SHORT COURSE IN BAYESIAN INFERENCE: THEORY AND METHODS: PART I PROBABILITY AND STATISTICAL INFERENCE

DAVID A. STEPHENS

July 2002

Contents

0.1 SHORT COURSE IN BAYESIAN INFERENCE: THEORY AND METHODS vi

1 Elementary Probability **1**

1.1 Definitions, Terminology and Notation 1
1.1.1 Events and the Sample space 1
1.1.2 Operations in set theory 1
1.1.3 The Probability Function 2
1.1.4 Properties of P() : The axioms of Probability 3
1.1.5 Conditional Probability 3
1.1.6 The Theorem of Total Probability 4
1.1.7 Bayes Theorem 5
1.2 Random variables & Probability Distributions 6
1.3 Discrete Random Variables 6
1.3.1 Continuous Random Variables 8
1.4 Expectations 11
1.4.1 Properties of Expectations 11
1.4.2 Approximations to mean and variance 12
1.5 Transformations of Random Variables 13
1.5.1 1-1 Transformations 13
1.6 Joint Probability Distributions 14
1.6.1 Conditional Probability Distributions 16
1.6.2 Independence for random variables 19
1.6.3 Conditional Expectation 20
1.7 Multivariate Transformations 20
1.8 Multivariate Expectations and Covariance 22
1.8.1 Sums of Random Variables 24
1.9 Generating Functions 24
1.9.1 Properties of MGFs 24
1.10 Discrete Probability Distributions 26
1.10.1 THE DISCRETE UNIFORM DISTRIBUTION $X \sim Uniform(N)$ 26
1.10.2 THE BERNOULLI DISTRIBUTION $X \sim Bernoulli(\theta)$ 27
1.10.3 THE BINOMIAL DISTRIBUTION $X \sim Binomial(n, \theta)$ 27
1.10.4 THE GEOMETRIC DISTRIBUTION $X \sim Geometric(\theta)$ 27
1.10.5 THE NEGATIVE BINOMIAL DISTRIBUTION $X \sim NegBin(n, \theta)$ 28
1.10.6 THE POISSON DISTRIBUTION $X \sim Poisson(\lambda)$ 28
1.11 SPECIAL CONTINUOUS DISTRIBUTIONS 29
1.11.1 THE CONTINUOUS UNIFORM DISTRIBUTION $X \sim Uniform(a, b)$ 29
1.11.2 THE EXPONENTIAL DISTRIBUTION $X \sim Exponential(\lambda)$ 29
1.11.3 THE GAMMA DISTRIBUTION $X \sim Gamma(\alpha, \beta)$ 30
1.11.4 THE BETA DISTRIBUTION $X \sim Beta(\alpha, \beta)$ 31
1.11.5 THE NORMAL DISTRIBUTION $X \sim N(\mu, \sigma^2)$ 31
1.12 Probability Results & Limit Theorems 32
1.12.1 Convergence in Distribution 33
1.12.2 Convergence In Probability 34
1.12.3 The Central Limit Theorem 35

2 Statistical Inference **37**
2.1 Introduction . . . 37
2.2 Statistical Summaries . . . 37
2.3 Sampling Distributions . . . 39
2.4 Score Functions . . . 41
2.5 The Exponential Family . . . 43
2.6 Multiparameter Problems . . . 44
2.7 Properties of Estimators . . . 44
2.7.1 Asymptotic Properties of Estimators . . . 44
2.8 Variance of Estimators . . . 45
2.8.1 The Cramer-Rao Lower Bound . . . 46
2.9 Sufficiency . . . 48
2.9.1 Introduction: Concepts of Sufficiency . . . 48
2.9.2 The Neyman Factorization Theorem . . . 48
2.9.3 Minimal Sufficiency . . . 49
2.9.4 Ancillarity . . . 50
2.9.5 Sufficiency: Examples . . . 50
2.10 Minimum Variance Unbiased Estimation . . . 51
2.10.1 The Rao-Blackwell Theorem . . . 51
2.10.2 Completeness . . . 52
2.11 Methods of Estimation . . . 53
2.11.1 Maximum likelihood estimators . . . 54
2.11.2 Properties of M.L.E.s . . . 54
2.11.3 Large sample properties of MLEs . . . 55
2.11.4 MLEs and Sufficiency . . . 59
2.12 Hypothesis Testing . . . 59
2.12.1 Introduction . . . 59
2.12.2 Classical Neyman-Pearson Theory . . . 59
2.12.3 Composite Hypotheses. . . . 62
2.12.4 Local Power . . . 62
2.12.5 Composite Null Hypotheses . . . 62
2.12.6 The Likelihood Ratio Test . . . 64

3 Practical Classical Statistical Inference **65**
3.1 General Framework . . . 65
3.1.1 Objectives of a statistical analysis . . . 66
3.1.2 Exploratory data analysis . . . 66
3.1.3 Graphical Summaries . . . 68
3.1.4 Outliers . . . 68
3.2 Practical Estimation . . . 68
3.3 Practical Hypothesis Testing . . . 70
3.3.1 Hypothesis Testing for Normal data I - The Z-test (σ Known) . . . 70
3.3.2 Hypothesis testing terminology . . . 71
3.3.3 Hypothesis Testing for Normal data II - The T-test (σ Unknown) . . . 72
3.3.4 Hypothesis Testing for Normal data II - testing σ. . . . 73
3.3.5 Two sample tests . . . 74
3.3.6 Confidence Intervals . . . 75
3.4 Linear Regression Analysis . . . 77
3.4.1 Terminology . . . 77
3.4.2 Least-Squares Estimation . . . 77
3.4.3 Relationship between least-squares and maximum likelihood . . . 79
3.4.4 Estimates of Error Variance and Residuals . . . 80
3.4.5 Prediction for a new covariate value . . . 80
3.4.6 Standard Errors of Estimators and t-statistics . . . 80

3.4.7 Hypothesis Tests and Confidence Intervals for Parameters . . . 80
3.4.8 Multiple Linear Regression . . . 81
3.5 Model Validation . . . 83
3.5.1 Probability Plots . . . 83
3.5.2 The Chi-Squared Goodness-of-Fit Test . . . 84
3.5.3 Model Validation in Linear Regression . . . 86

4 Bayesian Inference and Decision Making **87**
4.1 Introduction and Terminology . . . 87
4.2 The Role of Sufficiency . . . 88
4.3 Construction of Prior Distributions . . . 89
4.4 Scale and Location Parameters . . . 91
4.5 Bayesian Estimation . . . 91
4.6 Bayesian Hypothesis Testing . . . 92
4.7 Bayesian Inference and Decision Making . . . 92
4.8 Applications of Decision Theory to Estimation . . . 93

Appendix A: Probability Distributions **95**
A.1 **DISCRETE PROBABILITY DISTRIBUTIONS** . . . 95
A.1.1 DISCRETE UNIFORM DISTRIBUTION . . . 95
A.1.2 BERNOULLI DISTRIBUTION . . . 96
A.1.3 BINOMIAL DISTRIBUTION . . . 96
A.1.4 POISSON DISTRIBUTION . . . 97
A.1.5 GEOMETRIC DISTRIBUTION . . . 100
A.1.6 NEGATIVE BINOMIAL DISTRIBUTION . . . 101
A.2 CONTINUOUS PROBABILITY DISTRIBUTIONS . . . 102
A.2.1 CONTINUOUS UNIFORM DISTRIBUTION . . . 102
A.2.2 EXPONENTIAL DISTRIBUTION . . . 102
A.2.3 GAMMA DISTRIBUTION . . . 104
A.2.4 BETA DISTRIBUTION . . . 105
A.2.5 NORMAL DISTRIBUTION . . . 106
A.3 MULTIVARIATE PROBABILITY DISTRIBUTIONS . . . 108
A.3.1 MULTINOMIAL DISTRIBUTION . . . 108
A.3.2 DIRICHLET DISTRIBUTION . . . 108
A.3.3 THE MULTIVARIATE NORMAL DISTRIBUTION . . . 109
A.4 EXPECTATION AND VARIANCE RESULTS FOR STANDARD DISTRIBUTIONS . . . 109

0.1 SHORT COURSE IN BAYESIAN INFERENCE: THEORY AND METHODS

This 3-day course will provide a comprehensive introduction to Bayesian statistical inference from both a theoretical and applied viewpoint. The objectives are to introduce participants to the probabilistic basis of Bayesian methods, to compare and contrast the Bayesian approach to Classical statistical methods, to introduce computational techniques (specifically, simulation-based methods and Markov chain Monte Carlo) that are used routinely in Bayesian statistics, and to illustrate the application of the Bayesian approach in a range of practical examples. The course will assume a basic level of mathematical training. Course material will be provided. If feasible, the course will contain practical sessions using the statistical package SPLUS.

The first section will provide an introduction of the terminology and notation of probability theory, motivated by the need for a mathematical language to express ideas of uncertainty, variation and randomness, including random variables and standard probability distributions. The purpose of this section is to introduce the necessary framework within which Bayesian and classical frameworks can be compared. . Finally, the key objectives of inference, namely estimation, hypothesis testing, model selection and comparison, prediction, will be outlined.

In order to justify the Bayesian approach, it is first important to understand the classical/frequentist approach to inference, and to discover the advantages and limitations of the such an approach. Examples of inference problems studied from the classical perspective will be given. Then a complete outline of the Bayesian approach to inference, describe the advantages of taking a Bayesian viewpoint, and contrast the Bayesian solution with classical methods. Simple problems in which analytical calculations are feasible will be studied; conjugate modelling and the classical linear model will be studied in detail. Bayesian alternatives to classical statistical techniques will be outlined. Finally, the role of decision theory as a basis for statistical inference and optimal decision making will form part of a natural justification for the Bayesian paradigm.

One commonly held reason for mistrust of the Bayesian approach is its subjectivity, and the need for the specification of a potentially controversial prior distribution. This issue will be addressed and hopefully dispelled from both a theoretical and a practical viewpoint. One of the major issues raised in previous sessions will be Bayesian computation.

To carry out Bayesian inference in complex, high-dimensional problems (that is, to produce Bayesian estimates or model summaries), it is usually necessary to carry out multidimensional integrals. This was initially a considerable problem that prevented routine implementation of Bayesian methods. Standard approaches to Bayesian computation will be studied. In particular, analytic and numerical calculation of posterior quantities of interest, using numerical integration rules (quadrature) and simulation-based methods (pseudorandom and quasirandom Monte Carlo approaches and importance sampling) will be introduced and their feasibility and ease of implementation studied. Examples of the Monte Carlo approach to integration will be given.

The key development in Bayesian computation in the last fifteen years is the use of Markov chain Monte Carlo (MCMC) methods. In this session, an introduction to MCMC methods (its history and development) in general will be given. The most important components of the mathematical background to MCMC will be introduced, and the key algorithms (Metropolis-Hastings algorithm and the Gibbs sampler) will be studied in detail. Elementary and more sophisticated examples of the use of the MCMC approach will be given, along with a selection of "Tricks and Tips" gleaned from many years of practical experience with the technique. Finally, algorithmic and conceptual extensions to the core MCMC technique, such as Variable Dimension MCMC and perfect simulation, will be outlined.

The final sessions on the last day will be given over to a range of practical examples. In this first session, topics will include Bayesian hierarchical modelling, time series analysis and state space models, survival and reliability modelling, image analysis and spatial modelling, medical statistics, pharmacokinetics etc. Where possible, this session will involve practical examples including implementation using the statistical package SPLUS. Finally, the current state and status of Bayesian inference and inferential techniques will be appraised. In particular, extensions toward what might be termed "Modern Bayesian Statistical Methods" (the extended linear model, wavelet and other basis function representations, flexible modelling, non-linear classification and discrimination) will be outlined and other topics (Bayesian networks) and applications (econometrics, finance, bioinformatics and statistical genetics) will be discussed.

David Stephens.

Dr D A Stephens
Department of Mathematics
Imperial College London

1. Elementary Probability

1.1 Definitions, Terminology and Notation

1.1.1 Events and the Sample space

Definition 1.1.1 An **experiment** is a one-off or repeatable process or procedure for which
(a) there is a well-defined set of *possible* outcomes
(b) the *actual* outcome is not known with certainty

Definition 1.1.2 A **sample outcome**, ω, is precisely one of the possible outcomes of an experiment.

Definition 1.1.3 The **sample space**, Ω, of an experiment is the set of all possible outcomes.
NOTE : Ω is a set in the mathematical sense, so set theory notation can be used. For example, if the sample outcomes are denoted $\omega_1, ..., \omega_k$, say, then

$$\Omega = \{\omega_1, ..., \omega_k\} = \{\omega_i : i = 1, ..., k\},$$

and $\omega_i \in \Omega$ for $i = 1, ..., k$.

The sample space of an experiment can be

- a FINITE list of sample outcomes, $\{\omega_1, ..., \omega_k\}$
- an INFINITE list of sample outcomes, $\{\omega_1, \omega_2, ...\}$
- an INTERVAL or REGION of a real space, $\left\{\omega \ : \ \omega \in A \subseteq \mathbb{R}^d\right\}$

Definition 1.1.4 An **event**, E, is a designated collection of sample outcomes. Event E **occurs** if the actual outcome of the experiment is one of this collection.

Special Cases of Events

Definition 1.1.5 The event corresponding to collection of *all* sample outcomes is Ω, termed the **certain** event.
The event corresponding to a collection of *none* of the sample outcomes is denoted $\emptyset$, the **impossible** event.
i.e. The sets $\emptyset$ and Ω are also events.

1.1.2 Operations in set theory

Set theory operations can be used to manipulate events in probability theory. Consider events $E, F \subseteq \Omega$. Then the three basic operations are

UNION	$E \cup F$	"E or F or both occur"
INTERSECTION	$E \cap F$	"both E and F occur"
COMPLEMENT	$E^{'}$	"E does not occur"

Properties of Union/Intersection operators
Consider events $E, F, G \subseteq \Omega$.

COMMUTATIVITY	$E \cup F = F \cup E$ $E \cap F = F \cap E$
ASSOCIATIVITY	$E \cup (F \cup G) = (E \cup F) \cup G$ $E \cap (F \cap G) = (E \cap F) \cap G$
DISTRIBUTIVITY	$E \cup (F \cap G) = (E \cup F) \cap (E \cup G)$ $E \cap (F \cup G) = (E \cap F) \cup (E \cap G)$
DE MORGAN'S LAWS	$(E \cup F)^{'} = E^{'} \cap F^{'}$ $(E \cap F)^{'} = E^{'} \cup F^{'}$

NOTE: Union and intersection are *binary* operators, that is, they take only two arguments, and thus the bracketing in the above equations is necessary. For more than two events, $E_1, E_2, ..., E_k$, say, write

$$\bigcup_{i=1}^{k} E_i = E_1 \cup ... \cup E_k \qquad \text{and} \qquad \bigcap_{i=1}^{k} E_i = E_1 \cap ... \cap E_k$$

for the union and intersection of $E_1, E_2, ..., E_k$. The extension to an *infinite* collection of events is immediate.

Definition 1.1.6 Events E and F are **mutually exclusive** if $E \cap F = \emptyset$, that is, if events E and F cannot both occur.
If the sets of sample outcomes represented by E and F are **disjoint** (have no common element), then E and F are mutually exclusive.

Definition 1.1.7 Events $E_1, ..., E_k \subseteq \Omega$ form a **partition** of event $F \subseteq \Omega$

$$(a) E_i \cap E_j = \emptyset \text{ for } i \neq j, i, j = 1, ..., k \qquad (b) \bigcup_{i=1}^{k} E_i = F.$$

so that each element of the collection of sample outcomes corresponding to event F is in *one and only one* of the collections corresponding to events $E_1, ...E_k$.

1.1.3 The Probability Function

For an event $E \subseteq \Omega$, the **probability that E occurs** is written P(E).

Interpretation : P(.) is a *set-function* that assigns "weight" to collections of possible outcomes of an experiment. There are many ways to think about precisely how this assignment is achieved;

CLASSICAL : "Consider equally likely sample outcomes ..."
FREQUENTIST : "Consider long-run *relative frequencies* ..."
SUBJECTIVE : "Consider personal degree of belief ..."

or merely think of P(.) as a set-function.

1.1.4 Properties of P() : The axioms of Probability

AXIOM 1 Let $E \subseteq \Omega$. Then $0 \leq \mathrm{P}(E) \leq 1$.

AXIOM 2 $\mathrm{P}(\Omega) = 1$.

AXIOM 3 If $E, F \subseteq \Omega$, with $E \cap F = \emptyset$, then $\mathrm{P}(E \cup F) = \mathrm{P}(E) + \mathrm{P}(F)$.

Corollaries : For events $E, F \subseteq \Omega$

$\mathrm{P}(E^{'}) = 1 - \mathrm{P}(E)$.
$\mathrm{P}(\emptyset) = 0$.
If $E \subseteq F$, then $\mathrm{P}(E) \leq \mathrm{P}(F)$.
In general, $\mathrm{P}(E \cup F) = \mathrm{P}(E) + \mathrm{P}(F) - \mathrm{P}(E \cap F)$.

Extension : If events $E_1, ..., E_k$ are **pairwise** mutually exclusive, then

$$\mathrm{P}\left(\bigcup_{i=1}^{k} E_i\right) = \sum_{i=1}^{k} \mathrm{P}(E_i)$$

1.1.5 Conditional Probability

Definition 1.1.8 For events $E, F \subseteq \Omega$ the **conditional probability** that F occurs **given** that E has occurred is written $\mathrm{P}(F|E)$, and is defined by

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

if $\mathrm{P}(E) > 0$.

