STATISTICAL ANALYSIS AND MODELLING

David A. Stephens

Department of Mathematics, Imperial College

d.stephens@imperial.ac.uk
`stats.ma.ic.ac.uk/~das01/EPSCourse/`

4th March 2004

WEEK 1: Statistical Summaries & Statistical Testing

- Types of Study and their Statistical Analysis
- Motivation Elementary numerical and graphical summary methods
- Representing Uncertainty: Standard Deviations and Standard Errors
- Basic elements and logic of Probability Theory
- Statistical Hypothesis Testing: Introduction; One and Two sample tests for Normal samples, Analysis of Variance
- Non-normal/integer valued data
- Non-parametric Tests
- Simulation-based methods
- Bayesian Methods

SECTION 1.

STATISTICAL ANALYSIS

Statistical analysis involves the informal/formal comparison of hypothetical or predicted behaviour with experimental results. For example, we wish to be able to compare the predicted outcomes of an experiment, and the corresponding probability model, with a data histogram.

We will use both *qualitative* and *quantitative* approaches.

Broadly, the “*Scientific Process*” involves several different stages:

THEORETICAL MODELLING

↓

MATHEMATICAL/PROBABILISTIC MODELLING

↓

PREDICTION

↓

EXPERIMENTATION/OBSERVATION

↓

VALIDATION

Mathematical/Probabilistic modelling facilitates **PREDICTION**; *Statistical Analysis* provides the means of **VALIDATION** of predicted behaviour.

1.1 PRELIMINARIES

Suppose that an experiment or **trial** is to be repeated n times under identical conditions. This will result in n data points, possibly representing multiple observations on the same individual, or one observation on many individuals. The data may be

- **univariate** (single variable)
- **multivariate** (several variables)

Let

- X_i denote the result of experiment i **before** it is known
- x_i denote the **observed** result for experiment i

Eventually, we will build **probability models** for the X_i in order to facilitate **inference** (estimation, hypothesis testing, prediction, verification/model validation).

1.1.1 STATISTICAL OBJECTIVES

Suppose that we have observed experimental outcomes

- $x_1, ..., x_n$ on the n trials
- that is, we have observed $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$, termed a **random sample**.

This sample can be used to answer qualitative and quantitative questions about the nature of the experiment being carried out.

The objectives of a statistical analysis can be summarized as follows. We want to, for example,

- **SUMMARY : Describe** and **summarize** the sample $\{x_1, ..., x_n\}$ in such a way that allows a specific probability model to be proposed.

- **INFERENCE : Deduce** and **make inference about** the parameter(s) of the probability model θ.
- **TESTING : Test** whether θ is "**significantly**" larger/smaller/different from some specified value.
- **GOODNESS OF FIT : Test** whether the probability model encapsulated in the mass/density function f, and the other model assumptions are **adequate** to explain the experimental results.

The first objective can be viewed as an **exploratory** data analysis exercise. It is crucially important to understand whether a proposed probability distribution is suitable for modelling the observed data, otherwise the subsequent formal inference procedures (estimation, hypothesis testing, model checking) cannot be used.

In any case it is often useful to a reader to see summary measures of the data, irrespective of any subsequent formal analysis.

1.2 TYPES OF STUDY

The data in an experimental study can be obtained in a number of different situations that can be classified as follows:

- one sample
- two independent samples
- two related samples (“within individuals”)
- two related samples (predictor and response)
- k independent samples
- k related samples (multivariable, within individuals)

1.2.1 ONE SAMPLE

- repeated, independent observations of some phenomenon
- aim to summarize "location/scale" of sample
- test hypothesized *target* values
- test distributional summaries

ONE SAMPLE ANALYSIS

1.2.2 TWO INDEPENDENT SAMPLES

- repeated, independent observations under different conditions *(fixed effects)*
- control/treatment
- healthy/affected
- aim to compare two samples
- same mean level ?
- same variability ?
- same distribution ?

TWO SAMPLE ANALYSIS

1.2.3 TWO RELATED SAMPLES I : PAIRED ANALYSIS

- two repeated observations on same experimental units
- two observations on different but related *(matched)* experimental units
- start/end of trial
- matched/paired analysis
- any change in mean level ?

TWO SAMPLE PAIRED ANALYSIS

1.2.4 TWO RELATED SAMPLES II : PREDICTOR AND RESPONSE

- two related observations on different features of same experimental units
- predictor/response
- objective is to predict response
- normal data/non-normal data
- correlation ?
- any predictive ability ?
- classification ?

REGRESSION ANALYSIS

1.2.5 k INDEPENDENT SAMPLES

- $k \geq 2$ sets of independent observations (fixed effects)
- different experimental conditions (control, level 1,...,level $k - 1$)
- ordered levels ?
- normal/non-normal data ?
- any change in mean measure across treatment levels ?

ANOVA ANALYSIS

1.2.6 k RELATED SAMPLES

- $k \geq 2$ sets of observations (on same experimental units)
- time dependent
- same feature, different experimental conditions (fixed effects)
- different (related) features
- normal/non-normal data ?
- regression/correlation ?
- comparison of fixed effects ?

REPEATED MEASURES/MULTIVARIATE ANALYSIS

1.3 KEY CONSIDERATIONS

ANALYTICAL

- what is the key outcome of interest ?
- can some variables be omitted from the analysis ?
- are all experimental units acceptable for the study ?
- are there biases in the study design ?
- are all sources of variability being acknowledged ?

STATISTICAL

- summary
- inference
- testing
- distributional assumptions
- goodness of fit
- prediction
- study design

1.4 EXPLORATORY DATA ANALYSIS

We wish first to produce summaries of the data in order to convey general trends or features that are present in the sample. Secondly, in order to propose an appropriate probability model, we seek to **match** features in the observed data to features of one of the conventional probability distributions that may be used in more formal analysis. The four principal features that we need to assess in the data sample are

(1) The **location**, or "average value" in the sample.
(2) The **mode**, or "most common" value in the sample.
(3) The **scale** or **spread** in the sample.
(4) The **skewness** or **asymmetry** in the sample.

These features of the sample are important because we can relate them **directly** to features of probability distributions.

1.4.1 NUMERICAL SUMMARIES

The following quantities are useful numerical summary quantities

- Sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample variance: either (S^2 or s^2 may be used)

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Sample quantiles: sort data into ascending order and re-labelled

$$x_{(1)} < ... < x_{(n)}$$

then $x_{(i)}$ is the i/n^{th} sample quantile

$$\begin{array}{ccl} \text{Median} & m & = x^{(50)},\ \text{the 50th quantile} \\ \text{Lower quartile} & q_{25} & = x^{(25)},\ \text{the 25th quantile} \\ \text{Upper quartile} & q_{75} & = x^{(75)},\ \text{the 75th quantile} \\ & & \\ \text{Inter-quartile range} & IQR & = q_{75} - q_{25} \\ & & \\ \text{Sample minimum} & x_{\min} & = x_{(1)} \\ \text{Sample maximum} & x_{\max} & = x_{(n)} \\ \text{Sample range} & R & = x_{(n)} - x_{(1)} \end{array}$$

- Sample skewness

$$\kappa = \frac{1}{nS^2} \sum_{i=1}^{n} (x_i - \bar{x})^3$$

NOTE: Key aspects of the sample can be summarized using the first four **sample moments** and their transformations

- 1st Moment⇀**LOCATION** : $\frac{1}{n}\sum_{i=1}^{n} x_i$
- 2nd Moment⇀**SCALE** : $\frac{1}{n}\sum_{i=1}^{n} x_i^2$
- 3rd Moment⇀**SKEWNESS** : $\frac{1}{n}\sum_{i=1}^{n} x_i^3$
- 4th Moment⇀**KURTOSIS** (“heavy-tailedness”) : $\frac{1}{n}\sum_{i=1}^{n} x_i^4$

1.4.2 REPORTING UNCERTAINTY

It is common to report a sample mean and variance, $\overline{x}$,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

and, in addition, a **standard error of the mean**

$$SEM = \frac{s}{\sqrt{n}}.$$

But

- what is this quantity ?
- why this formula ?
- what if the data are **proportions**, or **counts out of** m **?**

For proportions, with x positive results out of n, then the estimate of the proportion is

$$\frac{x}{n}$$

and the **standard error** of this estimate is

$$\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} = \sqrt{\frac{x\,(n-x)}{n^3}}$$

Note that

- this is strictly an **estimated standard error**
- **all** statistics (sample median, sample skewness, sample standard deviation etc.) have an associated standard error !

It is common to report

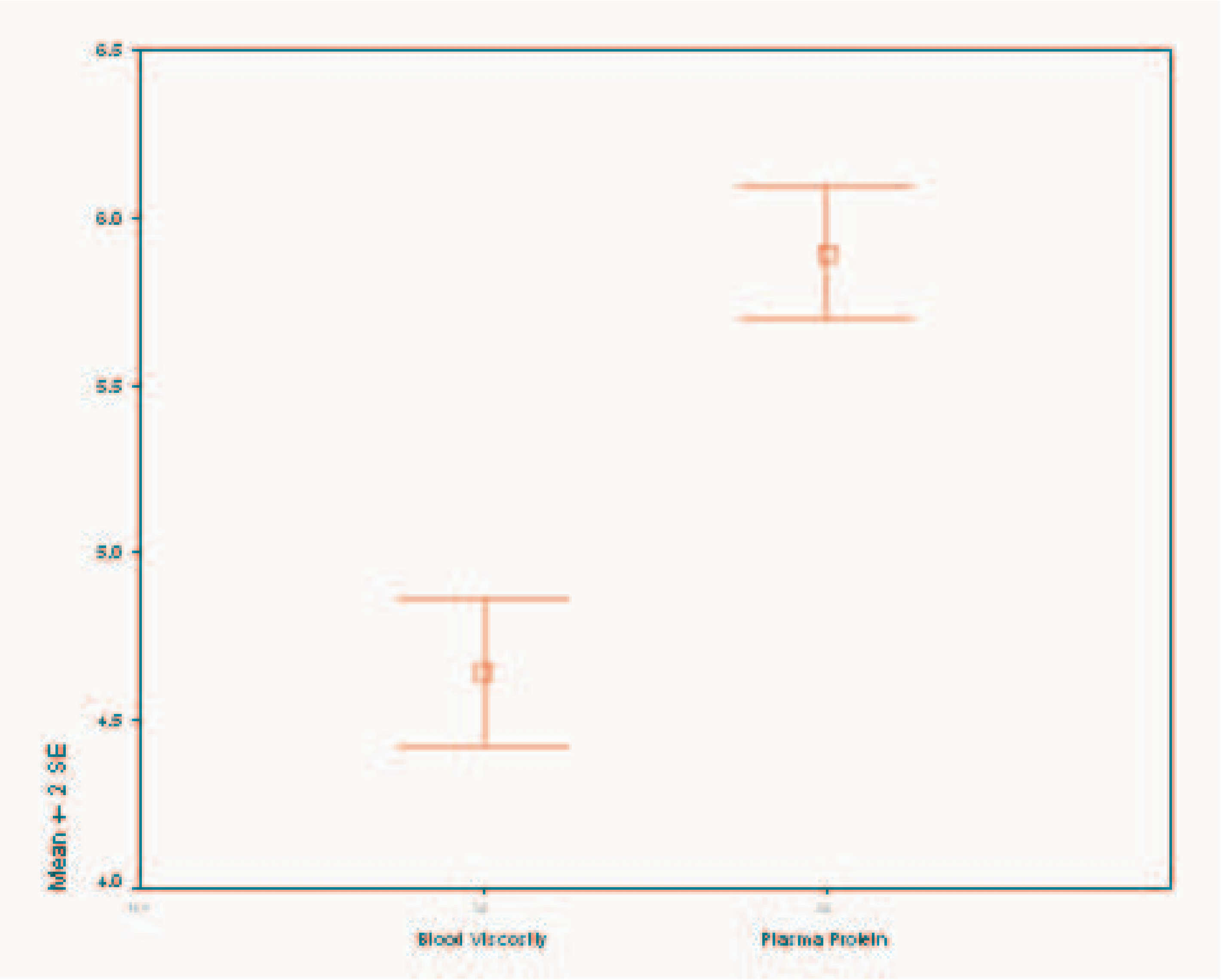
$$\bar{x} \pm SEM$$

as a sample summary. However, it might be more appropriate to report a **confidence interval**

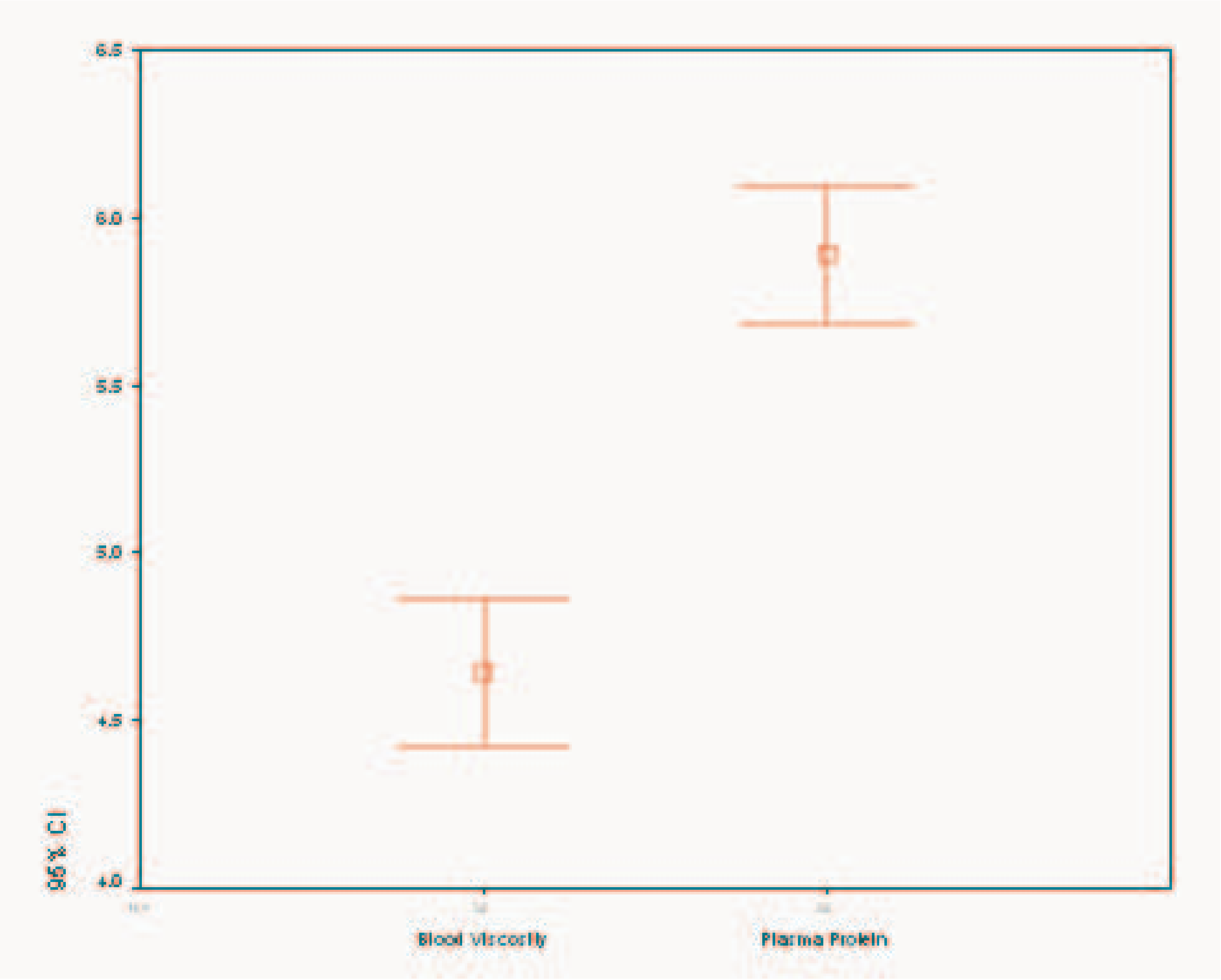
$$\bar{x} \pm 1.96 \times SEM$$

- what is the difference ?
- when is this formula valid ?
- why this formula ?

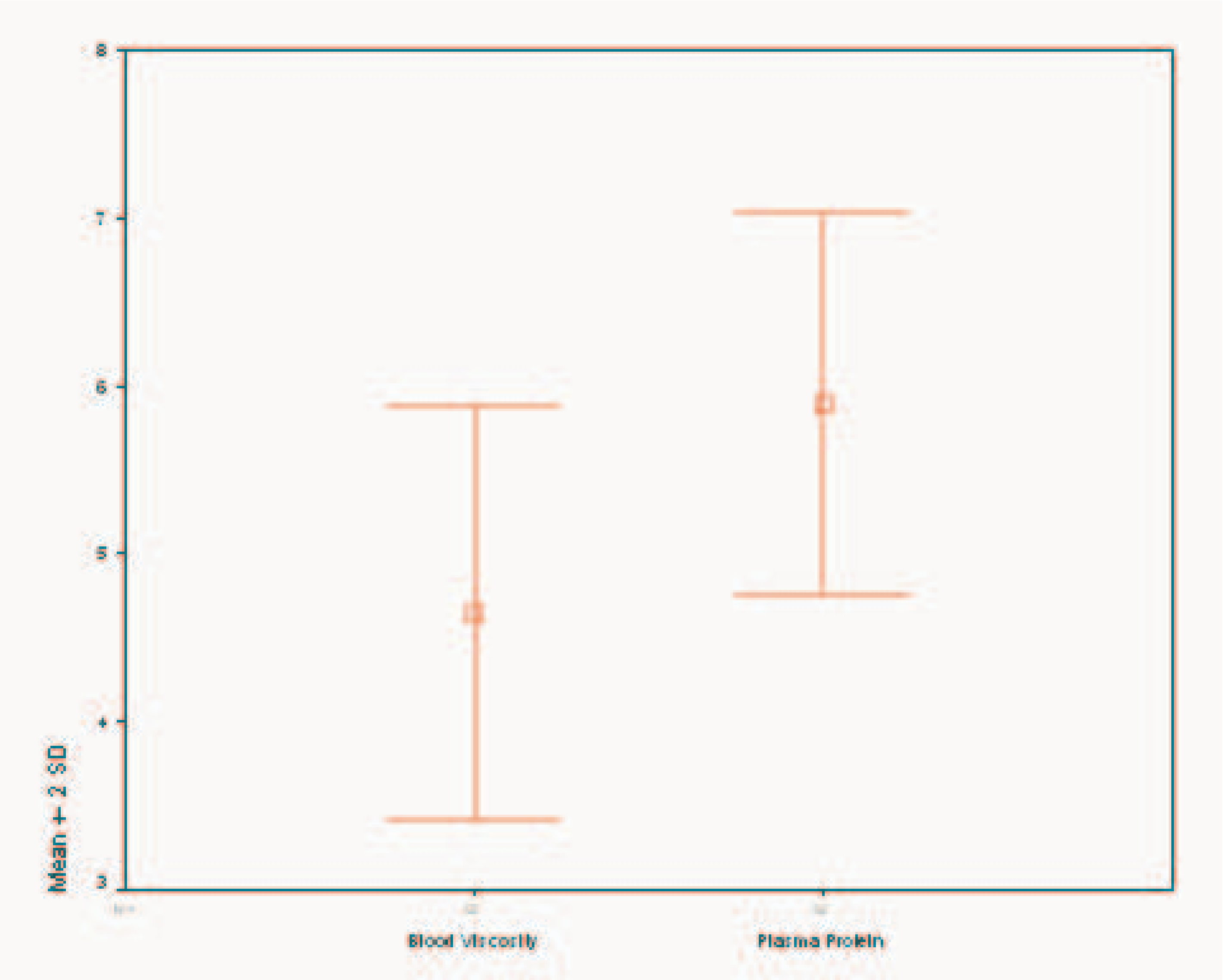
To understand the distinction, some results from probability theory are needed (see section 2.7)



MEAN±2×S.E.



