USING STATISTICS IN RESEARCH

David A. Stephens

Department of Mathematics, Imperial College

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

25th February 2004

SCHEDULE

Module 1 : 25 February

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
- Data Transformations
- Basic elements and logic of Probability Theory
- Statistical Hypothesis Testing: Introduction; One and Two sample tests for Normal samples

READING LIST:

- *Practical Statistics for Medical Research*, DG Altman
- *Schaum's Elements of Statistics II*, S. Bernstein & R. Bernstein
- *An Introduction to Medical Statistics*, M. Bland
- *SPSS For Windows Made Simple*, P. R. Kinnear & C. D. Gray

MODULE 1: CONTENTS

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: suppose that the sample has been sorted into ascending order and re-labelled $x_{(1)} < ... < x_{(n)}$. Then the pth quantile, $0 < p < 100$,is given by

$$x^{(p)} = x_{(k)}$$

where k is the nearest integer to $pn/100$. Special cases include

$$\begin{array}{ccl} \text{Median} & m & = x^{(50)}, \text{ the 50th quantile} \\ \text{Lower quartile} & q_{25} & = x^{(25)}, \text{ the 25th quantile} \\ \text{Upper quartile} & q_{75} & = x^{(75)}, \text{ the 75th quantile} \end{array}$$

$$\text{Inter-quartile range} \quad IQR = q_{75} - q_{25}$$

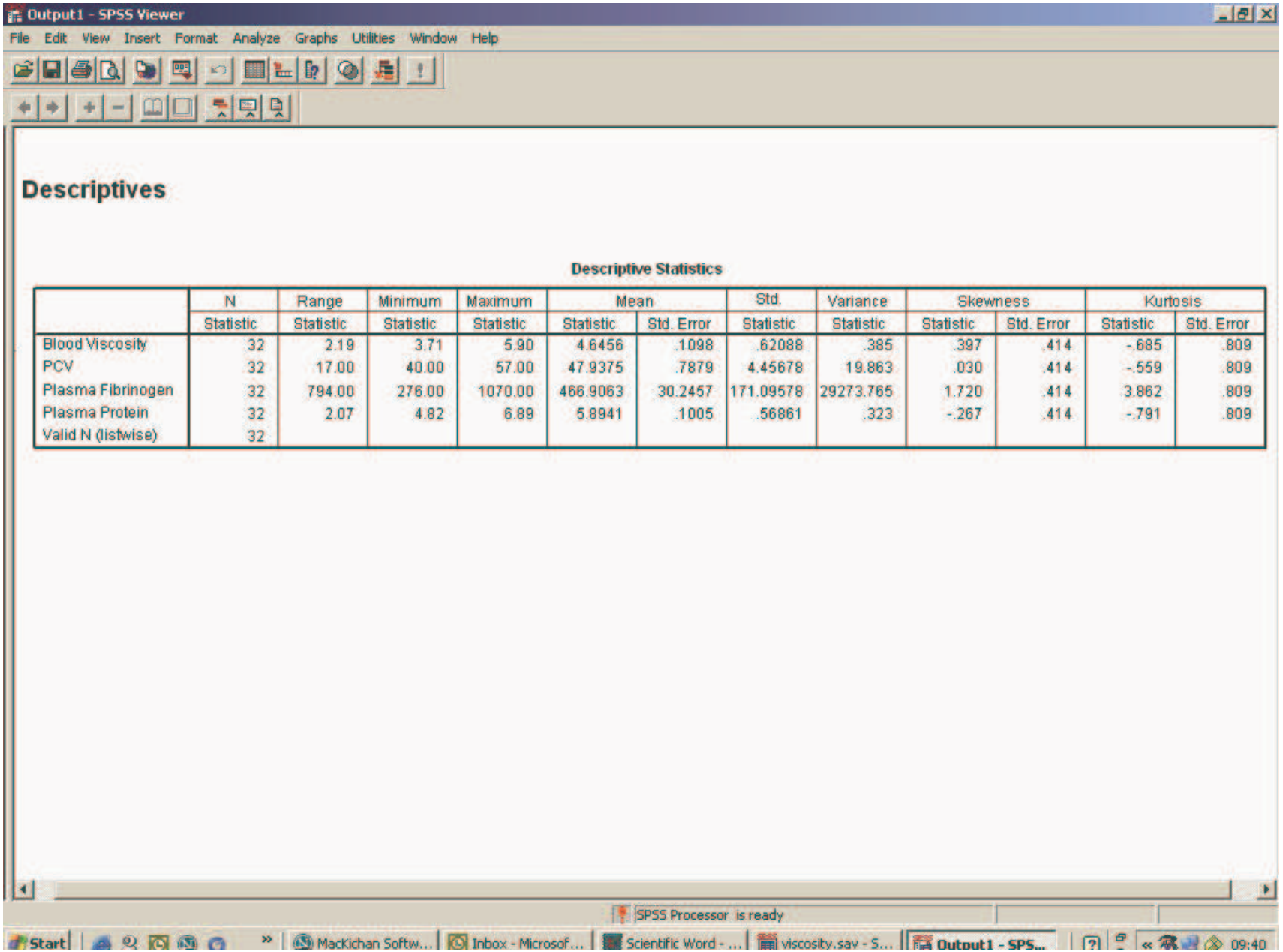
$$\begin{array}{ccl} \text{Sample minimum} & x_{\min} & = x_{(1)} \\ \text{Sample maximum} & x_{\max} & = x_{(n)} \\ \text{Sample range} & R & = x_{(n)} - x_{(1)} \end{array}$$

- 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") : $\frac{1}{n}\sum_{i=1}^{n} x_i^4$



	N	Range	Minimum	Maximum	Mean		Std.	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Blood Viscosity	32	2.19	3.71	5.90	4.6456	.1098	.62088	.385	.397	.414	-.685	.809
PCV	32	17.00	40.00	57.00	47.9375	.7879	4.45678	19.863	.030	.414	-.559	.809
Plasma Fibrinogen	32	794.00	276.00	1070.00	466.9063	30.2457	171.09578	29273.765	1.720	.414	3.862	.809
Plasma Protein	32	2.07	4.82	6.89	5.8941	.1005	.56861	.323	-.267	.414	-.791	.809
Valid N (listwise)	32											