NOTE: $\mathrm{P}(E \cap F) = \mathrm{P}(E)\mathrm{P}(F|E)$, and in general, for events $E_1, ..., E_k$,

$$\mathrm{P}\left(\bigcap_{i=1}^{k} E_i\right) = \mathrm{P}(E_1)\mathrm{P}(E_2|E_1)\mathrm{P}(E_2|E_1 \cap E_2)...\mathrm{P}(E_k|E_1 \cap E_2 \cap ... \cap E_{k-1}).$$

This result is known as the CHAIN or MULTIPLICATION RULE. This is the Mathematical representation of a **Probability Tree**

Definition 1.1.9 Events E and F are **independent** if

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

so that

$$\mathrm{P}(E \cap F) = \mathrm{P}(E)\mathrm{P}(F)$$

If events $E_1, ..., E_k$ are independent if, for **every** subset of events of size $l \leq k$, indexed by $\{i_1, ..., i_l\}$, say,

$$\mathrm{P}\left(\bigcap_{j=1}^{l} E_{i_j}\right) = \prod_{j=1}^{l} \mathrm{P}(E_{i_j}).$$

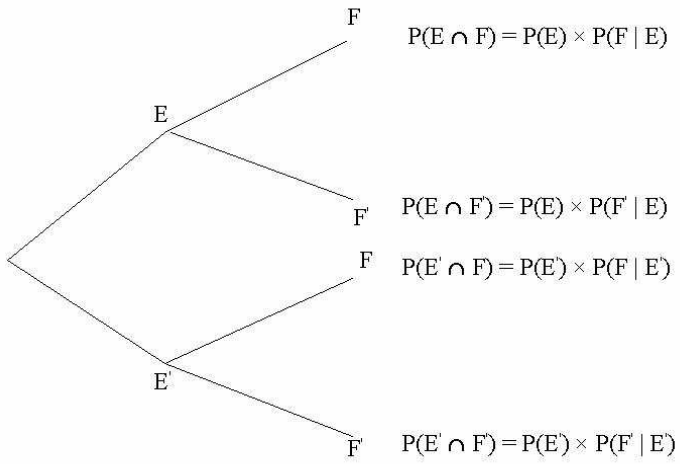


Figure 1.1: A two-step probability tree

1.1.6 The Theorem of Total Probability

Theorem 1.1.1 *Let $E_1, ..., E_k$ be a partition of Ω, and let $F \subseteq \Omega$. Then*

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

Proof. $E_1, ..., E_k$ form a partition of Ω, and $F \subseteq \Omega$, so

$$\begin{aligned} F &= (F \cap E_1) \cup ... \cup (F \cap E_k) \\ \Rightarrow \quad \mathrm{P}(F) &= \sum_{i=1}^{k} \mathrm{P}(F \cap E_i) \qquad \text{(by AXIOM3, as } E_i \cap E_j = \emptyset) \\ &= \sum_{i=1}^{k} \mathrm{P}(F|E_i)\mathrm{P}(E_i) \end{aligned}$$

■

Example 1.1.1 Simpsons Paradox - 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*	*PASSRATE*
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*	*PASSRATE*
FEMALE	120	180	0.4
MALE	30	70	0.3

so for the essay paper, the pass rate for FEMALES is higher than that for MALES. For the multiple choice test, the results were as follows;

	$PASS$	$FAIL$	$PASSRATE$
$FEMALE$	80	20	0.8
$MALE$	210	90	0.7

so for the multiple choice paper, 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. 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. The probabilities that really have been 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.

1.1.7 Bayes Theorem

Theorem 1.1.2 *Suppose $E, F \subseteq \Omega$, with $P(E), P(F) > 0$. Then*

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

Proof. By the conditional probability definition

$$\mathrm{P}(E \cap F) = \mathrm{P}(F|E)\,\mathrm{P}(E) \quad \text{and } \mathrm{P}(E \cap F) = \mathrm{P}(E|F)\,\mathrm{P}(F)$$

■

This theorem is very important because, in general,

$$\mathrm{P}(E|F) \neq \mathrm{P}(F|E)$$

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

Example 1.1.2 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

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

$$P[\text{ Suspect Passes Test } | \text{ Suspect is Not Guilty }] = P(T'|G') = 0.99 = \beta, \; say.$$

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.

1.2 Random variables & Probability Distributions

Definition 1.2.1 A **random variable** X is a function defined on a sample space Ω that associates a *real number* $X(\omega) = x$ with each possible outcome $\omega \in \Omega$. Formally, we regard X as a (possibly many-to-one) mapping from Ω to $\mathbb{R}$.

$$\begin{aligned} X: \quad & \Omega \longrightarrow \mathbb{R} \\ & \omega \longmapsto x \end{aligned}$$

Implication : we can associate any sample space Ω (for any experiment) with a sample space that is a set of real numbers, in which the events are subsets.

1.3 Discrete Random Variables

Definition 1.3.1 A random variable X is **discrete** if the set of all possible values of X (that is, the **range** of the function represented by X), denoted $\mathbb{X}$, is **countable**,

$$i.e. \quad \mathbb{X} = \{x_1, x_2, ..., x_n\} \quad \text{[FINITE]} \qquad or \qquad \mathbb{X} = \{x_1, x_2, ...\} \quad \text{[INFINITE]}$$

Definition 1.3.2 The function f_X, defined on $\mathbb{X}$ by

$$f_X(x) = \mathrm{P}[X = x] \qquad x \in \mathbb{X}$$

that assigns probability to each $x \in \mathbb{X}$ is the (discrete) **probability mass function**.

Note. Define

$$f_X(x) = 0 \qquad x \notin \mathbb{X}$$

so that f_X is defined for all $x \in \mathbb{R}$. Furthermore we will regard $\mathbb{X}$ as the *support* of random variable X, that is, the set of $x \in \mathbb{R}$ such that

$$f_X(x) > 0$$

■

Theorem 1.3.1 *A function f_X is a probability mass function for discrete random variable X with range $\mathbb{X}$ of the form $\{x_1, x_2, ...\}$ if and only if*

$$(i) f_X(x_i) \geq 0 \qquad (ii) \sum f_X(x_i) = 1$$

Definition 1.3.3 The **cumulative distribution function**, or **cdf**, F_X of a discrete random variable X is defined by

$$F_X(x) = \mathrm{P}[X \leq x] \quad \text{for} \quad x \in \mathbb{R}.$$

Theorem 1.3.2 *Let X be a discrete random variable with range $\mathbb{X} = \{x_1, x_2, ...\}$, where $x_1 < x_2 < ...$, and probability mass function f_X and cdf F_X. Then for any real value x, if $x < x_1$, then $F_X(x) = 0$, and for $x \geq x_1$,*

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i)$$

and hence $f_X(x_1) = F_X(x_1)$ and

$$f_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \qquad i = 2, 3, ...$$

Properties of F_X

(i) In the limiting cases,

$$\lim_{x \to -\infty} F_X(x) = 0 \qquad \lim_{x \to \infty} F_X(x) = 1.$$

(ii) F_X is **continuous from the right** (but not continuous) on $\mathbb{R}$, that is, for $x \in \mathbb{R}$,

$$\lim_{h \to 0^+} F_X(x + h) = F_X(x)$$

(iii) F_X is **non-decreasing**, that is

$$a < b \Longrightarrow F_X(a) \leq F_X(b)$$

(iv) $\mathrm{P}[a < X \leq b] = F_X(b) - F_X(a)$

The functions f_X and/or F_X can be used to describe the **probability distribution** of random variable X. A graph of the function f_X is non-zero only at the elements of $\mathbb{X}$. A graph of the function F_X is a **step-function** which takes the value zero at minus infinity, the value one at infinity, and is non-decreasing with points of discontinuity at the elements of $\mathbb{X}$.

1.3.1 Continuous Random Variables

Definition 1.3.4 A random variable X is **continuous** if the range of X, $\mathbb{X}$, is **uncountable**, and the function F_X defined on $\mathbb{R}$ by

$$F_X(x) = \mathrm{P}[X \leq x]$$

for $x \in \mathbb{R}$ is a **continuous** function on $\mathbb{R}$, that is, for $x \in \mathbb{R}$,

$$\lim_{h \to 0} F_X(x+h) = F_X(x).$$

Definition 1.3.5 The **cumulative distribution function**, or **cdf**, F_X of a continuous random variable X is defined by

$$F_X(x) = \mathrm{P}[X \leq x] \qquad \text{for} \quad x \in \mathbb{R}.$$

Definition 1.3.6 The **probability density function**, or **pdf**, f_X of a continuous random variable X is defined **in terms of** F_X by

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt$$

NOTES:

1. Such a function f_X need not exist but continuous random variables where f_X cannot be defined in this way will be ignored. The function f_X can be defined piecewise on intervals of $\mathbb{R}$.

2. Directly from the definition, at values of x where F_X is differentiable x,

$$f_X(x) = \frac{d}{dt}\{F_X(t)\}_{t=x}$$

3. If X is continuous,

$$f_X(x) \neq \mathrm{P}[X = x] = \lim_{h \to 0}[F_X(x+h) - F_X(x)] = 0.$$

4. For any interval $(a, b]$,

$$\mathrm{P}[a < X \leq b] = \mathrm{P}[a \leq X < b] = \mathrm{P}[a \leq X \leq b] = \mathrm{P}[a < X < b] = F_X(b) - F_X(a).$$

Theorem 1.3.3 *A function f_X is a pdf for a continuous random variable X if and only if*

$$(i) f_X(x) \geq 0 \text{ for all } x \in \mathbb{R} \qquad (ii) \int_{-\infty}^{\infty} f_X(x)\ dx = 1$$

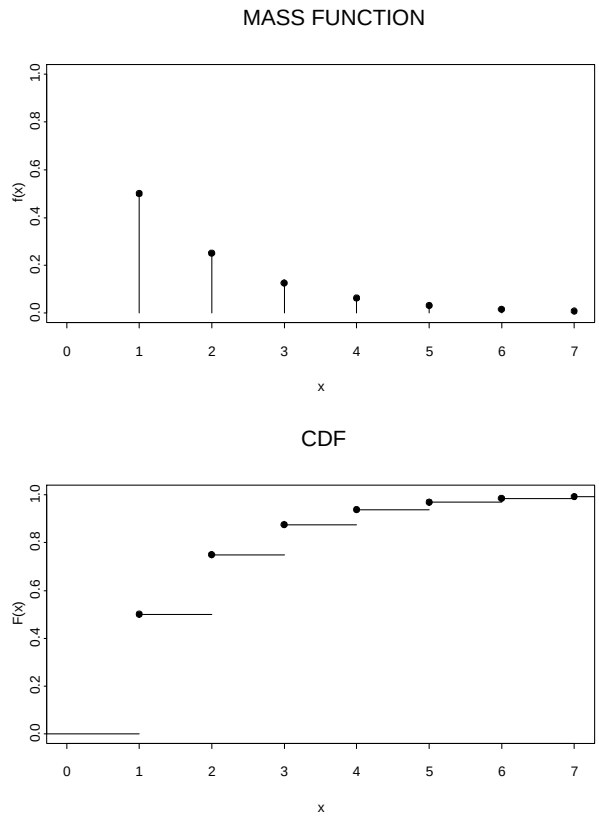


Figure 1.2: Discrete mass function and CDF.

Example 1.3.1 Consider a coin tossing experiment where a fair coin is tossed repeatedly under identical experimental conditions, with the sequence of tosses independent, until a Head is obtained. For this experiment, the sample space, Ω is then the set of sequences $(\{H\}, \{TH\}, \{TTH\}, \{TTTH\} ...)$ with associated probabilities 1/2, 1/4, 1/8, 1/16, ...

Define **discrete** random variable $X : \Omega \longrightarrow \mathbb{N} \subset \mathbb{R}$, by $X(\omega) = x \Leftrightarrow$first H on toss x. Then

$$f_X(x) = \mathrm{P}[X = x] = \left(\frac{1}{2}\right)^x \qquad x = 1, 2, 3, ...$$

and zero otherwise. For $x \geq 1$, let $k(x)$ be the largest integer not greater than x, then

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i) = \sum_{i=1}^{k(x)} f_X(i) = 1 - \left(\frac{1}{2}\right)^{k(x)}$$

and $F_X(x) = 0$ for $x < 1$.

Graphs of the probability mass function (left) and cumulative distribution function (right) are shown in Figure 1.2. Note that the mass function is only non-zero at points that are elements of $\mathbb{X}$, and that the cdf is defined for all real values of x, but is only continuous from the right. F_X is therefore a step-function.

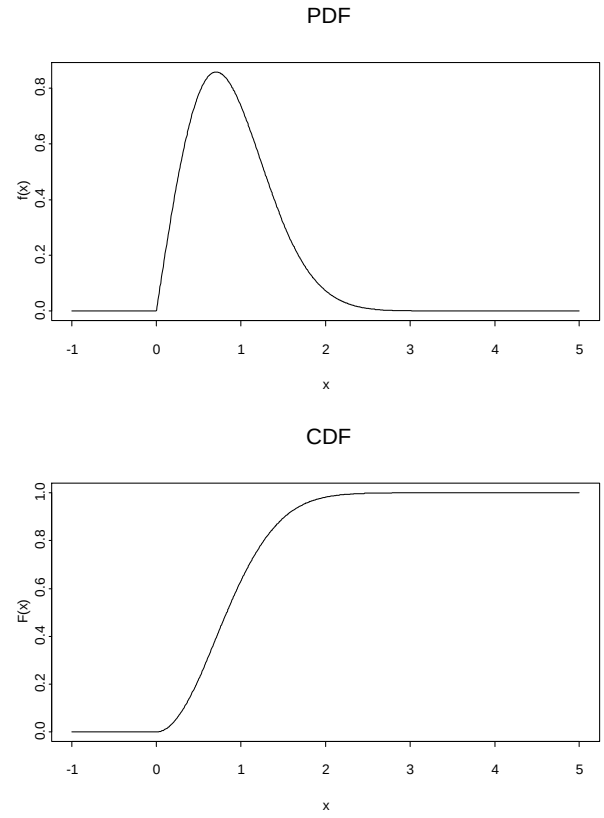


Figure 1.3: Continuous PDF and CDF

Example 1.3.2 Consider an experiment to measure the length of time that an electrical component functions before failure. The sample space of outcomes of the experiment, Ω is $\mathbb{R}^+$, and if A_x is the event that the component functions for longer than $x > 0$ time units, suppose that $\mathrm{P}(A_x) = \exp\left\{-x^2\right\}$.

Define **continuous** random variable $X : \Omega \longrightarrow \mathbb{R}^+ \subset \mathbb{R}$, by $X(\omega) = x \iff$ component fails at time x. Then, if $x > 0$,

$$F_X(x) = \mathrm{P}[X \leq x] = 1 - \mathrm{P}(A_x) = 1 - \exp\left\{-x^2\right\}$$

and $F_X(x) = 0$ if $x \leq 0$. Hence if $x > 0$,

$$f_X(x) = \frac{d}{dt}\left\{F_X(t)\right\}_{t=x} = 2x \exp\left\{-x^2\right\}$$

and zero otherwise.

Graphs of the probability density function (left) and cumulative distribution function (right) are shown in Figure 1.3. Note that both the pdf and cdf are defined for all real values of x, and that both are continuous functions.

Note that here

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt = \int_{0}^{x} f_X(t)\ dt$$

as $f_X(x) = 0$ for $x \leq 0$, and also that

$$\int_{-\infty}^{\infty} f_X(x)\ dx = \int_{0}^{\infty} f_X(x)\ dx = 1$$

Note that in both examples we have that

$$\lim_{x \to -\infty} F_X(x) = 0 \qquad \lim_{x \to \infty} F_X(x) = 1$$

1.4 Expectations

Definition 1.4.1 For a **discrete** random variable X with range $\mathbb{X}$ with probability mass function f_X, the **expectation** or **expected value** of X with respect to f_X is defined by

$$\mathrm{E}_{f_X}[X] = \sum_{x=-\infty}^{\infty} x f_X(x) = \sum_{x \in \mathbb{X}} x f_X(x)$$

For a **continuous** random variable X with range $\mathbb{X}$ and pdf f_X, the **expectation** or **expected value** of X with respect to f_X is defined by

$$\mathrm{E}_{f_X}[X] = \int_{-\infty}^{\infty} x f_X(x)\ dx = \int_{\mathbb{X}} x f_X(x)\ dx$$

The expected value is a weighted average of the values of X, but the sum/integral may not be convergent, and hence the expected value may be infinite.

