-tests and T-tests.

- 10–4: **The paired t-test** – This test comes together with an experimental design techniques that builds on the idea of coupling observations from two treatment (two populations) together so as to reduce variability. For example if doing road testing for wear-and-tear of tires, it means putting on the same car tires of one type on the left side and another type on the right side. This then potentially reduces the variability in differences of wear-and-tear between cars, because variability due to driver/car attributes are cancelled. Mathematically the calculations are identical to a single sample t-test, but conceptually the paired t-test is different. It is an important thing to know for basic experimental design.
- 10–5: **Inferences on the variances of two Normal distributions** – This is a suite of hypothesis tests (and confidence intervals) asking if two populations have the same variance or not. From a model-selection perspective, using these tests can be sometimes useful for choosing a specific model for comparing population means (as covered in the two sample inference, Unit 8).
- 10–6: **Inferences on two population proportions** – This is a suite of hypothesis tests (and confidence intervals) comparing population proportions in two populations.
- 11–9: **Regression on Transformed Variables** – Linear regression on the original variables is often not a sensible assumption. E.g. when there exists a logarithmic or quadratic relationship. By transforming variables, linear regression may sometimes be applied to other models. Another reason for (sometimes) transforming variables is to match the data to the model assumptions (constant variance and Normal residuals).

Material from Further Chapters

- Chapter 12 – **Multiple Linear Regression** – Linear regression goes well beyond finding the relationship of Y on (the scalar) X. In many practical situations, X is taken as a vector and the relationship of different components of this vector of Y are of interest. Multiple linear regression is then used.
- Chapter 13 – **Design and Analysis of Single-Factor Experiments: The Analysis of Variance** – The Analysis of Variance (ANOVA) is a very common method for comparing means of two or more populations under the assumption that variances are equal. It is key to know that this is a method for comparing means, even though the name is variance. The variability between sample means and within sample means is used to test the hypothesis that means are different, or the same. Under the null-hypothesis, stating all means are equal, the test statistic follows the well known F-**distribution**. ANOVA is often taught as part of a basic statistics course and is a very common method in practice. Understanding ANOVA, constitutes the basic understanding leading towards design of experiments. Note the the case of two samples, was covered in Unit 8 of the course (two sample inference). In that case, the T-test described is closely related to ANOVA.
- Chapter 14 – **Design of Experiments with Several Factors** – This is a more advanced aspect of scientific statistics. Often there are several factors that are postulated and trying out all possible combinations is difficult. This branch of design of experiments suggests sampling methods that allow to efficiently (from a statistical perspective) determine the validity of hypothesis. One aspect here is **two-way ANOVA** and other more advanced aspects include factorial designs of experiments.
- Chapter 15 – **Statistical Quality Control** – This branch of industrial statistics is often used in manufacturing/mining industries. The essence is to observe a time-series of statistics such as means, variances and ranges. A key goal is to have criteria for indicating where the "process is out of control" - indicating that some change has happened in the system and it should be further investigated.