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

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

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

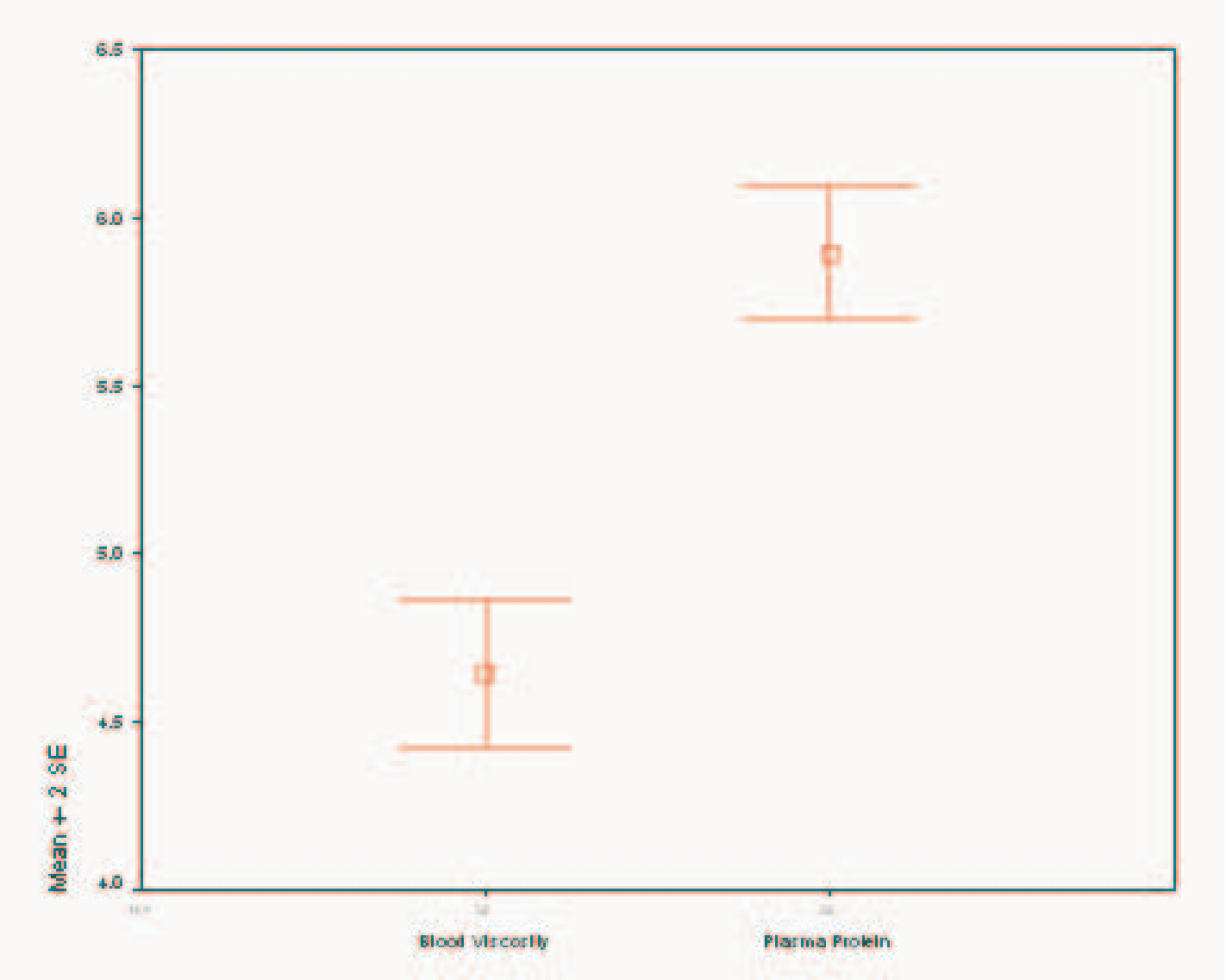
$$\bar{x} \pm SEM$$

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

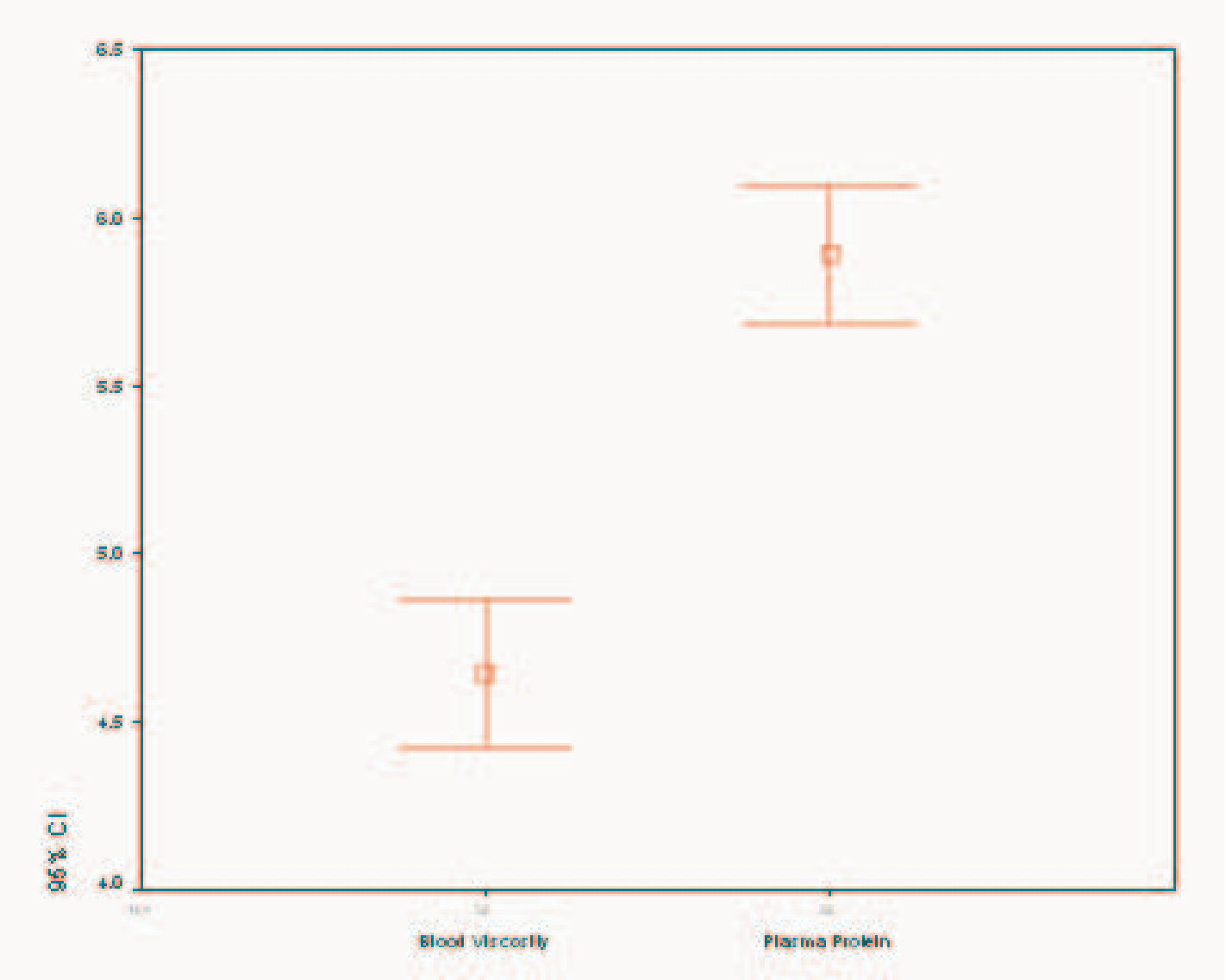
$$\bar{x} \pm 1.96 \times SEM$$

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

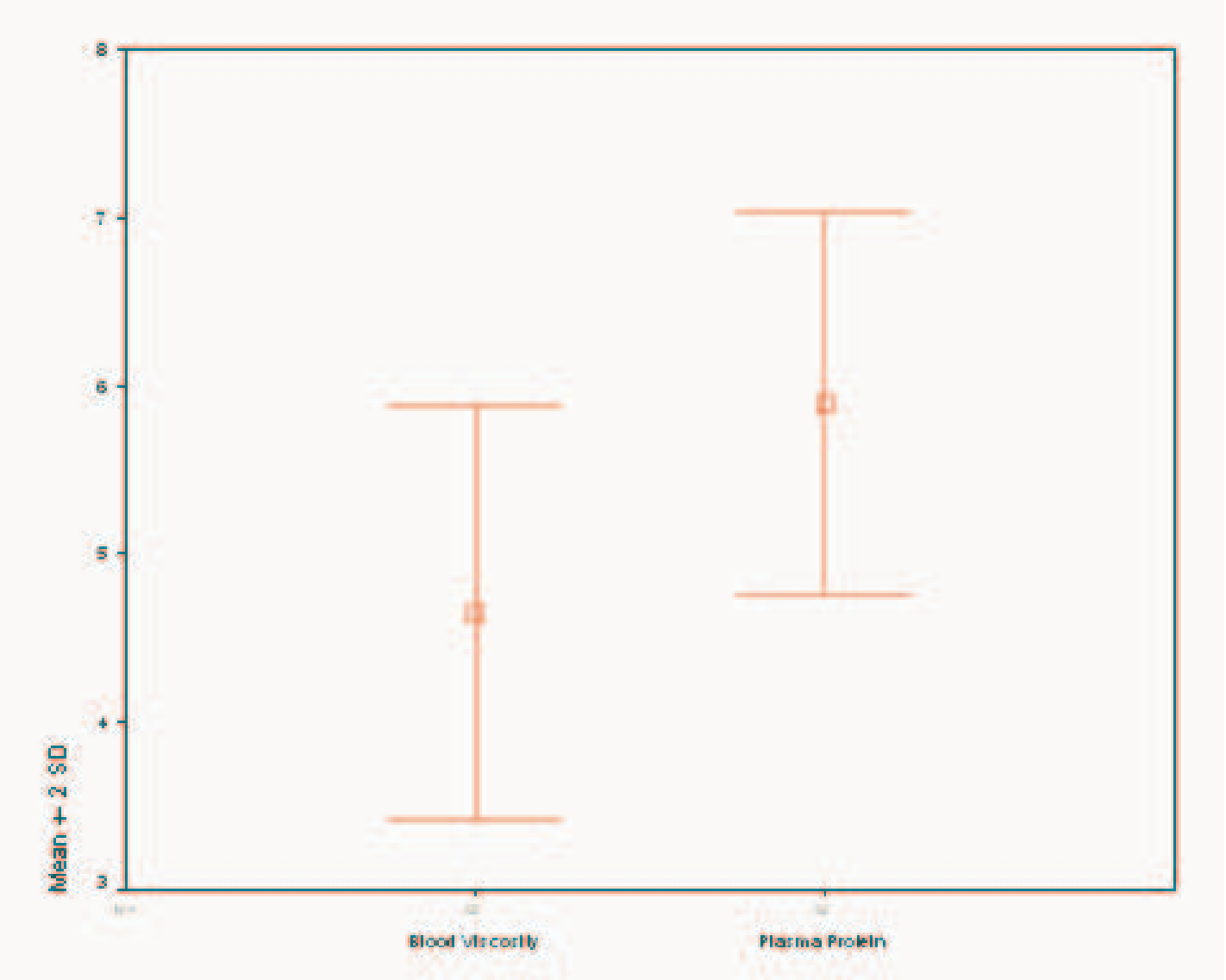
To understand the distinction, some results from probability theory are needed (see section 2.7)



MEAN±2×S.E.