Extension Let g be a real-valued function whose domain includes $\mathbb{X}$. Then

$$\mathrm{E}_{f_X}[g(X)] = \begin{cases} \sum_{x=-\infty}^{\infty} g(x) f_X(x) & \text{if } X \text{ is DISCRETE} \\ \int_{-\infty}^{\infty} g(x) f_X(x)\ dx & \text{if } X \text{ is CONTINUOUS} \end{cases}$$

1.4.1 Properties of Expectations

Let X be a random variable with mass function/pdf f_X. Let g and h be real-valued functions whose domains include $\mathbb{X}$, and let a and b be constants. Then

$$\mathrm{E}_{f_X}[ag(X) + bh(X)] = a\mathrm{E}_{f_X}[g(X)] + b\mathrm{E}_{f_X}[h(X)]$$

as (in the continuous case)

$$\begin{aligned} \mathrm{E}_{f_X}[ag(X) + bh(X)] &= \int [ag(x) + bh(x)] f_X(x)\ dx \\ &= a \int g(x) f_X(x)\ dx + b \int h(x) f_X(x)\ dx \\ &= a\mathrm{E}_{f_X}[g(X)] + b\mathrm{E}_{f_X}[h(X)] \end{aligned}$$

Special Cases :

(i) $\mathrm{E}_{f_X}[aX + b] = a\mathrm{E}_{f_X}[X] + b$

(ii) Consider $g(x) = (x - \mathrm{E}_{f_X}[X])^2$. Write $\mu = \mathrm{E}_{f_X}[X]$ (a constant that does not depend on x). Then

$$\begin{aligned}\mathrm{E}_{f_X}[g(X)] &= \int (x-\mu)^2 f_X(x)\, dx = \int x^2 f_X(x)\, dx - 2\mu \int x f_X(x)\, dx + \mu^2 \int f_X(x)\, dx \\ &= \int x^2 f_X(x)\, dx - 2\mu^2 + \mu^2 = \int x^2 f_X(x)\, dx - \mu^2 = \mathrm{E}_{f_X}[X^2] - \{\mathrm{E}_{f_X}[X]\}^2\end{aligned}$$

(iii) If $\mathrm{E}_{f_X}[X] = \mathrm{E}_{f_X}[X^2] - \{\mathrm{E}_{f_X}[X]\}^2$. Then $\mathrm{E}_{f_X}[X]$ is the **variance** of the distribution. Similarly, $\sqrt{\mathrm{E}_{f_X}[X]}$ is the **standard deviation** of the distribution.

(iv) Consider $g(x) = x^k$ for $k = 1, 2, ...$ Then in the continuous case

$$\mathrm{E}_{f_X}[g(X)] = \mathrm{E}_{f_X}[X^k] = \int x^k f_X(x)\, dx,$$

and $\mathrm{E}_{f_X}[X^k]$ is the kth **moment** of the distribution.

(v) Consider $g(x) = (x - \mu)^k$ for $k = 1, 2, ...$ Then

$$\mathrm{E}_{f_X}[g(X)] = \mathrm{E}_{f_X}[(X-\mu)^k] = \int (x-\mu)^k f_X(x)\, dx,$$

and $\mathrm{E}_{f_X}[(X - \mu)^k]$ is the kth **central moment** of the distribution.

(vi) Consider $g(x) = aX + b$. Then

$$\mathrm{E}_{f_X}[aX + b] = a^2_{f_X} \mathrm{E}[X]$$

as

$$\begin{aligned}\mathrm{E}_{f_X}[g(X)] &= \mathrm{E}_{f_X}[(aX + b - \mathrm{E}_{f_X}[aX + b])^2] \\ &= \mathrm{E}_{f_X}[(aX + b - a\mathrm{E}_{f_X}[X] - b)^2] \\ &= \mathrm{E}_{f_X}[(a^2(X - \mathrm{E}_{f_X}[X])^2] \\ &= a^2_{f_X} \mathrm{E}[X]\end{aligned}$$

1.4.2 Approximations to mean and variance

A Taylor series expansion method can be used to obtain approximations to expectations of functions of a random variable. Let X be a continuous random variable, with range $\mathbb{X}$ and pdf f_X. Suppose that the expectation and variance of X with respect to f_X are denoted μ and σ^2 respectively, and let g be a real-valued function whose domain includes $\mathbb{X}$. Then a Taylor approximation of g around μ is given for real-value x by,

$$g(x) \approx g(\mu) + (x - \mu)g'(\mu) + \frac{1}{2}(x - \mu)^2 g''(\mu)$$

where g' and g'' are the first and second derivatives of g respectively. Using the Taylor approximation, and ignoring terms in $(x-\mu)^k$ for $k = 3, 4, ...$, the expectation of $g(X)$ with respect to f_X is given approximately by

$$E_{f_X}[g(X)] \approx g(\mu) + \frac{1}{2}\sigma^2 g''(\mu).$$

Ignoring terms in $(x - \mu)^2$ and higher, the variance of $g(X)$ with respect to f_X is given approximately by

$$\mathrm{Var}_{f_X}[g(X)] \approx \sigma^2 \{g'(\mu)\}^2$$

1.5 Transformations of Random Variables

Consider a discrete/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 real-valued function whose domain includes $\mathbb{X}$, and suppose that

$$\begin{array}{rl} g: & \mathbb{X} \longrightarrow \mathbb{Y} \\ & x \longmapsto y \end{array}$$

Then $Y = g(X)$ is also a random variable as Y is a function from Ω to $\mathbb{R}$. We wish to derive the probability distribution of random variable Y; in order to do this, we first consider the inverse transformation g^{-1} from $\mathbb{Y}$ to $\mathbb{X}$

Consider first the cdf of Y, F_Y, evaluated at a point $y \in \mathbb{R}$. We have

$$F_Y(y) = \mathrm{P}[\ Y \leq y\ ] = \ \mathrm{P}[\ g(X) \leq y\ ] = \begin{cases} \displaystyle\sum_{x \in A_y} f_X(x) & \text{if } X \text{ is DISCRETE} \\ \\ \displaystyle\int_{A_y} f_X(x)\ dx & \text{if } X \text{ is CONTINUOUS} \end{cases}$$

where $A_y = \{\ x \in \mathbb{X}\ :\ g(x) \leq y\}$. Often, the set is A_y is easy to identify for a given y, that is, $F_Y(y) = \mathrm{P}[\ g(X) \leq y\ ] = \mathrm{P}[\ x_1 \leq X \leq\ x_2]$, where x_1 and x_2 depend on y and g or g^{-1}. The main objective is therefore to identify the set A_y.

1.5.1 1-1 Transformations

The mapping $g(X)$ is a function of X from $\mathbb{X}$ which is *1-1 and onto* $\mathbb{Y}$ if,
(i) for each $x \in \mathbb{X}$, there exists one and only one y such that $y = g(x)$, and
(ii) for each $y \in \mathbb{Y}$, there exists an $x \in \mathbb{X}$ such that $g(x) = y$.

The following theorem gives the distribution for random variable $Y = g(X)$ when g is 1-1.

Theorem 1.5.1 *Let X be a random variable with mass/density function f_X and support $\mathbb{X}$. Let g be a 1-1 function from $\mathbb{X}$ onto $\mathbb{Y}$ with inverse g^{-1}. Then $Y = g(X)$ is a random variable with support $\mathbb{Y}$ and*
***Discrete Case** : The mass function of random variable Y is given by*

$$f_Y(y) = f_X(g^{-1}(y)) \qquad y \in \mathbb{Y} = \{\ y \mid f_Y(y) > 0\ \}$$

.where x is the unique solution of $y = g(x)$ (so that $x = g^{-1}(y)$).
***Continuous Case** : The pdf of random variable Y is given by*

$$f_Y(y) = f_X(g^{-1}(y))\ \left|\frac{d}{dt}\left\{g^{-1}(t)\right\}_{t=y}\right| \qquad y \in \mathbb{Y} = \{\ y \mid f_Y(y) > 0\ \}$$

where $y = g(x)$, provided that the derivative

$$\frac{d}{dt}\left\{g^{-1}(t)\right\}$$

is continuous *and* non-zero *on $\mathbb{Y}$.*

Proof. Discrete case: by direct calculation,

$$f_Y(y) = \mathrm{P}[\ Y = y\ ] = \mathrm{P}[\ g(X) = y\ ] = \mathrm{P}[\ X = g^{-1}(y)\ ] = f_X(x)$$

. where $x = g^{-1}(y)$, and hence $f_Y(y) > 0 \Longleftrightarrow f_X(x) > 0$.

Continuous case: function g is either (I) a monotonic increasing, or (II) a monotonic decreasing function.

Case (I): If g is *increasing*, then for $x \in \mathbb{X}$ and $y \in \mathbb{Y}$, we have that $g(x) \leq y \Longleftrightarrow x \leq g^{-1}(y)$. Therefore, for $y \in \mathbb{Y}$,

$$F_Y(y) = \mathrm{P}[\ Y \leq y\ ] = \mathrm{P}[\ g(X) \leq y\ ] = \mathrm{P}[\ X \leq g^{-1}(y)\ ] = F_X(g^{-1}(y))$$

and, by differentiation, because g is monotonic increasing,

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dt}\left\{g^{-1}(t)\right\}_{t=y} = f_X(g^{-1}(y))\left|\frac{d}{dy}\left\{g^{-1}(y)\right\}_{t=y}\right| \qquad \frac{d}{dt}\left\{g^{-1}(t)\right\} > 0.$$

Case (II): If g is *decreasing*, then for $x \in \mathbb{X}$ and $y \in \mathbb{Y}$ we have $g(x) \leq y \Longleftrightarrow x \geq g^{-1}(y)$. Therefore, for $y \in \mathbb{Y}$,

$$F_Y(y) = \mathrm{P}[\ Y \leq y\ ] = \mathrm{P}[\ g(X) \leq y\ ] = \mathrm{P}[\ X \geq g^{-1}(y)\ ] = 1 - F_X(g^{-1}(y))$$

so

$$f_Y(y) = -f_X(g^{-1}(y))\frac{d}{dt}\left\{g^{-1}(y)\right\} = f_X(g^{-1}(y))\left|\frac{d}{dt}\left\{g^{-1}(t)\right\}_{t=y}\right| \qquad \frac{d}{dt}\left\{g^{-1}(t)\right\} < 0.$$

■

Definition 1.5.1 Suppose transformation $g : \mathbb{X} \longrightarrow \mathbb{Y}$ is 1-1, and is defined by $g(x) = y$ for $x \in \mathbb{X}$. Then the **Jacobian** of the transformation, denoted $J(y)$, is given by

$$J(y) = \left|\frac{d}{dt}\left\{g^{-1}(t)\right\}_{t=y}\right|$$

that is, the absolute value of first derivative of g^{-1} evaluated at $y = g(x)$. Note that the inverse transformation $g^{-1} : \mathbb{Y} \longrightarrow \mathbb{X}$ has Jacobian $1/J(x)$.

1.6 Joint Probability Distributions

Consider a collection of random variables $X_1, ..., X_k$, with X_i a function from sample space Ω_i to $\mathbb{X}_i \subseteq \mathbb{R}$. Then the **vector** of random variables $\mathbf{X} = (X_1, ..., X_k)$ is a **vector function** from the Cartesian product of the sample spaces to $\mathbb{X}^{(k)}$

$$\Omega_1 \times ... \times \Omega_k \longrightarrow \mathbb{X}^{(k)} \subseteq \mathbb{X}_1 \times ... \times \mathbb{X}_k \subseteq \mathbb{R}^k$$

The **vector random variable** $\mathbf{X} = (X_1, ..., X_k)$ could correspond to the outcomes of k different experiments carried out once, or a single experiment carried out k times.

Definition 1.6.1 JOINT MASS FUNCTION
The **joint probability mass function** of the k dimensional discrete random variable, that is, $\mathbf{X} = (X_1, ..., X_k)$ is denoted $f_{X_1,...,X_k}$ and is defined by

$$f_{X_1,...,X_k}(x_1, ..., x_k) = \mathrm{P}[\ (X_1 = x_1) \cap ... \cap (X_k = x_k)\ ] \ = \ \mathrm{P}[\ X_1 = x_1, ..., X_k = x_k\ ]$$

for all possible values of the random vector, $\mathbf{x} = (x_1, ..., x_k) \in \mathbb{X}^{(k)}$. As for the single variable case, the joint mass function must be **non-negative** for all $\mathbf{x} \in \mathbb{X}^{(k)}$, and the sum of the joint mass function evaluated for all $\mathbf{x} \in \mathbb{X}^{(k)}$ must be 1.

Definition 1.6.2 JOINT DISCRETE cdf
The **joint cumulative distribution function** of k dimensional discrete random variable $\mathbf{X} = (X_1, ..., X_k)$ is denoted $F_{X_1,...,X_k}$ and is defined by for **any** real vector $(x_1, ..., x_k)$ by

$$F_{X_1,...,X_k}(x_1, ..., x_k) = \mathrm{P}[\ X_1 \leq x_1, ..., X_k \leq x_k\ ]$$

Note: The joint cdf must satisfy certain properties (i.e. behaviour for limiting values of $x_1, ..., x_k$ individually and jointly) analogous to the univariate case.

Definition 1.6.3 MARGINAL MASS FUNCTION
The **marginal mass function** of random variable X_i, denoted f_{X_i}, is defined in terms of the joint mass function $f_{X_1,...,X_k}$ for $x_i \in \mathbb{X}_i$ by

$$f_{X_i}(x_i) = \sum_{\mathbb{X}_1} ... \sum_{\mathbb{X}_{i-1}} \sum_{\mathbb{X}_{i+1}} ... \sum_{\mathbb{X}_k} f_{X_1,...,X_k}(x_1, ..., x_k)$$

that is, the summation of joint mass function evaluated at $(x_1, ..., x_k)$ for **all** values of $x_j \in \mathbb{X}_j$ for $j \neq i$.

Definition 1.6.4 JOINT CONTINUOUS cdf
The **joint cumulative distribution function** of k dimensional continuous random variable, that is, $\mathbf{X} = (X_1, ..., X_k)$ is denoted $F_{X_1,...,X_k}$ and is defined for any real vector $(x_1, ..., x_k)$ by

$$F_{X_1,...,X_k}(x_1, ..., x_k) = \mathrm{P}[\ X_1 \leq x_1, ..., X_k \leq x_k\ ]$$

Note: The joint cdf must satisfy certain properties (i.e. behaviour for limiting values of $x_1, ..., x_k$ individually and jointly) analogous to the univariate case.

Definition 1.6.5 JOINT pdf
The **joint probability density function** of k dimensional continuous random variable $(X_1, ..., X_k)$ is denoted $f_{X_1,...,X_k}$, and is defined in terms of the joint cdf $F_{X_1,...,X_k}$ for vector $(x_1, ..., x_k) \in \mathbb{X}^{(k)}$ by

$$F_{X_1,...,X_k}(x_1, ..., x_k) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_k} f_{X_1,...,X_k}(t_1, ..., t_k)\ dt_1 ... dt_k$$

so that

$$f_{X_1,...,X_k}(x_1, ..., x_k) \ = \ \frac{\partial^k}{\partial t_1 ... \partial t_k} \{\{F_{X_1,...,X_k}(t_1, ..., t_k)\}\}_{t_1=x_1,...,t_k=x_k}$$

Note : As for the single variable case, the joint pdf need not exist, but if it does exist, the joint pdf must be take non-negative values for all $(x_1, ..., x_k) \in \mathbb{X}^{(k)}$, and

$$\int_{\mathbb{X}_1} ... \int_{\mathbb{X}_k} f_{X_1,...,X_k}(x_1, ..., x_k)\ dx_1 ... dx_k = 1$$

Definition 1.6.6 MARGINAL pdf
The **marginal probability density function** of random variable X_i, denoted f_{X_i}, is defined in terms of the joint pdf $f_{X_1,...,X_k}$ for $x_i \in \mathbb{X}_i$ by

$$f_{X_i}(x_i) = \int_{\mathbb{X}_1} ... \int_{\mathbb{X}_{i-1}} \int_{\mathbb{X}_{i+1}} ... \int_{\mathbb{X}_k} f_{X_1,...,X_k}(x_1, ..., x_k)\ dx_1 ... dx_{i-1} dx_{i+1} ... dx_k$$

that is, the joint pdf integrated out over the ranges of the remaining $k-1$ variables $X_j, j \neq i$.

Note In both discrete and continuous cases, the concept of **marginalization** can be extended from consideration of the marginal probability distribution of a single variable. For example, consider the pair of variables (X_i, X_j) for $i \neq j$; the **joint marginal mass function/pdf** can be obtained by summation/integration over the remaining $k-2$ variables. This can be further extended to consideration of more than two variables.

1.6.1 Conditional Probability Distributions

Definition 1.6.7 The **conditional probability mass/density function**of random variable X_i, given that $X_1 = x_1, , ..., X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, ..., X_k = x_k$ denoted

$$f_{X_i|X_1,...,X_{i-1},X_{i+1},...,X_k},$$

is defined by

$$f_{X_i|X_1,...,X_{i-1},X_{i+1},...,X_k}(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_k) = \frac{f_{X_1,...,X_k}(x_1, ..., x_k)}{f_{X_1,...,X_{i-1},X_{i+1},...,X_k}(x_1, ..., x_{i-1}, x_{i+1}, ..., x_k)}$$

if the denominator is strictly positive, that is, the ratio of the (full) joint mass function/pdf for the k variables $X_1, ..., X_k$ to the (marginal) joint mass function/pdf for the $k-1$ variables $X_1, ..., X_{i-1}, X_{i+1}, ..., X_k$. The form of the denominator is obtained by summing/integrating the (full) joint mass function/pdf out over the variable X_i.

Note : This definition is directly related to the definition of **conditional probability** given earlier, and can be extended to the definition of conditional distribution of two variables, given the values of the remaining $k-2$, and so on.