CONFIDENCE INTERVAL

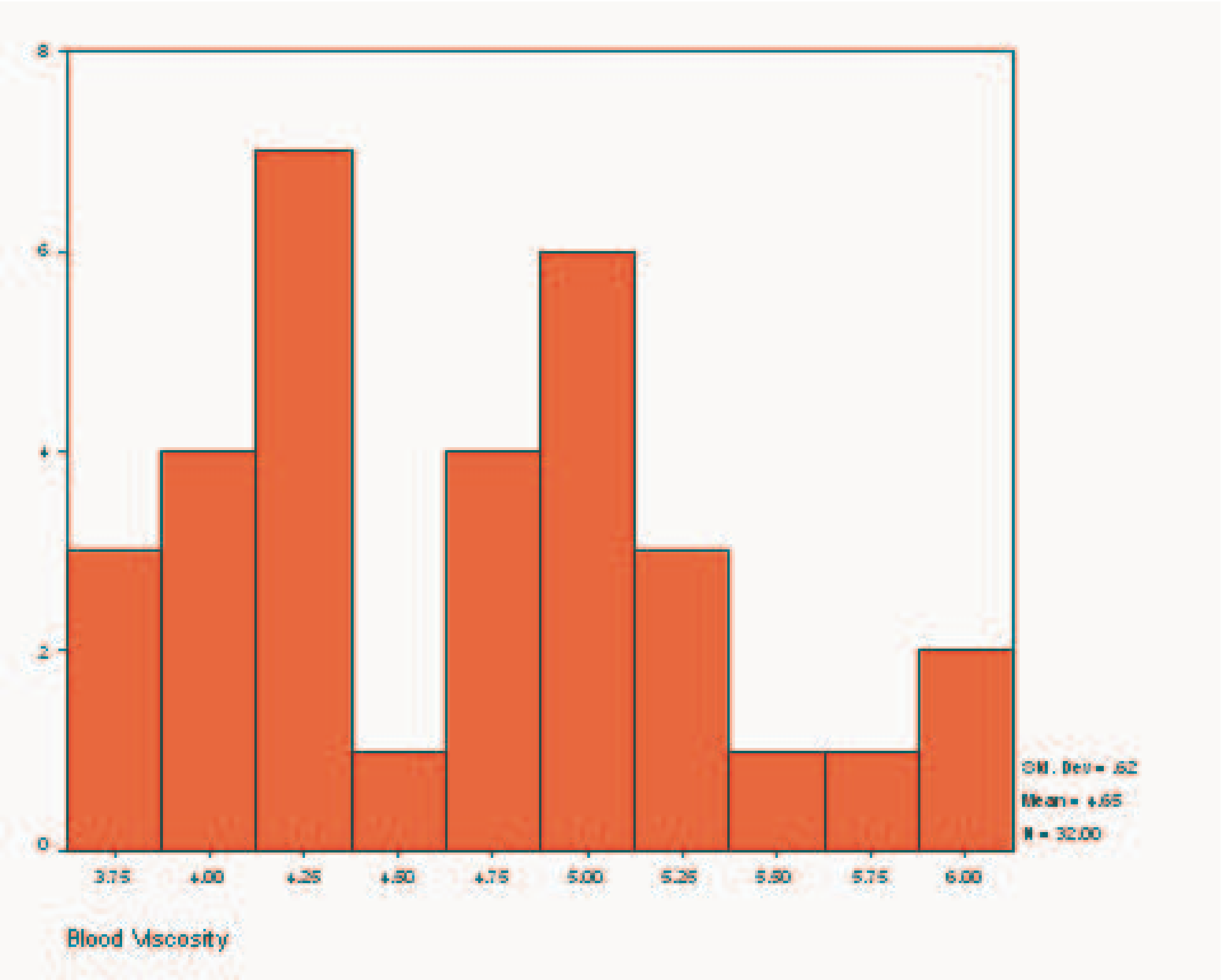


MEAN±2×S.D.

1.4.3 GRAPHICAL SUMMARIES

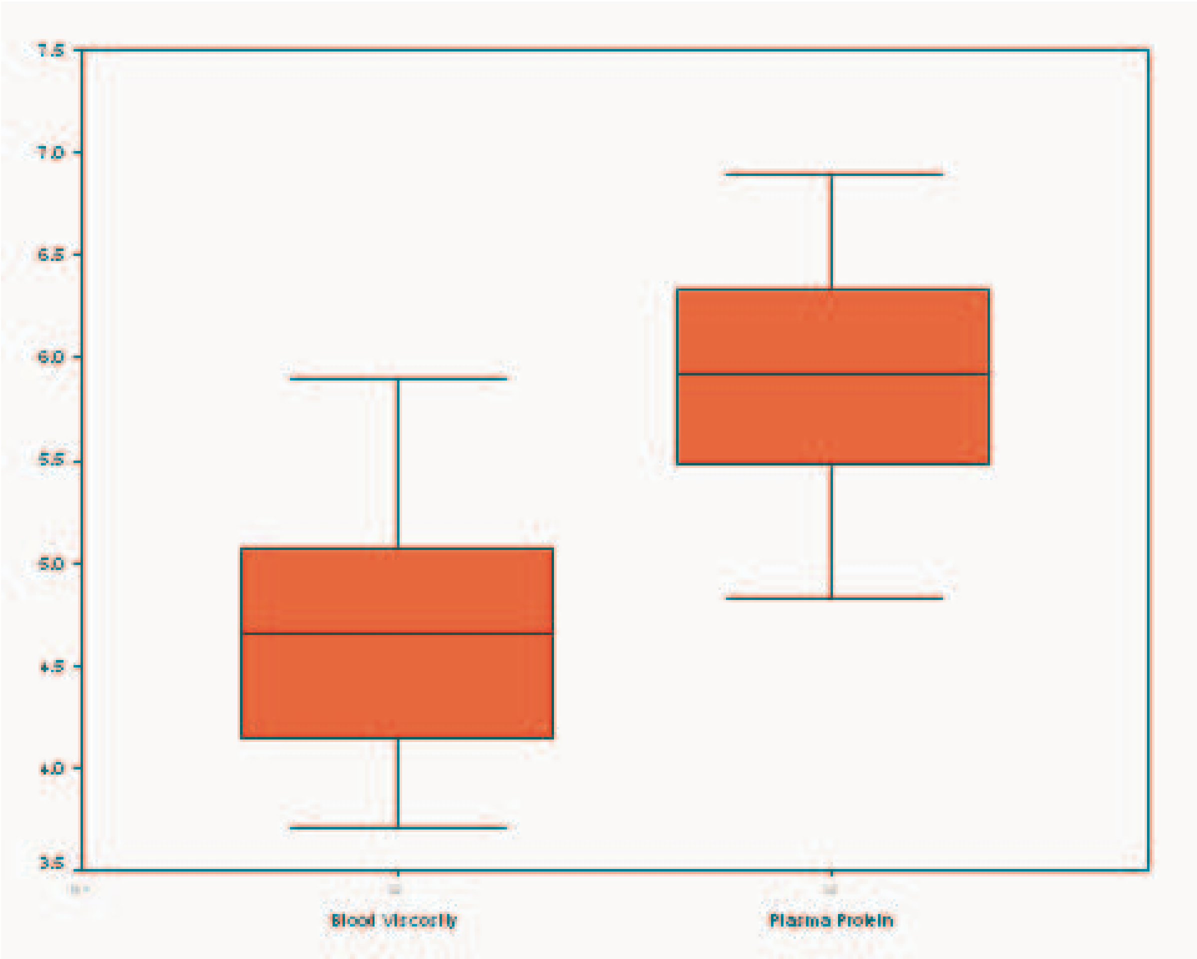
- **HISTOGRAMS**: The most common graphical summary technique is the **histogram**. Typically, the observation range, $\mathbb{X}$, is divided into a number of **bins**, $\mathbb{X}_1, ..., \mathbb{X}_H$ say, and the **frequency** with which a data value in the sample is observed to lie in subset $h = 1, ..., H$ is noted. This procedure leads to a set of counts $n_1, ..., n_H$ (where $n_1 + ... + n_H = n$) which are then plotted on a graph as **bars**, where the hth bar has height n_h and occupies the region of $\mathbb{X}$ corresponding to $\mathbb{X}_h$.

  The histogram again aims to approximate the "true" probability distribution generating the data by the observed sample distribution. It illustrates the concepts of **location**, **mode**, **spread** and **skewness** and general shape features that have been recognized as important features of probability distributions.


8
6
4
2
0
3.75
4.00
4.25
4.50
4.75
5.00
5.25
5.50
5.75
6.00
Blood Viscosity
Std. Dev = .62
Mean = 4.65
N = 32.00

HISTOGRAM

- **BOXPLOTS:** A **boxplot** is a simple way of displaying the variation in a number of data subgroups, or a mean/sem range, or a confidence interval. Typically, a **three point** (min, median, max) or **five point** (min, lower quartile, median, upper quartile, max) summary is used, and often outlying observations are included. The exact form varies from package to package; in SPSS, the following features are plotted
  - The **median** (horizontal line)
  - The **box (**the lower and upper quartiles, or **hinges**)
  - The **whiskers**
  - The **fences (**lower and upper horizontal lines, the smallest and largest values that are not **outliers** or **extreme values**)
  - **outliers** (plotted as circles, more than 1.5 **box** lengths above the **box**)
  - **extreme values** (plotted as asterisks, more than 3.0 **box** lengths above the **box**)



BOXPLOT for two groups of observations

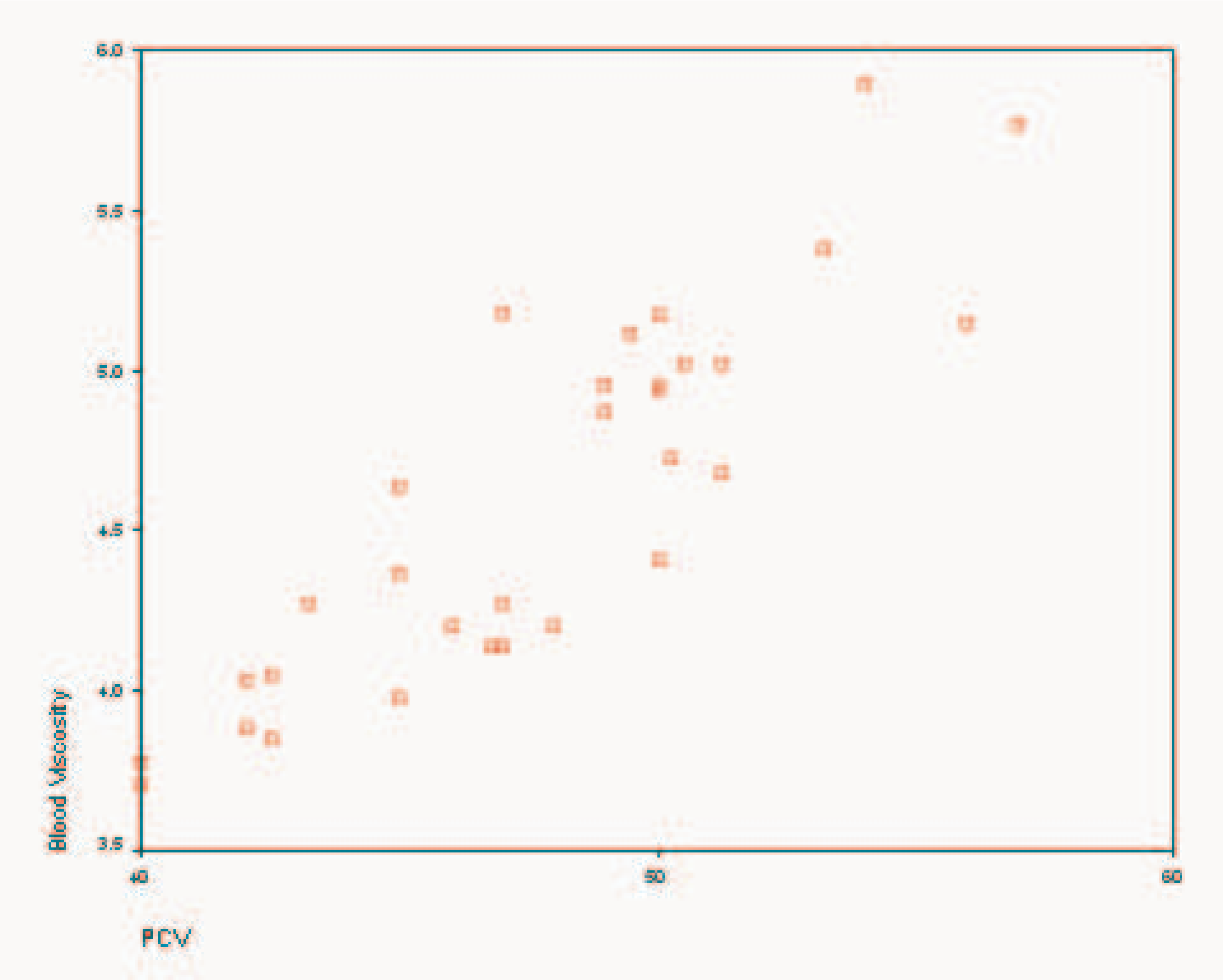
- **SCATTERPLOTS**: Scatterplots are used to illustrate the relationships between variables, and can be useful in discovering

  CORRELATION

  DEPENDENCE

  ASSOCIATION

  between variables



SCATTER PLOT

1.4.4 OUTLIERS

Sometimes, for example due to slight variation in experimental conditions, one or two values in the sample may be much larger or much smaller in magnitude than the remainder of the sample. Such observations are termed **outliers** and must be treated with care, as they can distort the impression given by some of the summary statistics.

For example, the **sample mean** and **sample variance** are extremely sensitive to the presence of outliers in the sample. Other summary statistics, for example those based on sample percentiles (median, quartiles) are less sensitive to outliers. Outliers can usually be identified by inspection of the raw data, or from careful plotting of histograms, or using boxplots.

1.5 TRANSFORMATIONS

It may be necessary or advantageous to consider data **transformations;**

- $y_i = \log_{10} x_i$
- $y_i = \log x_i = \ln x_i$
- $y_i = \sqrt{x_i} = x_i^{1/2}$
- $y_i = x_i^{\alpha}$ some α
- $y_i = \log\left(\dfrac{x_i}{1 - x_i}\right)$

NOTE: This is not any form of statistical trickery, but may be necessary to allow formal statistical assessment

SECTION 2.
PROBABILITY THEORY

2.1 MOTIVATION

The random variation associated with “measurement” procedures in a scientific analysis requires a framework in which the **uncertainty** and **variability** that are inherent in the procedure can be handled. The key goal of Probability and Statistical modelling are to establish a mathematical framework within which *random* variation (due to, for example, experimental error or natural variation) can be quantified so that *systematic* variation (arising due to potentially important biological differences) can be studied.

KEY QUESTION:

Is the result we observe the result of a **genuine, systematic** phenomenon, or is it the product of entirely **random** variation ?

To explain the variation in observed data, we need to introduce the concept of a *probability distribution*. Essentially we need to be able to model, or specify, or compute the "chance" of observing the data that we collect or expect to collect. This will then allow us to assess how likely the data were to occur by chance alone, that is, how "surprising" the observed data are in light of an assumed theoretical model.

2.2 BASIC PROBABILITY CONCEPTS

EXPERIMENTS AND EVENTS

An **experiment** is any procedure

(a) with a well-defined **set** of possible outcomes - the **sample space**, S.
(b) whose **actual** outcome is not known in advance.

A **sample outcome**, s, is precisely one of the possible outcomes of the experiment.

The **sample space**, S, is the entire set of possible outcomes.

Probability Theory is concerned with assigning "weights" or "probabilities" to sets of possible outcomes.

SIMPLE EXAMPLES:

(a) Coin tossing: $S = \{H, T\}$.

(b) Dice : $S = \{1, 2, 3, 4, 5, 6\}$.

(c) Proportions: $S = \{x : 0 \leq x \leq 1\}$

(d) Time measurement: $S = \{x : x > 0\} = \mathbb{R}^+$

(e) Temperature measurement: $S = \{x : a \leq x \leq b\} \subseteq \mathbb{R}$

There are two basic types of experiment, namely

COUNTING

and

MEASUREMENT

- we shall see that these two types lead to two distinct ways of specifying probability distributions.

The collection of sample outcomes is a **set (**a collection of items) , so we write

$$s \in S$$

if *s is a member of the set S.*

DEFINITION

An **event** E is a set of the possible outcomes of the experiment, that is E is a **subset** of S, $E \subseteq S$, E **occurs** if the actual outcome is in this set.

NOTE: the sets S and E can be either be written as a list of items, for example,

$$E = \{s_1, s_2, ..., s_n, ...\}$$

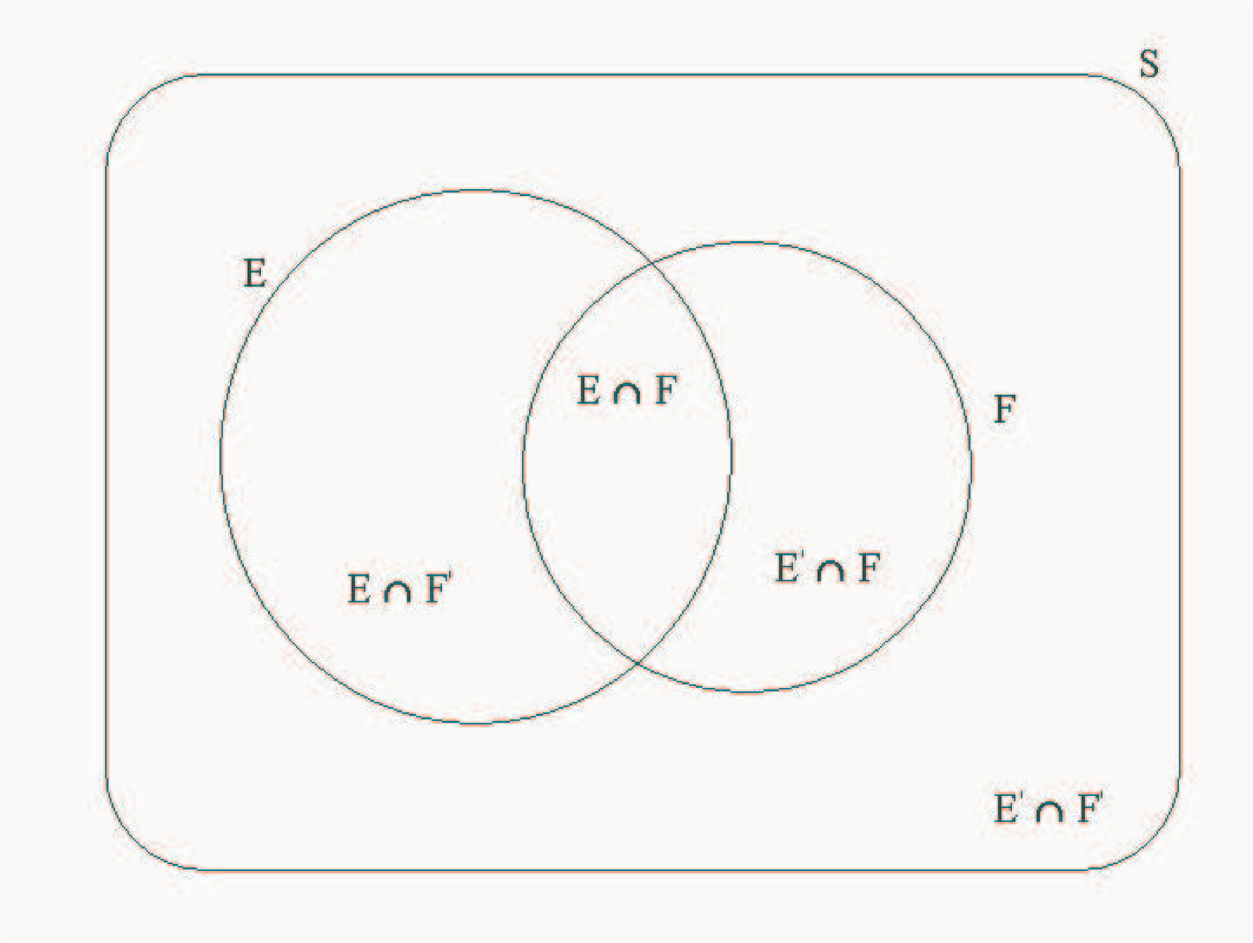
which may a finite or infinite list, or can only be represented by a continuum of outcomes, for example

$$E = \{x : 0.6 < x \leq 2.3\}$$

Events are manipulated using **set theory** notation; if E, F are two events, $E, F \subseteq S$,

Union	$E \cup F$	"E **or** F **or both** occurs"
Intersection	$E \cap F$	"E **and** F occur"
Complement	$E^{'}$	"E **does not** occur"

We can interpret the events $E \cup F$, $E \cap F$, and $E^{'}$ in terms of collections of sample outcomes, and use **Venn Diagrams** to represent these concepts.



Venn Diagram

Another representation for this two event situation is given by the following table:

	E	E'
F	$E \cap F$	$E' \cap F$
F'	$E \cap F'$	$E' \cap F'$

so that, taking unions in the columns

$$(E \cap F) \cup (E \cap F') \equiv E$$

$$(E' \cap F) \cup (E' \cap F') \equiv E'$$

and, taking unions in the rows

$$(E \cap F) \cup (E' \cap F) = F$$

$$(E \cap F') \cup (E' \cap F') = F'$$

Special cases of events:

THE IMPOSSIBLE EVENT – $\emptyset$
the empty set, the collection of sample outcomes with zero elements
THE CERTAIN EVENT – Ω
the collection of all sample outcomes

DEFINITION

Events E and F are **mutually exclusive** if

$$E \cap F = \emptyset$$

that is, the collections of sample outcomes E and F have no element in common.
Mutually exclusive events are very important in probability and statistics, as they allow complicated events to be simplified in such a way as to allow straightforward probability calculations to be made.

2.3 THE RULES OF PROBABILITY

We require that the probability function $P(.)$ must satisfy the following properties:

For any events E and F in sample space S,

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

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

(3) If $E \cap F = \emptyset$, then $P(E \cup F) = P(E) + P(F)$

For the general two event situation:

	E	E'	Sum
F	$P(E \cap F)$	$P(E' \cap F)$	$P(F)$
F'	$P(E \cap F')$	$P(E' \cap F')$	$P(F')$
Sum	$P(E)$	$P(E')$	

so that, summing in the columns

$$P(E \cap F) + P(E \cap F') = P(E)$$

$$P(E' \cap F) + P(E' \cap F') = P(E')$$

and summing in the rows

$$P(E \cap F) + P(E' \cap F) = P(F)$$

$$P(E \cap F') + P(E' \cap F') = P(F')$$

A common type of statistical analysis investigates the analysis of 2 × 2 tables, where the entries in a table correspond to counts of occurrences of particular cross-classified observations.