CONFIDENCE INTERVAL

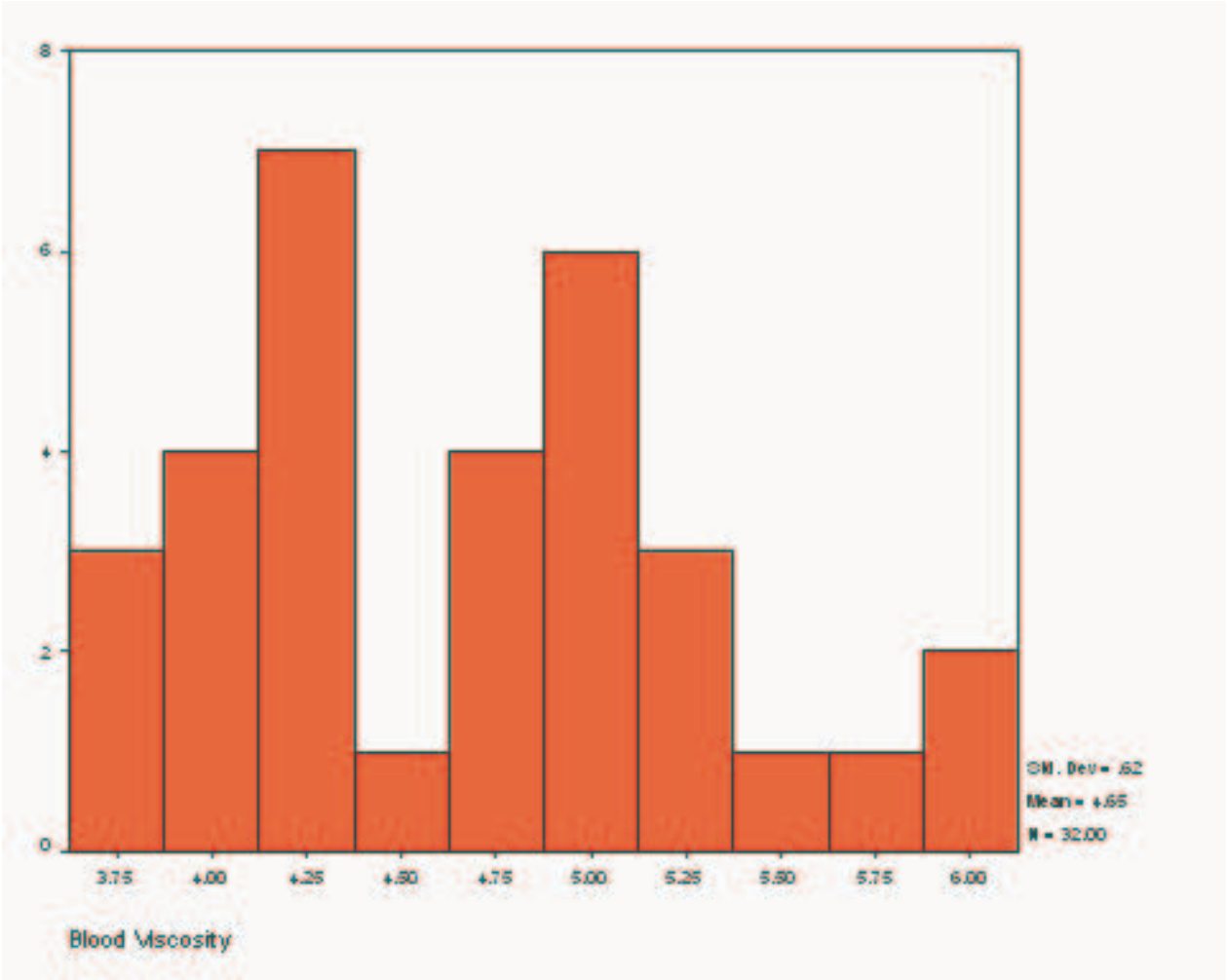


MEAN$\pm 2\times$S.D.

1.4.3 GRAPHICAL SUMMARIES

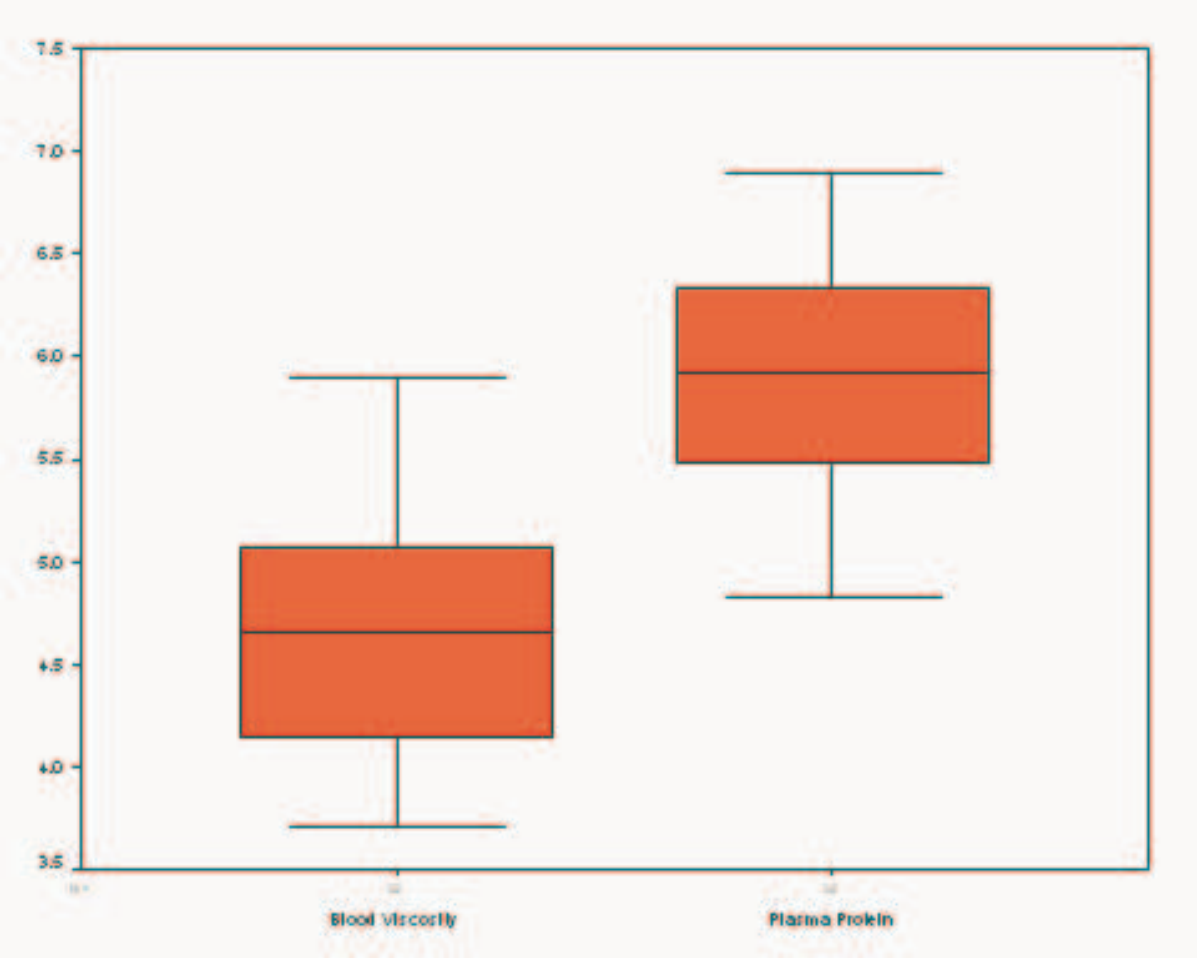
- **HISTOGRAMS**: The most common graphical summary technique is the **histogram**. Typically, the observation range, $\mathbb{X}$, is divided into a number of **bins**, $\mathbb{X}_1, ..., \mathbb{X}_H$ say, and the **frequency** with which a data value in the sample is observed to lie in subset $h = 1, ..., H$ 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