Example 1.6.1 Consider the particular case $k = 2$, $\mathbf{X} = (X_1, X_2)$. Then the following functions are used to specify the probability distribution of $\mathbf{X}$; The joint mass/density function is

$$f_{X_1,X_2}(x_1, x_2)$$

that is, a function of **two** variables, which in the discrete case reduces to

$$f_{X_1,X_2}(x_1, x_2) = P[(X_1 = x_1) \cap (X_2 = x_2)] = P[X_1 = x_1, X_2 = x_2].$$

It follows that $f_{X_1,X_2}(x_1, x_2) \geq 0$ for all possible outcomes x_1, x_2, and

$$\sum_{x_1} \sum_{x_2} f_{X_1,X_2}(x_1, x_2) = 1$$

where the double summation/integration is over all possible values of (x_1, x_2).
Typically, such a specification is represented by a **probability table**; for example for **discrete** random variables X_1 and X_2, we could have

		X_1			
		1	2	3	4
X_2	1	0.100	0.200	0.000	0.000
	2	0.200	0.250	0.050	0.000
	3	0.000	0.050	0.050	0.025
	4	0.000	0.000	0.025	0.050

where the entry in column i, row j is

$$f_{X_1,X_2}(i, j) = P[X_1 = i) \cap (X_2 = j)] = P[X_1 = i, X_2 = j],$$

The joint mass function automatically defines the probability distribution of the individual random variables. For example, when $k = 2$, then there are two marginal mass/density functions are $f_{X_1}(x_1)$ and

$f_{X_2}(x_2)$
in the discrete case,

$$\begin{aligned} f_{X_1}(x_1) &= \sum_{x_2} f_{X_1,X_2}(x_1,x_2) \\ f_{X_2}(x_2) &= \sum_{x_1} f_{X_1,X_2}(x_1,x_2) \end{aligned}$$

in the continuous case

$$\begin{aligned} f_{X_1}(x_1) &= \int f_{X_1,X_2}(x_1,x_2)dx_2 \\ f_{X_2}(x_2) &= \int f_{X_1,X_2}(x_1,x_2)dx_1 \end{aligned}$$

so the marginal mass/density function for random variable X_1 is obtained by summing/integrating out the joint mass/density function for X_1 and X_2 over all possible values of random variable X_2. In the discrete case

$$\mathrm{P}[X_1 = x_1] = \sum_{x_2} \mathrm{P}[(X_1 = x_1) \cap (X_2 = x_2)]$$

which is a result that is justified by the Theorem of Total Probability.
In the table above, the marginal mass functions can be computed easily

		X_1				$\downarrow$
		1	2	3	4	$f_{X_2}(x_2)$
X_2	1	0.100	0.200	0.000	0.000	**0.300**
	2	0.200	0.250	0.050	0.000	**0.500**
	3	0.000	0.050	0.050	0.025	**0.125**
	4	0.000	0.000	0.025	0.050	**0.075**
$\rightarrow$	$f_{X_1}(x_1)$	**0.300**	**0.500**	**0.125**	**0.075**	

so that the marginal mass functions are formed by the column and row sums respectively. In this case, it turns out that

$$f_{X_1}(x) = f_{X_2}(x) \qquad x = 1,2,3,4$$

but this will not always be the case.

For the conditional mass/density functions, in the discrete two variable case, consider the probability

$$\mathrm{P}[X_1 = x_1 | X_2 = x_2]$$

that is, the conditional probability distribution of X_1, given that $X_2 = x_2$. This conditional distribution is easily computed from the conditional probability definition, that is

$$\mathrm{P}[X_1 = x_1 | X_2 = x_2] = \frac{\mathrm{P}[\ X_1 = x_1, X_2 = x_2\ ]}{\mathrm{P}[\ X_2 = x_2\ ]} = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}$$

that is, proportional to the x_2 row of the table. By extending these concepts, we may define the conditional probability distributions for both variables in the discrete and continuous cases; The two

conditional mass/density functions are $f_{X_1|X_2}(x_1|x_2)$ and $f_{X_2|X_1}(x_2|x_1)$

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} \qquad f_{X_2}(x_2) > 0$$

$$f_{X_2|X_1}(x_2|x_1) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_1}(x_1)} \qquad f_{X_1}(x_1) > 0$$

In the discrete case, this result becomes

$$f_{X_1|X_2}(x_1|x_2) = \mathrm{P}[X_1 = x_1|X_2 = x_2] = \frac{\mathrm{P}[(X_1 = x_1) \cap (X_2 = x_2)]}{\mathrm{P}[X_2 = x_2]}$$

if $\mathrm{P}[X_2 = x_2] > 0$, which is justified by the definition of conditional probability. For example, consider the table

		X_1				
		1	2	3	4	$f_{X_2}(x_2)$
X_2	1	0.100	**0.200**	0.000	0.000	0.300
	2	0.200	**0.250**	0.050	0.000	0.500
	3	0.000	**0.050**	0.050	0.025	0.125
	4	0.000	**0.000**	0.025	0.050	0.075
	$f_{X_1}(x_1)$	0.300	**0.500**	0.125	0.075	

The highlighted column gives the conditional mass function for X_2 given that $X_1 = 2$; from the definition,

$$f_{X_2|X_1}(1|2) = \frac{0.200}{0.500} = 0.400 \qquad f_{X_2|X_1}(1|2) = \frac{0.250}{0.500} = 0.500$$

$$f_{X_2|X_1}(3|2) = \frac{0.050}{0.500} = 0.100 \qquad f_{X_2|X_1}(4|2) = \frac{0.000}{0.500} = 0.000$$

Note that

$$\sum_{x=1}^{4} f_{X_2|X_1}(x|2) = 0.400 + 0.500 + 0.100 + 0.000 = 1$$

which we must have for a conditional mass function. Note that, in general, the conditional mass functions will be **different** for different values of the conditioning variable.

In summary, suppose that X_1 and X_2 are discrete random variables that take values $\{1, 2, ..., n\}$ and $\{1, 2, ..., m\}$ respectively. Then the joint mass function can be displayed as a table with n columns and m rows, where
- the (i, j)th cell contains $\mathrm{P}[(X_1 = i) \cap (X_2 = j)]$
- the marginal mass function for X_1 is given by the **column** totals
- the marginal mass function for X_2 is given by the **row** totals
- the conditional mass function for X_1 given $X_2 = j$ is given by the jth row divided by the sum of the jth row
- the conditional mass function for X_2 given $X_1 = i$ is given by the ith column divided by the sum of the ith column

for $i = 1, ..., n$ and $j = 1, ..., m$.

Example 1.6.2 If the joint density of continuous variables X_1 and X_2 is given by

$$f_{X_1,X_2}(x_1,x_2) = x_2^2 e^{-x_2(1+x_1)}$$

for $x_1, x_2 \geq 0$ and zero otherwise. It can be shown that

$$\int_{-\infty}^{infty} \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\ dx_1 dx_2 = 1$$

and that the marginal pdf for X_1 is given by

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)\ dx_2 = \int_0^{\infty} x_2^2 e^{-x_2(1+x_1)}\ dx_2 = \frac{2}{(1+x_1)^3}$$

for $x_1 \geq 0$, and zero otherwise.

1.6.2 Independence for random variables

Definition 1.6.8 Random variables $X_1, ...X_k$ are **independent** if, for $a_i < b_i$, $i = 1, ...k$,

$$\mathrm{P}[\ a_1 < X_1 \leq b_1, ..., a_k < X_k \leq b_k\ ] = \prod_{i=1}^{k} \mathrm{P}[\ a_i < X_i \leq b_i\ ]$$

In fact, $X_1, ...X_k$ are independent **if and only if**

$$F_{X_1,...,X_k}(x_1,...,x_k) = \prod_{i=1}^{k} F_{X_i}(x_i) \qquad \text{or} \qquad f_{X_1,...,X_k}(x_1,...,x_k) = \prod_{i=1}^{k} f_{X_i}(x_i)$$

that is, if and only if the joint cdf and the joint mass function/pdf factorizes into the k marginal cdfs and mass function/pdfs. Sufficient conditions for variables $X_1, ..., X_k$ to be independent are

(i) the **support** of the joint mass function/pdf of $\mathbf{X}^{(k)}$ is a Cartesian product, that is,

$$\mathbf{X}^{(k)} = \mathbb{X}_1 \times ... \times \mathbb{X}_k = \{(x_1,...,x_k)\ :\ x_1 \in \mathbb{X}_1, ..., x_k \in \mathbb{X}_k\}\,.$$

(ii) the joint mass function/pdf factorizes into a product of marginal functions.

The chain rule for random variables

There is an explicit relationship between joint, marginal, and conditional mass/density functions. For example, consider three continuous random variables X_1, X_2, X_3, with joint pdf f_{X_1,X_2,X_3}. Then,

$$f_{X_1,X_2,X_3}(x_1,x_2,x_3) = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_1,X_2}(x_3|x_1,x_2) B$$

so that, for example,

$$f_{X_1}(x_1) = \int_{\mathbb{X}_2} \int_{\mathbb{X}_3} f_{X_1,X_2,X_3}(x_1,x_2,x_3)\ dx_2 dx_3 = \int_{\mathbb{X}_2} \int_{\mathbb{X}_3} f_{X_1|X_2,X_3}(x_1|x_2,x_3) f_{X_2,X_3}(x_2,x_3) dx_2 dx_3$$

Equivalent relationships hold in the discrete case, can be extended to determine the explicit relationship between joint, marginal, and conditional mass/density functions for any number of random variables.

1.6.3 Conditional Expectation

Definition 1.6.9 Consider two discrete/continuous random variables X_1 and X_2 with joint mass function/pdf f_{X_1,X_2}, and the conditional mass function/pdf of X_1 given $X_2 = x_2$, defined in the usual way by

$$f_{X_1|X_2}(x_1|x_2) \;=\; \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}$$

Then the **conditional expectation** of X_1 given $X_2 = x_2$ is defined by

$$\mathrm{E}_{f_{X_1|X_2}}[X_1|X_2 = x_2] \;=\; \begin{cases} \displaystyle\sum_{x_1 \in \mathbb{X}_1} x_1 f_{X_1|X_2}(x_1|x_2) & X_1 \text{ DISCRETE} \\ \\ \displaystyle\int_{\mathbb{X}_1} x_1 f_{X_1|X_2}(x_1|x_2)\, dx_1 & X_1 \text{ CONTINUOUS} \end{cases}$$

i.e. the expectation of X_1 *with respect to* the conditional density of X_1 given $X_2 = x_2$, (possibly a function of x_2).

Result.The Law of Iterated Expectation

For two continuous random variables X_1 and X_2 with joint pdf f_{X_1,X_2},

$$\begin{aligned} \mathrm{E}_{f_{X_1}}[X_1] &= \int_{\mathbb{X}_1} x_1 f_{X_1}(x_1)\, dx_1 = \int_{\mathbb{X}_1} x_1 \left\{ \int_{\mathbb{X}_2} f_{X_1,X_2}(x_1,x_2)\, dx_2 \right\} dx_1 \\ &= \int_{\mathbb{X}_1} x_1 \left\{ \int_{\mathbb{X}_2} f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2)\, dx_2 \right\} dx_1 \\ &= \int_{\mathbb{X}_1} \int_{\mathbb{X}_2} x_1 f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2)\, dx_2 \\ &= \int_{\mathbb{X}_2} \left\{ \int_{\mathbb{X}_1} x_1 f_{X_1|X_2}(x_1|x_2)\, dx_1 \right\} f_{X_2}(x_2)\, dx_2 \\ &= \int_{\mathbb{X}_2} \left\{ \mathrm{E}_{f_{X_1|X_2}}[X_1|X_2 = x_2] \right\} f_{X_2}(x_2)\, dx_2 = \mathrm{E}_{f_{X_2}} \left[\mathrm{E}_{f_{X_1|X_2}}[X_1|X_2 = x_2] \right] \end{aligned}$$

so the expectation of X_1 can be calculated by finding the conditional expectation of X_1 given $X_2 = x_2$, giving a function of x_2, and then taking the expectation of this function with respect to the marginal density for X_2. Note that this proof only works if the conditional expectation and the marginal expectation are finite. This results extends naturally to k variables.

1.7 Multivariate Transformations

Theorem 1.7.1 *Let $\mathbf{X} = (X_1, ..., X_k)$ be a vector of random variables, with joint mass/density function $f_{X_1,...,X_k}$. Let $\mathbf{Y} = (Y_1, ..., Y_k)$ be a vector of random variables defined by $Y_i = g_i(X_1, ..., X_k)$ for some functions $g_i, i = 1, ..., k$, where the vector function $\mathbf{g}$ mapping $(X_1, ..., X_k)$ to $(Y_1, ..., Y_k)$ is a 1-1 transformation. Then the joint mass/density function of $(Y_1, ..., Y_k)$ is given by*

$$f_{Y_1,...,Y_k}(y_1, ..., y_k) = f_{X_1,...,X_k}(x_1, ..., x_k)$$

in the discrete case, and

$$f_{Y_1,...,Y_k}(y_1, ..., y_k) = f_{X_1,...,X_k}(x_1, ..., x_k) J(y_1, ..., y_k)$$

where $\mathbf{x} = (x_1, ..., x_k)$ *is the unique solution of* $\mathbf{y} = g(\mathbf{x})$*, so that* $\mathbf{x} = \mathbf{g}^{-1}(\mathbf{y})$*. and where* $J(y_1, ..., y_k)$ *is the* ***Jacobian****, of the transformation, that is, the modulus of the determinant of the* $k \times k$ *matrix whose* (i, j)*th element is*

$$\frac{\partial}{\partial t_j} \left\{ {g_i}^{-1}(\mathbf{t}) \right\}_{t_1 = y_1, ..., t_k = y_k}$$

where g_i^{-1} *is the inverse function uniquely defined by* $X_i = g_i^{-1}(Y_1, ..., Y_k)$*.*

Example 1.7.1 Consider the case $k = 2$, and suppose that X_1 and X_2 are *independent* continuous random variables with ranges $\mathbb{X}_1 = \mathbb{X}_2 = [0, 1]$ and pdfs given respectively by

$$\begin{array}{ll} f_{X_1}(x_1) & = 6x_1(1 - x_1) \quad 0 \leq x_1 \leq 1 \\ f_{X_2}(x_2) & = 3x_2^2 \qquad\quad 0 \leq x_2 \leq 1 \end{array}$$

and zero elsewhere. In order to calculate the pdf of random variable Y_1 defined

$$Y_1 = X_1 X_2$$

using the transformation result, consider the additional random variable Y_2, where $Y_2 = X_1$ (note, as X_1 and X_2 take values on $[0, 1]$, $X_1 \geq X_1 X_2$ so $Y_1 \leq Y_2$).
The transformation $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ is then specified by the two functions

$$g_1(t_1, t_2) = t_1 t_2 \qquad g_2(t_1, t_2) = t_1$$

and the inverse transformation $\mathbf{X} = \mathbf{g}^{-1}(\mathbf{Y})$ (i.e. $\mathbf{X}$ in terms of $\mathbf{Y}$) is

$$X_1 = Y_2 \qquad X_2 = Y_1 / Y_2$$

giving

$$g_1^{-1}(t_1, t_2) = t_2 \qquad g_2^{-1}(t_1, t_2) = t_1 / t_2.$$

Hence

$$\frac{\partial}{\partial t_1} \left\{ {g_1}^{-1}(t_1, t_2) \right\} = 0 \qquad \frac{\partial}{\partial t_2} \left\{ {g_1}^{-1}(t_1, t_2) \right\} = 1$$

$$\frac{\partial}{\partial t_1} \left\{ {g_2}^{-1}(t_1, t_2) \right\} = 1/t_2 \quad \frac{\partial}{\partial t_2} \left\{ {g_2}^{-1}(t_1, t_2) \right\} = -t_1/t_2^2$$

and so the Jacobian $J(y_1, y_2)$ of the transformation is given by the modulus of

$$\begin{vmatrix} 0 & 1 \\ 1/y_2 & -y_1/y_2^2 \end{vmatrix}$$

so that $J(y_1, y_2) = 1/y_2$. Hence, using the theorem

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= f_{X_1, X_2}(y_2, y_1/y_2) J(y_1, y_2) \\ &= 6y_2(1 - y_2) \,.\, 3(y_1/y_2)^2 \,.\, 1/y_2 \\ &= 18y_1^2(1 - y_2)/y_2^2 \end{aligned}$$

on the set $\mathbb{Y}^{(2)} = \{(y_1, y_2) \,:\, 0 \leq y_1 \leq y_2 \leq 1\}$, and zero otherwise. Hence

$$\begin{aligned} f_{Y_1}(y_1) &= \int_{y_1}^{1} 18y_1^2(1 - y_2)/y_2^2 \, dy_2 \\ &= 18y_1^2 \left[-1/y_2 - \log y_2 \right]_{y_1}^{1} \\ &= 18y_1^2(-1 + 1/y_1 + \log y_1) \\ &= 18y_1(1 - y_1 + y_1 \log y_1) \end{aligned}$$

for $0 \leq y_1 \leq 1$, and zero otherwise.

1.8 Multivariate Expectations and Covariance

Consider vector random variable $\mathbf{X} = (X_1, ..., X_k)$ with range $\mathbb{X}^{(k)}$ and joint mass or density function $f_{X_1,...,X_k}$, and a real-valued scalar function g whose domain includes $\mathbb{X}^{(k)}$.