COLUMNS: Treatment 1/Treatment 2

ROWS : Outcome 1/Outcome 2

	TMT 1	TMT 2
OUTCOME 1	n_{11}	n_{12}
OUTCOME 2	n_{21}	n_{22}

There are many types of analysis that can be performed on these data,

EXAMPLE CALCULATION Examination Pass Rates

The examination performance of students in a year of eight hundred students is to be studied: a student either chooses an essay paper or a multiple choice test. The pass figures and rates are given in the table below:

	PASS	FAIL	PASS RATE
FEMALE	200	200	0.5
MALE	240	160	0.6

The result of this study is clear: the pass rate for MALES is higher than that for FEMALES.

Further investigation revealed a more complex result: for the essay paper, the results were as follows;

	PASS	FAIL	PASS RATE
FEMALE	120	180	0.4
MALE	30	70	0.3

so the pass rate for FEMALES is higher than that for MALES.

For the multiple choice test, the results were as follows;

	PASS	FAIL	PASS RATE
FEMALE	80	20	0.8
MALE	210	90	0.7

so, again, the pass rate for FEMALES is higher than that for MALES.

Hence we conclude that FEMALES have a higher pass rate on the essay paper, and FEMALES have a higher pass rate on the multiple choice test, but MALES have a higher pass rate overall.

IS THIS A CONTRADICTORY RESULT ?

In fact, this apparent contradiction can be resolved by careful use of the probability definitions. First introduce notation; let E be the event that the student chooses an essay, F be the event that the student is female, and G be the event that the student passes the selected paper.

A REAL EXAMPLE: Reintjes R., de Boer A, van Pelt W, Mintjes-de Groot J. Simpson's Paradox: an example from hospital epidemiology. *Epidemiology* 2000; **11**: 81-83

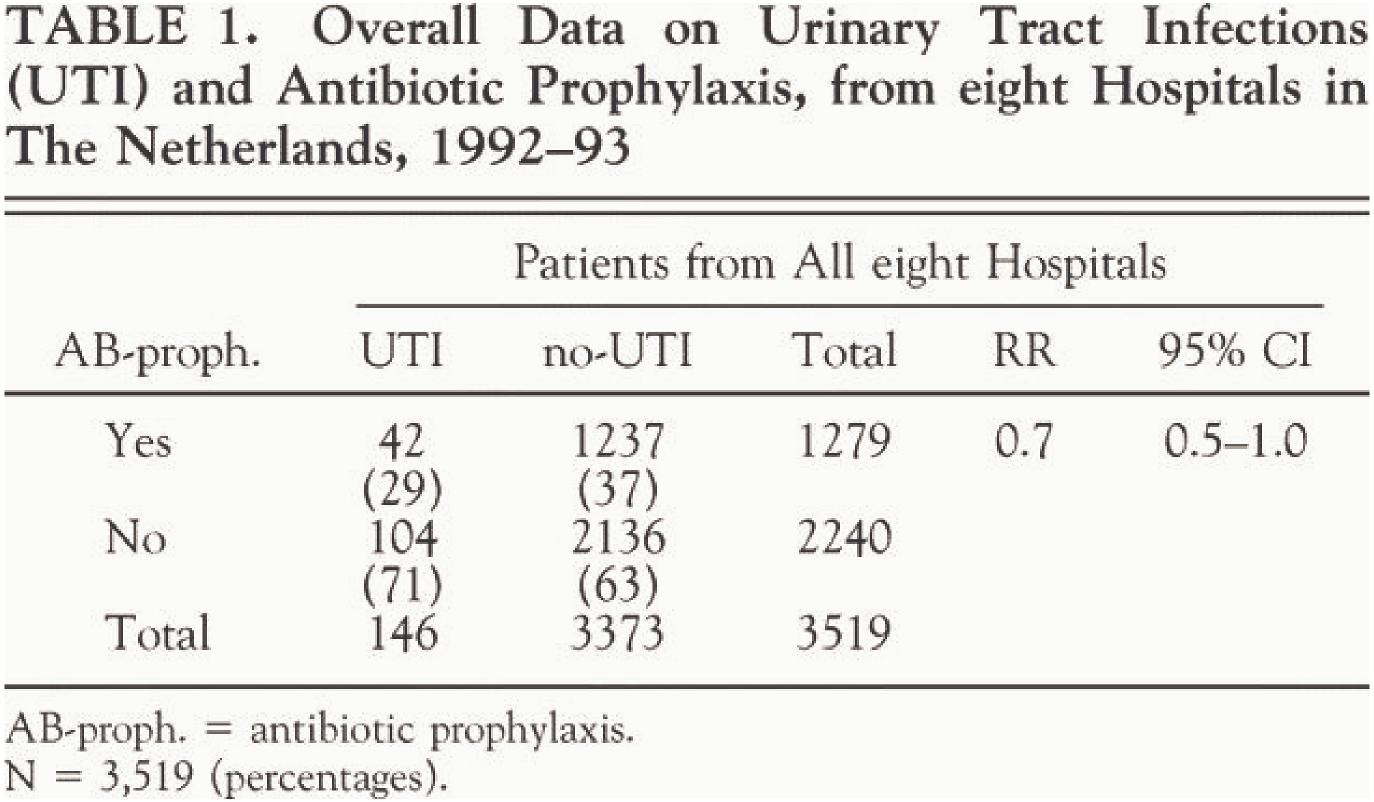


TABLE 1. Overall Data on Urinary Tract Infections (UTI) and Antibiotic Prophylaxis, from eight Hospitals in The Netherlands, 1992–93

	Patients from All eight Hospitals				
AB-proph.	UTI	no-UTI	Total	RR	95% CI
Yes	42 (29)	1237 (37)	1279	0.7	0.5–1.0
No	104 (71)	2136 (63)	2240		
Total	146	3373	3519		

AB-proph. = antibiotic prophylaxis.
N = 3,519 (percentages).

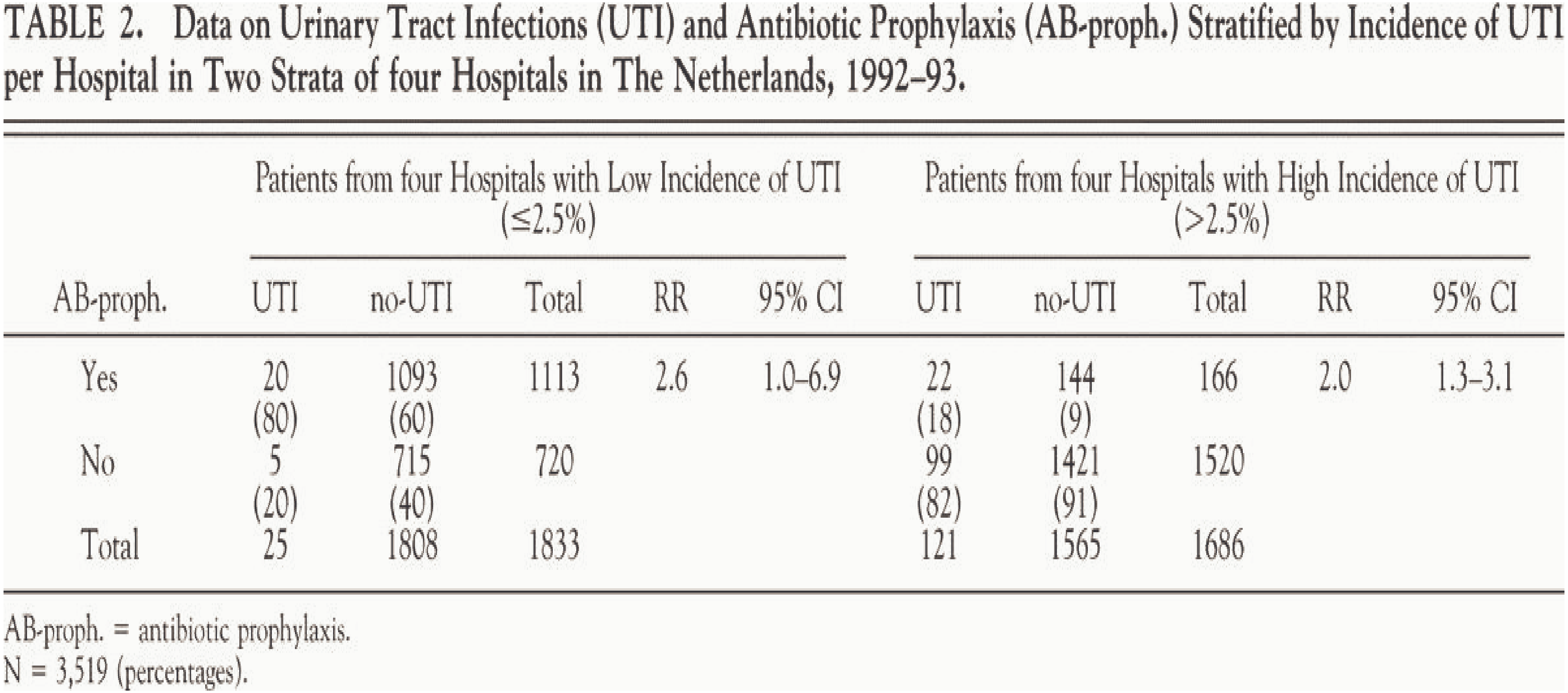


TABLE 2. Data on Urinary Tract Infections (UTI) and Antibiotic Prophylaxis (AB-proph.) Stratified by Incidence of UTI per Hospital in Two Strata of four Hospitals in The Netherlands, 1992–93.

	Patients from four Hospitals with Low Incidence of UTI (≤2.5%)					Patients from four Hospitals with High Incidence of UTI (>2.5%)				
AB-proph.	UTI	no-UTI	Total	RR	95% CI	UTI	no-UTI	Total	RR	95% CI
Yes	20 (80)	1093 (60)	1113	2.6	1.0–6.9	22 (18)	144 (9)	166	2.0	1.3–3.1
No	5 (20)	715 (40)	720			99 (82)	1421 (91)	1520		
Total	25	1808	1833			121	1565	1686		

AB-proph. = antibiotic prophylaxis.
N = 3,519 (percentages).

THIS RESULT IS IMPORTANT FOR MANY TYPES OF STATISTICAL ANALYSIS.

WE MUST TAKE CARE TO ENSURE THAT ANY REPORTED SYSTEMATIC VARIATION IS DUE TO THE SOURCE TO WHICH IT IS ATTRIBUTED, AND NOT DUE TO HIDDEN, CONFOUNDING FACTORS.

2.4 CONDITIONAL PROBABILITY

DEFINITION

For two events E and F with $P(F) > 0$, the **conditional probability** that E occurs, **given** that F occurs, is written $P(E|F)$, and is defined by

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \qquad \text{so that} \qquad P(E \cap F) = P(E|F)P(F)$$

It is easy to show that this new probability operator $P(\ .\ |\ .\ )$ satisfies the probability axioms.

[In the exam results problem, what we really have specified are conditional probabilities. From the pooled table, we have

$$P(G|F) = 0.5 \qquad P(G|F^{'}) = 0.6,$$

from the essay results table, we have

$$P(G|E \cap F) = 0.4 \qquad P(G|E \cap F^{'}) = 0.3,$$

and from the multiple choice table, we have

$$P(G|E^{'} \cap F) = 0.8 \qquad P(G|E^{'} \cap F^{'}) = 0.7$$

and so interpretation is more complicated than originally thought.]

The probability of the **intersection** of events $E_1, ..., E_k$ is given by the **chain rule**

$$P(E_1 \cap ... \cap E_k) \;=\; P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2)...P(E_k|E_1 \cap E_2 \cap ... \cap E_{k-1})$$

Special Case: Independence

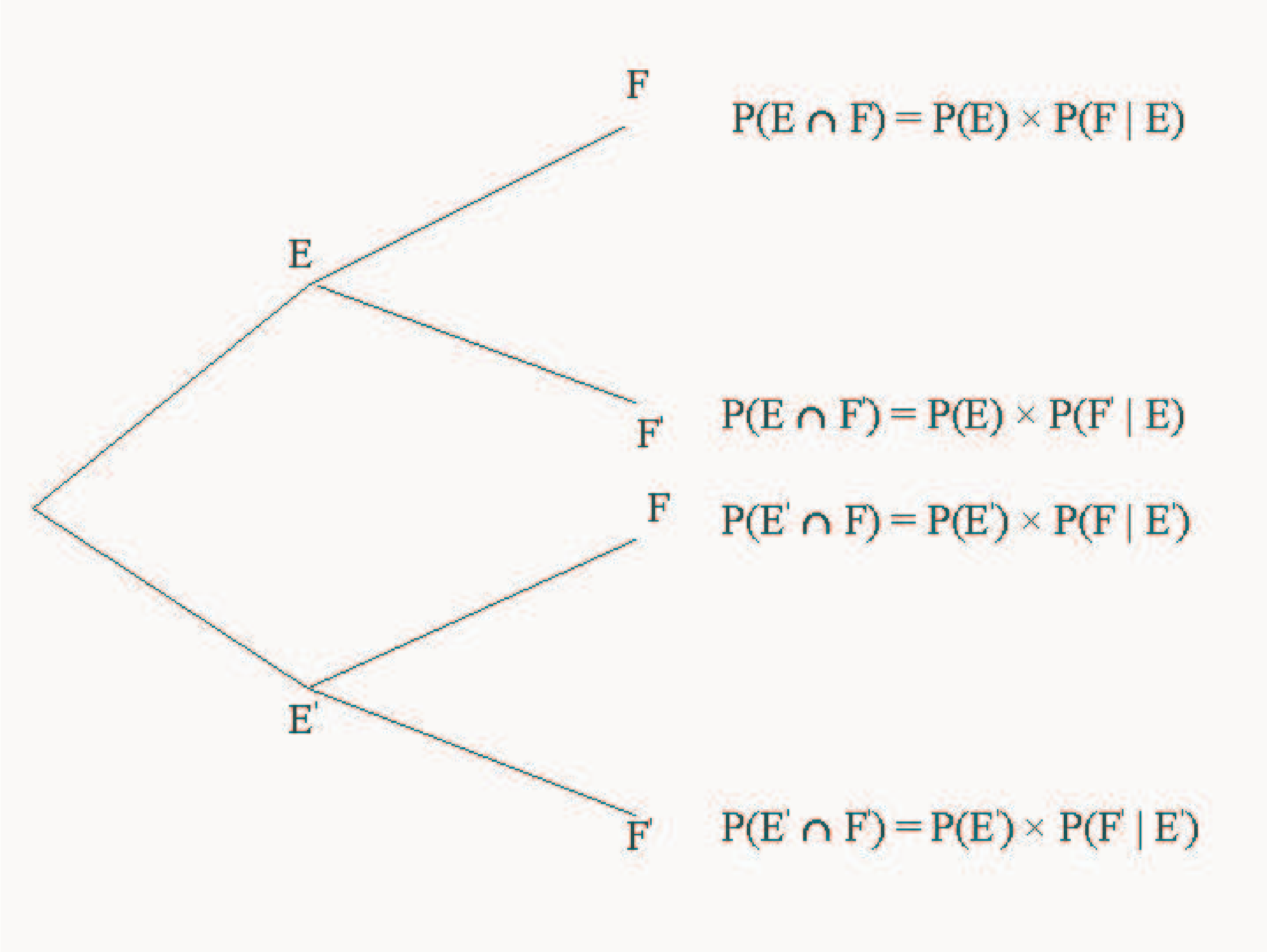
Events F and F are **independent** if

$$P(E|F) \;=\; P(E) \text{ so that } P(E \cap F) \;=\; P(E)P(F)$$

and so if $E_1, ..., E_k$ are independent events, then

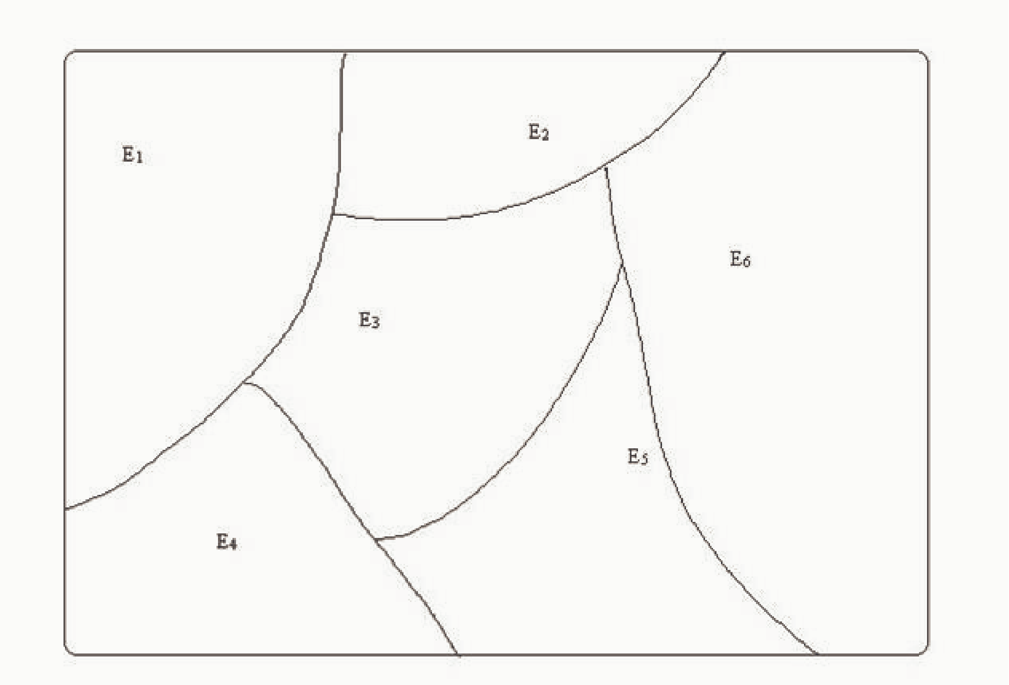
$$P(E_1 \cap ... \cap E_k) = \prod_{i=1}^{k} P(E_i) = P(E_1)...P(E_k)$$

A simple way to think about joint and conditional probability is via a probability tree:

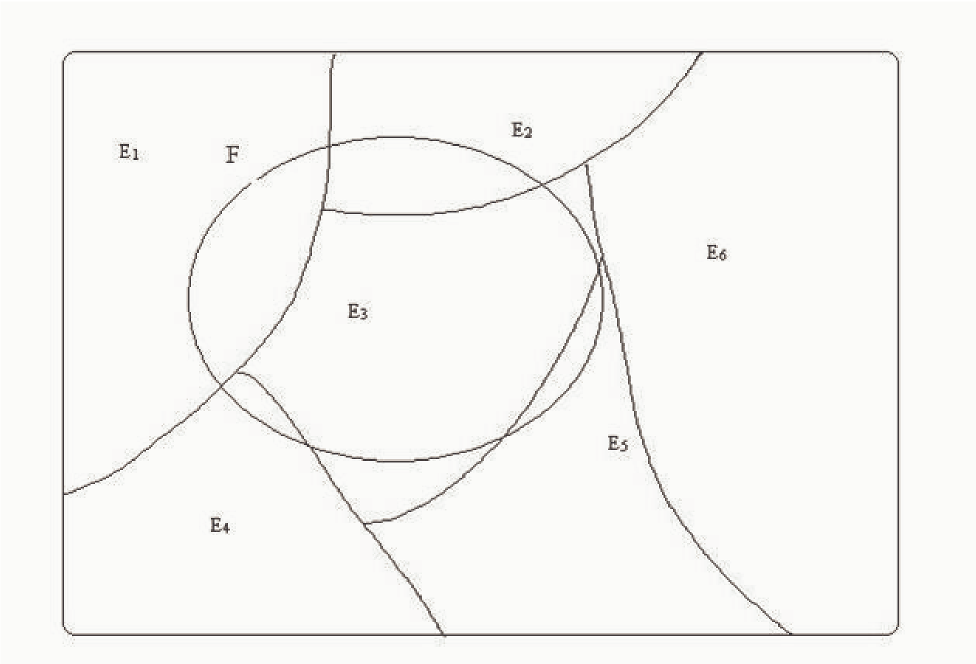

E
E'
F
F'
F
F'
$P(E \cap F) = P(E) \times P(F \mid E)$
$P(E \cap F') = P(E) \times P(F' \mid E)$
$P(E' \cap F) = P(E') \times P(F \mid E')$
$P(E' \cap F') = P(E') \times P(F' \mid E')$

Probability Tree for the Theorem of Total Probability

2.5 PARTITIONS



A partition of S



A partition of $F \subset S$ implied by the partition of S

2.6 TOTAL PROBABILITY

If events $E_1, ..., E_k$ form a **partition** of event $F \subseteq S$, and event $G \subseteq S$ is such that $\mathrm{P}(G) > 0$, then

$$\begin{aligned} P(F) &= \sum_{i=1}^{k} P(F|E_i)P(E_i) \\ P(F|G) &= \sum_{i=1}^{k} P(F|E_i \cap G)P(E_i|G) \end{aligned}$$

The results follows as

$$F = \bigcup_{i=1}^{k}(E_i \cap F) \Longrightarrow P(F) = \sum_{i=1}^{k} P(E_i \cap F) = \sum_{i=1}^{k} P(F|E_i)P(E_i)$$

2.7 BAYES THEOREM

For events E and F such that P(E), P(F) > 0,

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

If events $E_1, ..., E_k$ form a partition of S, with $P(E_i) > 0$ for all i, then

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{P(F)} = \frac{P(F|E_i)P(E_i)}{\displaystyle\sum_{j=1}^{k} P(F|E_j)P(E_j)}$$

This result follows immediately from the conditional probability definition:

$$P(E \cap F) = P(E|F)P(F) \qquad \text{and} \qquad P(E \cap F) = P(F|E)P(E)$$

Note that in the second part of the theorem,

$$P(E_i|F) = \frac{P(F|E_i)P(E_i)}{P(F)} = \frac{P(F|E_i)}{P(F)}\ P(E_i)$$

so the probabilities $P(E_i)$ are re-scaled to $P(E_i|F)$ by conditioning on F. Note that

$$\sum_{i=1}^{k} P(E_i|F) = 1$$

This theorem is very important because, in general,

$$P(E|F) \neq P(F|E)$$

and it is crucial to condition on the correct event in a conditional probability calculation.