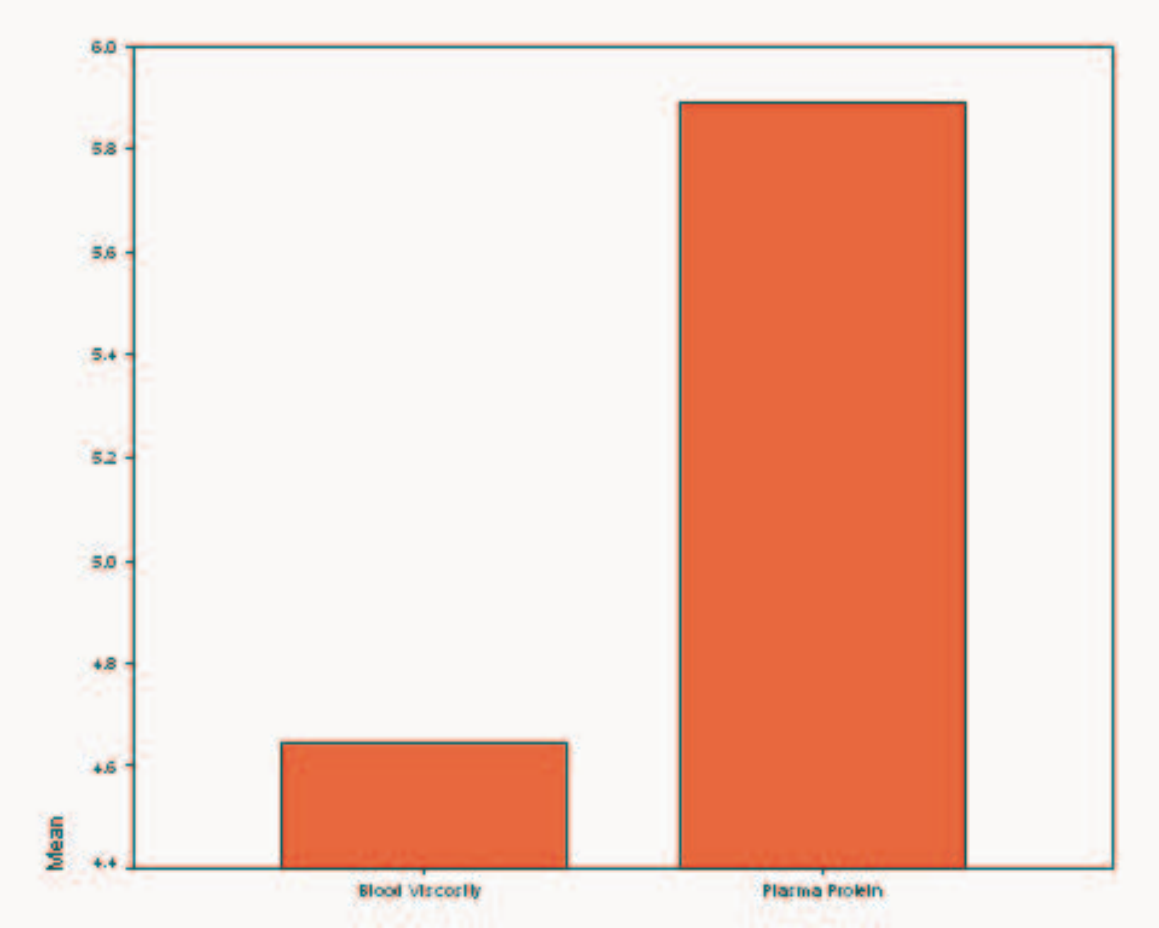
- **BOXPLOTS:** A **boxplot** is a simple way of displaying the variation in a number of data subgroups, or a mean/sem range, or a confidence interval. Typically, a **three point** (min, median, max) or **five point** (min, lower quartile, median, upper quartile, max) summary is used, and often outlying 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

PERSONAL NOTE: histogram and mean-level/boxplot plots

- typical example "BAR-CHART" (available in Excel, lesser packages...)
- often have error bars added to bars
- don't really add anything to mean/se plots
- difficult to interpret ?
- misleading ?



BAR PLOT

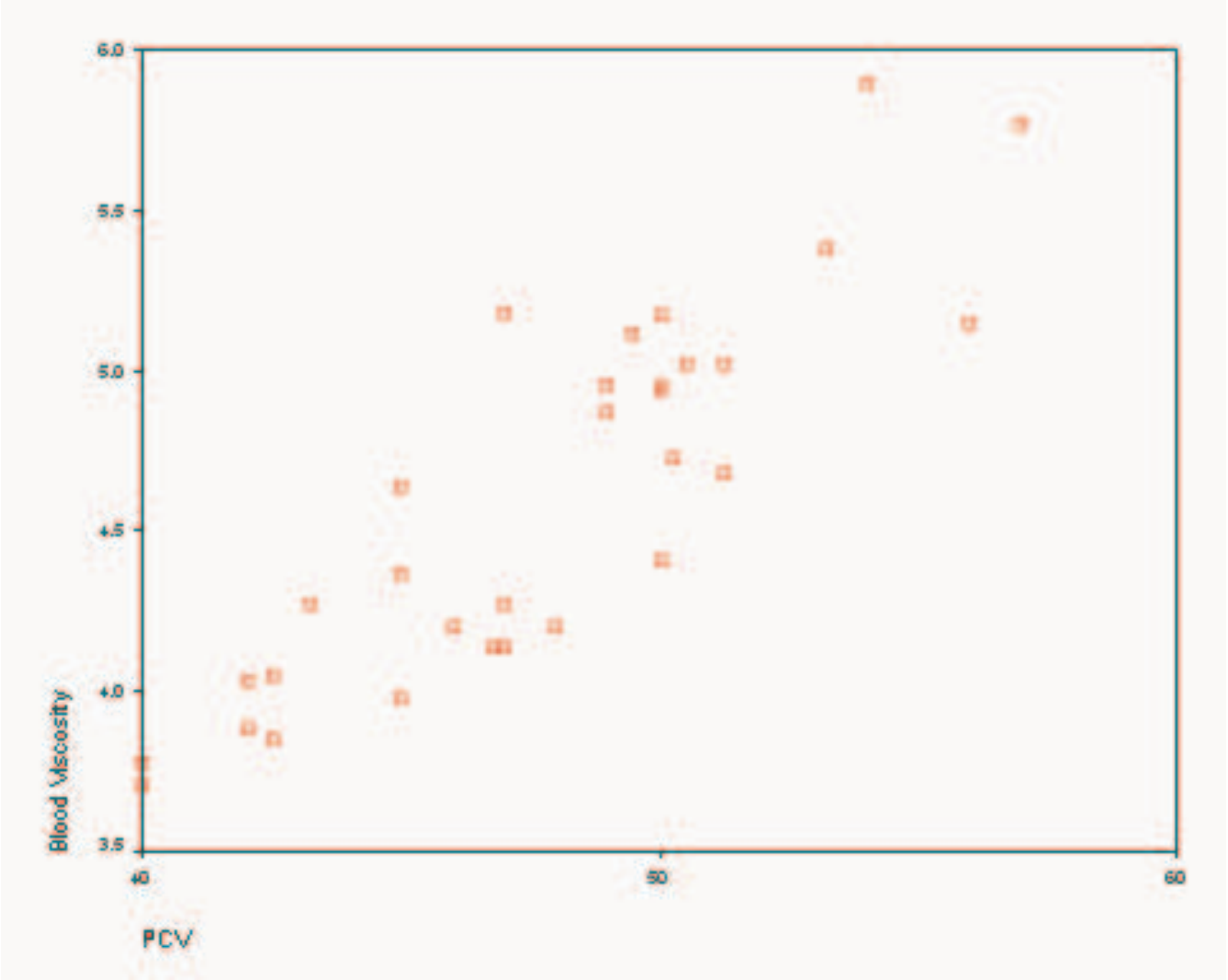
- **SCATTERPLOTS**: Scatterplots are used to illustrate the relationships between variables, and can be useful in discovering