Definition 1.8.1 For random variables $X_1, ..., X_k$ with range $\mathbb{X}^{(k)}$ with mass/density function $f_{X_1,...,X_k}$, the **expectation** of $g(X_1, ..., X_k)$ is defined in the discrete and continuous cases by

$$\mathrm{E}_{f_{X_1,...,X_k}}[\ g(X_1, ..., X_k)\ ] = \begin{cases} \displaystyle\sum_{x_1} ... \sum_{x_k} g(x_1, ..., x_k) f_{X_1,...,X_k}(x_1, ..., x_k) \\ \\ \displaystyle\int_{\mathbb{X}_1} ... \int_{\mathbb{X}_k} g(x_1, ..., x_k) f_{X_1,...,X_k}(x_1, ..., x_k)\ dx_1 ... dx_k \end{cases}$$

Properties of Multivariate Expectations :

(i) Let $X_1, ...X_k$ be random variables with joint mass function/pdf $f_{X_1,...,X_k}$. Let g and h be real-valued functions whose domains include $\mathbb{X}^{(k)}$, and let a and b be constants. Then, if $f_{\mathbf{X}} \equiv f_{X_1,...,X_k}$,

$$\mathrm{E}_{f_{\mathbf{X}}}[\ ag(X_1, ..., X_k) + bh(X_1, ..., X_k)\ ] = a\mathrm{E}_{f_{\mathbf{X}}}[\ g(X_1, ..., X_k)\ ] + b\mathrm{E}_{f_{\mathbf{X}}}[\ h(X_1, ..., X_k)\ ].$$

(ii) Let $X_1, ...X_k$ be independent random variables with mass functions/pdfs $f_{X_1}, ..., f_{X_k}$ respectively. Let $g_1, ..., g_k$ be scalar functions of $X_1, ..., X_k$ respectively (that is, g_i is a function of X_i *only* for $i = 1, ..., k$). If $g(X_1, ..., X_k) = g_1(X_1)...g_k(X_k)$, then

$$\mathrm{E}_{f_{\mathbf{X}}}[\ g(X_1, ..., X_k)\ ] = \prod_{i=1}^{k} \mathrm{E}_{f_{X_i}}[\ g_i(X_i)\ ]$$

where $\mathrm{E}_{f_{X_i}}[\ g_i(X_i)\ ]$ is the marginal expectation of $g_i(X_i)$ with respect to f_{X_i}.

Covariance and Correlation

Definition 1.8.2 If X_1 and X_2 are random variables with joint mass function/pdf f_{X_1,X_2}, then the **covariance** of X_1 and X_2 is denoted $\mathrm{Cov}_{f_{X_1,X_2}}[X_1, X_2]$, and is defined by

$$\mathrm{Cov}_{f_{X_1,X_2}}[X_1, X_2] = \mathrm{E}_{f_{X_1,X_2}}[(X_1 - \mu_1)(X_2 - \mu_2)] = \mathrm{E}_{f_{X_1,X_2}}[X_1 X_2] - \mu_1\mu_2$$

where

$$\mu_i = \mathrm{E}_{f_{X_i}}[X_i]$$

is the marginal expectation of X_i, for $i = 1, 2$, and where $\mathrm{E}_{f_{X_1,X_2}}[X_1 X_2]$ is the expectation of function $g(x_1, x_2) = x_1 x_2$ with respect to the joint distribution f_{X_1,X_2}.

Definition 1.8.3 If X_1 and X_2 are random variables with joint mass function/pdf f_{X_1,X_2}, then the **correlation** of X_1 and X_2 is denoted $\mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2]$, and is defined by

$$\mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2] = \frac{\mathrm{Cov}_{f_{X_1,X_2}}[X_1, X_2]}{\sqrt{\mathrm{Var}_{f_{X_1}}[X_1]\ \mathrm{Var}_{f_{X_2}}[X_2]}}$$

If $\mathrm{Cov}_{f_{X_1,X_2}}[X_1, X_2] = \mathrm{Corr}_{f_{X_1,X_2}}[X_1, X_2] = 0.$ then variables X_1 and X_2 are **uncorrelated**.

Result. If random variables X_1 and X_2 are *independent* then

$$\text{Corr}_{f_{X_1,X_2}}[X_1, X_2] = 0$$

and X_1 and X_2 are uncorrelated,.as if X_1 and X_2 are independent, then

$$\text{E}_{f_{X_1,X_2}}[X_1X_2] = \text{E}_{f_{X_1}}[X_1]\text{E}_{f_{X_2}}[X_2]$$

and hence the result follows.

NOTES:

(i) For random variables X_1 and X_2, with (marginal) expectations μ_1 and μ_2 respectively, and (marginal) variances σ_1^2 and σ_2^2 respectively, if random variables Z_1 and Z_2 are defined

$$Z_1 = (X_1 - \mu_1)/\sigma_1 \qquad Z_2 = (X_2 - \mu_2)/\sigma_2$$

that is, Z_1 and Z_2 are **standardized** variables. Then $\text{E}_{f_{X_1,X_2}}[X_1, X_2]$ =$\text{E}_{f_{Z_1,Z_2}}[Z_1, Z_2]$.

(ii) Extension to k variables: covariances can only be calculated for **pairs** of random variables, but if k variables have a joint probability structure it is possible to construct a $k \times k$ **matrix**, $\mathbf{C}$ say, of covariance values, whose (i, j)th element is

$$\text{E}_{f_{X_i,X_j}}[X_i, X_j]$$

for $i, j = 1, .., k$, that captures the complete covariance structure in the joint distribution. If $i \neq j$, then

$$\text{E}_{f_{X_j,X_i}}[X_j, X_i] = \text{E}_{f_{X_i,X_j}}[X_i, X_j]$$

so $\mathbf{C}$ is *symmetric*, and if $i = j$,

$$\text{E}_{f_{X_i,X_i}}[X_i, X_i] \equiv \text{E}_{f_{X_i}}[X_i]$$

The matrix $\mathbf{C}$ is referred to as the **variance-covariance** matrix.

(iii) If random variable X is defined by

$$X = \sum_{i=1}^{k} a_i X_i$$

for random variables $X_1, ..., X_k$ and constants $a_1, ..., a_k$, then

$$\begin{aligned} \text{Var}_{f_X}[X] &= \sum_{i=1}^{k} a_i \text{Var}_{f_{X_i}}[X_i] \\ &= \sum_{i=1}^{k} a_i^2 \text{Var}_{f_{X_i}}[X_i] + 2\sum_{i=1}^{k}\sum_{j=1}^{i-1} a_i a_j \text{Cov}_{f_{X_i,X_j}}[X_i, X_j] \end{aligned}$$

1. (iv) Combining (i) and (iii) when $k = 2$, and defining standardized variables Z_1 and Z_2,

$$\begin{aligned} 0 \leq \text{Var}_{f_{Z_1,Z_2}}[Z_1 \pm Z_2] &= \text{Var}_{f_{Z_1}}[Z_1] + \text{Var}_{f_{Z_2}}[Z_2] \pm 2\ \text{Cov}_{f_{Z_1,Z_2}}[Z_1, Z_2] \\ &= 1 + 1 \pm 2\ \text{Cov}_{f_{X_1,X_2}}[X_1, X_2] \\ &= 2(1 \pm \text{Cov}_{f_{X_1,X_2}}[X_1, X_2]) \end{aligned}$$

and hence $-1 \leq$Cov$_{f_{X_1,X_2}}[X_1, X_2] \leq 1$.

1.8.1 Sums of Random Variables

Theorem 1.8.1 ***The Convolution Theorem***
Let X_1 and X_2 be continuous random variables taking values on $\mathbb{X}_1$ and $\mathbb{X}_2$ respectively with joint density function f_{X_1,X_2}. Then, if random variable Y is defined by $Y = X_1 + X_2$, the marginal density of Y is given by

$$f_Y(y) = \int f_{X_1,X_2}(x_1, y - x_1)dx_1$$

*If X_1 and X_2 are **independent**, then*

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

so if $Y = X_1 + X_2$, then Y has pdf given by

$$f_Y(y) = \int_{\mathbb{X}_1} f_{X_1}(x_1)f_{X_2}(y - x_1)\ dx_1$$

1.9 Generating Functions

Definition 1.9.1 For random variable X with mass/density function f_X, the **moment generating function**, or **mgf**, of X, M_X, is defined by

$$M_X(t) = \mathrm{E}_{f_X}\left[e^{tX}\right]$$

if this expectation exists for all values of $t \in (-h, h)$ for some $h > 0$, that is,

$$\mathrm{E}_{f_X}[g(X)] = \begin{cases} \sum_{x=-\infty}^{\infty} e^{tx} f_X(x) & \text{if } X \text{ is DISCRETE} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\ dx & \text{if } X \text{ is CONTINUOUS} \end{cases}$$

It can be shown that if X_1 and X_2 are random variables taking values on $\mathbb{X}$ with mass/density functions f_{X_1} and f_{X_2}, and mgfs M_{X_1} and M_{X_2} respectively, then

$$f_{X_1}(x) \equiv f_{X_2}(x),\ x \in \mathbb{X} \quad \Longleftrightarrow \quad M_{X_1}(t) \equiv M_{X_2}(t),\ t \in (-h, h)$$

1.9.1 Properties of MGFs

(i) If X is a **discrete** random variable, the rth derivative of M_X evaluated at t, $M_X^{(r)}(t)$, is given by

$$\begin{aligned} M_X^{(r)}(t) &= \frac{d^r}{ds^r}\left\{M_X(s)\right\}_{s=t} = \frac{d^r}{ds^r}\left\{\sum_{x\in\mathbb{X}} e^{sx} f_X(x)\right\}_{s=t} \\ &= \sum_{x\in\mathbb{X}} x^r e^{tx} f_X(x) \end{aligned}$$

and hence

$$M_X^{(r)}(0) = \sum_{x\in\mathbb{X}} x^r f_X(x) = \mathrm{E}_{f_X}[X^r]$$

Similarly, if X is a **continuous** random variable, the rth derivative of M_X is given by

$$\begin{aligned} M_X^{(r)}(t) &= \frac{d^r}{ds^r}\left\{\int_{x\in\mathbb{X}} e^{sx} f_X(x)\ d_x\right\}_{s=t} \\ &= \int_{x\in\mathbb{X}} x^r e^{tx} f_X(x)\ dx \end{aligned}$$

and hence

$$M_X^{(r)}(0) = \int_{x\in\mathbb{X}} x^r f_X(x)\ dx = \mathrm{E}_{f_X}[X^r]$$

(ii) If X is a **discrete** random variable, then

$$\begin{aligned} M_X(t) &= \sum_{x\in\mathbb{X}} e^{tx} f_X(x) \\ &= \sum_{x\in\mathbb{X}} \left\{\sum_{r=0}^{\infty} \frac{(tx)^r}{r!}\right\} f_X(x) \\ &= 1 + \sum_{r=1}^{\infty} \frac{t^r}{r!} \left\{\sum_{x\in\mathbb{X}} x^r f_X(x)\right\} \\ &= 1 + \sum_{r=1}^{\infty} \frac{t^r}{r!} \mathrm{E}_{f_X}[X^r] \end{aligned}$$

with a similar result for continuous variables.

(iii) From the general result for expectations of functions of random variables

$$\mathrm{E}_{f_Y}[e^{tY}] \equiv \mathrm{E}_{f_X}[e^{t(aX+b)}] \Longrightarrow M_Y(t) = \mathrm{E}_{f_X}[e^{t(aX+b)}] = e^{bt}_{f_X}\mathrm{E}[e^{atX}] = e^{bt} M_X(at).$$

Therefore, if $Y = aX + b$,

$$M_Y(t) = e^{bt} M_X(at)$$

Theorem 1.9.1 *Let $X_1, ..., X_k$ be independent random variables with mgfs $M_{X_1}, ..., M_{X_k}$ respectively. Then if random variable Y is defined by $Y = X_1 + ... + X_k$,*

$$M_Y(t) = \prod_{i=1}^{k} M_{X_i}(t)$$

Proof. Using the general result for expectations of functions of independent random variables,

$$M_Y(t) = \mathrm{E}_{f_Y}[e^{tY}] = \mathrm{E}_{f_{X_1,...,X_k}}[e^{t(X_1+...+X_k)}] = \prod_{i=1}^{k} \mathrm{E}_{f_{X_i}}[e^{tX_i}] = \prod_{i=1}^{k} M_{X_i}(t).$$

Special Case If $X_1, ..., X_k$ are **identically distributed**, then $M_{X_i}(t) \equiv M_X(t)$, say, for all i, so

$$M_Y(t) = \prod_{i=1}^{k} M_X(t) = \{M_X(t)\}^k$$

■

Definition 1.9.2 For random variable X, with mass/density function f_X, the **factorial moment** or **probability generating function**, **fmgf** or **pgf** , of X, G_X, is defined by

$$G_X(t) = \mathrm{E}_{f_X}[t^X] = \mathrm{E}_{f_X}[e^{X \log t}] = M_X(\log t)$$

if this expectation exists for all values of $t \in (1-h, 1+h)$ for some $h > 0$.
Properties :
(i) Using similar techniques to those used for the mgf, it can be shown that

$$G_X^{(r)}(t) = \frac{d^r}{ds^r}\{G_X(s)\}_{s=t} = \mathrm{E}_{f_X}[X(X-1)...(X-r+1)t^{X-r}] \Longrightarrow G_X^{(r)}(1) = \mathrm{E}_{f_X}[X(X-1)...(X-r+1)]$$

where $\mathrm{E}_{f_X}[X(X-1)...(X-r+1)]$ is the rth **factorial moment**.

(ii) For **discrete** random variables, it can be shown by using a Taylor series expansion of G_X that, for $r = 1, 2, ...,$

$$\frac{G_X^{(r)}(0)}{r!} = \mathrm{P}[X = r]$$

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

1.10.1 THE DISCRETE UNIFORM DISTRIBUTION $X \sim Uniform(N)$

Range : $\mathbb{X} = \{1, 2, ..., N\}$
Parameter : $N \in \mathbf{Z}^+$
Mass function :

$$f_X(x) = \frac{1}{N} \qquad x \in \{1, 2, ..., N\}$$

Interpretation : This model gives **equal** probability to each of the N possible values of the random variable

1.10.2 THE BERNOULLI DISTRIBUTION $X \sim Bernoulli(\theta)$

Range : $\mathbb{X} = \{0, 1\}$
Parameter : $\theta \in [0, 1]$
Mass function :

$$f_X(x) = \theta^x (1-\theta)^{1-x} \qquad x \in \{0, 1\}$$

Interpretation : A model for "the number of successes in a single 0-1 trial" where the probability of success is θ.

1.10.3 THE BINOMIAL DISTRIBUTION $X \sim Binomial(n, \theta)$

Range : $\mathbb{X} = \{0, 1, ..., n\}$
Parameters : $n \in \mathbf{Z}^+$, $\theta \in [0, 1]$
Mass function :

$$f_X(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x} = \frac{n!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} \qquad x \in \{0, 1, ..., n\}$$

Recall that $f_X(x) = \mathrm{P}[\ X = x\ ]$. Now if $X = x$, we must have a sequence of n 0-1 trials which contains x 1s and $n - x$ 0s. As the trials are independent, with probability θ of obtaining a 1 on any trial, all such sequences have probability $\theta^x (1-\theta)^{n-x}$ and there are ${}^nC_x$ (that is "n choose x") such sequences, where

$${}^nC_x = \binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{n \times (n-1) \times (n-2) \times ... \times (n-x+1)}{x \times (x-1) \times (x-2) \times ... \times 2 \times 1}$$

(that is we must choose the x positions in the sequence to place the ones). Hence the total probability of observing the event $X = x$ is

$$\binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Interpretation : A model for "the number of successes in n i.i.d. 0-1 trials" where the probability of success on each trial is θ. Note that the sum of n independent and identically distributed (i.i.d.) $Bernoulli(\theta)$ random variables has a Binomial distribution, that is,

$$\text{If } X_1, ..., X_n \sim Bernoulli(\theta) \text{ } i.i.d., \text{ then } X = \sum_{i=1}^{n} X_i \sim Binomial(n, \theta)$$

EXAMPLE : In a multiple choice exam, each of the ten questions has five possible answers. For any single question, the probability of guessing the right answer is 1/5. Therefore the random variable X corresponding to the total number of correct answers assuming that each question is answered by guesswork has a binomial distribution, $X \sim Binomial(10, 1/5)$.

1.10.4 THE GEOMETRIC DISTRIBUTION $X \sim Geometric(\theta)$

Range : $\mathbb{X} = \{1, 2, ...\}$
Parameter : $\theta \in (0, 1]$
Mass function :

$$f_X(x) = (1-\theta)^{x-1} \theta. \qquad x \in \{1, 2, ...\}$$

Interpretation : A model for "the number of 0-1 trials until the first success is obtained " - where the probability of success on each trial is θ. a **discrete waiting-time** model.

Recall that $f_X(x) = \mathrm{P}[\ X = x\ ]$. Now if $X = x$, we must have a sequence of x 0-1 trials which contains $x-1$ 0s and then a 1. As the trials are independent, with probability θ of obtaining a 1 on any trial, such a sequence has probability $(1-\theta)^{x-1}\theta$ which immediately gives the form of the mass function.

EXAMPLE : A coin is tossed until the first head is obtained. The random variable X corresponding to the number of tosses required to obtain a the first head has a geometric distribution, $X \sim Geometric(1/2)$

1.10.5 THE NEGATIVE BINOMIAL DISTRIBUTION $X \sim NegBin(n, \theta)$

Range : $\mathbb{X} = \{n, n+1, n+2, ...\}$
Parameter : $n \in Z^+$, $\theta \in (0, 1]$
Mass function :

$$f_X(x) = \binom{x-1}{n-1}\theta^n(1-\theta)^{x-n} \qquad x \in \{n, n+1, n+2, ...\}\,.$$

Interpretation : A model for "the number of 0-1 trials until the nth success is obtained" where the probability of success on each trial is θ.