EXAMPLE Lie-detector test.

In an attempt to achieve a criminal conviction, a lie-detector test is used to determine the guilt of a suspect. Let G be the event that the suspect is guilty, and let T be the event that the suspect fails the test.

The test is regarded as a good way of determining guilt, because laboratory testing indicate that the detection rates are high; for example it is known that

$$\begin{aligned} P[\text{ Suspect Fails Test } | \text{ Suspect is Guilty }] &= P(T|G) \\ &= 0.95 = 1 - \alpha, \text{ say} \end{aligned}$$

$$\begin{aligned} P[\text{ Suspect Passes Test } | \text{ Suspect is Not Guilty }] &= P(T^{'}|G^{'}) \\ &= 0.99 = \beta, \text{ say} \end{aligned}$$

Suppose that the suspect fails the test. What can be concluded ?

The probability of real interest is $P(G|T)$; we do not have this probability but can compute it using Bayes Theorem. For example, we have

$$P(G|T) = \frac{P(T|G)P(G)}{P(T)}$$

where $P(G)$ is not yet specified, but $P(T)$ can be computed using the Theorem of Total probability, that is,

$$P(T) = P(T|G)P(G) + P(T|G^{'})P(G^{'})$$

so that

$$P(G|T) = \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|G')P(G')}$$

Clearly, the probability $P(G)$, the probability that the suspect is guilty *before* the test is carried out, plays a crucial role. Suppose, that $P(G) = p = 0.005$, so that only 1 in 200 suspects taking the test are guilty. Then

$$P(T) = 0.95 \times 0.005 + 0.01 \times 0.995 = 0.0147$$

so that

$$P(G|T) = \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995} = 0.323$$

which is still relatively small. So, as a result of the lie-detector test being failed, the probability of guilt of the suspect has increased from 0.005 to 0.323.

More extreme examples can be found by altering the values of α, β and p.

EXAMPLE Diagnostic Testing.

A diagnostic test for a disease is to be given to each of the 100000 people in a city. Let S be the event that an individual actually has the disease, and let T be the event that the individual tests positive for the disease.

	S	$S^{'}$	TOTAL
T	4950	15000	19950
$T^{'}$	50	80000	80050
TOTAL	5000	95000	100000

What can be concluded if an individual, selected at random from the city population, admits to having tested positive ?

SECTION 3.

RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

3.1 RANDOM VARIABLES

A **random variable** X is a function from experimental sample space S to some set of real numbers $\mathbb{X}$ that maps $s \in S$ to a unique $x \in \mathbb{X}$

$$\begin{aligned} X : \quad S &\longrightarrow \mathbb{X} \subseteq \mathbb{R} \\ s &\longmapsto x \end{aligned}$$

Interpretation A random variable is a way of describing the outcome of an experiment in terms of real numbers.

RANDOM VARIABLE

EXAMPLE 1 $X =$“No. days in Feb. with zero precipitation”

EXAMPLE 2 $X =$“No. goals in a football match”

EXAMPLE 3 $X =$“the measured operating temperature”

Therefore X is merely the count/number/measured value corresponding to the outcome of the experiment.

Depending on the type of experiment being carried out, there are two possible forms for the set of values that X can take:

- A random variable is **DISCRETE** if the set $\mathbb{X}$ is a finite or infinite set of **distinct** values $x_1, x_2, ..., x_n, ...$ Discrete random variables are used to describe the outcomes of experiments that involve **counting** or **classification**.

- A random variable is **CONTINUOUS** if the set $\mathbb{X}$ is the union of **intervals** in $\mathbb{R}$. Continuous random variables are used to describe the outcomes of experiments that involve **measurement**.

3.2 PROBABILITY DISTRIBUTIONS

We will specify two mathematical functions to describe the distribution of probability across the possible values of the random variables:

- For **DISCRETE** random variables
  - the **probability mass function (pmf)** $f(x) = P[X = x]$
  - the **cumulative distribution function (cdf)** $F(x) = P[X \leq x]$
- For **CONTINUOUS** random variables
  - the **probability density function (pdf)** $f(x)$
  - the **cumulative distribution function (cdf)** $F(x) = P[X \leq x]$

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)dt$$

Most commonly, we deal with the “little-f” function.

3.3 EXPECTATION AND VARIANCE

The expectation and variance of a probability distribution can be used to aid description, or to characterize the distribution;

- the **EXPECTATION** is a measure of **location** (that is, the "centre of mass" of the probability distribution.
- the **VARIANCE** is a measure of **scale** or **spread** of the distribution (how widely the probability is distributed) .

Note : The **expectation** and **variance** of a **probability distribution** are entirely different quantities from the **sample mean** and **sample variance** derived from a sample of **data**

In the discrete case, the **expectation** is defined by

$$E_{f_X}[X] = \sum_x x f_X(x)$$

and in the continuous case

$$E_{f_X}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

whenever the sum or integral is finite. The **variance** is defined by

$$Var_{f_x}[X] = \int_{-\infty}^{\infty} (x - E_{f_X}[X])^2 f_X(x)\, dx = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx - (E_{f_X}[X])^2$$

3.3.1 SUMS OF RANDOM VARIABLES:

Suppose that X_1 and X_2 are independent random variables, and a_1 and a_2 are constants. Then if $Y = a_1X_1 + a_2X_2$, it can be shown that

$$\begin{aligned} \mathrm{E}_{f_Y}[Y] &= a_1\mathrm{E}_{f_{X_1}}[X_1] + a_2\mathrm{E}_{f_{X_2}}[X_2] \\ \mathrm{Var}_{f_Y}[Y] &= a_1^2\mathrm{Var}_{f_{X_1}}[X_1] + a_2^2\mathrm{Var}_{f_{X_2}}[X_2] \end{aligned}$$

so that, in particular (when $a_1 = a_2 = 1$) we have

$$\begin{aligned} \mathrm{E}_{f_Y}[Y] &= \mathrm{E}_{f_{X_1}}[X_1] + \mathrm{E}_{f_{X_2}}[X_2] \\ \mathrm{Var}_{f_Y}[Y] &= \mathrm{Var}_{f_{X_1}}[X_1] + \mathrm{Var}_{f_{X_2}}[X_2] \end{aligned}$$

so we have a simple additive property for expectations and variances. Note also that if $a_1 = 1, a_2 = -1$, then

$$\begin{aligned} \mathrm{E}_{f_Y}[Y] &= \mathrm{E}_{f_{X_1}}[X_1] - \mathrm{E}_{f_{X_2}}[X_2] \\ \mathrm{Var}_{f_Y}[Y] &= \mathrm{Var}_{f_{X_1}}[X_1] + \mathrm{Var}_{f_{X_2}}[X_2] \end{aligned}$$

Sums of random variables crop up naturally in many statistical calculations. Often we are interested in a random variable Y that is defined as the sum of some other **independent and identically distributed** (i.i.d) random variables, $X_1, ..., X_n$. If

$$Y = \sum_{i=1}^{n} X_i \qquad \text{with} \qquad \mathrm{E}_{f_{X_i}}[X_i] = \mu \quad \text{and} \quad \mathrm{Var}_{f_{X_i}}[X_i] = \sigma^2$$

we have

$$\begin{aligned}
\mathrm{E}_{f_Y}[Y] &= \sum_{i=1}^{n} \mathrm{E}_{f_{X_i}}[X_i] = \sum_{i=1}^{n} \mu = n\mu \\
\mathrm{Var}_{f_Y}[Y] &= \sum_{i=1}^{n} \mathrm{Var}_{f_{X_i}}[X_i] = \sum_{i=1}^{n} \sigma^2 = n\sigma^2
\end{aligned}$$

and ao, if

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{is the } \textbf{sample mean} \text{ random variable}$$

then, using the properties listed above

$$\mathrm{E}_{f_{\overline{X}}}[\overline{X}] = \frac{1}{n}\mathrm{E}_{f_Y}[Y] = \frac{1}{n}n\mu = \mu \qquad \text{and} \qquad \mathrm{Var}_{f_Y}[Y] = \frac{1}{n^2}\mathrm{Var}_{f_Y}[Y] = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

3.3.2 SOME SPECIAL DISCRETE PROBABILITY DISTRIBUTIONS

Discrete probability models are used to model the outcomes of counting experiments. Depending on the experimental situation, it is often possible to justify the use of one of a class of "Special" discrete probability distributions. These are listed in this chapter, and are all motivated from the central concept of a *binary* or 0-1 trial, where the random variable concerned has range consisting of only two values with associated probabilities θ and $1 - \theta$ respectively; typically we think of the possible outcomes as "successes" and "failures". All of the distributions in this section are derived by making different modelling assumptions about sequences of 0-1 trials.

Single 0-1 trial - count number of 1s	$\Longrightarrow$ BERNOULLI
n independent 0-1 trials - count number of 1s	$\Longrightarrow$ BINOMIAL
Sequence of independent 0-1 trials - count number of trials until first 1	$\Longrightarrow$ GEOMETRIC
Sequence of independent 0-1 trials - count number of trials until nth 1	$\Longrightarrow$ NEGATIVE BINOMIAL
Limiting case of binomial distribution	$\Longrightarrow$ POISSON

3.3.3 SOME SPECIAL CONTINUOUS DISTRIBUTIONS

Here is a list of probability models are used in standard modelling situations. Unlike the discrete case, there are not really any explicit links between most of them, although some connections can be made by means of "transformation" from one variable to another.

UNIFORM DISTRIBUTION
EXPONENTIAL DISTRIBUTION
GAMMA DISTRIBUTION
BETA DISTRIBUTION
NORMAL DISTRIBUTION
STUDENT-T DISTRIBUTION
FISHER-F DISTRIBUTION

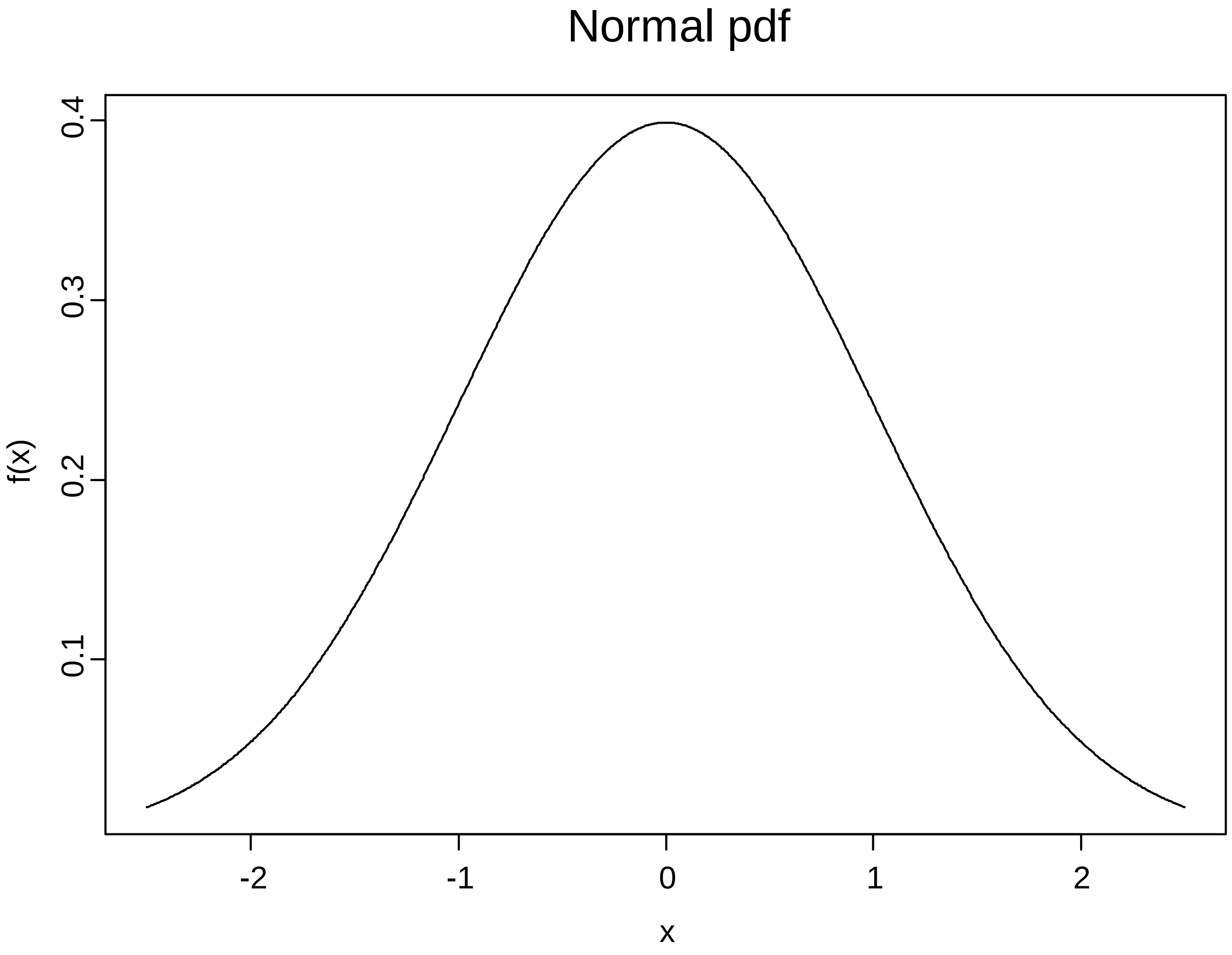
3.3.4 THE NORMAL DISTRIBUTION $X \sim N(\mu, \sigma^2)$

Range : $\mathbb{X} = \mathbb{R}$
Parameters : $\mu \in \mathbb{R}, \sigma \in \mathbb{R}^+$
Density function :

$$f_X(x) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \qquad x \in \mathbb{R}.$$

Interpretation : A probability model that reflects observed (**empirical**) behaviour of data samples; this distribution is often observed in practice.

The pdf is symmetric about μ, and hence μ is controls the *location* of the distribution and σ^2 controls the *spread* or *scale* of the distribution.



NORMAL PDF

NOTES

(1) The Normal density function is justified by the **Central Limit Theorem**.

(2) Special case: $\mu = 0, \sigma^2 = 1$ - the **standard** or **unit** normal distribution. In this case, the density function is denoted $\phi(x)$, and the cdf is denoted $\Phi(x)$ so that

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\ dt = \int_{-\infty}^{x} \left(\frac{1}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}t^2\right\}\ dt.$$

This integral can only be calculated numerically.

(3) If $X \sim N(0,1)$, and $Y = \sigma X + \mu$, then $Y \sim N(\mu, \sigma^2)$.

(4) If $X \sim N(0,1)$, and $Y = X^2$, then

$$Y \sim Gamma(1/2, 1/2) = \chi_1^2$$

This is the **Chi-squared distribution** with 1 **degree of freedom**..

The Chi-squared distribution is another continuous probability distribution; its most general version is the **Chi-squared distribution** with α **degrees of freedom**, where α is some non-negative whole number.

(5) If $X \sim N(0,1)$ and $Y \sim \chi_\alpha^2$ are independent random variables, then random variable T, defined by

$$T = \frac{X}{\sqrt{Y/\alpha}}$$

has a **Student-t distribution** with α **degrees of freedom**.

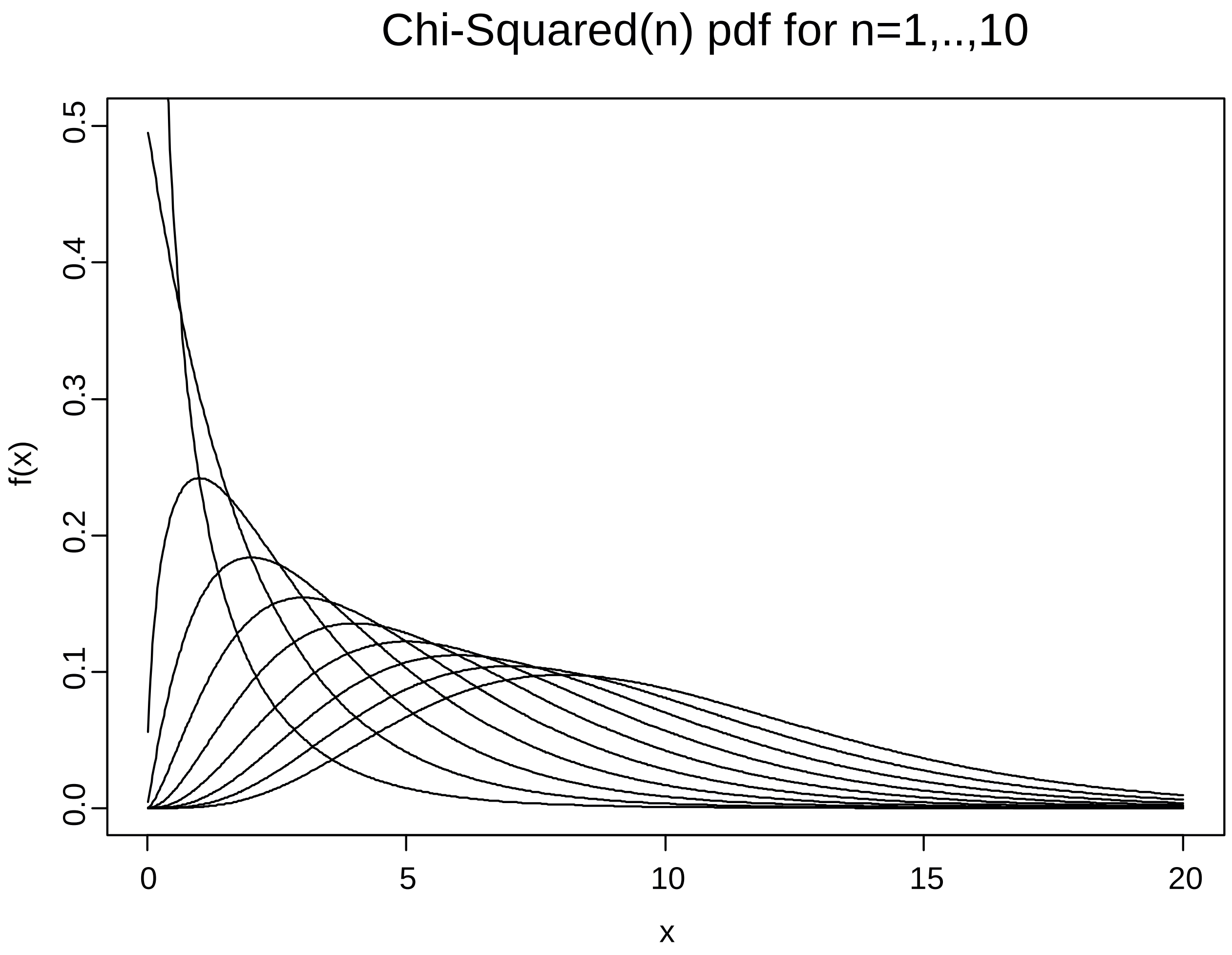
The Student-t distribution plays an important role in certain statistical testing procedures.