  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,

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

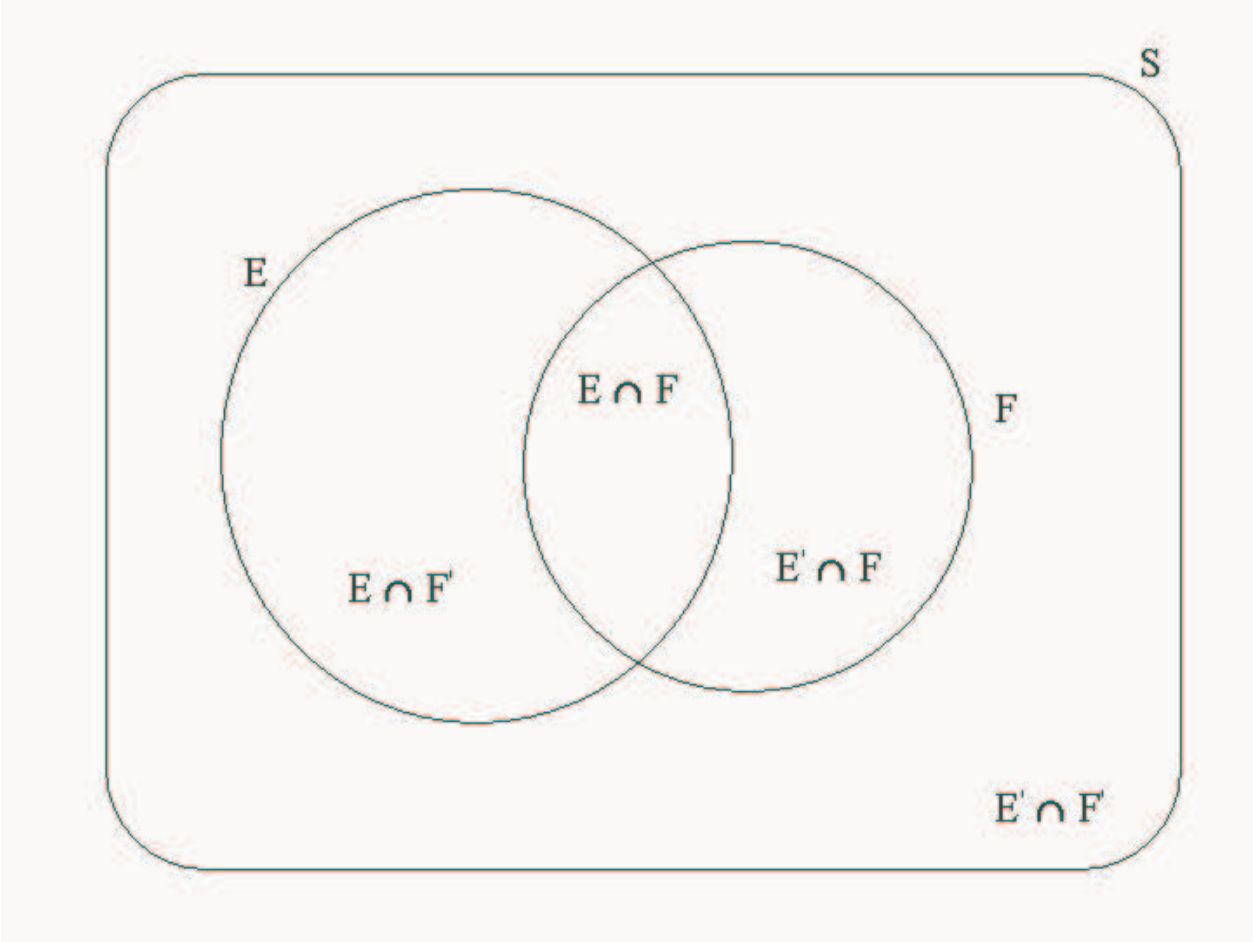
which may a finite or infinite list, or can only be represented by a continuum of outcomes, for example

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

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

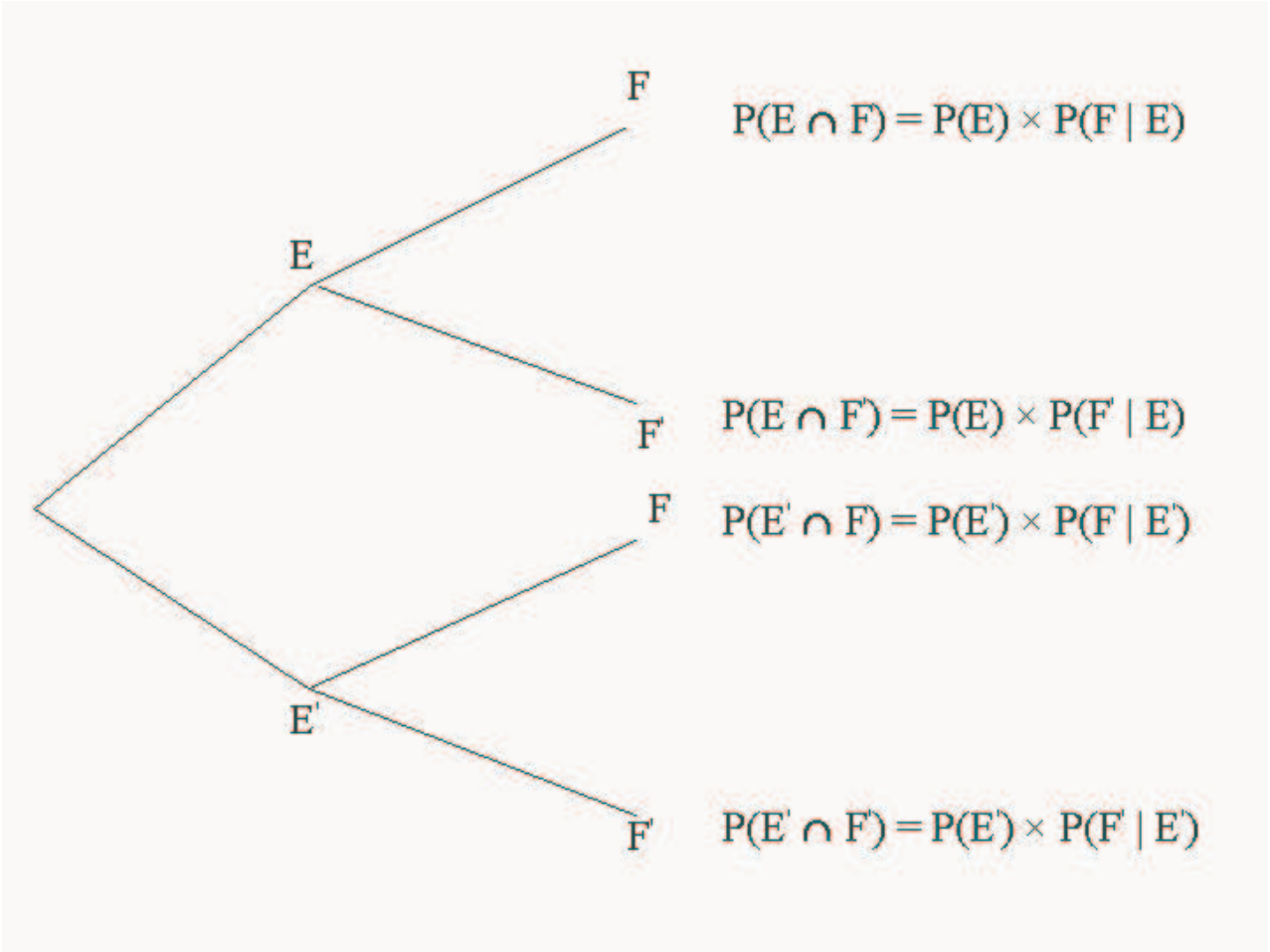
Events F and F are **independent** if

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

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

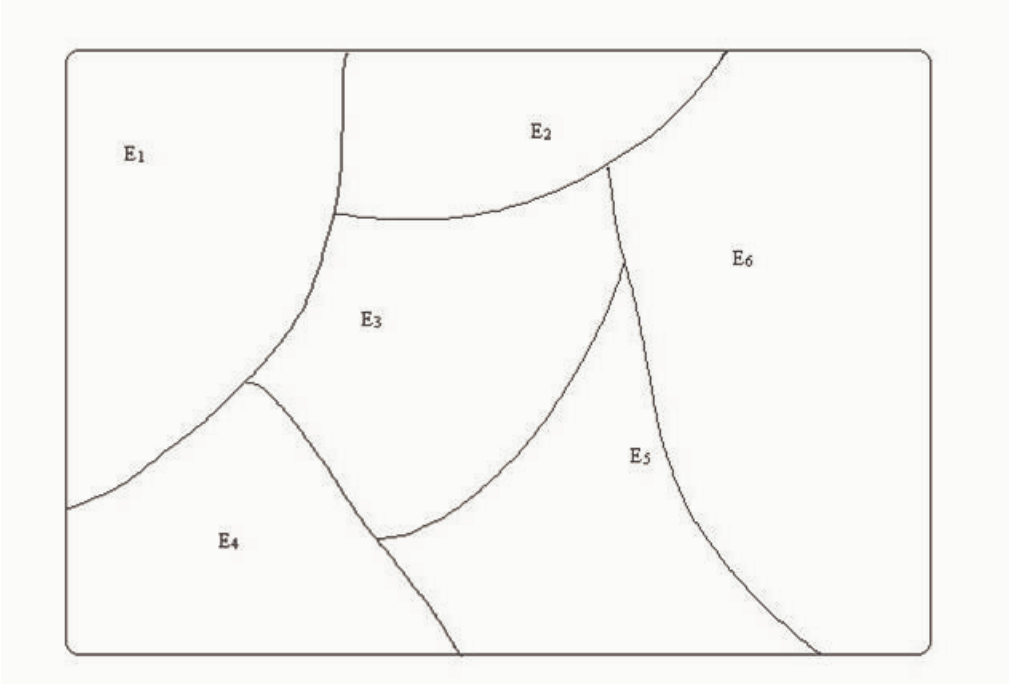
$$P(E_1 \cap ... \cap E_k) = \prod_{i=1}^{k} P(E_i) = P(E_1)...P(E_k)$$