To compute the mass function, note that if the nth success occurs on the xth trial, then the sequence of 0-1 trials must comprise a sequence of $x-1$ trials containing $n-1$ 1s and $x-n$ 0s, followed by a final 1. Thus the mass function for a negative binomial distribution is given by

$$\binom{x-1}{n-1}\theta^{n-1}(1-\theta)^{x-n} \times \theta$$

that is a binomial probability multiplied by a single θ. as all trials are independent.

Note: The sum of n i.i.d. $Geometric(\theta)$ random variables has a Negative Binomial distribution, that is,

$$\text{If } X_1, ..., X_n \sim Geometric(\theta) \qquad \text{with } X_1, ..., X_n \text{ i.i.d, then} \qquad X = \sum_{i=1}^{n} X_i \sim NegBin(n, \theta)$$

that is, the number of trials until the nth 1 is the sum of the number of trials until the first 1, plus the number of trials between the first and second 1, etc. For this reason, the negative binomial distribution is also known as the **GENERALIZED GEOMETRIC** distribution.

EXAMPLE : A coin is tossed until the tenth head is obtained. The random variable X corresponding to the number of tosses required to obtain a the tenth head has a negative binomial distribution, $X \sim NegBin(10, 1/2)$

1.10.6 THE POISSON DISTRIBUTION $X \sim Poisson(\lambda)$

Range : $\mathbb{X} = \{0, 1, 2, ...\}$
Parameter : $\lambda \in \mathbb{R}^+$
Mass function :

$$f_X(x) = \frac{\lambda^x}{x!}e^{-\lambda} \qquad x \in \{0, 1, 2, ...\}$$

Note : The Poisson is the limiting case of the binomial distribution, that is, if we write $\theta = \lambda/n$, and then consider a limiting case as $n \longrightarrow \infty$. Specifically, if $X \sim Bin(n, \lambda/n)$, then

$$f_X(x) = \binom{n}{x}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} = \frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\frac{n!}{(n-x)!}\left(\frac{1}{n-\lambda}\right)^x \longrightarrow \frac{\lambda^x}{x!}e^{-\lambda}$$

as $n \longrightarrow \infty$ with λ constant. That is, we have that

$$Binomial(n, \theta) \longrightarrow Poisson(\lambda).$$

in the limiting case.
Interpretation : The Poisson model is appropriate for count data, where the number of events (accidents, breakdowns etc) that occur in a given time period are being counted.

Alternative derivation : Consider a sequence of events occurring **independently** and **at random** in time at a **constant rate** λ per unit time. Let $X(t)$ be the random variable defined for $t > 0$ by

"$X(t) = x$" if and only if "x events occur in time interval $[0, t)$".

Then $X(t) \sim Poisson(\lambda t)$, so that

$$f_{X(t)}(x) = \mathrm{P}[X(t) = x] = \frac{(\lambda t)^x}{x!} e^{-(\lambda t)} \qquad x \in \{0, 1, 2, ...\}\,.$$

EXAMPLE : The reliability of a machine is to be monitored. The random variable X corresponding to the number of breakdowns that the machine suffers in one month can be modelled using a Poisson distribution.

1.11 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.

1.11.1 THE CONTINUOUS UNIFORM DISTRIBUTION $X \sim Uniform(a, b)$

Range : $\mathbb{X} = (a, b)$
Parameters : $a, b \in \mathbb{R}$
Density function :

$$f_X(x) = \frac{1}{b-a} \qquad a < x < b$$

Interpretation : A model with **constant** probability density on a region.

1.11.2 THE EXPONENTIAL DISTRIBUTION $X \sim Exponential(\lambda)$

Range : $\mathbb{X} = \mathbb{R}^+$
Parameter : $\lambda \in \mathbb{R}^+$
Density function :

$$f_X(x) = \lambda e^{-\lambda x} \qquad x \in \mathbb{R}^+$$

Interpretation : A **continuous** waiting-time model (that is, a useful model for data that arise at the end of a measurement experiment).

The cdf for the exponential distribution can be calculated easily;

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\ dt = \int_0^x \lambda e^{-\lambda t}\ dt = 1 - e^{-\lambda x} \qquad x \geq 0.$$

and note that

$$\mathrm{P}[\ X > x\ ] = 1 - \mathrm{P}[\ X \leq x\ ] = 1 - F_X(x) = e^{-\lambda x}$$

which may give some motivation for using the Exponential model in practice.

Note: Suppose that a mechanical component fails in such a way that the number of failures in a given month has a Poisson distribution with parameter λ, that is, the failure events occur repeatedly through time at random, but failures are relatively rare. Suppose that a component is installed at time $x = 0$. Then the random variable X corresponding to the *time until the first failure* has an exponential distribution, $X \sim Exponential(\lambda)$.

In fact, it can be shown that *all* the inter-failure event times have an exponential distribution, and that the inter-event times are probabilistically independent. One important property of the Exponential distribution linked to this result is the **lack of memory** property which is defined as follows; consider the **conditional** probability that random variable X is greater than $x_0 + x$, given that X is greater than x_0; such a conditional probability is important if we are trying to assess, say, the probability that, if a component functions without failure for a time x_0, it continues to function without failure until $x_0 + x$. From the conditional probability definition, we have

$$\mathrm{P}\left[X > x_0 + x | X > x_0\right] = \frac{\mathrm{P}\left[(X > x_0 + x) \cap (X > x_0)\right]}{\mathrm{P}\left[X > x_0\right]} = \frac{\mathrm{P}\left[X > x_0 + x\right]}{\mathrm{P}\left[X > x_0\right]} = \frac{e^{-\lambda(x_0+x)}}{e^{-\lambda x_0}} = e^{-\lambda x} = \mathrm{P}\left[X > x\right]$$

so that the component has "forgotten" the time period up to x_0.

1.11.3 THE GAMMA DISTRIBUTION $X \sim Gamma(\alpha, \beta)$

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

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x \in \mathbb{R}^+$$

where

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\ dt \qquad \alpha > 0.$$

is known as the **Gamma Function**. This integral cannot by calculated analytically, but can be computed using numerical integration for any $\alpha > 0$.

Interpretation : The Gamma distribution **continuous** waiting-time model. It can be shown the sum of i.i.d. Exponential random variables has a Gamma distribution, that is, if $X_1, X_2, ..., X_n$ are independent and identically distributed $Exponential(\lambda)$ random variables, then

$$X = \sum_{i=1}^{n} X_i \sim Gamma(n, \lambda)$$

1. If $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$.
2. If $\alpha = 1, 2, ...$, $\Gamma(\alpha) = (\alpha - 1)!$.

3. $\Gamma(1/2) = \sqrt{\pi}$.

4. If $\alpha = 1, 2, ...$, then the $Gamma(\alpha/2, 1/2)$ distribution is known as the **Chi-squared distribution** with α **degrees of freedom**, denoted χ^2_α.

EXAMPLE: Suppose that a mechanical component fails in such a way that the number of failures in a given month has a Poisson distribution with parameter λ. Suppose that a component is installed at time $x = 0$. The random variable X corresponding to the time until the nth ($n = 1, 2, ...$) failure has a gamma distribution, $X \sim Gamma(n, \lambda)$.

1.11.4 THE BETA DISTRIBUTION $X \sim Beta(\alpha, \beta)$

Range : $\mathbb{X} = (0, 1)$
Parameters : $\alpha, \beta \in \mathbb{R}^+$
Density function :

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \qquad x \in (0, 1).$$

Interpretation : A model for experiments whose outcomes lie in a bounded interval.

EXAMPLE: The percentage content of a particular element in a material is to be measured. The random variable corresponding to the observed content can be modelled using a beta distribution.

1.11.5 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.

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

1.12 Probability Results & Limit Theorems

Theorem 1.12.1 *If X is a random variable, and h is a non-negative real function, then for any $c > 0$,*

$$\mathrm{P}\left[h(X) \geq c\right] \leq \frac{\mathrm{E}_{f_X}\left[h(X)\right]}{c}$$

Proof.Proof (continuous case) : Suppose that X has density function f_X which is positive for $x \in \mathbb{X}$. Let $\mathcal{A} = \{\ x \in \mathbb{X} :\ h(x) \geq c\} \subseteq \mathbb{X}$. Then, as $h(x) \geq c$ on$\mathcal{A}$,

$$\begin{aligned}
\mathrm{E}_{f_X}\left[\ h(X)\ \right] &= \int_{\mathbb{X}} h(x) f_X(x)\ dx \\
&= \int_{\mathcal{A}} h(x) f_X(x)\ dx + \int_{\mathcal{A}'} h(x) f_X(x)\ dx \\
&\geq \int_{\mathcal{A}} h(x) f_X(x)\ dx \\
&\geq \int_{\mathcal{A}} c f_X(x)\ dx \\
&= c\mathrm{P}\left[\ X \in \mathcal{A}\ \right] \ = c\mathrm{P}\left[h(X) \geq c\right]
\end{aligned}$$

Special Case : The **Markov Property** : $h(x) = |x|^r$ for $r > 0$, so

$$\mathrm{P}\left[|X|^r \geq c\right] \leq \frac{\mathrm{E}_{f_X}\left[|X|^r\right]}{c}$$

■

Theorem 1.12.2 *The **Chebychev Inequality***
Suppose that X is a random variable with expectation μ and variance σ^2. Then for any $k > 0$,

$$\mathrm{P}\left[|X - \mu| \geq k\sigma\right] \leq \frac{1}{k^2}$$

Proof. Put $h(x) = (x - \mu)^2$ and $c = k^2\sigma^2$ in the previous theorem. ■

Corollary 1.1 *For $\epsilon > 0$,*

$$P\left[|X - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{\epsilon^2} \qquad P\left[|X - \mu| < \epsilon\right] \geq 1 - \frac{\sigma^2}{\epsilon^2}$$

Definition 1.12.1 A function g is **convex** if

$$\frac{d^2}{dt^2}\left\{g(t)\right\}_{t=x} = g^{''}(x) > 0 \text{ for all } x$$

.

Theorem 1.12.3 ***Jensen's Inequality***
Suppose that X is a random variable, and function g is convex. Then

$$\mathrm{E}_{f_X}\left[g(X)\right] \geq g(\mathrm{E}_{f_X}\left[X\right])$$

Proof.A Taylor expansion of $g(x)$ around $x = \mu$ gives $g(x) = g(\mu) + (x-\mu)g^{'}(\mu) + \frac{1}{2}(x-\mu)^2 g^{''}(x_0)$ for some x_0 such that $x < x_0 < \mu$. Thus, taking expectations,

$$\begin{aligned}\mathrm{E}_{f_X}\left[g(X)\right] &= g(\mu) + \mathrm{E}_{f_X}\left[(X-\mu)\right] g^{'}(\mu) + \frac{1}{2}\mathrm{E}_{f_X}\left[(X-\mu)^2\right] g^{''}(x_0) \\ &\geq g(\mu) = g(\mathrm{E}_{f_X}\left[X\right])\end{aligned}$$

as $\mathrm{E}_{f_X}\left[(X-\mu)\right] = 0$, and $\mathrm{E}_{f_X}\left[(X-\mu)^2\right], g''(x_0) \geq 0$. ■

1.12.1 Convergence in Distribution

Definition 1.12.2 Consider a sequence of random variables $X_1, X_2, ...$ and a corresponding sequence of cdfs, $F_{X_1}, F_{X_2}, ...$ so that for $n = 1, 2, ..$ $F_{X_n}(x) = \mathrm{P}\left[X_n \leq x\right]$. Suppose that there exists a cdf, F_X, such that **for all x at which F_X is continuous**,

$$\lim_{n \longrightarrow \infty} F_{X_n}(x) = F_X(x).$$

Then $X_1, ..., X_n$ **converges in distribution** to random variable X with cdf F_X, denoted

$$X_n \xrightarrow{d} X$$

and F_X is the **limiting distribution**.
Note: Convergence of a sequence of mgfs also indicates convergence in distribution, that is, if for all t at which $M_X(t)$ is defined, if as $n \longrightarrow \infty$, we have

$$M_{X_i}(t) \longrightarrow M_X(t) \qquad \Longrightarrow \qquad X_n \xrightarrow{d} X$$

- this result was used to prove that the Poisson distribution arose as the limiting case of the Binomial distribution.

Definition 1.12.3 The sequence of random variables $X_1, ..., X_n$ converges in distribution to constant c if the limiting distribution of $X_1, ..., X_n$ is **degenerate at** c, that is,

$$X_n \xrightarrow{d} X$$

and $\mathrm{P}\left[X = c\right] = 1$, so that

$$F_X(x) = \begin{cases} 0 & x < c \\ 1 & x \geq c \end{cases}$$

Interpretation A *special case* of convergence in distribution occurs when the limiting distribution is *discrete*, with the probability mass function only being non-zero at a *single* value, that is, if the limiting random variable is X, then

$$f_X(x) = 1 \qquad x = c,$$

and zero otherwise. The following theorem illustrates another aspect of convergence in distribution.

Theorem 1.12.4 *The sequence of random variables $X_1, ..., X_n$ converges in distribution to c if and only if, for all $\epsilon > 0$,*

$$\lim_{n \longrightarrow \infty} \mathrm{P}\left[|X_n - c| < \epsilon\right] = 1$$

This theorem indicates that convergence in distribution to a constant c occurs if and only if the probability becomes increasingly concentrated around c as $n \longrightarrow \infty$.

1.12.2 Convergence In Probability

Definition 1.12.4 The sequence of random variables $X_1, ..., X_n$ **converges in probability** to **constant** c, denoted

$$X_n \xrightarrow{p} c$$

if

$$\lim_{n \longrightarrow \infty} \mathrm{P}\left[|X_n - c| < \epsilon\right] = 1 \qquad \text{or equivalently} \qquad \lim_{n \longrightarrow \infty} \mathrm{P}\left[|X_n - c| \geq \epsilon\right] = 0$$

that is, if the limiting distribution of $X_1, ..., X_n$ is degenerate at c.

Interpretation Convergence in probability to a constant is *precisely* equivalent to convergence in distribution to a constant.

Theorem 1.12.5 *The **Weak Law of Large Numbers***
Suppose that $X_1, ..., X_n$ is a sequence of i.i.d. random variables with expectation μ and variance σ^2. Let Y_n be defined by

$$Y_n = \frac{1}{n} \sum_{i=1}^{n} X$$

then, for all $\epsilon > 0$,

$$\lim_{n \longrightarrow \infty} P[|Y_n - \mu| < \epsilon] = 1,$$

*that is, $Y_n \xrightarrow{p} \mu$, and thus the **mean** of $X_1, ..., X_n$ converges in probability to μ.*

Proof. Using the properties of expectation, it can be shown that Y_n has expectation μ and variance σ^2/n, and hence by the **Chebychev Inequality**,

$$\mathrm{P}\left[|Y_n - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0 \text{ as } n \longrightarrow \infty$$

for all $\epsilon > 0$. Hence

$$\mathrm{P}\left[|Y_n - \mu| < \epsilon\right] \longrightarrow 1 \qquad n \longrightarrow \infty$$

■

1.12.3 The Central Limit Theorem

Theorem 1.12.6 *Suppose $X_1, ..., X_n$ are i.i.d. random variables with mgf M_X, with*

$$E_{f_X}[X_i] = \mu \qquad Var_{f_X}[X_i] = \sigma^2$$

both finite. Let the random variable Z_n be defined by

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

and let Z_n have mgf M_{Z_n}. Then, as $n \longrightarrow \infty$,

$$M_{Z_n}(t) \longrightarrow \exp\left\{\frac{t^2}{2}\right\}$$

irrespective *of the form of M_X.*

Proof. First, let $Y_i = (X_i - \mu)/\sigma$ for $i = 1, ..., n$. Then $Y_1, ..., Y_n$ are i.i.d. with mgf M_Y say, and by the elementary properties of expectation, $\mathrm{E}_{f_Y}[Y_i] = 0$,$\mathrm{Var}_{f_Y}[Y_i] = 1$ for each i. Using the power series expansion result for mgfs, we have that

$$\begin{aligned} M_Y(t) &= 1 + tE_{f_Y}[Y] + \frac{t^2}{2!}E_{f_Y}[Y^2] + \frac{t^3}{3!}E_{f_Y}[Y^3] + ... \\ &= 1 + \frac{t^2}{2!} + E_{f_Y}[Y^3] + ... \end{aligned}$$

Now, the random variable Z_i can be rewritten

$$Z_i = \frac{1}{\sqrt{n}}Y_i$$

and thus, again by a standard mgf result, as $Y_1, ..., Y_n$ are independent, we have that

$$M_{Z_n}(t) = \prod_{i=1}^{n}\left\{M_Y(t/\sqrt{n})\right\} = \left\{1 + \frac{t^2}{2n} + \frac{t^3}{6n^{3/2}}E_{f_Y}[Y^3] + ...\right\}^n$$

Taking logs, and using the expansion $\log(1+s) = s - s^2/2 + s^3/3 - ...$ we have that

$$\log M_{Z_n}(t) = n\left[\left(\frac{t^2}{2n} + \frac{t^3}{6n^{3/2}}E_{f_Y}[Y^3] + ...\right) - \frac{1}{2}\left(\frac{t^2}{2n} + \frac{t^3}{6n^{3/2}}E_{f_Y}[Y^3] + ...\right)^2 + ...\right]$$

Thus, as $n \longrightarrow \infty$, only the very first term is non-zero, and

$$\log M_{Z_n}(t) \longrightarrow \frac{t^2}{2} \qquad \text{so that} \qquad M_{Z_n}(t) \longrightarrow \exp\left\{\frac{t^2}{2}\right\}$$

■

2. Statistical Inference

2.1 Introduction

Suppose that $y = (y_1, ..., y_n)$ are observed values of random variables $Y = (Y_1, ..., Y_n)$ that constitute a random (independent and identically distributed sample from a probability distribution with joint density $f_{Y|\theta}(y;\theta)$ depending on a parameter θ. Let $\mathbb{Y}$ denote the *sample space*, that is

$$\mathbb{Y} = \left\{y : f_{Y|\theta}(y;\theta) > 0\right\}$$

and let Θ denote the *parameter space*

$$\Theta = \left\{\theta : f_{Y|\theta}(y;\theta) \text{ is a density function in } y\right\}. \tag{2.1}$$

The following objectives will be the focus of inference: we wish to

1. report an *estimate* of the parameter θ in light of the data y
2. report an *uncertainty interval* within which the "true" value of θ lies
3. answer specific questions about θ

2.2 Statistical Summaries

Definition 2.2.1 A collection of i.i.d. random variables $Y_1, ..., Y_n$ each of which has distribution defined by cdf F_Y (or mass/density function f_Y) is a **random sample** of size n from F_Y (or f_Y).