3.3.5 The Chi-Squared Distribution

For the **Chi-squared distribution** with α **degrees of freedom**,

$$f(x) = \frac{\left(\frac{1}{2}\right)^{\alpha/2}}{\Gamma\left(\frac{1}{2}\right)} x^{\alpha/2-1} \exp\left\{-\frac{x}{2}\right\} \qquad x > 0$$

- **EXPECTATION** is α
- **VARIANCE** is 2α



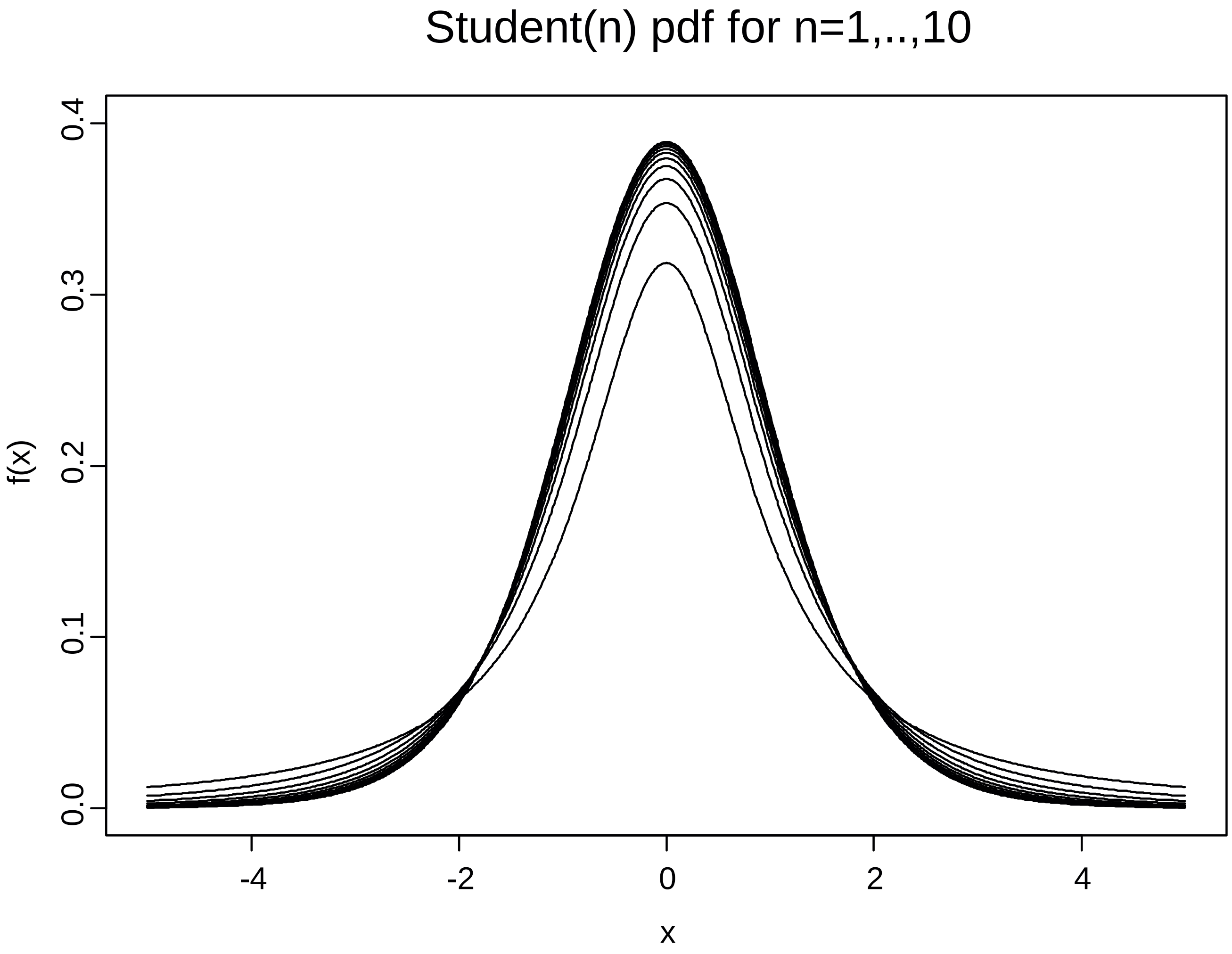
CHI-SQUARED PDF with n DEGREES OF FREEDOM

3.3.6 The Student-t Distribution

For the **Student-t distribution** with α **degrees of freedom**,

$$f(x) = \frac{\Gamma\left(\dfrac{\alpha+1}{2}\right)}{\sqrt{\pi\alpha}\,\Gamma\left(\dfrac{\alpha}{2}\right)} \frac{1}{\left\{1+\dfrac{x^2}{\alpha}\right\}^{(\alpha+1)/2}} \qquad x > 0$$

- **EXPECTATION** is 0 (if $\alpha > 1$)
- **VARIANCE** is $\alpha - 2$ (if $\alpha > 2$)
- $\Gamma(.)$ is a special function known as the **Gamma Function**



STUDENT-T PDF with n DEGREES OF FREEDOM

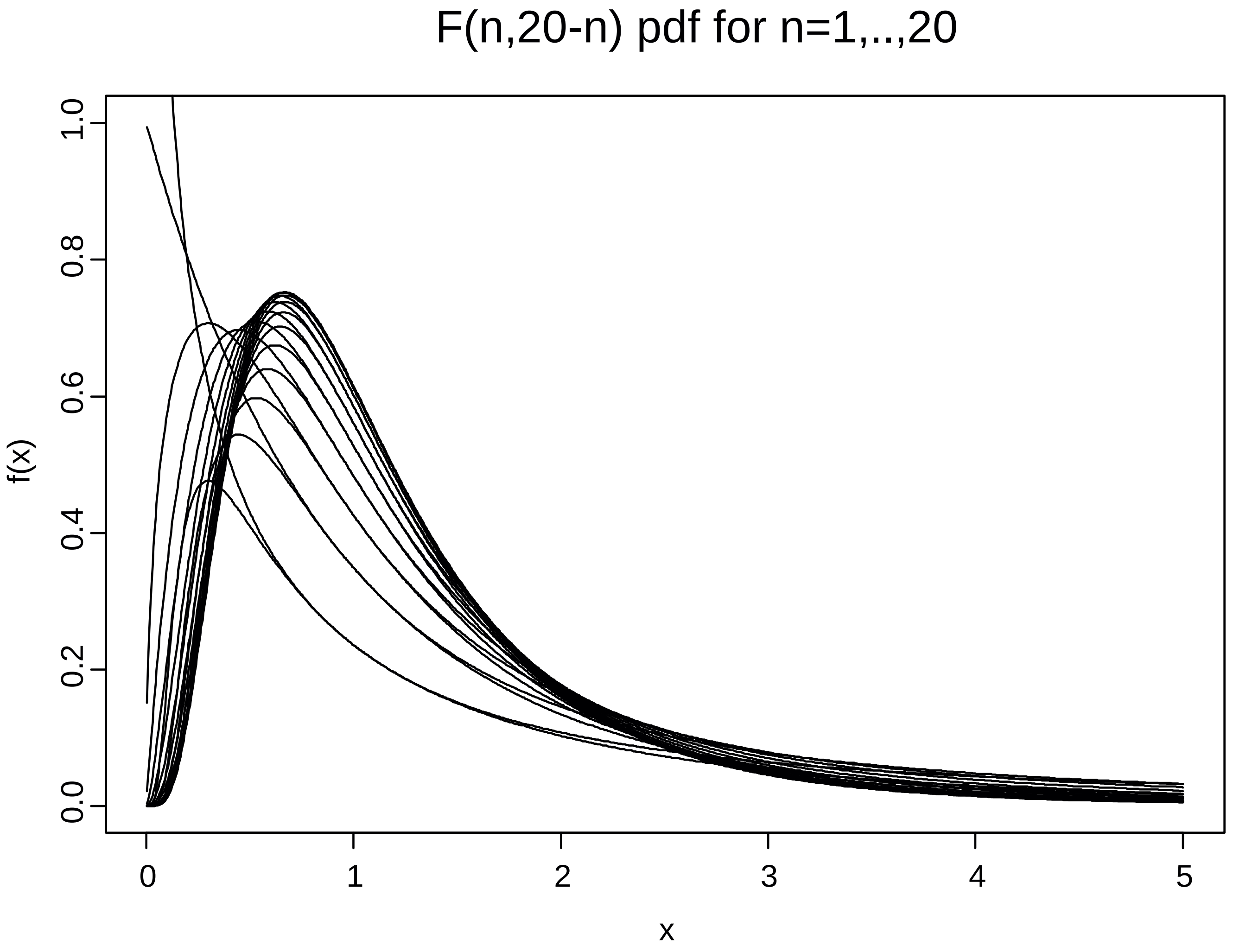
3.3.7 The Fisher-F Distribution

For the **Fisher-F distribution** with ν_1 and ν_2 **degrees of freedom**,

$$f(x) = \frac{\Gamma\left(\dfrac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\dfrac{\nu_1}{2}\right)\Gamma\left(\dfrac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{(\nu_1/2)-1}}{\left\{1+\dfrac{\nu_1}{\nu_2}x\right\}^{(\nu_1+\nu_2)/2}} \qquad x > 0$$

- **EXPECTATION** is $\nu_2/\left(\nu_2-2\right)$ (if $\nu_2 > 2$)
- **VARIANCE** is

$$2\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{(\nu_1+\nu_2-2)}{\nu_1\left(\nu_2-4\right)} \qquad \text{if } \nu_2 > 4$$



FISHER-F distribution with $(n, 20-n)$ degrees of freedom

3.4 TRANSFORMATIONS

Consider a discrete or continuous random variable X with range $\mathbb{X}$ and probability distribution described by mass/pdf f_X, or cdf F_X. Suppose g is a function. Then $Y = g(X)$ is also a random variable as Y and typically we wish to derive the probability distribution of random variable Y.

Most transformations are 1-1 transformations (the exceptions being transformations involving powers of X, like $g(x) = x^2$, or $g(x) = x(1-x)$. The following result gives the distribution for random variable $Y = g(X)$ when g is 1-1.

Some of the continuous distributions that we have studied are directly connected by transformations

Distribution of X	Transformation	Distribution of Y
$X \sim Uniform(0,1)$	$Y = -\frac{1}{\lambda}\log X$	$Y \sim Exponential(\lambda)$
$X \sim Gamma(\alpha,1)$	$Y = X/\beta$	$Y \sim Gamma(\alpha,\beta)$
$X \sim Normal(0,1)$	$Y = \mu + \sigma X$	$Y \sim Normal(\mu,\sigma^2)$
$X \sim Normal(0,1)$	$Y = X^2$	$Y \sim Gamma\left(\frac{1}{2},\frac{1}{2}\right) \equiv \chi_1$

3.5 JOINT PROBABILITY DISTRIBUTIONS

Consider a vector of k random variables, $\mathbf{X} = (X_1, ..., X_k)$, (representing the outcomes of k different experiments carried out once each, or of one experiment carried out k times). The probability distribution of $\mathbf{X}$ is described by a **joint** probability mass or density function.

Two concepts are key:

- **COVARIANCE :** The co-variability of two variables

$$Cov\left[X,Y\right] = E\left[XY\right] - E\left[X\right]E\left[Y\right]$$

- **CORRELATION:** A scaled version of covariance

$$Corr\left[X,Y\right] = \frac{Cov\left[X,Y\right]}{\sqrt{Var\left[X\right]Var\left[Y\right]}} \qquad \text{so that } -1 \leq Corr\left[X,Y\right] \leq 1$$

Key interpretation

COVARIANCE AND CORRELATION ARE MEASURES OF THE DEGREE OF $\boxed{\textbf{ASSOCIATION}}$ BETWEEN VARIABLES

that is, two variables for which the correlation is **large** in magnitude are strongly associated, whereas variables that have low correlation are **weakly** associated.

Note : The **correlation** between two **random variables** is a different quantity from the **sample correlation** derived from a sample of **data.**

SECTION 4.

STATISTICAL INFERENCE

It is often of interest to draw inference from data regarding the parameters of the proposed probability distribution; recall that many aspects of the standard distributions studied are controlled by the distribution parameters.

It is therefore important to find a simple and yet general technique for parameter estimation

4.1 MAXIMUM LIKELIHOOD ESTIMATION

Maximum Likelihood Estimation is a systematic technique for estimating parameters in a probability model from a data. Suppose a sample $x_1, ..., x_n$ has been obtained from a probability model specified by mass or density function $f(x;\theta)$ depending on parameter(s) θ lying in parameter space Θ. The **maximum likelihood estimate** or **m.l.e.** is produced as follows;

STEP 1 Write down the **likelihood function**, $L(\theta)$, where

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

that is, the product of the n mass/density function terms (where the ith term is the mass/density function evaluated at x_i) viewed as a function of θ.

STEP 2 Take the natural log of the likelihood, and collect terms involving θ.

STEP 3 Find the value of $\theta \in \Theta$, $\hat{\theta}$, for which $\log L(\theta)$ is maximized, for example by differentiation. If θ is a single parameter, find $\hat{\theta}$ by solving

$$\frac{d}{d\theta}\ \{\log L(\theta)\} = 0$$

in the parameter space Θ. If θ is vector-valued, say $\theta = (\theta_1, ..., \theta_d)$, then find $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_d)$ by simultaneously solving the d eqnarray*s given by

$$\frac{\partial}{\partial\theta_j}\ \{\log L(\theta)\} = 0 \qquad j = 1, ..., d$$

in parameter space Θ.

Note that, if parameter space Θ is a bounded interval, then the maximum likelihood estimate may lie on the boundary of Θ.

STEP 4 Check that the estimate $\hat{\theta}$ obtained in STEP 3 truly corresponds to a maximum in the (log) likelihood function by inspecting the second derivative of $\log L(\theta)$ with respect to θ. If

$$\frac{d^2}{d\theta^2}\ \{\log L(\theta)\} < 0$$

at $\theta = \hat{\theta}$, then $\hat{\theta}$ is confirmed as the m.l.e. of θ (other techniques may be used to verify that the likelihood is maximized at $\hat{\theta}$).

This procedure is a systematic way of producing parameter estimates from sample data and a probability model; it can be shown that such an approach produces estimates that have good properties. After they have been obtained, the estimates can be used to carry out *prediction* of behaviour for future samples.

EXAMPLE A sample $x_1, ..., x_n$ is modelled by a Poisson distribution with parameter denoted λ; hence

$$f(x;\theta) \equiv f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \qquad x = 0, 1, 2, ...$$

for some $\lambda > 0$.

STEP 1 Calculate the likelihood function $L(\lambda)$. For $\lambda > 0$,

$$L(\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{x_i}}{x_i!}e^{-\lambda} \right\} = \frac{\lambda^{x_1+...+x_n}}{x_1!....x_n!}e^{-n\lambda}$$

STEP 2 Calculate the log-likelihood $\log L(\lambda)$.

$$\log L(\lambda) = \sum_{i=1}^{n} x_i \log\lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

STEP 3 Differentiate $\log L(\lambda)$ with respect to λ, and equate the derivative to zero.

$$\frac{d}{d\lambda}\{\log L(\lambda)\} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0 \Longrightarrow \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Thus the maximum likelihood estimate of λ is $\hat{\lambda} = \bar{x}$

STEP 4 Check that the second derivative of $\log L(\lambda)$ with respect to λ is negative at $\lambda = \hat{\lambda}$.

$$\frac{d^2}{d\lambda^2}\{\log L(\lambda)\} = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i < 0 \text{ at} \lambda = \hat{\lambda}$$

4.2 SAMPLING DISTRIBUTIONS

EXAMPLE : Suppose $x_1, ..., x_n$ have a Normal distribution with parameters μ and σ^2, then the maximum likelihood estimates are

$$\hat{\mu} = \bar{x} \qquad \hat{\sigma}^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

If five samples (from five different labs) of eight observations are collected, the estimate $\hat{\mu}$ of μ is different each time:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	$\bar{x}$
10.4	11.2	9.8	10.2	10.5	8.9	11.0	10.3	10.29
9.7	12.2	10.4	11.1	10.3	10.2	10.4	11.1	10.66
12.1	7.9	8.6	9.6	11.0	11.1	8.8	11.7	10.10
10.0	9.2	11.1	10.8	9.1	12.3	10.3	9.7	10.31
9.2	9.7	10.8	10.3	8.9	10.1	9.7	10.4	9.89

We attempt to understand how $\bar{x}$ varies by calculating the **probability distribution** of the corresponding **estimator**, $\bar{X}$.

- The estimator $\bar{X}$ is a **random variable**, the value of which is **unknown** *before* the experiment is carried out.
- As a random variable, $\bar{X}$ has a probability distribution, known as the **sampling distribution**.
- The form of this distribution can often be calculated, and used to understand how $\bar{x}$ varies.
- In the case where the sample data have a Normal distribution

The following theorem gives the sampling distributions of the maximum likelihood estimators;

THEOREM If $X_1, ..., X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables, then

(1) $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$,

(2) $\displaystyle \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{nS^2}{\sigma^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$,

(3) $\bar{X}$ and S^2 are **independent** random variables.

This theorem tells us how we expect the sample mean and sample variance to behave. In particular, it tells us that

$$E[\bar{X}] = \mu \qquad E\left[S^2\right] = \frac{n-1}{n}\sigma^2 \qquad E\left[s^2\right] = \sigma^2$$

Interpretation : This theorem tells us how the sample mean and variance will behave if the original random sample is assumed to come from a Normal distribution.

For example, if we believe that $X_1, ..., X_{10}$ are i.i.d random variables from a Normal distribution with parameters $\mu = 10.0$ and $\sigma^2 = 25$, then $\bar{X}$ has a Normal distribution with parameters $\mu = 10.0$ and $\sigma^2 = 25/10 = 2.5$.

The result will be used to facilitate formal tests about model parameters.

For example, given a sample of experimental, we wish to answer **specific** questions about parameters in a proposed probability model.

4.3 HYPOTHESIS TESTING

Given a sample $x_1, ..., x_n$ from a probability model $f(x; \theta)$ depending on parameter θ, we can produce an estimate $\hat{\theta}$ of θ, and in some circumstances understand how $\hat{\theta}$ varies for repeated samples. Now we might want to test, say, whether or not there is evidence from the sample that true (but unobserved) value of θ is not equal to a specified value. To do this, we use estimate of θ, and the corresponding estimator and its sampling distribution, to quantify this evidence.

In particular, we concentrate on data samples that we can presume to have a normal distribution, and utilize the Theorem from the previous section. We will look at two situations, namely **one sample** and **two sample** experiments.

- **ONE SAMPLE**

Random variables	$X_1, ..., X_n \sim N(\mu, \sigma^2)$
sample observations	$x_1, ...x_n$
Possible Models	$\mu = \mu_0 \qquad \sigma = \sigma_0$

- **TWO SAMPLE**

Random variables	$X_1, ..., X_n \sim N(\mu_X, \sigma_X^2)$
sample 1 observations	$x_1, ...x_n$
Random variables	$Y_1, ..., Y_n \sim N(\mu_Y, \sigma_Y^2)$
sample 2 observations	$y_1, ...y_n$
Possible Models :	$\mu_X = \mu_Y \qquad \sigma_X = \sigma_Y$

4.3.1 HYPOTHESIS TESTS FOR NORMAL DATA I - THE Z-TEST (σ KNOWN)

If $X_1, ..., X_n \sim N(\mu, \sigma^2)$ are the i.i.d. outcome random variables of n experimental trials, then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{and} \qquad \frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$

with $\bar{X}$ and S^2 statistically independent. Suppose we want to test the **hypothesis** that $\mu = \mu_0$, for some specified constant μ_0, (where, for example, $\mu_0 = 20.0$) is a plausible model; more specifically, we want to test the hypothesis $H_0 : \mu = \mu_0$ against the hypothesis $H_1 : \mu \neq \mu_0$, that is, we want to test whether H_0 is true, or whether H_1 is true. From the results above, the distribution of the estimator $\bar{X}$ is Normal, and

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Longrightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

where Z is a **random variable**. Now, when we have observed the data sample, we can calculate $\bar{x}$, and therefore we have a way of testing whether $\mu = \mu_0$ is a plausible model; we calculate $\bar{x}$ from $x_1, ..., x_n$, and then calculate

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

If H_0 is true, and $\mu = \mu_0$, then the **observed** z should be an observation from an $N(0,1)$ distribution (as $Z \sim N(0,1)$), that is, it should be near zero with high probability. In fact, z should lie between -1.96 and 1.96 with probability $1 - \alpha = 0.95$, say, as

$$P[-1.96 \leq Z < 1.96] = \Phi(1.96) - \Phi(-1.96) = 0.975 - 0.025 = 0.95.$$

If we observe z to be outside of this range, then there is evidence that H_0 is **not true**.

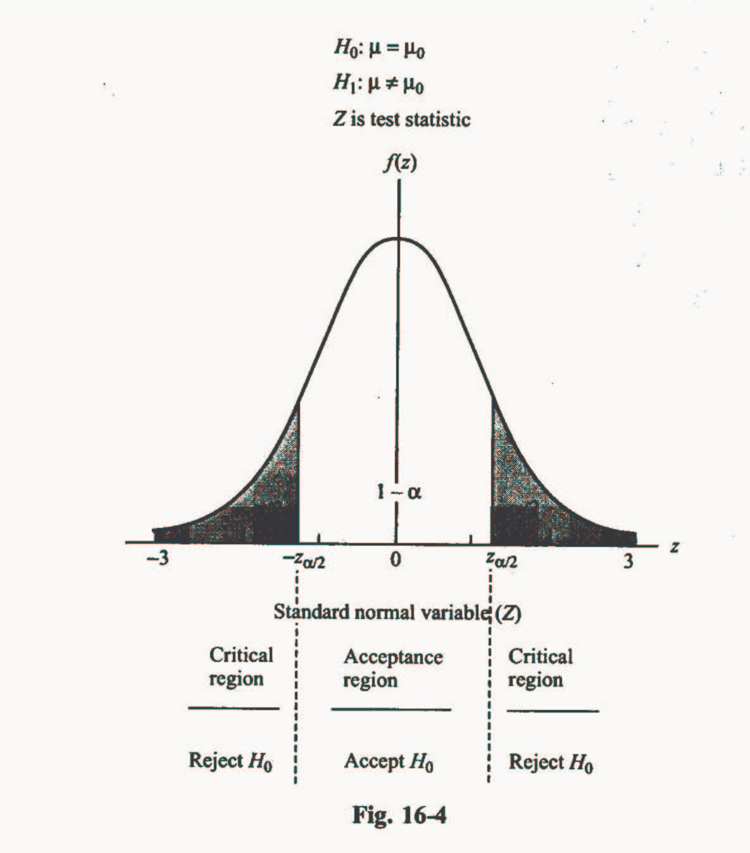


Fig. 16-4

CRITICAL REGIONS IN A Z-TEST (taken from *Schaum's ELEMENTS OF STATISTICS II, Bernstein & Bernstein)*

Alternatively, we could calculate the probability p of observing a z value that is **more extreme** than the z we did observe; this probability is given by

$$p = \begin{cases} 2\Phi(z) & z < 0 \\ 2(1 - \Phi(z)) & z \geq 0 \end{cases}$$

If p is very small, say $p \leq \alpha = 0.05$, then again. there is evidence that H_0 is **not true**. In summary, we need to assess whether z is a **surprising** observation from an $N(0, 1)$ distribution - if it is, then we can **reject** H_0.

p is probably the most important and widely-used quantity that is computed during the hypothesis test; it quantifies the amount of "suprisingness" in the observed data by reporting how likely we were to observe a **more extreme** test statistic than the one we did observe **IF THE MODEL REPRESENTED BY H_0 IS TRUE.**

It is termed the **p-value** or **achieved significance level** of the test

4.3.2 HYPOTHESIS TESTING TERMINOLOGY

There are five crucial components to a hypothesis test, namely

- **TEST STATISTIC**
- **NULL DISTRIBUTION**
- **SIGNIFICANCE LEVEL**, denoted α
- **P-VALUE**, denoted p.
- **CRITICAL VALUE(S)**

In the Normal example given above, we have that

z is the **test statistic**

The distribution of random variable Z if H_0 is true is the **null distribution**

$\alpha = 0.05$ is the **significance level** of the test (we could use $\alpha = 0.01$ if we require a "stronger" test).

p is the **p-value** of the test statistic under the null distribution

The solution C_R of $\Phi(C_R) = 1 - \alpha/2$ ($C_R = 1.96$ above) gives the **critical values** of the test $\pm C_R$.