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

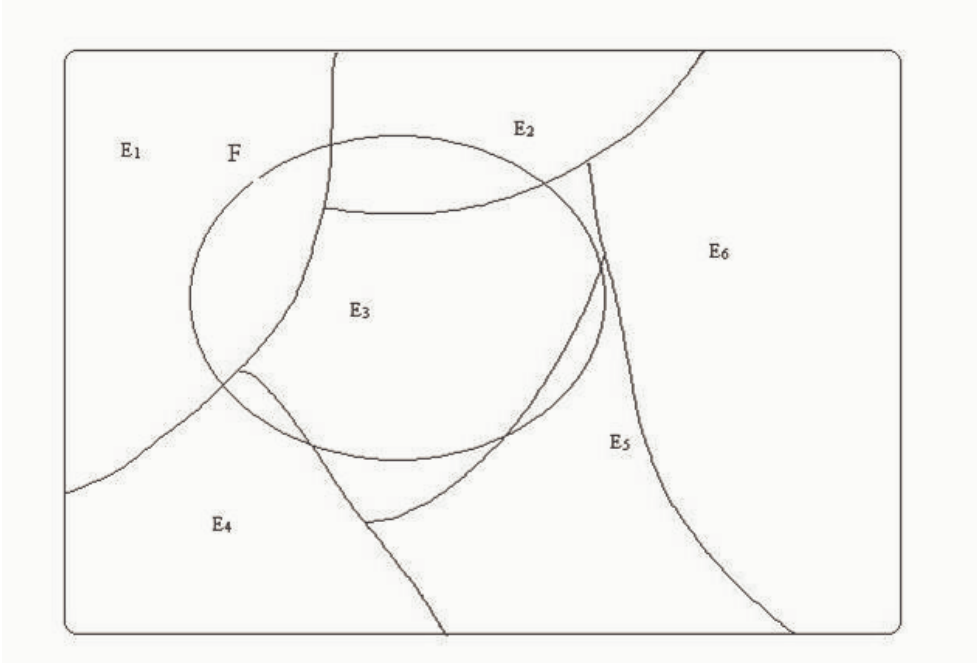


Probability Tree for the Theorem of Total Probability

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

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

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

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 $\mathrm{P}(E)$, $\mathrm{P}(F) > 0$,

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

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

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

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.

3.2.1 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.**

Note : Some simple results for sums of random variables: if $X_1, ...X_n$ are independent, identically distributed random variables, and

$$Y = \sum_{i=1}^{n} X_i \qquad \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

then

$$\begin{aligned} E_{f_{\overline{X}}}[\overline{X}] &= \frac{1}{n} E_{f_Y}[Y] = \frac{1}{n} n\mu = \mu \\ Var_{f_Y}[\overline{X}] &= \frac{1}{n^2} Var_{f_Y}[Y] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

This latter result justifies the form of the standard error formula.

3.2.2 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

3.2.3 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

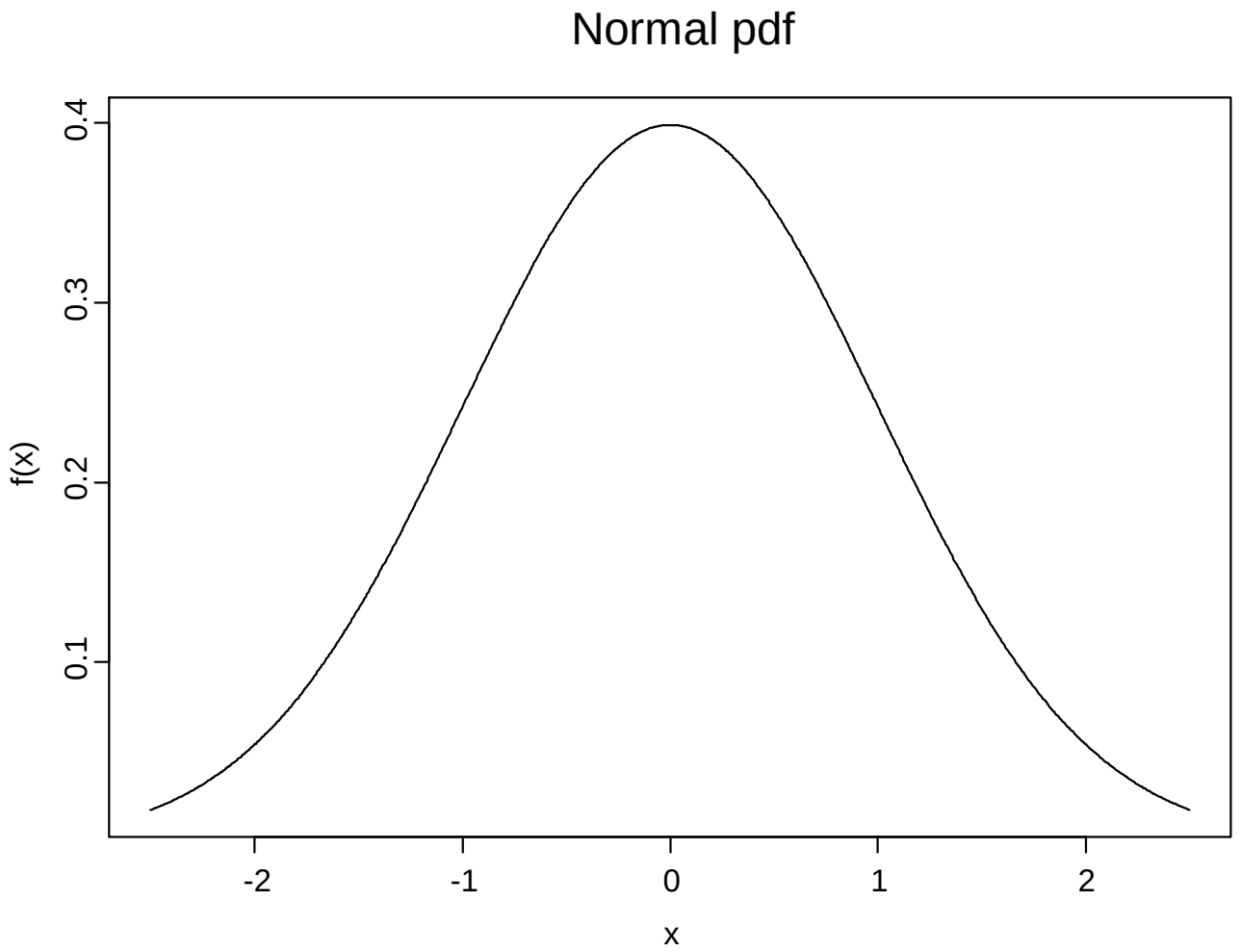
3.2.4 THE NORMAL DISTRIBUTION $X \sim N(\mu, \sigma^2)$

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

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

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

$$Y \sim Gamma(1/2, 1/2) = \chi^2_1$$

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 $Y \sim \chi^2_\alpha$ 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**.