Definition 2.2.2 A function, T, of a random sample, $Y_1, ..., Y_n$, that is, $T = t(Y_1, ..., Y_n)$ that depends **only** on $Y_1, ..., Y_n$ is a **statistic**. A statistic is a *random variable*.

Example 2.2.1 The **sample mean**

$$Y, Y = \frac{1}{n}\sum_{i=1}^{n} Y$$

is a statistic.

Definition 2.2.3 The **likelihood function** for parameter θ and data y, $l(\theta; y)$, is defined by

$$l(\theta; y) = f_{Y|\theta}(y;\theta)$$

viewed as a function of θ. If the data $y = (y_1, ..., y_n)$ are **independent**, then

$$l(\theta; y) = f_{Y|\theta}(y;\theta) = \prod_{i=1}^{n} f_{Y|\theta}(y_i;\theta)$$

Note. The likelihood function is **NOT** a probability density in θ. ■

One key objective is to construct a procedure that takes in the data values y and outputs an estimate of the unknown parameter θ. The following definitions and considerations will be required.

Definition 2.2.4 An **estimate** of θ, denoted $\widehat{\theta}$, is any function t of the data y

$$\widehat{\theta} = t(y).$$

The corresponding random variable

$$T = t(Y)$$

formed as a function of the random variables $Y_1, ..., Y_n$ is termed an **estimator**

Example 2.2.2 Define

$$t = t(y) = t(y_1, ..., y_n) = \frac{y_1 + ... + y_n}{n} \qquad T = t(Y) = t(Y_1, ..., Y_n) = \frac{Y_1 + ... + Y_n}{n}$$

The distribution of T may depend on θ; the distribution of T will be of interest.

Definition 2.2.5 The distribution of a statistic T=$t(Y)$ is termed the **sampling distribution**. Denote the sampling distribution (density) by $f_{T|\theta}$.

Definition 2.2.6 An estimator T of θ is **unbiased** if

$$\mathrm{E}_{T|\theta}[T] = \theta \qquad \forall \theta \in \Theta$$

that is, if the expected value of T with respect to its sampling distribution (given θ) is equal to θ.
The **bias** of T is defined by

$$b = b(\theta) = \mathrm{E}_{T|\theta}[T] - \theta.$$

Definition 2.2.7 An estimator T_{MV} of θ is a **minimum variance unbiased estimator (MVUE)** if T_{MV} is unbiased for θ and

$$\mathrm{Var}_{T_{MV}|\theta}[T_{MV}] \leq \mathrm{Var}_{T|\theta}[T] \qquad \forall \theta \in \Theta$$

that is, if the expected value of T with respect to its sampling distribution (given θ) is equal to θ.

Definition 2.2.8 The **mean-square error (MSE)** of estimator T of θ is defined by

$$MSE = \mathrm{E}_{T|\theta}\left[(T-\theta)^2\right] = \mathrm{Var}_{T|\theta}[T] + b^2$$

Definition 2.2.9 An estimation procedure is termed **regular** if $f_{T|\theta}$
(i) there is no discontinuity in $f_{Y|\theta}$ that depends on θ
(ii) $f_{Y|\theta}$ satisfies the conditions of Fubini's theorem in y and θ (that is, the order of differentiation and intergration can be interchanged).

2.3 Sampling Distributions

Recall that a sampling distribution is the distributioon of a statistic.

Example 2.3.1 If $X_1, ..., X_n$ are independent random variables, with $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, ..., n$, and $a_1, ..., a_n$ are constants, consider the distribution of random variable Y defined by

$$Y = \sum_{i=1}^{n} a_i X_i$$

Using standard mgf results, the distribution of Y is derived to be normal with parameters

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i \qquad \sigma_Y^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

Now consider the special case of this result when $X_1, ..., X_n$ are i.i.d. with $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2$, and where $a_i = 1/n$ for $i = 1, ..., n$. Then

$$Y = \sum_{i=1}^{n} \frac{1}{n} X_i = \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Definition 2.3.1 For a random sample $X_1, ..., X_n$ from a probability distribution, then the **sample variance**, s^2, is a statistic defined by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Theorem 2.3.1 *Suppose that $X_1, ..., X_n$ is a random sample from a normal distribution, say $X_i \sim N(\mu, \sigma^2)$. Then*
*(1) $\bar{X}$ is **independent** of $\left\{X_i - \bar{X}, i = 1, ..., n\right\}$*
*(2) $\bar{X}$ and s^2 are **independent** random variables*
(3) The random variable

$$\frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{\sigma^2}$$

*has a **chi-squared** distribution with $n-1$ degrees of freedom.*

Proof.(1) First note that

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} \left(X_i - \bar{X} + \bar{X} - \mu\right)^2 = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 + n\left(\bar{X} - \mu\right)^2$$

The joint pdf of $X_1, ..., X_n$ is readily obtained as the variables are independent

$$\begin{aligned} f_{X_1,..,X_n}(x_1,..,x_n) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp \quad \left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp \quad \left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right]\right\} \end{aligned} \tag{2.2}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Now consider the multivariate transformation to $Y_1, ..., Y_n$ where

$$Y_1 = \bar{X} \qquad Y_i = X_i - \bar{X} \quad i = 2, ..., n \qquad \Longleftrightarrow \qquad X_1 = Y_1 - \sum_{i=2}^{n} Y_i \qquad X_i = Y_i + Y_1.$$

Note also that

$$X_1 - \bar{X} = -\sum_{i=2}^{n}\left(X_i - \bar{X}\right) = -\sum_{i=2}^{n} Y_i,$$

and so

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 = \left(X_1-\bar{X}\right)^2 + \sum_{i=2}^{n}\left(X_i-\bar{X}\right)^2 = \left(-\sum_{i=2}^{n} Y_i\right)^2 + \sum_{i=2}^{n} Y_i^2$$

The Jacobian of this transformation is n, so the joint density of $Y_1, ..., Y_n$ is given by

$$f_{Y_1,..,Y_n}(y_1,..,y_n) = n\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp \quad \left\{-\frac{1}{2\sigma^2}\left[\left(-\sum_{i=2}^{n} y_i\right)^2 + \sum_{i=2}^{n} y_i^2 + n(y_1-\mu)^2\right]\right\}$$

so that

$$f_{Y_1,..,Y_n}(y_1,..,y_n) = f_{Y_2,..,Y_n}(y_2,..,y_n) f_{Y_1}(y_1).$$

Therefore Y_1 is independent of $Y_2, ..., Y_n$, and hence $\bar{X}$ is independent of the random variables terms

$$\left\{X_i - \bar{X}, i = 2, ..., n\right\}$$

Finally, as X

$$_1 - \bar{X} = -\sum_{i=2}^{n}\left(X_i - \bar{X}\right)$$

, $\bar{X}$ is also independent of $X_1 - \bar{X}$.

(2) s^2 is a function only of $\left\{X_i - \bar{X}, i = 1, ..., n\right\}$. As $\bar{X}$ is independent of these variables, $\bar{X}$ and s^2 are also independent.

(3) From above

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sigma^2} + \frac{n\left(\bar{X}-\mu\right)^2}{\sigma^2}$$

or $V_1 = V_2 + V_3$, say. Now, $X_i \sim N(\mu, \sigma^2)$, so therefore $(X_i - \mu)/\sigma \sim N(0, 1)$, and hence

$$(X_i-\mu)^2/\sigma^2 \sim \chi_1^2 \equiv Ga(1/2, 1/2) \qquad \Longrightarrow \qquad \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = V_1 \sim \chi_n^2$$

as the X_is are **independent**, and the sum of n independent $Ga(1/2, 1/2)$ variables has a $Ga(n/2, 1/2)$ distribution. Similarly, as $\bar{X} \sim N(\mu, \sigma^2/n)$, $V_3 \sim \chi^2_1$. By part (2), V_2 and V_3 are **independent**, and so the mgfs of V_1, V_2 and V_3 are related by

$$M_{V_1}(t) = M_{V_2}(t)M_{V_3}(t) \implies M_{V_2}(t) = \frac{M_{V_1}(t)}{M_{V_3}(t)}.$$

As V_1 and V_3 are Gamma random variables, M_{V_1} and M_{V_3} are given by

$$M_{V_1}(t) = \left(\frac{1/2}{1/2-t}\right)^{n/2}, \; M_{V_3}(t) = \left(\frac{1/2}{1/2-t}\right)^{1/2} \qquad \Longrightarrow \qquad M_{V_2}(t) = \left(\frac{1/2}{1/2-t}\right)^{(n-1)/2}$$

which is also the mgf of a Gamma random variable, and hence

$$V_2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

. ■

Theorem 2.3.2 *Suppose that $X_1, ..., X_n$ is a random sample from a normal distribution, say $X_i \sim N(\mu, \sigma^2)$. Then the random variable*

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

*has a **Student-t** distribution with $n-1$ degrees of freedom.*

Proof. Consider the random variables $Z = \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$ and $V = (n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$, and use the properties of the normal distribution and related quantities to find the distribution of $Z/\sqrt{V/n-1}$. ■

2.4 Score Functions

Consider a regular estimation problem for single parameter θ.

Definition 2.4.1 The **efficient score function** , denoted $U_\theta(y)$, is defined by

$$U_\theta(y) = \frac{\partial}{\partial\theta} \log f_{Y|\theta}(y; \theta)$$

Note: $U_\theta(Y)$ is a random variable.

Lemma 2.4.1 If $\phi = \phi(\theta)$ is a differentiable function of θ. Then

$$U_\phi(y) = \frac{\partial}{\partial\phi} \log f_{Y|\phi}(y; \phi) = \frac{1}{\phi'(\theta)} U_\theta(y)$$

Lemma 2.4.2 For efficient score $U_\theta(Y)$

$$\mathrm{E}_{U|\theta}[U_\theta] = 0$$

Proof. From first principles,

$$\int_{\mathbb{Y}} f_{Y|\theta}(y;\theta)\,dy = 1 \Longrightarrow \int_{\mathbb{Y}} \frac{\partial}{\partial\theta} f_{Y|\theta}(y;\theta)\,dy = 0$$

exchanging the order of integration and differentiation. But

$$\int_{\mathbb{Y}} \frac{\partial}{\partial\theta} f_{Y|\theta}(y;\theta)\,dy = \int_{\mathbb{Y}} \left\{\frac{\partial}{\partial\theta}\log f_{Y|\theta}(y;\theta)\right\} f_{Y|\theta}(y;\theta)\,dy = \mathrm{E}_{U|\theta}[U_\theta]$$

and hence

$$\mathrm{E}_{U|\theta}[U_\theta] = 0.$$

■

Definition 2.4.2 The **Fisher Information** about θ is denoted $I(\theta)$, and defined as

$$I(\theta) = \mathrm{Var}_{U|\theta}[U_\theta] = \mathrm{E}_{U|\theta}\left[U_\theta^2\right]$$

Lemma 2.4.3 In a regular estimation problem

$$I(\theta) = \mathrm{E}_{U|\theta}\left[U_\theta^2\right] = -\mathrm{E}_{Y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f_{Y|\theta}(Y;\theta)\right]$$

Proof. From above

$$\mathrm{E}_{U|\theta}[U_\theta] = 0 = \int_{\mathbb{Y}} \left\{\frac{\partial}{\partial\theta}\log f_{Y|\theta}(Y;\theta)\right\} f_{Y|\theta}(Y;\theta)\,dy$$

and therefore, differentiating partially with respect to θ under the integral,

$$0 = \int_{\mathbb{Y}} \left[\left\{\frac{\partial}{\partial\theta} f_{Y|\theta}(Y;\theta)\right\}\left\{\frac{\partial}{\partial\theta}\log f_{Y|\theta}(Y;\theta)\right\} + f_{Y|\theta}(y;\theta)\left\{\frac{\partial^2}{\partial\theta^2}\log f_{Y|\theta}(Y;\theta)\right\}\right] dy$$

But the first term

$$\int_{\mathbb{Y}} \left\{\frac{\partial}{\partial\theta} f_{Y|\theta}(Y;\theta)\right\}\left\{\frac{\partial}{\partial\theta}\log f_{Y|\theta}(Y;\theta)\right\} dy = \int_{\mathbb{Y}} \left\{\frac{\partial}{\partial\theta}\log f_{Y|\theta}(Y;\theta)\right\}^2 f_{Y|\theta}(Y;\theta)\,dy = \mathrm{E}_{U|\theta}\left[U_\theta^2\right]$$

and hence

$$\mathrm{E}_{U|\theta}\left[U_\theta^2\right] + \int_{\mathbb{Y}} \left\{\frac{\partial^2}{\partial\theta^2}\log f_{Y|\theta}(y;\theta)\right\} f_{Y|\theta}(y;\theta)\,dy = 0$$

and the result follows. ■

Lemma 2.4.4 For a differentiable function $\phi = \phi(\theta)$,

$$I(\phi) = \frac{I(\theta)}{\left(\phi'(\theta)\right)^2}$$

The Fisher information measures the curvature of the likelihood function near its maximum value.

2.5 The Exponential Family

Definition 2.5.1 Consider a random variable Y. The density function of Y is a member of the **one-parameter exponential family** if

$$f_{Y|\theta}(y;\theta) = \exp\{a(\theta)b(y) + c(\theta) + d(y)\}$$

for known analytic functions a, b, c, d.

Note. For $f_{Y|\theta}$ in a one-parameter exponential family,

$$U_\theta = \frac{\partial}{\partial\theta}\left[a(\theta)b(y) + c(\theta) + d(y)\right] = a'(\theta)b(y) + c'(\theta)$$

But

$$E_{U|\theta}[U_\theta] = 0 \Longrightarrow E_{Y|\theta}[b(Y)] = -\frac{c'(\theta)}{a'(\theta)}.$$

Also

$$\begin{aligned} I(\theta) = \mathrm{E}_{U|\theta}\left[U_\theta^2\right] &= -\mathrm{E}_{Y|\theta}\left[\frac{\partial}{\partial\theta}U_\theta\right] \\ &= -a''(\theta)E_{Y|\theta}[b(Y)] - c''(\theta) \\ &= a''(\theta)\frac{c'(\theta)}{a'(\theta)} - c''(\theta) \\ &= -a'(\theta)\frac{\partial}{\partial\theta}\left\{\frac{c'(\theta)}{a'(\theta)}\right\} \end{aligned}$$

■

Note. Note that if $Z = b(Y)$ is a 1-1 transformation, and $\phi = a(\theta)$, then under this transformation/reparameterization

$$f_{Z|\phi}(z|\phi) = \frac{f_{Y|\theta}(b^{-1}(z)|\theta)}{b'(b^{-1}(z))} = \exp\{z\phi + c^*(\phi) + d^*(z)\}.$$

Hence, for a 1-1 transformation $Y \to b(Y)$, changing notation and the definitions of functions c and d,

$$f_{Y|\theta}(y|\theta) = \exp\{y\theta + c(\theta) + d(y)\}$$

Then

$$\mathrm{E}_{Y|\theta}[Y] = -c'(\theta) \qquad I(\theta) = \mathrm{Var}_{Y|\theta}[Y] = -c''(\theta)$$

■

Definition 2.5.2 The **canonical form** of a one-parameter distribution is

$$f_{Y|\theta}(y;\theta) = \exp\{y\theta + c(\theta) + d(y)\}$$

Note.

Corollary 2.1 *For a random sample* $y = (y_1, ..., y_n)$

$$f_{Y|\theta}(y;\theta) = \prod_{i=1}^{n} f_{Y|\theta}(y_i;\theta) = \exp\left\{\theta\sum_{i=1}^{n} y_i + nc(\theta) + \sum_{i=1}^{n} d(y)\right\}$$

and hence, using the previous result,

$$I(\theta) = -nc''(\theta)$$

■

2.6 Multiparameter Problems

Definition 2.6.1 For a vector-valued parameter $\theta^T = (\theta_1, ..., \theta_q)$, the efficient score function, is defined by

$$U_\theta^T(y) = \left[\frac{\partial}{\partial\theta_1}\log f_{Y|\theta}(y;\theta), ..., \frac{\partial}{\partial\theta_q}\log f_{Y|\theta}(y;\theta)\right].$$

The Fisher Information about θ is defined as

$$I(\theta) = \mathrm{Var}_{U|\theta}[U_\theta] = \left[\mathrm{E}_{Y|\theta}\left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f_{Y|\theta}(y;\theta)\right]\right]_{ij}$$

(that is, the $q \times q$ matrix of expected values of 2nd order partial derivatives of the log-likelihood)

2.7 Properties of Estimators

Definition 2.7.1 An estimator, T, is an **invariant** estimator of parameter θ if, for a function $\tau(\theta)$ of θ, $\tau(T)$ is an estimator of $\tau(\theta)$.