EXAMPLE : A sample of size 10 has sample mean $\bar{x} = 19.7$. Suppose we want to test the hypothesis

$$\begin{aligned} H_0 &: \mu = 20.0 \\ H_1 &: \mu \neq 20.0 \end{aligned}$$

under the assumption that the data follow a Normal distribution with $\sigma = 1.0$.

We have

$$z = \frac{19.7 - 20.0}{1/\sqrt{10}} = -0.95$$

which lies between the critical values ±1.96, and therefore we have no reason to reject H_0. Also, the p-value is given by $p = 2\Phi(-0.95) = 0.342$, which is greater than $\alpha = 0.05$, which confirms that we have no reason to reject H_0.

4.3.3 HYPOTHESIS TESTS FOR NORMAL DATA II - THE T-TEST (σ UNKNOWN)

In practice, we will often want to test hypotheses about μ when σ is unknown. We cannot perform the Z-test, as this requires knowledge of σ to calculate the z statistic. We proceed as follows; recall that we know the sampling distributions of $\bar{X}$ and s^2, and that the two estimators are statistically independent. Now, from the properties of the Normal distribution, if we have independent random variables $Z \sim N(0,1)$ and $Y \sim \chi^2_\nu$, then we know that random variable T defined by

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

has a Student-t distribution with ν degrees of freedom.

Using the previous results, we can derive the null distribution of T.

$$T = \frac{\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)s^2/\sigma^2}{(n-1)}}} = \frac{(\bar{X}-\mu)}{s/\sqrt{n}} \sim t_{n-1}$$

and T has a Student-t distribution with $n-1$ degrees of freedom, denoted $St(n-1)$. Thus we can repeat the procedure used in the σ known case, but use the sampling distribution of T rather than that of Z to assess whether the test statistic is “surprising” or not. Specifically, we calculate

$$t = \frac{(\bar{x}-\mu)}{s/\sqrt{n}}$$

and find the critical values for a $\alpha = 0.05$ significance test by finding the ordinates corresponding to the 0.025 and 0.975 percentiles of a Student-t distribution, $St(n-1)$ (rather than a $N(0,1)$) distribution.

EXAMPLE : A sample of size 10 has sample mean $\bar{x} = 19.7.$ and $s^2 = 0.78^2$. Suppose we want to carry out a test of the hypotheses

$$\begin{aligned} H_0 &: \mu = 20.0 \\ H_1 &: \mu \neq 20.0 \end{aligned}$$

under the assumption that the data follow a Normal distribution with σ unknown.

We have test statistic t given by

$$t = \frac{19.7 - 20.0}{0.78/\sqrt{10}} = -1.22.$$

The upper critical value C_R is obtained by solving

$$F_{t_{n-1}}(C_R) = 0.975$$

where $F_{St(n-1)}$ is the c.d.f. of a Student-t distribution with $n-1$ degrees of freedom.

Here $n = 10$, so we can use the statistical tables to find $C_R = 2.262$, and not that, as Student-t distributions are symmetric the lower critical value is $-C_R$.

Thus t lies between the critical values, and therefore we have **no reason to reject** H_0.

The p-value is given by

$$p = \begin{cases} 2F_{t_{n-1}}(t) & t < 0 \\ 2(1 - F_{t_{n-1}}(t)) & t \geq 0 \end{cases}$$

so here, $p = 2F_{t_{n-1}}(-1.22)$ which we can find to give $p = 0.253$; this confirms that we have no reason to reject H_0.

4.3.4 HYPOTHESIS TESTS FOR NORMAL DATA III - TESTING σ.

The Z-test and T-test are both tests for the parameter μ. Suppose that we wish to test a hypothesis about σ, for example

$$\begin{aligned} H_0 &: \sigma^2 = \sigma_0 \\ H_1 &: \sigma^2 \neq \sigma_0 \end{aligned}$$

We construct a test based on the estimate of variance, s^2. In particular, the random variable Q, defined by

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

if the data have an $N(\mu, \sigma^2)$ distribution.

Hence if we define test statistic q by

$$q = \frac{(n-1)s^2}{\sigma_0^2}$$

then we can compare q with the critical values derived from a χ^2_{n-1} distribution; we look for the 0.025 and 0.975 quantiles - note that the Chi-squared distribution is not symmetric, so we need two distinct critical values.

In the above example, to test

$$H_0 : \sigma^2 = 1.0$$
$$H_1 : \sigma^2 \neq 1.0$$

we compute test statistic

$$q = \frac{(n-1)s^2}{\sigma_0^2} = \frac{90.78^2}{1.0} = 5.4375$$

We compare this with

$$\begin{aligned} C_{R_1} &= F_{\chi^2_{n-1}}(0.025) & \Longrightarrow C_{R_1} &= 2.700 \\ C_{R_2} &= F_{\chi^2_{n-1}}(0.975) & \Longrightarrow C_{R_2} &= 19.022 \end{aligned}$$

so q is not a surprising observation from a χ^2_{n-1} distribution, and hence we cannot reject H_0.

We can compute the p-value by inspection of the cumulative distribution function of the χ^2_{n-1} distribution. However, we know already that this p-value will not be smaller than the significance level, as the test-statistic does not lie in the critical region.

4.3.5 TWO SAMPLE TESTS

It is straightforward to extend the ideas from the previous sections to two sample situations where we wish to compare the distributions underlying two data samples. Suppose that the sample mean and sample variance for samples one and two are denoted $(\bar{x}, s_X^2)$ and $(\bar{y}, s_Y^2)$ respectively.

First, consider testing the hypothesis

$$\begin{array}{c} H_0 : \mu_X = \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{array}$$

when $\sigma_X = \sigma_Y = \sigma$ is known. Now, we have from the sampling distributions theorem we have

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{n_X}\right) \qquad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{n_Y}\right) \Longrightarrow \bar{X} - \bar{Y} \sim N\left(0, \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}\right)$$

and hence

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}} \sim N(0,1)$$

giving us a test statistic z defined by

$$z = \frac{\bar{x} - \bar{y}}{\sigma\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$$

which we can compare with the standard normal distribution; if z is a surprising observation from $N(0,1)$, and lies outside of the critical region, then we can reject H_0. This procedure is the Two Sample Z-Test.

If $\sigma_X = \sigma_Y = \sigma$ is unknown, we parallel the one sample T-test by replacing σ by an estimate in the two sample Z-test. First, we obtain an estimate of σ by "pooling" the two samples; our estimate is the **pooled estimate**, s_P^2, defined by

$$s_P^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}$$

which we then use to form the test statistic t defined by

$$t = \frac{\bar{x} - \bar{y}}{s_P \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

It can be shown that, if H_0 is true then t should be an observation from a Student-t distribution with $n_X + n_Y - 2$ degrees of freedom. Hence we can derive the critical values from the tables of the Student-t distribution.

If $\sigma_X \neq \sigma_Y$, but both parameters are known, we can use a similar approach to the one above to derive test statistic z defined by

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}}}$$

which has an $N(0, 1)$ distribution if H_0 is true.

If $\sigma_X \neq \sigma_Y$, but both parameters are unknown, we can use a similar approach to the one above to derive test statistic t defined by

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\dfrac{s_X^2}{n_X} + \dfrac{s_Y^2}{n_Y}}}$$

This statistic has an approximate Student-t distribution if H_0 is true.

Clearly, the choice of test depends on whether $\sigma_X = \sigma_Y$ or otherwise; we may test this hypothesis formally; to test

$$\begin{aligned} H_0 &: \sigma_X = \sigma_Y \\ H_1 &: \sigma_X \neq \sigma_Y \end{aligned}$$

We compute the test statistic

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

which has a null distribution known as the **Fisher** or F distribution with $(n_X - 1, n_Y - 1)$ degrees of freedom; this distribution can be denoted $F(n_X - 1, n_Y - 1)$, and its quantiles are tabulated.

We can find the 0.025 and 0.975 quantiles of the $F(n_X - 1, n_Y - 1)$ distribution and define the critical region; if the test statistic q is very different from 1, then it is a surprising observation from the F distribution, and we reject the hypothesis of equal variances.

4.3.6 ONE-SIDED AND TWO-SIDED TESTS

So far we have considered hypothesis tests of the form

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned}$$

which is referred to as a **two-sided test**, that is, the alternative hypothesis is supported by an extreme test statistic in **either** tail of the distribution. We may also consider a **one-sided test** of the form

$$\begin{array}{c} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{array} \quad \text{or} \quad \begin{array}{c} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{array} .$$

Such a test proceeds exactly as the two-sided test, except that a significant result can only occur in the right (or left) tail of the null distribution, and there is a single critical value, placed, for example, at the 0.95 (or 0.05) probability point of the null distribution.

4.3.7 CONFIDENCE INTERVALS

The procedures above allow us to test specific hypothesis about the parameters of probability models. We may complement such tests by reporting a **confidence interval**, which is an interval in which we believe the “true” parameter lies with high probability. Essentially, we use the sampling distribution to derive such intervals. For example, in a one sample Z-test, we saw that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

that is, that, for critical values $\pm C_R$ in the test at the 5 % significance level

$$\mathrm{P}\left[-C_R \leq Z \leq C_R\right] = \mathrm{P}\left[-C_R \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq C_R\right] = 0.95$$

Now, from tables we have $C_R = 1.96$, so re-arranging this expression we obtain

$$\mathrm{P}\left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right] = 0.95$$

from which we deduce a **95 % Confidence Interval** for μ based on the sample mean $\bar{x}$ of

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

We can derive other confidence intervals (corresponding to different significance levels in the equivalent tests) by looking up the appropriate values of the critical values. The general approach for construction of confidence interval for generic parameter θ proceeds as follows. >From the modelling assumptions, we derive a **pivotal quantity**, that is, a statistic, T_{PQ}, say, (usually the test statistic random variable) that depends on θ, but whose sampling distribution is “parameter-free” (that is, does not depend on θ). We then look up the critical values C_{R_1} and C_{R_2}, such that

$$\mathrm{P}\left[C_{R_1} \leq T_{PQ} \leq C_{R_2}\right] = 1 - \alpha$$

where α is the significance level of the corresponding test. We then rearrange this expression to the form

$$\mathrm{P}\left[c_1 \leq \theta \leq c_2\right] = 1 - \alpha$$

where c_1 and c_2 are functions of C_{R_1} and C_{R_2} respectively. Then a $1 - \alpha$ % Confidence Interval for θ is $[c_1, c_2]$.

4.3.8 PAIRED TESTS

In a two-sample testing situation, we may have data that are **paired**, in the sense that each observation in one sample has a corresponding sample in the other sample; this could arise if two measurements (pre-treatment/post treatment) are available on a set of individuals, denoted (x_{i1}, x_{i2}). In a paired t-test, the assumption of normality is necessary for the **differences** in the measurements

$$z_i = x_{i1} - x_{i2}$$

but **not** for the individual observations. Hence the paired sample gives rise to a single sample of differences $\{z_i = x_{i1} - x_{i2}, i = 1, ..., n\}$ that can be tested using

- one sample Z-tests or one sample T-tests

depending on whether the variance of the differenced sample is to be presumed known or unknown respectively.

SECTION 5.

HYPOTHESIS TESTING EXTENSIONS

1. consider a pair of competing **hypotheses**, H_0 and H_1
2. define a suitable test statistic random variable $T = T(X_1, ..., X_n)$ (that is, some function of the original random variables)
3. **assume that** H_0 **is true**, and compute the sampling distribution of T, f_T or F_T; this is the **null distribution**
4. compute the **observed** value of T, $t = T(x_1, ..., x_n)$; this is the **observed test statistic**
5. assess whether t is a surprising observation from the distribution f_T. If it **is** surprising, we have evidence to **reject** H_0; if it is not surprising, we **cannot reject** H_0

KEY POINT:

This is a **generic** approach that we have seen applied in the normal, one and two-sample case. Effectively, for the *p*-value, we have computed

$$P\left[\text{Data at least } \textbf{as extreme as} \text{ the data we did observe} \mid \text{Null model is } \textbf{TRUE}\right]$$

or,

$$P\left[T(X) \geq t \mid H_0\textbf{TRUE}\right]$$

This strategy can be applied to more complicated normal examples, and also non-normal and non-parametric testing situations. It is a general strategy for assessing the statistical evidence for or against a hypothesis.

Note that we only have to compute the probability conditional on the “null” hypothesis being **true**.

5.1 ANALYSIS OF VARIANCE

The first extension we consider still presumes a normality assumption for the data, but extends the ideas from Z and T tests, which compare at most two samples, to allow for the analysis of any number of samples.

Analysis of variance or **ANOVA** is used to display the sources of variability in a collection of data groups.

The ANOVA F-test compares variability **between** groups with the variability **within** groups.

5.1.1 ONE-WAY ANOVA

The T-test can be extended to allow a test for differences between more than two data samples. Suppose there are K groups of sizes $n_1, ..., n_K$ (let $n = n_1 + ... + n_K$) from different populations. Let y_{kj} be the jth observation in the kth group, then

$$y_{kj} = \mu_k + \varepsilon_{kj}$$

for $k = 1, ..., K$, and $\varepsilon_{kj} \sim N\left(0, \sigma^2\right)$. This model assumes that

$$Y_{kj} \sim N\left(\mu_k, \sigma^2\right)$$

and that the expectations for the different groups are different. We can view the data as a table comprising K columns, with each column corresponding to a sample.

The groups are commonly referred to as **FACTORS.**

EXAMPLE: ANTIBIOTIC/SERUM PROTEIN BINDING

	Penicillin G	Tetra-cyclin	Strepto-mycin	Erythro-mycin	Chloram-phenicol
	29.6	27.3	5.8	21.6	29.2
	24.3	32.6	6.2	17.4	32.8
	28.5	30.8	11.0	18.3	25.0
	32.0	34.8	8.3	19.0	24.2
Mean	28.6	31.4	7.8	19.1	27.8

Is there any evidence that the amount of serum-binding differs across antibioitics ?

To test the hypothesis that each column (or "population") has the same mean, that is, the hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = ... = \mu_K \\ H_1 &: \text{not } H_0 \end{aligned}$$

an **Analysis of Variance (ANOVA)** F-test may be carried out.

The alternative hypothesis H_1 corresponds to the model where **at least one** of the μ parameters, the mean levels for the factors, is different from the others.

To carry out a test of the hypothesis, the following **ANOVA table** should be completed;

Source	D.F.	Sum of squares	Mean square	$ANOVA-F$ F_a
Between Samples	$K-1$	FSS	$FSS/(K-1)$	$\dfrac{FSS/(K-1)}{RSS/(n-K)}$
Within Samples	$n-K$	RSS	$RSS/(n-K)$	
Total	$n-1$	TSS		

The test is completed by evaluating a p-value using the **observed** $ANOVA-F$ statistic, f_a, that is, the probability

$$P\left[F_a \geq f_a | F \text{ has a } Fisher-F\,(K-1, n-K) \text{ distribution }\right]$$

where

$$TSS = \sum_{k=1}^{K} \sum_{j=1}^{n_k} \left(y_{kj} - \overline{y}_{..}\right)^2 \quad RSS = \sum_{k=1}^{K} \sum_{j=1}^{n_k} \left(y_{kj} - \overline{y}_{k}\right)^2$$

$$FSS = \sum_{k=1}^{K} n_k \left(\overline{y}_k - \overline{y}_{..}\right)^2$$

where

- TSS is the **total** sum-of-squares (i.e. total deviation from the overall data mean $\overline{y}_{.}$)
- RSS is the **residual** sum-of-squares (i.e. sum of deviations from individual group means $\overline{y}_k$, $k = 1, ..., K$) and
- FSS is the **fitted** sum-of-squares (i.e. weighted sum of deviations of group means from the overall data mean, with weights equal to number of data points in the individual samples)

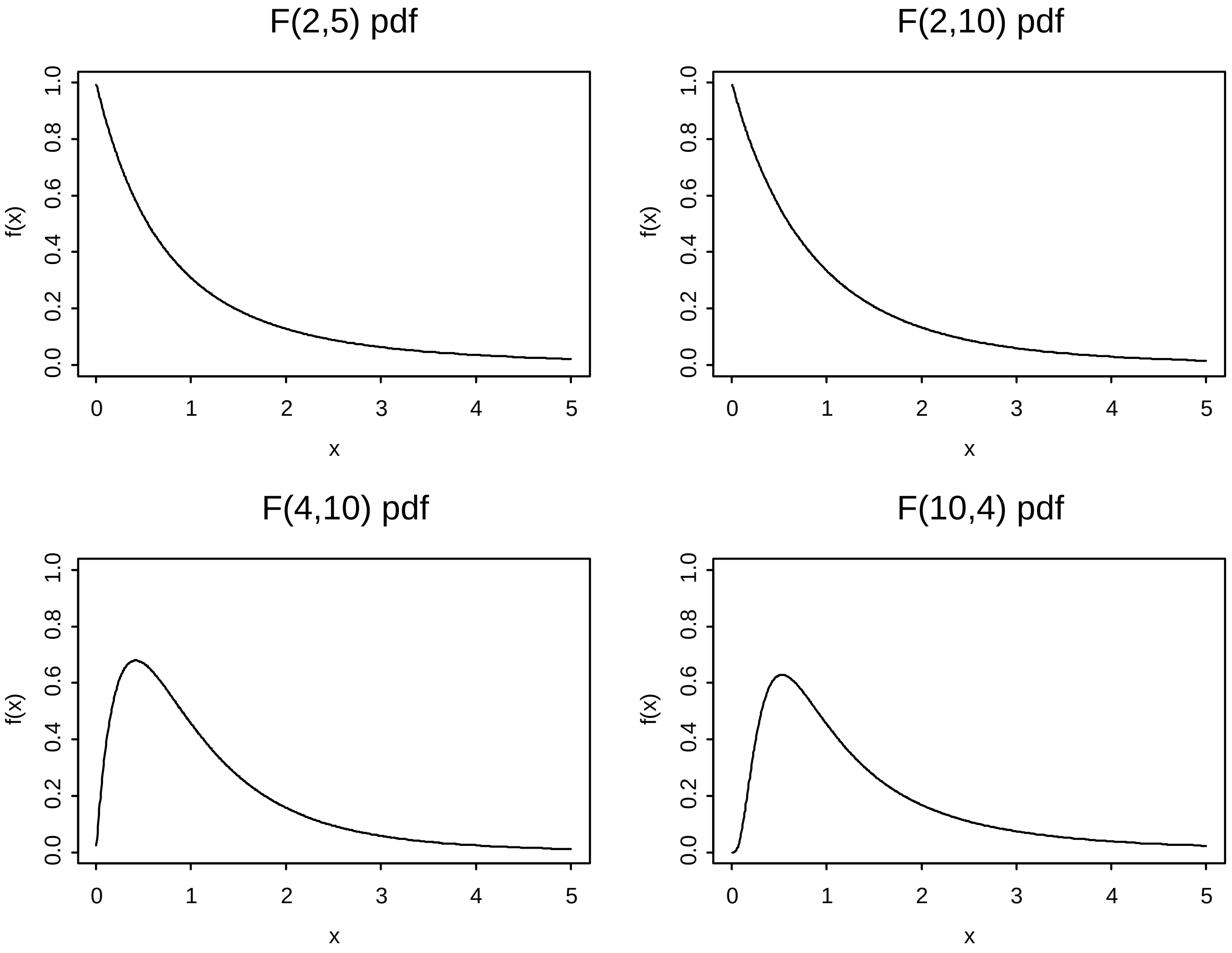
Note:that

$$TSS = FSS + RSS$$

The definitions of these three sums of squares quantities gives insight into how ANOVA works by decomposing the total variation in the observed data

- TSS is the **overall** variation
- FSS is the variation caused by the **systematic** component (that is, the differences in group means)
- RSS is the **random** variation

If the F statistic is calculated in this way, and compared with an F distribution with parameters $K-1$, $n-K$, the hypothesis that all the individual samples have the same mean can be tested. We write $F_{K-1,n-K}$ for this Fisher-F distribution.


F(2,5) pdf
F(2,10) pdf
F(4,10) pdf
F(10,4) pdf
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
x

EXAMPLE: ANTIBIOTIC/SERUM PROTEIN BINDING

Source	D.F.	Sum of squares	Mean square	F
SERUM	4	1480.82	370.21	40.88
Residual	15	135.82	9.05	
Total	19	1616.64		

which gives a p-value (of 6.74×10^{-8}) in comparison with a $Fisher\ F_{4,15}$ distribution)

This is a highly statistically significant result, and thus there is strong evidence to **reject** the null hypothesis that the mean serum protein binding is equal for all antibiotics (under the ANOVA assumptions).