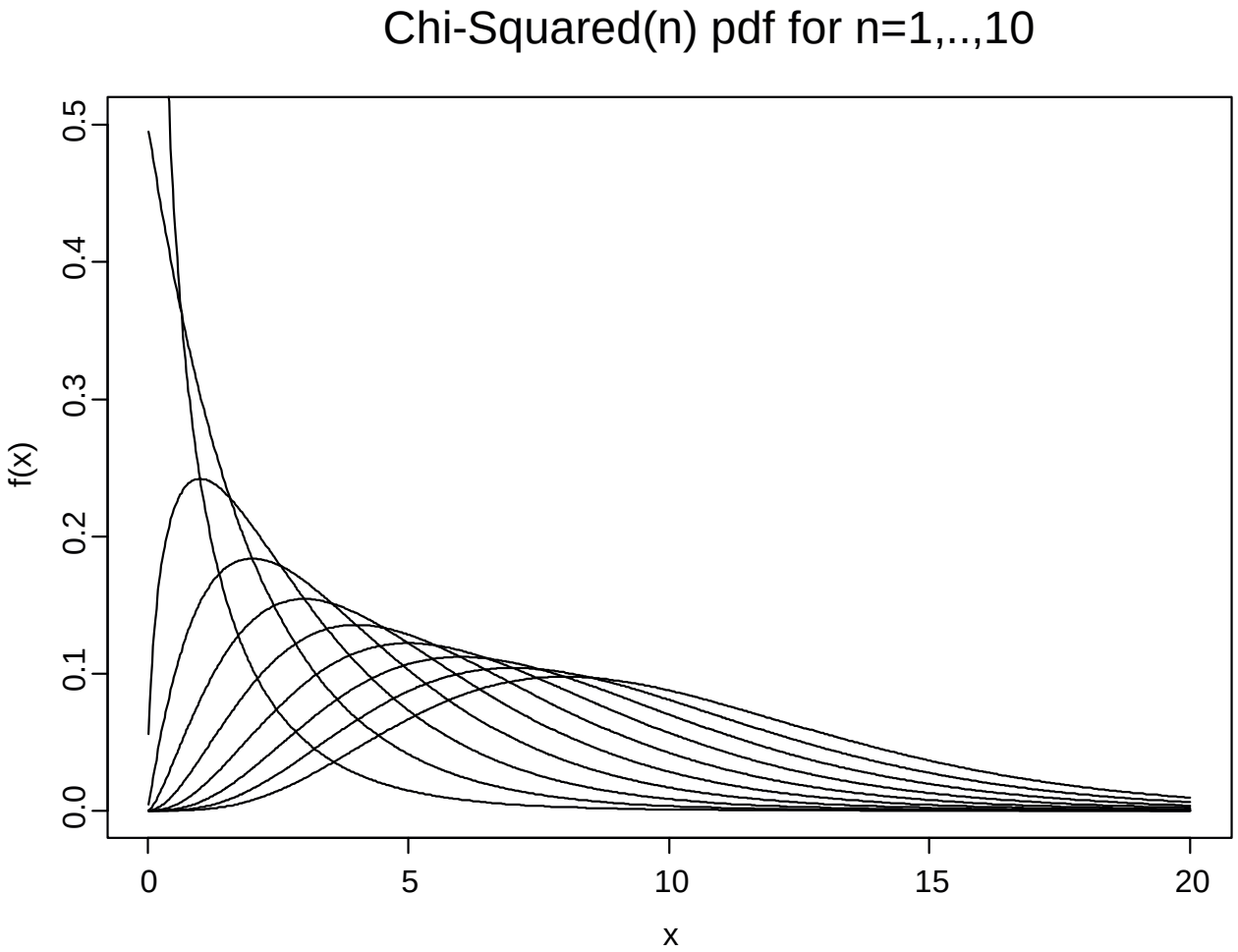
The Student-t distribution plays an important role in certain statistical testing procedures.

3.2.5 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



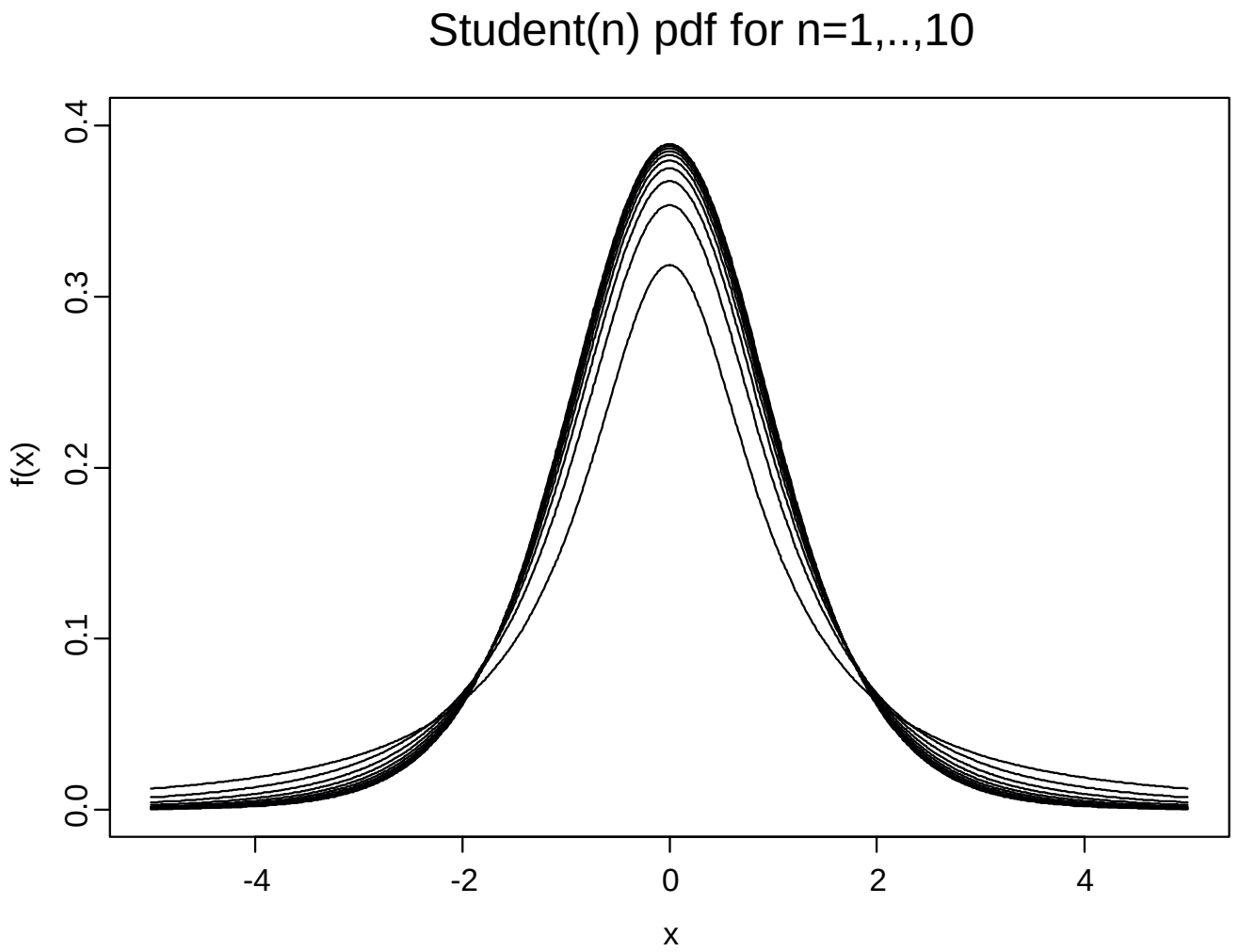
CHI-SQUARED PDF with n DEGREES OF FREEDOM