EXAMPLE Three genomic segments were used to studied in order to discover whether the distances (in kB) between successive occurrences of a particular motif were substantially different. Several measurements were taken using for each segment;

	Method		
	SEGMENT A	SEGMENT B	SEGMENT C
	42.7	44.9	41.9
	45.6	48.3	44.2
	43.1	46.2	40.5
	41.6		43.7
			41.0
Mean	43.25	46.47	42.26
Variance	2.86	2.94	2.66

For these data, the ANOVA table is as follows;

Source	D.F.	Sum of squares	Mean square	F
SEGMENTS	2	34.1005	17.0503	6.11
Residual	9	25.1087	2.7899	
Total	11	59.2092		

and the F statistic must be compared with an $F_{2,9}$ distribution. For a significance test at the 0.05 level, F must be compared with the 95th percentile (in a **one-sided** test) of the $F_{2,9}$ distribution. This value is 4.26. Therefore, the F statistic **is** surprising, given the hypothesized model, and therefore there is evidence to reject the hypothesis that the segments are identical.

5.1.2 POST-HOC TESTS

The hypothesis of equal means across all groups is not necessarily the only hypothesis of interest. We may wish to test, for example

$$H_0 : \mu_r = \mu_s$$

against the general alternative for any possible pair of columns r and s, even if the null hypothesis of equal means in all columns is not rejected.

Pairwise tests for equality of column means that are carried out after an F-test has led to the rejection of the ANOVA null hypothesis are referred to as **post-hoc** tests. The key consideration for such tests is the appropriate correction for **multiple testing;** a number of methods have been proposed.

5.1.3 TWO-WAY ANOVA

One-way ANOVA can be used to test whether the underlying means of several groups of observations are equal Now consider the following data collection situation Suppose there are K treatments, and L groups of observations that are believed to have different responses, that all treatments are administered to all groups, and measurement samples of size n are made for each of the $K \times L$ combinations of treatments × groups. The experiment can be represented as follows: let y_{klj} be the jth observation in the kth treatment on the lth group, then

$$y_{klj} = \mu_k + \delta_l + \varepsilon_{klj}$$

for $k = 1, ..., K$, $l = 1, ..., L$, and again $\varepsilon_{klj} \sim N\left(0, \sigma^2\right)$.

This model assumes that $Y_{kj} \sim N\left(\mu_k + \delta_l, \sigma^2\right)$ and that the expectations for the different samples are different. We can view the data as a 3 dimensional-table comprising K columns and L rows, with n observations for each column × row combination, corresponding to a sample.

It is possible to test the hypothesis that each **treatment**, and/or that each **group** has the same mean, that is, the two null hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = ... = \mu_K \\ H_0 &: \delta_1 = \delta_2 = ... = \delta_L \end{aligned}$$

against the alternative H_1 not H_0 in each case.

For these tests, a **Two-way Analysis of Variance (ANOVA)** F-test may be carried out.

The Two-Way ANOVA table is computed as follows

Source	D.F.	Sum of squares	Mean square	F
TREATMENTS	$K-1$	FSS_1	$FSS_1/(K-1)$	$\frac{FSS_1/(K-1)}{RSS/(R+1)}$
GROUPS	$L-1$	FSS_2	$FSS_2/(L-1)$	$\frac{FSS_2/(L-1)}{RSS/(R+1)}$
Residual	$R+1$	RSS	$RSS/(R+1)$	
Total	$N-1$	TSS		

where $N = K \times L \times n$, $R = N - L - K$. and again

$$TSS = FSS_1 + FSS_2 + RSS.$$

In the table below, there are $K = 6$ Treatments, and $L = 3$ Groups, and $n = 1$

	I	II	III	GROUP totals
1	0.96	0.94	0.98	2.88
2	0.96	0.98	1.01	2.95
3	0.85	0.87	0.86	2.58
4	0.86	0.84	0.90	2.60
5	0.86	0.87	0.89	2.62
6	0.89	0.93	0.92	2.74
TREATMENT totals	5.38	5.43	5.56	16.37

There are two natural hypotheses to test; first, do the TREATMENTS differ, and second, do the GROUPS differ ?

Two-way analysis of variance can be used to analyze such data. Given two sources of variation the data can be thought of as a table with the rows and columns representing these two sources . Two-way analysis of variance studies the variability due to

- the GROUP effect (here, variability between the columns),
- and the variability due to the TREATMENT effect (variability between the rows)

and calibrates them against the average level of variability in the data overall. Having performed the appropriate calculations, the results are displayed in an ANOVA table.

For example, for the data above

Source	D.F.	Sum of squares	Mean square	F
TREATMENT	5	0.040828	0.0081656	31.54
GROUP	2	0.002878	0.001439	5.57
Residual	10	0.002589	0.0002589	
Total	17	0.046295		

The two F statistics can be interpreted as follows;

- the first ($F = 31.54$) is the test statistic for the test of equal means in the **rows**, that is, that there is no difference between **TREATMENTS**. This statistic must be compared with an

  $$F_{5,10}$$

  distribution (the two degrees of freedom being the entries in the degrees of freedom column in the specimens and residual rows of the ANOVA table). The 95th percentile of the $F_{5,10}$ distribution is 3.33, and thus the test statistic is **more extreme** than this critical value, and thus the hypothesis that each specimen has the same mean can be **rejected**.

- The second F statistic, ($F = 5.57$), is the test statistic for the test of equal means in the **columns**, that is, that there is no difference between **GROUPS**. This statistic must be compared with an

  $$F_{2,10}$$

  distribution (the two degrees of freedom being the entries in the degrees of freedom column in the methods and residual rows of the ANOVA table). The 95th percentile of the $F_{2,10}$ distribution is 4.10, and thus the test statistic is **more extreme** than this critical value, and thus the hypothesis that each method has the same mean can be **rejected**.

Note: In this example, we do **not** have replicate data; this limits the complexity of the model that we can fit. Ideally we would like to be able to fit an **interaction** between the two-factors.

Mean parameters in two-way cross classification (full models):

No Interaction model

	I	II	III	IV	V
1	$\mu_1 + \delta_1$	$\mu_1 + \delta_2$	$\mu_1 + \delta_3$	$\mu_1 + \delta_4$	$\mu_1 + \delta_5$
2	$\mu_2 + \delta_1$	$\mu_2 + \delta_2$	$\mu_2 + \delta_3$	$\mu_2 + \delta_4$	$\mu_2 + \delta_5$
3	$\mu_3 + \delta_1$	$\mu_3 + \delta_2$	$\mu_3 + \delta_3$	$\mu_3 + \delta_4$	$\mu_3 + \delta_5$

Interaction model

	I	II	III	IV	V
1	γ_{11}	γ_{12}	γ_{13}	γ_{14}	γ_{15}
2	γ_{21}	γ_{22}	γ_{23}	γ_{24}	γ_{25}
3	γ_{31}	γ_{32}	γ_{33}	γ_{34}	γ_{35}

No Interaction Model: $8 = 3 + 5$ parameters
Interaction Model: $15 = 3 \times 5$ parameters; can be fit if replicate data are available.

5.1.4 ANOVA: KEY ASSUMPTIONS

In ANOVA, there are three key assumptions

(i) all data are **independent**

(ii) the data are **normally** distributed

(iii) the data subgroups (defined by the cross classification by factors) have **equal variances.**

Of these three points, (i) can be assessed by consideration of the study design, (ii) can be be tested formally using methods that will be described in later sections, and (iii) can be tested using statistical hypothesis testing in the following way using **Levene's Test**

LEVENE'S TEST FOR HOMOGENEITY OF VARIANCE

The Levene test is defined for the two hypotheses as follows: suppose that the data Y of size n is partitioned into K subgroups of sizes $n_1, ..., n_K$ where $n = n_1 + ... + n_K$. It is of interest to test whether the subgroups have the same variance, that is the hypothesis

$$\begin{aligned} H_0 &: \sigma_1 = \sigma_2 = ... = \sigma_K \\ H_1 &: \sigma_i \neq \sigma_j \text{ for at least one pair } (i, j). \end{aligned}$$

Test Statistic

$$W = \frac{(n-K)\sum_{i=1}^{K} n_i \left(\overline{Z}_i - \overline{Z}\right)^2}{(K-1)\sum_{i=1}^{K}\sum_{j=1}^{n_k} n_i \left(Z_{ij} - \overline{Z}_i\right)^2}$$

where Z_{ij} can have one of the following three definitions:

1. $Z_{ij} = |Y_{ij} - \overline{Y_i}|$, where $\overline{Y_i}$ is the mean of the ith subgroup.
2. $Z_{ij} = |Y_{ij} - Y_i^{(MEDIAN)}|$ where $Y_i^{(MEDIAN)}$ is the median of the ith subgroup.
3. $Z_{ij} = |Y_{ij} - Y_i^{(TRIMMED)}|$ where $Y_i^{(TRIMMED)}$ is the 10% trimmed mean of the ith subgroup.

The three choices for defining Z_{ij} determine the **robustness** (to not falsely detect unequal variances when the underlying data are not normally distributed) and **power** (to detect accurately unequal variances) of Levene's test. The Levene test rejects the hypothesis that the variances are equal at significance level α (typically, $\alpha = 0.05$) if

$$W > F_{K-1,n-K}(1-\alpha)$$

where $F_{K-1,n-K}(1-\alpha)$ is the $(1-\alpha)\%$ quantile of the Fisher F distribution with $K-1$ and $n-K$ degrees of freedom.

5.2 NON-NORMAL DATA

5.2.1 COUNTS AND PROPORTIONS

The one and two sample tests described in earlier sections can also be applied to non-normal data. A common form of non-normal data arise when the counts of numbers of "successes" or "failures" that arise in a fixed number of trials.

In this case, the Binomial distribution model is appropriate; in a one sample testing, we model the number of successes, X, by assuming

$$X \sim Binomial(n, \theta)$$

and test hypotheses about θ.

In the two sample case, we assume that the number of successes in the two samples are random variables X_1 and X_2, where

$$\begin{aligned} X_1 &\sim Binomial(n_1, \theta_1) \\ X_2 &\sim Binomial(n_2, \theta_2), \end{aligned}$$

and perhaps test the null hypothesis

$$H_0 : \theta_1 = \theta_2$$

against some alternative hypothesis ($\theta_1 \neq \theta_2, \theta_1 > \theta_2$ or $\theta_1 < \theta_2$)

5.2.2 ONE-SAMPLE TESTING

In the one sample case, two alternative approaches can be adopted:

- an **exact** test, where the distribution of the chosen test statistic under $H_0 : \theta = \theta_0$ is computed exactly, giving exact critical values and p-values
- an approximate test based on a Normal approximation to the binomial distribution.

For the exact test, we note that, **if H_0 is true**, and $\theta = \theta_0$, then $X \sim Binomial(n, \theta_0)$ so the critical values in a two-sided test can be computed directly by inspection of the $Binomial(n, \theta_0)$ c.d.f; that is

$$F_{BIN}\left(C_{R_1}; n, \theta = \theta_0\right) = 0.025 \qquad\qquad C_{R_2} = F_{BIN}\left(0.975; n, \theta = \theta_0\right)$$

where $F_{BIN}\left(-; n, \theta\right)$ is the c.d.f. of the $Binomial(n, \theta)$ distribution

$$F_{BIN}\left(x; n, \theta\right) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} \theta^{i} \left(1 - \theta\right)^{n-i}$$

where

$$\lfloor x \rfloor \text{ is largest whole number} \leq x.$$

For the approximate test, we use the fact that

$$X \sim Binomial(n, \theta) \approx Normal\left(n\theta, n\theta\left(1-\theta\right)\right)$$

and hence random variable Z

$$Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

is approximately distributed as $Normal(0,1)$. For the approximate test of $H_0 : \theta = \theta_0$, we therefore use the test statistic

$$z = \frac{x - n\theta_0}{\sqrt{n\theta_0(1-\theta_0)}}$$

(x is the actual, observed count) and compare this with the standard normal c.d.f.. This test is virtually equivalent to the one-sample t-test.

5.2.3 TWO SAMPLE TESTING

For a two sample test of $H_0 : \theta_1 = \theta_2$, we use a similar normal approximation to the one-sample case. If H_0 is true, then there is a common probability θ determining the success frequency in both samples, and the maximum likelihood estimate of θ is

$$\widehat{\theta} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{x}{n}, \text{ say}$$

and thus it can be shown that the test statistic.

$$z = \frac{\dfrac{x_1}{n_1} - \dfrac{x_2}{n_2}}{\sqrt{\dfrac{(n_1 + n_2)}{n_1 n_2} \left(\dfrac{x_1 + x_2}{n_1 + n_2} \right) \left(1 - \dfrac{x_1 + x_2}{n_1 + n_2} \right)}}$$

has an approximate standard Normal distribution.

5.2.4 CONTINGENCY TABLES

Contingency tables are constructed when a sample of data of size n are classified according to D factors, with each factor having k_d levels or categories, for $d = 1, ..., D$. When the classification is complete, the result can be represented by a D-way table of $k_1 \times k_2 \times ... \times k_D$ "cells", with each cell containing a fraction of the original data. For example, if $D = 2$, the table consists of k_1 rows and k_2 columns, and the number data in cell (i, j) is denoted n_{ij} for $i = 1, ..., k_1$ and $j = 1, ..., k_2$, where

$$\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} n_{ij} = n$$

Such a table when $D = 2$, $k_1 = 4$ and $k_2 = 6$ is illustrated below

		COLUMN						
		1	2	3	4	5	6	Total
ROW	1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}	$n_{1.}$
	2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_{26}	$n_{2.}$
	3	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	n_{36}	$n_{3.}$
	4	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_{46}	$n_{4.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	$n_{.6}$	n

This is a **cross-classification** table; it says that n_{ij} out of a total of n individuals had

- row classification i
- column classification j

5.2.5 CHI-SQUARED GOODNESS-OF-FIT TEST

It is often of interest to test whether row classification is **independent** of column classification, as this would indicate **independence** between row and column factors. An approximate test can be carried out using a **Chi-Squared Goodness-of-Fit** statistic; if the independence model is correct, the expected cell frequencies $\hat{n}_{ij}$ can be calculated as

$$\hat{n}_{ij} = \frac{n_{i.}n_{.j}}{n} \qquad i = 1, ..., k_1,\ j = 1, ..., k_2$$

where $n_{i.}$ is the *total* of cell counts in row i and $n_{.j}$ is the *total* of cell counts in column j, and that, under independence, the χ^2 test statistic

$$\chi^2 = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}$$

has an approximate chi-squared distribution with $(k_1 - 1)(k_2 - 1)$ degrees of freedom.

5.2.6 LIKELIHOOD RATIO TEST

Another approximate test is based on a **Likelihood Ratio** (**LR**) statistic

$$LR = 2 \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}}$$

This statistic also has an approximate Chi-squared distribution

$$\chi^2_{(k_1-1)(k_2-1)}$$

again given that H_0 **is true.**

It compares the “likelihood” under the independence model with the likelihood of the “saturated” model that fits a parameter for each cell in the table;

- the independence model has

$$1 + (k_1 - 1) + (k_2 - 1) = k_1 + k_2 - 1$$

  parameters, and hence

$$(k_1 \times k_2) - (k_1 + k_2 - 1) = (k_1 - 1)(k_2 - 1)$$

  degrees of freedom

- the saturated model has $(k_1 \times k_2)$ parameters, and hence 0 degrees of freedom

- the difference in degrees of freedom is hence

$$(k_1 - 1)(k_2 - 1)$$

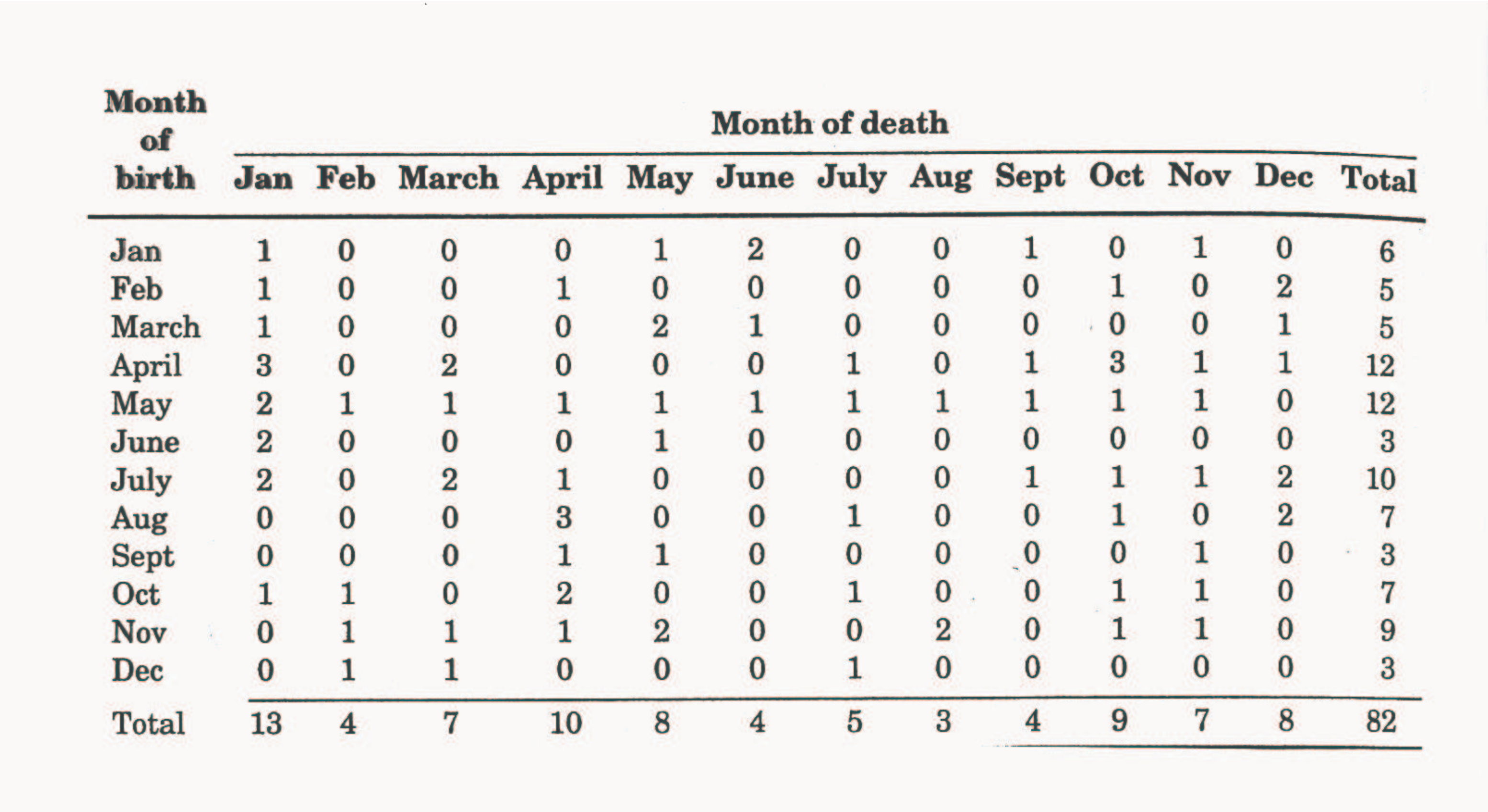
EXAMPLE: 4×4 TABLE PHENOTYPIC RELATIONSHIP

		Hair colour				
		Black	Brunette	Red	Blonde	Total
	Brown	68	119	26	7	220
	Blue	20	84	17	94	215
Eye color	Hazel	15	54	14	10	93
	Green	5	29	14	16	64
	Total	108	286	71	127	592

Number of tables: 1,225,914,276,276,768,514

Any evidence of **dependence/association** between traits ?

EXAMPLE : DESCENDENTS OF QUEEN VICTORIA



Month of birth	Month of death												
	Jan	Feb	March	April	May	June	July	Aug	Sept	Oct	Nov	Dec	Total
Jan	1	0	0	0	1	2	0	0	1	0	1	0	6
Feb	1	0	0	1	0	0	0	0	0	1	0	2	5
March	1	0	0	0	2	1	0	0	0	0	0	1	5
April	3	0	2	0	0	0	1	0	1	3	1	1	12
May	2	1	1	1	1	1	1	1	1	1	1	0	12
June	2	0	0	0	1	0	0	0	0	0	0	0	3
July	2	0	2	1	0	0	0	0	1	1	1	2	10
Aug	0	0	0	3	0	0	1	0	0	1	0	2	7
Sept	0	0	0	1	1	0	0	0	0	0	1	0	3
Oct	1	1	0	2	0	0	1	0	0	1	1	0	7
Nov	0	1	1	1	2	0	0	2	0	1	1	0	9
Dec	0	1	1	0	0	0	1	0	0	0	0	0	3
Total	13	4	7	10	8	4	5	3	4	9	7	8	82

Any evidence of **association** between birth/death months ?
Note: A very "sparse" table.

5.3 2×2 TABLES

When $k_1 = k_2 = 2$, the contingency table reduces to a two-way binary classification

		COLUMN		
		1	2	Total
	1	n_{11}	n_{12}	$n_{1.}$
ROW	2	n_{21}	n_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	n

In this case we can obtain some more explicit tests: one is again an **exact test**, the other is based on a normal approximation. The chi-squared test described above is feasible, but other tests may also be constructed:

- **FISHER'S EXACT TEST FOR INDEPENDENCE**

  Suppose we wish to test for **independence** between the row and column variables of a contingency table. When the data consist of two categorical variables, a contingency table can be constructed reflecting the number of occurrences of each factor combination. **Fisher's exact test** assesses whether the classification according to one factor is independent of the classification according to the other, that is the test is of the null hypothesis H_0 that the factors are independent, against the general alternative, **under the assumption that the row and column totals are fixed**.

- – The data for such a table comprises the row and column totals $(n_{1.}, n_{2.}, n_{.1}, n_{.2})$ and the cell entries

$$(n_{11}, n_{12}, n_{21}, n_{22})$$

The test statistic can be defined as the upper left cell entry n_{11}; for the null distribution, we compute the probability of the observing **all possible tables** with these row and column totals.. Under H_0 this distribution is **hypergeometric** and the probability of observing the table $(n_{11}, n_{12}, n_{21}, n_{22})$ is

$$p\left(n_{11}\right) = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n}{n_{.1}}} = \frac{n_{1.}!n_{.1}!n_{2.}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

where $n! = 1 \times 2 \times 3 \times .. \times (n-1) \times n$.

- – For the p-value, we need to assess the whether or not the observed table is surprising under this null distribution; suppose we observe $n_{11} = x$, then we can compare $p(x)$ with all $p(y)$ for all feasible y, that is y in the range $\max\{0, n_{1.} - (n - n_{.1})\} \leq y \leq \min\{n, n_{.1}\}$. We are thus calculating the null distribution **exactly** given the null distribution assumptions and the row and column totals; if the observed test statistic lies in the tail of the distribution, we can reject the null hypothesis of independent factors.

- **MANTEL-HAENSZEL TEST FOR INDEPENDENCE**

  This test allows you to test for independence between two factors in the presence of a third, and possibly related variable. It extends the two-way Chi-squared test of independence described above; the test statistic is a chi-squared type statistic, and the null distribution under independence is a Chi-squared distribution.

- **McNEMAR'S TEST FOR PAIRED SAMPLES**

In a 2×2 table representing paired data (where observations are, for example, matched in terms of medical history or genotype, or phenotype) the usual chi-squared test is not appropriate, and **McNemar's test** can instead be used. Consider the following table for a total of n matched pairs of observations, in which each individual in the pair has been classified (or randomized to class) A or B, with one A one B in each pair, and then the outcome (disease status, survival status) recorded.

		A		
		YES	NO	Total
	YES	n_{11}	n_{12}	$n_{1.}$
B	NO	n_{21}	n_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	n

that is, n_{11} pairs were observed for which both A and B classified individuals had disease/survival status YES, whereas n_{12} pairs were observed for which the A individual had status NO, but the B individual had status YES, and so on.

An appropriate test statistic here for a test of symmetry or "discordancy" in these results (that is, whether the two classifications are significantly different in terms of outcome) is

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

which effectively measures how different the off-diagonal entries in the table are. This statistic is an adjusted Chi-squared statistic, and has a χ_1^2 distribution under the null hypothesis that there is no asymmetry. Again a one-tailed test is carried out: "surprising" values of the test statistic are large.

SECTION 6.

NON-PARAMETRIC TESTS

The standard test for the equality of expectations of two samples is the two-sample T-test. This test is predicated on the assumption of normality of the underlying distributions. In many cases, such an assumption is inappropriate, possible due to distributional asymmetry or the presence of outliers, and thus other tests of the hypothesis of equality of population locations must be developed.

Some of the standard non-parametric tests used in statistical analysis are described below: we concentrate on two-sample tests for the most part. All of these tests can be found in good statistics packages.

References: Conover, *Practical Nonparametric Statistics*
Hollander and Wolfe, *Nonparametric Statistical Methods*

Non-parametric tests are usually based on the **ranks** of the data: typically, we

- sort the pooled data into ascending order (forming the **order statistics/empirical quantiles)**
- assign the ranks from 1 up to the total sample size to the data points
- examine statistics based on functions of the ranks (for example, the rank-sum) for data within the identified subgroups.
- base group comparison on differences in the rank statistics
- the rank statistics are used to construct a test statistic, whose distribution is typically approximated using a normal approximation.
- a “distribution-free” procedure.

6.1 THE MANN-WHITNEY-WILCOXON TEST

Consider two samples $x_1, ..., x_{n_1}$ and $y_1, ..., y_{n_2}$. The **Mann-Whitney-Wilcoxon** test proceeds as follows; first, sort the pooled sample into ascending order. Add up the ranks of the data from sample one to get u_1 say. Repeat for sample two to get u_2. Note that

$$u_1 + u_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

The Mann-Whitney-Wilcoxon statistic is u_1. It can be shown that, under the hypothesis that the data are from populations with the equal medians, then u_1 has an approximate normal distribution with mean and variance

$$\frac{n_1(n_1 + n_2 + 1)}{2} \qquad \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

This is the non-parametric alternative to the two sample t-test.

6.2 THE KOLMOGOROV-SMIRNOV TEST

The two-sample Kolmogorov-Smirnov test is a non-parametric test for comparing two samples via their **empirical cumulative distribution function**. For data $x_1, ..., x_n$, the empirical c.d.f. is the function $\widehat{F}$

$$\widehat{F}(x) = \frac{c(x)}{n} \qquad\qquad c(x) = \text{“Number of data } \leq x\text{”}$$

Thus, for two samples, we have two empirical c.d.f., and the (two-sided) **Kolmogorov-Smirnov** test that the two samples come from the same underlying distribution is based on the statistic

$$T = \max_x \left| \widehat{F}_1\,(x) - \widehat{F}_2(x) \right|.$$

It is easy to show that $0 \leq T \leq 1$, but the null distribution of T is not available in closed form. Fortunately, the p-value probability in the test for test statistic t, $p =$P$[T > t]$ can be obtained for various different sample sizes using statistical tables or packages.

NOTE : There is a one-sample version of the Kolmogorov-Smirnov test for testing whether a sample are well represented by a specified probability model with cdf F_0. It is based on the test statistic

$$T = \max_x \left| \widehat{F}_1\left(x\right) - F_0(x) \right| .$$

It can be used as a **goodness-of-fit** test, to test against a specific distribution.

6.3 TESTING NORMALITY

- **THE CHI-SQUARED GOODNESS-OF-FIT TEST**

  The **chi-squared goodness-of-fit test** is a non-parametric test for which the null distribution of the test statistic

  $$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  can be well approximated by a Chi-squared distribution. In this formula, k is the number of “bins” into which the range of the data is broken down, and

  - O_i is the number of observations **observed** to fall into bin i
  - E_i is the number of observations **expected** to fall into bin i under the normal model

Alternative tests/Assessments:

- **The Shapiro-Wilk Test:** The **Shapiro-Wilk** test can be used to test this hypothesis; the test statistic is commonly denoted W, and critical and p- values from its null distribution are available from tables or statistics packages.

- The **Kolmogorov-Smirnov one-sample** test can be used in a one-sample problem to test any distributional assumptions, including normality.

- **Probability Plotting** or **Quantile-Quantile** (QQ) plotting involves plotting empirical quantiles versus theoretical quantiles; a straight line in the QQ plot indicates that the distributional assumption is valid.

6.4 THE KRUSKAL-WALLIS TEST

The **Kruskal-Wallis rank test** is a nonparametric alternative to a *one-way analysis of variance.*

- The null hypothesis is that the true location parameter is the same in each of the samples.
- The alternative hypothesis is that at least one of the samples has a different location.
- Unlike one-way ANOVA, this test does not require normality

6.5 THE FRIEDMAN RANK SUM TEST

The **Friedman rank sum test** is a nonparametric alternative to a specific *two-way analysis of variance*

- It is appropriate for data arising from an experiment in which exactly one observation was collected from each experimental unit, or group, under each treatment.

- The elements of the samples are assumed to consist of a treatment effect, plus a group effect, plus independent and identically distributed residual errors

SECTION 7.

EXACT TESTS AND SIMULATION-BASED METHODS

Chi-squared tests involved the construction of a chi-squared statistic of the form

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

The distribution of the test-statistic is approximated by a suitable Chi-squared distribution. This approximation is

- **good** when the sample size is large
- **poor** when the table is "sparse", with some low (expected) cell entries (under the null hypothesis)

We have also seen two examples of **Exact Tests:** the exact binomial test in section (5.2.1) and Fisher's Exact Test in section (5.3). For these tests, we proceeded as follows, mimicking the general hypothesis strategy outlined at the start of the section.

1. Write down a null hypothesis H_0 and a suitable alternative hypothesis H_1
2. Construct a test statistic T deemed appropriate for the hypothesis under study
3. Compute the null distribution of T, that is the sampling distribution of T if H_0 is true, f_T
4. Compare the observed value of T, $t = T(x)$ for sample data $x = (x_1, ..., x_n)$ with the null distribution and assess whether the observed test statistic is a surprising observation from f_T; if it is reject H_0

Step 3 is crucial: for some tests (for example, one and two sample tests based on the Normal distribution assumption), it is possible to find f_T analytically for appropriate choices of T in Step 2. For others, such as the chi-squared goodness of fit and related tests, f_T is only available approximately.

However, the null distribution (and hence the critical regions and p-value) can, in theory, **always** be found : it is the probability distribution of the statistic T under the model restriction imposed by the null hypothesis.

We may not be able to compute the null distribution **analytically** (as for the tests for normal samples), but we can do it **numerically**, using **simulation**.

For most of the hypothesis tests above, we start with the assumptions and work forward to derive the sampling distribution of the test statistic under the null hypothesis.

- For **permutation tests,** we will reverse the procedure, since the sampling distribution involves the permutations which give the procedure its name and are the key theoretical issue in understanding the test.

- For **resampling** or **bootstrap** methods , we will resample the original data uniformly and randomly so as to explore the variability of a test statistic.

7.1 PERMUTATION TESTS

A permutation is a reordering of the numbers $1, ..., n$. For example, (1, 2, 3, 4, 5, 6), (1, 3, 2, 4, 5, 6), (4, 5, 2, 6, 1, 3) (3, 2, 1, 6, 4, 5) are all permutations of the numbers 1 through 6 (note that this includes the standard order in first line). There are $n! = 1 \times 2 \times 3 \times ... \times n$ permutations of n objects.

The central idea of permutation tests refers to rearrangements of the data. The null hypothesis of the test specifies that **the permutations are all equally likely**. The sampling distribution of the test statistic under the null hypothesis is computed by forming all (or many) of the permutations, calculating the test statistic for each and considering these values all equally likely.

Consider the following two group example, where we want to test for any significant difference between the groups.

$$\begin{aligned} \text{Group 1} &\quad : \quad 55, 58, 60 \\ \text{Group 2} &\quad : \quad 12, 22, 34 \end{aligned}$$

Here are the steps we will follow to use a permutation test to analyze the differences between the two groups. For the original order the sum for Group 1 is 173. In this example, if the groups were truly equal (**and the null hypothesis was true**) then randomly moving the observations among the groups would make no difference in the sum for Group 1. Some of the sums would be a little larger than the original sum and some would be a bit smaller. For the six observations there are 720 permutations of which there are 20 distinct combinations for which we can compute the sum of Group 1.

	GROUP 1	GROUP 2	SUM		GROUP 1	GROUP 2	SUM
1	55,58,60	12,22,34	173	11	12,22,60	55,58,34	94
2	55,58,12	60,22,34	125	12	12,58,22	55,60,34	92
3	55,58,22	12,60,34	135	13	55,12,22	12,55,58	89
4	55,58,34	12,22,34	148	14	12,34,60	55,58,34	106
5	55,12,60	58,22,34	127	15	12,58,34	55,22,60	104
6	55,22,60	12,58,34	137	16	55,12,34	12,58,60	101
7	55,34,60	12,22,58	149	17	22,34,60	55,58,34	116
8	12,58,60	55,22,34	130	18	22,58,34	55,22,60	114
9	22,58,60	12,55,34	140	19	55,22,34	12,58,60	111
10	34,58,60	12,22,55	152	20	12,22,34	55,58,60	68

Only **one** of the twenty orderings has a Group 1 sum that greater than that of the original ordering; thus the probability of a sum at least this large by chance alone is $1/20 = 0.05$; it can be considered statistically significant.

7.2 MONTE CARLO METHODS

Above, the permutation yielded an **exact test** because we were able to enumerate all of the possible combinations. In larger examples it will not be possible , so we will have to take a large number of random orderings, sampled uniformly from the permutation distribution.

Monte Carlo methods replace an **analytic** calculation of the probability function by a **numerical**, **simulation-based** one. The principal is that large **samples** from probability distributions can be used accurately to **approximate** the probability distribution itself.

A general **Monte Carlo** strategy for two sample testing is outlined below:

1. For two sample tests for samples of size n_1 and n_2, compute the value of the test statistic for the observed sample t^*

2. Randomly select one of the $(n_1 + n_2)!$ permutations, re-arrange the data according to this permutation, allocate the first n_1 to pseudo-sample 1 and the remaining n_2 to pseudo-sample 2, and then compute the test statistic t_1

3. Repeat 2. N times to obtain a random sample of $t_1, t_2, ..., t_N$ of test statistics from the TRUE null distribution.

4. Compute the p-value by reporting

$$\frac{\text{Number of } t_1, t_2, ..., t_N \text{ more extreme than } t^*}{N}$$

this value will be a good approximation to the true $p-$value if the Monte Carlo sample size N is large enough.

7.3 THE BOOTSTRAP AND JACKKNIFE

In statistical analysis, we usually interested in obtaining estimates of a parameter via some statistic, and also an estimate of the variability or uncertainty attached to this point estimate, and a confidence interval for the true value of the parameter.

Traditionally, researchers have relied on normal approximations to obtain standard errors and confidence intervals. These techniques are valid only if the statistic, or some known transformation of it, is asymptotically normally distributed. If the normality assumption does not hold, then the traditional methods should not be used to obtain confidence intervals. A major motivation for the traditional reliance on normal-theory methods has been computational tractability, computational methods remove the reliance on asymptotic theory to estimate the distribution of a statistic.

Resampling techniques such as the **bootstrap** and **jackknife** provide estimates of the standard error, confidence intervals, and distributions for any statistic. The fundamental assumption of bootstrapping is that the observed data are representative of the underlying population. By resampling observations from the observed data, the process of sampling observations from the population is mimicked. The key techniques are

- **THE BOOTSTRAP:** In bootstrap resampling, B new samples, each of the same size as the observed data, are drawn with replacement from the observed data. The statistic is first calculated using the observed data and then recalculated using each of the new samples, yielding a bootstrap distribution. The resulting replicates are used to calculate the bootstrap estimates of bias, mean, and standard error for the statistic.

- **THE JACKKNIFE:** In **jackknife** resampling, a statistic is calculated for the n possible samples of size $n-1$, each with one observation left out. The default sample size is $n - 1$, but more than one observation may be removed. Jackknife estimates of bias, mean, and standard error are available and are calculated differently than the equivalent bootstrap statistics.

Using the bootstrap and jackknife procedures, all informative summaries (mean, variance, quantiles etc) for the sample-based estimates' sampling distribution can be approximated.

This is vitally important if we want to compute **measures of uncertainty** (standard errors, confidence intervals) for parameters in the model, or statistics.

SECTION 8.

BAYESIAN INFERENCE

The classical (maximum-likelihood) view of Statistical Inference Theory contrasts with the alternative **Bayesian** approach.

- In Bayesian theory, the likelihood function still plays a central role,
- the likelihood is combined with a **prior** probability distribution to give a **posterior** distribution for the parameters in the model.
- Inference, estimation, uncertainty reporting and hypothesis testing can be carried out within the Bayesian framework.

8.1 PRIOR AND POSTERIOR DISTRIBUTIONS

In the Bayesian framework, inference about an unknown parameter θ is carried out via the **posterior probability distribution** that combines prior opinion about the parameter with the information contained in the likelihood $f_{X|\theta}(x;\theta)$ which represents the data contribution. In terms of events, Bayes Theorem says that

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

that is, it relates the two conditional probabilities $P(A|B)$ and $P(B|A)$.

Carrying this idea over to probability distributions, it follows that we can carry out inference via the conditional probability distribution for parameter θ **given** data $X = x$. Specifically for parameter θ, the **posterior probability distribution** for θ is denoted $p_{\theta|X}\left(\theta|x\right)$, and is calculated as

$$p_{\theta|X}\left(\theta|x\right) = \frac{f_{X|\theta}\left(x;\theta\right) p_{\theta}\left(\theta\right)}{\int f_{X|\theta}\left(x;\theta\right) p_{\theta}\left(\theta\right) d\theta} = c(x) f_{X|\theta}\left(x;\theta\right) p_{\theta}\left(\theta\right) \tag{1}$$

say, where $f_{X|\theta}\left(x;\theta\right)$ is the likelihood, and $p_{\theta}\left(\theta\right)$ is the **prior probability distribution** for θ. The denominator in (1) can be regarded as the **marginal distribution** (or **marginal likelihood**) for data X evaluated at the observed data x

$$f_X(x) = \int f_{X|\theta}\left(x;\theta\right) p_{\theta}\left(\theta\right) d\theta. \tag{2}$$

EXAMPLE: Binomial/Beta

- **Likelihood: BINOMIAL**

$$f_{X|\theta}\left(x;\theta\right) = \binom{n}{x}\theta^{x}\left(1-\theta\right)^{n-x}$$

- **Prior: BETA**

$$p_{\theta}\left(\theta\right) = \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}$$

- **Posterior :**

$$\begin{aligned} p_{\theta|X}\left(\theta|x\right) &\propto f_{X|\theta}\left(x;\theta\right)p_{\theta}\left(\theta\right) \\ &\propto \theta^{x}\left(1-\theta\right)^{n-x}\times\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1} = \theta^{x+\alpha-1}\left(1-\theta\right)^{n-x+\beta-1} \end{aligned}$$

so that the posterior is **BETA**

$$p_{\theta|X}\left(\theta|x\right) \equiv Beta\left(x+\alpha, n-x+\beta\right)$$

8.2 ESTIMATION AND UNCERTAINTY INTERVALS

Inference for the parameter θ via the posterior $\pi_{\theta|Y}(\theta|y)$ can be carried out once the posterior has been computed. Intuitively appealing methods rely on summaries of this probability distribution, that is, moments or quantiles.

For example, one Bayes estimate, $\widehat{\theta}_B$ of θ is the **posterior expectation**

$$\widehat{\theta}_B = E_{p_{\theta|X}}\left[\theta|X=x\right] = \int \theta p_{\theta|X}(\theta|x)d\theta$$

Another estimate is the **posterior mode,** $\hat{\theta}_{MODE}$

$$\hat{\theta}_{MODE} = \arg\max_{\theta} p_{\theta|X}(\theta|x)$$

Uncertainty intervals are also available (and more natural in the Bayesian setting)

- A **100(1 − $\boldsymbol{\alpha}$)%Bayesian Credible Interval** for θ is a subset C of Θ such that

$$\mathrm{P}\left[\theta \in C\right] \geq 1 - \alpha$$

- The **100(1 − $\boldsymbol{\alpha}$)%Highest Posterior Density Bayesian Credible Interval** for θ, subject to $\mathrm{P}[\theta \in C] \geq 1 - \alpha$ is a subset C of Θ such that

$$C = \left\{\theta \in \Theta : p_{\theta|X}(\theta|x) \geq k\right\}$$

  where k is the largest constant such that

$$\mathrm{P}\left[\theta \in C\right] \geq 1 - \alpha.$$

8.3 HYPOTHESIS TESTING

To mimic the Likelihood Ratio testing procedure outlined in previous sections. For two hypotheses H_0 and H_1 define

$$\alpha_0 = \mathrm{P}\left[H_0 | X = x\right] \qquad \alpha_1 = \mathrm{P}\left[H_1 | X = x\right]$$

For example,

$$\mathrm{P}\left[H_0 | X = x\right] = \int_R \pi_{\theta|X}(\theta|x) d\theta$$

where R is some region of Θ. Typically, the quantity

$$\frac{\mathrm{P}\left[H_0 | X = x\right]}{\mathrm{P}\left[H_1 | X = x\right]}$$

(the **posterior odds** on H_0) is examined.

EXAMPLE: To test two simple hypothesis

$$\begin{aligned} H_0 &: \quad \theta = \theta_0 \\ H_1 &: \quad \theta = \theta_1 \end{aligned}$$

define the prior probabilities of H_0 and H_1as p_0 and p_1 respectively. Then, by Bayes Theorem

$$\frac{\mathrm{P}\left[H_1|X=x\right]}{\mathrm{P}\left[H_0|X=x\right]} = \frac{\dfrac{f_{X|\theta}(x;\theta_1)p_1}{f_{X|\theta}(x;\theta_0)p_0 + f_{X|\theta}(x;\theta_1)p_1}}{\dfrac{f_{X|\theta}(x;\theta_0)p_0}{f_{X|\theta}(x;\theta_0)p_0 + f_{X|\theta}(x;\theta_1)p_1}} = \frac{f_{X|\theta}(x;\theta_1)p_1}{f_{X|\theta}(x;\theta_0)p_0}$$

More generally, two hypotheses or models can be compared via the observed marginal likelihood that appears in (2), that is if

$$\frac{f_X(x; \text{Model 1})}{f_X(x; \text{Model 0})} = \frac{\int f^{(1)}_{X|\theta}(x; \theta_1)\, p_{\theta_1}(\theta_1)\, d\theta_1}{\int f^{(0)}_{X|\theta}(x; \theta_0)\, p_{\theta_0}(\theta_0)\, d\theta_0}$$

is greater than one we would favour Model 1 (with likelihood $f^{(1)}_{X|\theta}$ and prior p_{θ_1}) over Model 0 (with likelihood $f^{(0)}_{X|\theta}$ and prior p_{θ_0}).