3.2.6 The Student-t Distribution

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

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

- **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

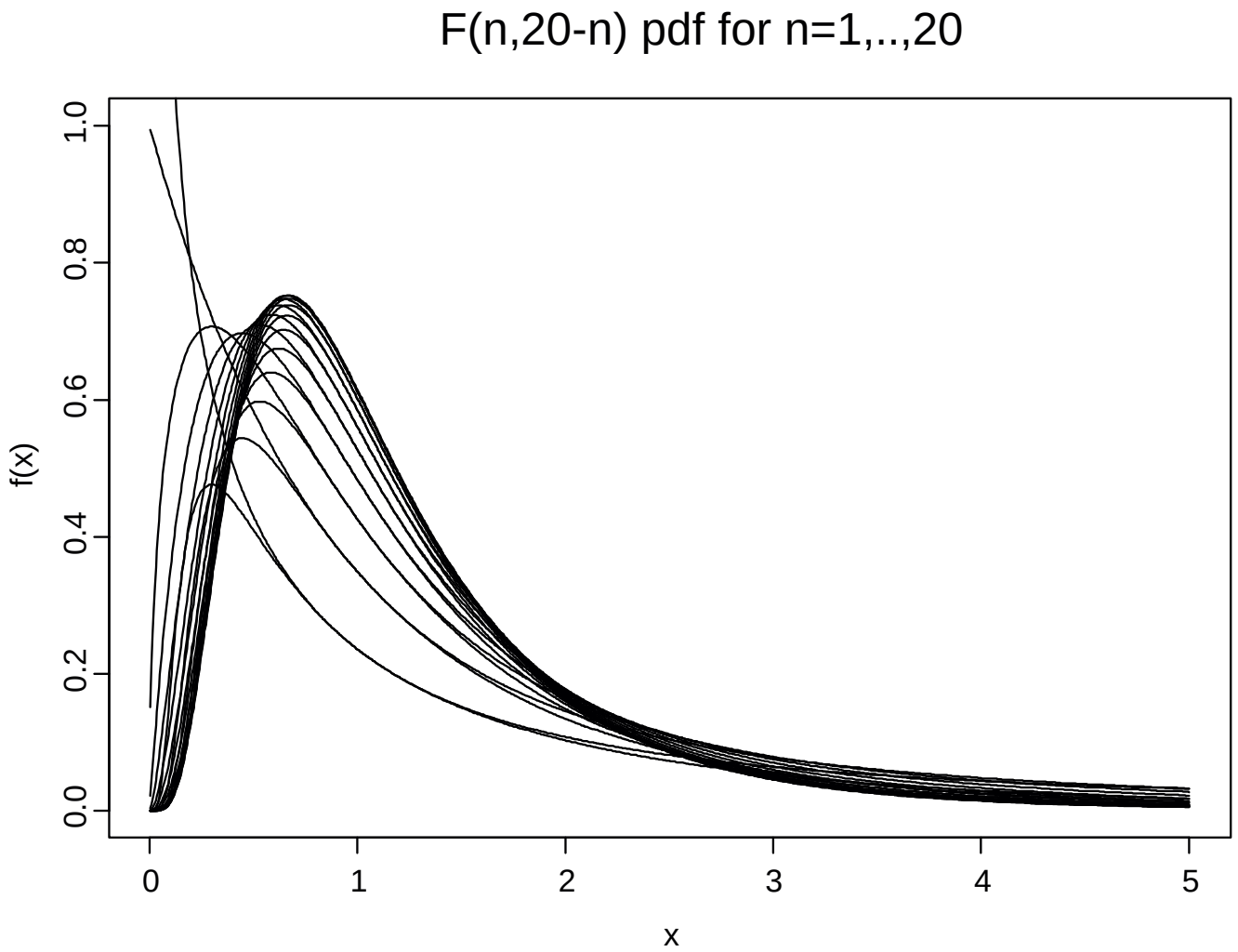
3.2.7 The Fisher-F Distribution

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

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

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

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



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

DISCRETE DISTRIBUTIONS: EXPECTATION AND VARIANCE

	Parameters	EXPECTATION	VARIANCE
$Bernoulli(\theta)$	θ	θ	$\theta(1-\theta)$
$Binomial(n,\theta)$	n,θ	$n\theta$	$n\theta(1-\theta)$
$Poisson(\lambda)$	λ	λ	λ
$Geometric(\theta)$	θ	$\frac{1}{\theta}$	$\frac{(1-\theta)}{\theta^2}$

CONTINUOUS DISTRIBUTIONS: EXPECTATION AND VARIANCE

	Parameters	EXPECTATION	VARIANCE
$Uniform(a,b)$	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Exponential(\lambda)$	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$Gamma(\alpha,\beta)$	α, β	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
$Normal(\mu,\sigma^2)$	μ, σ^2	μ	σ^2

3.3 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
$X \sim Uniform(0,1)$	$Y = -\frac{1}{\lambda}\log X$	$Y \sim Exponential(\lambda)$
$X \sim Gamma(\alpha,1)$	$Y = X/\beta$	$Y \sim Gamma(\alpha,\beta)$
$X \sim Normal(0,1)$	$Y = \mu + \sigma X$	$Y \sim Normal(\mu,\sigma^2)$
$X \sim Normal(0,1)$	$Y = X^2$	$Y \sim Gamma\left(\frac{1}{2},\frac{1}{2}\right) \equiv \chi_1$

3.4 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 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 equations 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.

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

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 $\lambda = \hat{\lambda}$.

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

4.2 SAMPLING DISTRIBUTIONS

Maximum likelihood can be used systematically to produce estimates from sample data.

EXAMPLE : If a sample of data $x_1, ..., x_n$ are believed to have a Normal distribution with parameters μ and σ^2, then the maximum likelihood estimates based on the sample are given by

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

If five samples of eight observations are collected, however, we might get

five different sample means

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

and so the estimate $\hat{\mu}$ of μ is different each time.

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

(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	$\mu = \mu_0 \qquad \sigma = \sigma_2$

- **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**.

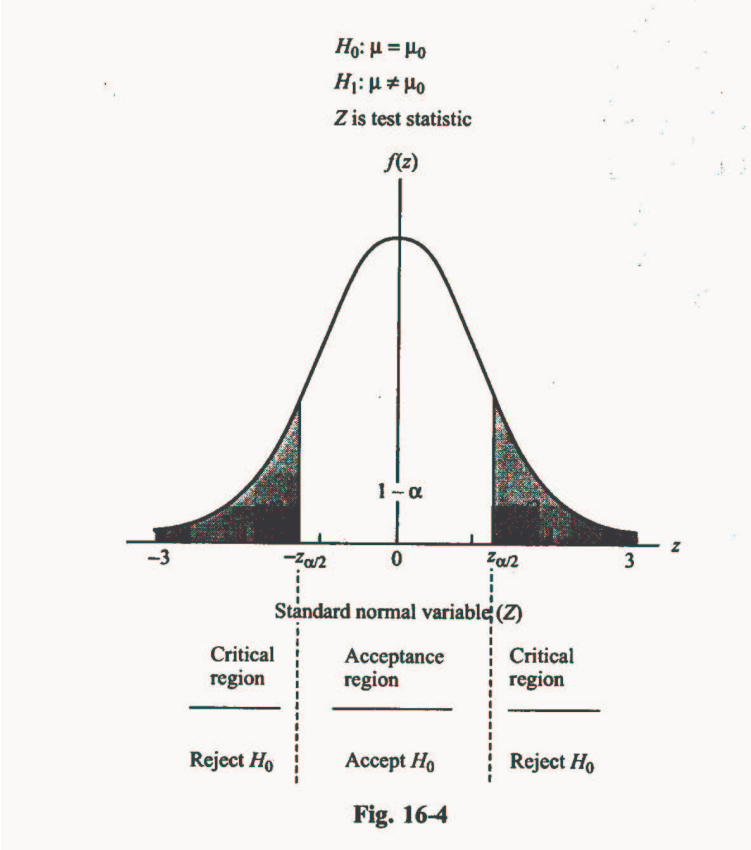


Fig. 16-4

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

$$H_0 : \mu = 20.0$$
$$H_1 : \mu \neq 20.0$$

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 this result, and recalling the sampling distributions of $\bar{X}$ and s^2, we see that

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

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

$$H_0 : \mu = 20.0$$
$$H_1 : \mu \neq 20.0$$

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

$$H_0 : \sigma^2 = \sigma_0$$
$$H_1 : \sigma^2 \neq \sigma_0$$

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

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

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. Typically, we consider sample one, $x_1, ..., x_{n_X}$, from a $N(\mu_X, \sigma_X^2)$ distribution, and sample two, $y_1, ..., y_{n_Y}$, independently from a $N(\mu_Y, \sigma_Y^2)$ distribution, and test the equality of the parameters in the two models. 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

$$H_0 : \mu_X = \mu_Y$$
$$H_1 : \mu_X \neq \mu_Y$$

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

tions 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

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

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

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

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

$$H_0 : \sigma_X = \sigma_Y$$
$$H_1 : \sigma_X \neq \sigma_Y$$

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

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

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

$$\begin{aligned} H_0 &: \mu = c \\ H_1 &: \mu > c \end{aligned} \qquad \text{or} \qquad \begin{aligned} H_0 &: \mu = c \\ H_1 &: \mu < c \end{aligned} .$$

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 $[c_1, c_2]$.

SUMMARY

For the tests discussed in previous sections, the calculation of the form of the confidence intervals is straightforward: in each case, C_{R_1} and C_{R_2} are the $\alpha/2$ and $1-\alpha/2$ quantiles of the distribution of the pivotal quantity.

ONE SAMPLE TESTS

Test	Pivotal Quantity T_{PQ}	Null Distribution	Parameter
Z-TEST	$Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$	$N(0,1)$	μ
T-TEST	$T = \dfrac{\bar{X} - \mu}{s/\sqrt{n}}$	$St(n-1)$	μ
Q-TEST	$Q = \dfrac{(n-1)s^2}{\sigma^2}$	χ^2_{n-1}	σ^2

TWO SAMPLE TESTS

Test	Pivotal Quantity T_{PQ}	Null Distribution	Parameter
Z-TEST(1)	$Z = \dfrac{(\bar{X} - \mu_X) - (\bar{Y} - \mu_Y)}{\sigma\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$	$N(0,1)$	$\mu_X - \mu_Y$
T-TEST	$T = \dfrac{(\bar{X} - \mu_X) - (\bar{Y} - \mu_Y)}{s_P\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$	$St(n_X + n_Y - 2)$	$\mu_X - \mu_Y$
Z-TEST(2)	$Z = \dfrac{(\bar{X} - \mu_X) - (\bar{Y} - \mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}}}$	$N(0,1)$	$\mu_X - \mu_Y$
Q-TEST	$Q = \dfrac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$	$F(n_X - 1, n_Y - 1)$	$\dfrac{\sigma_X^2}{\sigma_Y^2}$

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.