Example 2.7.1 If $Y_1, ..., Y_n$ is a random sample from an $Eyp(\lambda)$ distribution, then $T = 1/Y$ is the maximum likelihood estimator of λ. Suppose now that we wish to find the maximum likelihood estimator of

$$\tau = \tau(\lambda) = \mathrm{P}[Y > 1] = e^{-\lambda}$$

for a new random variable Y that is independent of, and has the same distribution as, $Y_1, ..., Y_n$. Then, we can re-write the likelihood function in terms of τ (that is, by setting $\lambda = -\log\tau$) and use the maximum likelihood technique to obtain the m.l.e. of τ. It can be shown that the maximum likelihood estimator of τ is given by

$$\hat{\tau} = e^{-1/Y} \Longrightarrow \hat{\tau}(\lambda) = \tau(\hat{\lambda}).$$

Definition 2.7.2 An estimator, T, is an **unbiased** estimator of function $\tau(\theta)$ of parameter θ if

$$\mathrm{E}_{f_T}[T] = \tau(\theta)$$

where f_T is the **sampling distribution** of T. The **bias**, $b(T)$, of an estimator T of $\tau(\theta)$ is defined by

$$b(T) = \mathrm{E}_{f_T}[T] - \tau(\theta)$$

and the **Mean Squared Error**, or **MSE**, of T is defined by

$$MSE(T) = \mathrm{E}_{f_T}[(T - \tau(\theta))^2]$$

2.7.1 Asymptotic Properties of Estimators

By looking at the sampling distribution of an estimator derived from a random sample of size n, it is possible to assess whether the estimator has "good" properties (e.g. unbiasedness, minimum variance). We now look for estimators which exhibit desirable behaviour as $n \longrightarrow \infty$.

Definition 2.7.3 Suppose that $T_1, ..., T_n, ...$ are a sequence of estimators of function $\tau(\theta)$ of parameter θ. The sequence of estimators are **consistent** estimators of $\tau(\theta)$ if, for all $\epsilon > 0$,

$$\lim_{n\longrightarrow\infty} \mathrm{P}[|T_n - \tau(\theta)| < \epsilon] = 1$$

for every $\theta \in \Theta$, that is, if

$$T_n \xrightarrow{P} \tau(\theta)$$

and T_n converges in probability to $\tau(\theta)$.

Definition 2.7.4 A sequence $T_1, ..., T_n, ...$ of estimators of function $\tau(\theta)$ of parameter θ are **mean squared error consistent** if if,

$$\lim_{n \longrightarrow \infty} \mathrm{E}_{f_{T_n}}[(T_n - \tau(\theta))^2] = 0$$

for every $\theta \in \Theta$.

Definition 2.7.5 A sequence $T_1, ..., T_n, ...$ of estimators of function $\tau(\theta)$ of parameter θ are **asymptotically unbiased** if,

$$\lim_{n \longrightarrow \infty} \mathrm{E}_{f_{T_n}}[T_n] = \tau(\theta)$$

for every $\theta \in \Theta$.

Example 2.7.2 If $Y_1, ..., Y_n$ is a random sample from an $N(\mu, \sigma^2)$ distribution, then $T = Y$ is an estimator of μ, and the sampling distribution of T, f_T, is a normal distribution with expectation μ and variance σ^2/n. Hence $\mathrm{E}_{f_T}[T] = \mu$, and thus Y is an **unbiased** estimator of μ. Similarly, the maximum likelihood estimator of σ^2 is $\hat{\sigma}^2 = s^2$, the sample variance. By a previous theorem, the random variable Y given by

$$Y = \frac{(n-1)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (Y_i - Y)^2}{\sigma^2} \sim \chi^2_{n-1}$$

and hence the expectation of Y is equal to $(n-1)/\sigma^2$ multiplied by the expectation of a χ^2_{n-1} random variable. Hence the expectation of statistic $T = s^2$ is given by $(n-1)\sigma^2/(n-1) = \sigma^2$, and therefore s^2 is an **unbiased** estimator of σ^2 Note: The method of moments estimator in this case is **not** an unbiased estimator of σ^2.)

2.8 Variance of Estimators

It is desirable that the sampling distribution of an estimator is "concentrated" around the true value of the parameter. For example, if T_1 and T_2 are two possible estimators of a function $\tau(\theta)$ of parameter θ, then we would prefer T_1 to T_2 if T_1 is "more concentrated" about $\tau(\theta)$ than is T_2, that is, if, for all $\epsilon > 0$,

$$\mathrm{P}[\tau(\theta) - \epsilon < T_1 < \tau(\theta) + \epsilon] \geq \mathrm{P}[\tau(\theta) - \epsilon < T_2 < \tau(\theta) + \epsilon]$$

where the left and right hand side depend on the sampling distributions of T_1 and T_2 respectively. Furthermore, if T is an **unbiased** estimator of $\tau(\theta)$, then the Chebychev Inequality suggests that

$$\mathrm{P}[\tau(\theta) - \epsilon < T < \tau(\theta) + \epsilon] \geq 1 - \mathrm{Var}_{f_T}[T]/\epsilon^2$$

for all $\epsilon > 0$, as $\mathrm{E}_{f_T}[T] = \tau(\theta)$. Hence an unbiased estimator whose sampling distribution has a small variance will be concentrated around $\tau(\theta)$, and thus will be preferable to estimators with larger variance. If this is true for **all** possible values of parameter θ, then this estimator might be regarded as the "best" estimator of $\tau(\theta)$.

Definition 2.8.1 Suppose that $Y_1, ..., Y_n$ are a random sample of size n from mass or density function $f_{Y|\theta}$, that depends on (possibly vector) parameter $\theta \in \Theta$. Then an estimator T^* of function $\tau(\theta)$ of θ is the **uniformly minimum variance unbiased estimator**, or **UMVUE**, of $\tau(\theta)$ if
(i) T^* is unbiased for $\tau(\theta)$, and
(ii) T is any other unbiased estimator of $\tau(\theta)$ then

$$\mathrm{Var}_{f_{T^*|\theta}}[T^*] \leq \mathrm{Var}_{f_{T|\theta}}[T] \qquad \forall \theta \in \Theta,$$

In the single parameter case, it can be shown that, if T is an unbiased estimator of $\tau(\theta)$ based on a random sample of size n from f_Y, then

$$\mathrm{Var}_{f_T}[T] \geq \frac{[\tau'(\theta)]^2}{n\mathrm{E}_{f_{Y|\theta}}\left[\left\{\frac{\partial}{\partial\theta}\{\log f_Y(Y;\theta)\}\right\}^2\right]} = \frac{[\tau'(\theta)]^2}{n\mathrm{E}_{f_Y}\left[\{\partial(Y;\theta)\}^2\right]}$$

where $\tau'(\theta)$ is the first partial derivative of $\tau(\theta)$ with respect to θ. This is the **Cramer-Rao Lower Bound** on the variance of an unbiased estimator. If an estimator can be found that has the minimum variance, then this estimator can be regarded as the "best" unbiased estimator.

We study this important issue further below:

2.8.1 The Cramer-Rao Lower Bound

A natural question relating to unbiased estimators and minimum variance is whether a lower bound for the variance in the class of unbiased estimators can be found. In fact, such a bound exists, and is given by the following results.

Lemma 2.8.1 For random variables X and Y

$$\{\mathrm{Cov}[X,Y]\}^2 \leq \mathrm{Var}[X]\,\mathrm{Var}[Y]$$

and

$$\{\mathrm{Cov}[X,Y]\}^2 = \mathrm{Var}[X]\,\mathrm{Var}[Y]$$

if and only if

$$Y = aX + b$$

Proof. For any a, b, $\mathrm{Var}[aX + bY] \geq 0$, and setting $a = -\mathrm{Cov}[X,Y], b = \mathrm{Var}[X]$, and hence

$$\begin{aligned}\mathrm{Var}[aX + bY] &= a^2\mathrm{Var}[X] + 2ab\mathrm{Cov}[X,Y] + b^2\mathrm{Var}[Y] \\ &= \{\mathrm{Cov}[X,Y]\}^2\,\mathrm{Var}[X] - 2\{\mathrm{Cov}[X,Y]\}^2\,\mathrm{Var}[X] + \{\mathrm{Var}[X]\}^2\,\mathrm{Var}[Y] \geq 0\end{aligned}$$

and hence

$$\{\mathrm{Cov}[X,Y]\}^2 \leq \mathrm{Var}[X]\,\mathrm{Var}[Y]$$

(provided $\mathrm{Var}[X] > 0$). Further,

$$\mathrm{Var}[aX + bY] = 0 \Leftrightarrow aX + bY \text{ is a constant random variable}$$

■

Lemma 2.8.2 If T is an estimator of θ with bias $b(\theta)$ then

$$\mathrm{Var}_{T|\theta}[T] \geq \frac{(1 + b'(\theta))^2}{I(\theta)}$$

Proof. By definition, $E_{T|\theta}[T] = \theta + b(\theta)$. But

$$E_{T|\theta}[T] = \int_{\mathbb{Y}} t(Y) f_{Y|\theta}(y;\theta) dY$$

and hence, differentiating partially left and right hand sides by θ

$$\begin{aligned} 1 + b'(\theta) &= \int_{\mathbb{Y}} t(Y) \frac{\partial}{\partial\theta} f_{Y|\theta}(y;\theta) dY \\ &= \int_{\mathbb{Y}} t(Y) \left\{ \frac{\partial}{\partial\theta} \log f_{Y|\theta}(y;\theta) \right\} f_{Y|\theta}(y;\theta) dY \\ &= \mathrm{E}_{Y|\theta}[t(Y).U_\theta] \\ &= \mathrm{Cov}_{Y|\theta}[t(Y).U_\theta] \end{aligned} \tag{2.3}$$

as $\mathrm{E}_{Y|\theta}[U_\theta] = 0$. But, from the previous Lemma

$$\left\{\mathrm{Cov}_{Y|\theta}[t(Y).U_\theta]\right\}^2 \leq \mathrm{Var}_{Y|\theta}[t(Y)] \mathrm{Var}_{Y|\theta}[U_\theta]$$

and hence

$$\mathrm{Var}_{T|\theta}[T] = \mathrm{Var}_{Y|\theta}[t(Y)] \geq \frac{\left\{\mathrm{Cov}_{Y|\theta}[t(Y), U_\theta]\right\}^2}{\mathrm{Var}_{Y|\theta}[U_\theta]} = \frac{(1 + b'(\theta))^2}{I(\theta)}.$$

■

Note. If

$$\mathrm{Var}_{T|\theta}[T] = \frac{\left\{\mathrm{Cov}_{Y|\theta}[t(Y), U_\theta]\right\}^2}{\mathrm{Var}_{Y|\theta}[U_\theta]} \tag{2.4}$$

then $T = t(Y)$ and U_θ must be linearly related. Hence

$$U_\theta = g(\theta)(T - \theta - b(\theta))$$

for some function $g(\theta)$. ■

Theorem 2.8.1 ***The Cramer-Rao Lower Bound***
If $T = t(Y)$ is unbiased for θ

$$Var_{T|\theta}[T] \geq \frac{1}{I(\theta)}$$

with equality if and only if

$$U_\theta = g(\theta)(T - \theta)$$

Proof. If T is unbiased for θ, $b(\theta) = 0$, and the result follows from the previous Lemmas.
Note also that, from (2.3)

$$\mathrm{Cov}_{Y|\theta}[T, U_\theta] = 1 + b'(\theta) = 1$$

and hence, from (2.4)

$$\left\{\mathrm{Cov}_{Y|\theta}[T, U_\theta]\right\}^2 = \mathrm{Var}_{Y|\theta}[U_\theta] \mathrm{Var}_{T|\theta}[T] \Leftrightarrow 1.$$

Therefore

$$1 = g(\theta)\mathrm{Var}_{T|\theta}[T] \Leftrightarrow g(\theta) = I(\theta) \text{ as } \mathrm{Var}_{T|\theta}[T] = \frac{1}{I(\theta)}$$

and thus

$$U_\theta = I(\theta)(T - \theta).$$

■

This important result says that if an unbiased estimator T has variance that attains the Cramer-Rao lower bound, then it is MVUE.

2.9 Sufficiency

2.9.1 Introduction: Concepts of Sufficiency

Definition 2.9.1 A **statistic** $t = t(y)$ is a function of the data alone. Note that any statistic partitions the sample space $\mathbb{Y}$; two data samples y_1 and y_2, are termed **equivalent** with respect to statistic t, denoted $y_1 \sim y_2$ if

$$t(y_1) = t(y_2).$$

Note that the random variable $T = t(Y)$ is also termed a statistic.

Definition 2.9.2 A partition of $\mathbb{Y}$ is a **sufficient partition** for θ if, for any set $A \subset X$ and $Y \in A$, the conditional distribution of Y given $Y \in A$ does not depend on θ.

Definition 2.9.3 A statistic $t(y)$ is **sufficient** for θ if its partition is sufficient for θ,that is, if the conditional distribution of Y given $t(Y) = r$ does not depend on θ.for any r.

Interpretation: A sufficient statistic is a data summary that contains all relevant information about θ. Note that if a statistic is sufficient for θ, it is also sufficient for any function $\phi(\theta)$ of θ.

2.9.2 The Neyman Factorization Theorem

The following theorem provides a recipe for finding sufficient statistics

Theorem 2.9.1 *A necessary and sufficient condition for $t(y)$ to be sufficient for θ is that the likelihood function for θ factorizes as*

$$f_{Y|\theta}(y;\theta) = g(t(y),\theta)h(y) \tag{2.5}$$

where g is a function of t and θ and h is a function of y alone.

Proof. In the discrete case, the conditional mass function for $y.$ given $T = r$ (and θ)

$$f_{Y|\theta,T}(y;\theta,r) = \frac{f_{Y,T|\theta}(y,r;\theta)}{f_{T|\theta}(r;\theta)} = \frac{f_{Y,T|\theta}(y,r;\theta)}{\sum\limits_{y':t(y')=r} f_{Y|\theta}(y';\theta)}.$$

Then if (2.5) holds

$$f_{Y|\theta,T}(y;\theta,r) = \frac{g(r,\theta)h(y)}{\sum\limits_{y':t(y')=r} g(r,\theta)h(y')} = \frac{h(y)}{\sum\limits_{y':t(y')=r} h(y')}$$

which is independent of θ. Hence t is sufficient for θ.

Conversely, by an elementary probability result,

$$f_{Y|\theta}(y;\theta) = f_{Y|\theta,T}(y;\theta,r)f_{T|\theta}(r;\theta) = f_{Y|T}(y;r)f_{T|\theta}(r;\theta) \qquad \text{when } t(y) = r$$

then, setting

$$g(t(y),\theta) = g(r,\theta) = f_{T|\theta}(r;\theta) = \mathrm{P}\left[t(y) = r|\theta\right]$$
$$h(y) = f_{Y|\theta,T}(y;\theta,r) = f_{Y|T}(y;r) \qquad \text{which is independent of } \theta$$

■

2.9.3 Minimal Sufficiency

For any data set y, the n-vector-valued statistic that contains each individual data point is a sufficient statistic. With the objective of data reduction, a new goal is to find a sufficient statistic of fewer dimensions.

Definition 2.9.4 A partition $\mathcal{P}_1$ of $\mathbb{Y}$ is **at least as coarse** as a partition $\mathcal{P}_2$ if every set of $\mathcal{P}_2$ is a subset of some set of $\mathcal{P}_1$, that is, $\mathcal{P}_2$ partitions $\mathcal{P}_1$.

Definition 2.9.5 A partition $\mathcal{P}$ of $\mathbb{Y}$ is **minimally sufficient** for θ if it is sufficient for θ and if it is at least as course as every other sufficient partition for θ.

Definition 2.9.6 A statistic $t(y)$ is **minimally sufficient** for θ if its partition is minimally sufficient for θ.

A minimal sufficient statistic represents the greatest possible reduction in data without losing information about θ. Two points of $\mathbb{Y}$ with the same value of a minimum sufficient statistic may have different values for another sufficient statistic.

Lemma 2.9.1 Minimum sufficient statistics are unique (up to 1-1 transformations)

Proof. Suppose there exists two minimal sufficient partitions $\mathcal{P}_1$ and $\mathcal{P}_2$. As $\mathcal{P}_1$ and $\mathcal{P}_2$ are minimal sufficient, for sets $A_1, A_2 \in \mathcal{P}_1$ and $B \in \mathcal{P}_2$ then

$$A_1 \subseteq B \qquad \text{and} \qquad B \subseteq A_2 \qquad \therefore A_1 \equiv A_2 \equiv B$$

and the two partitions are identical. ■

Lemma 2.9.2 A minimal sufficient statistic is a function any other sufficient statistic

Proof. Let $\mathcal{P}_1$ be minimal sufficient, and $\mathcal{P}_2$ be sufficient, with corresponding statistics t_1 and t_2. Suppose x is a point in the range of t_2, that is, there exists a y such that $t_2(y) = x$. Now y is not unique, but $t_1(y)$ is unique, because if $t_2(y_1) = t_2(y_2) = x$ then y_1 and y_2 are in the same set of $\mathcal{P}_2$ and hence in the same set of $\mathcal{P}_1$, and therefore $t_1(y_1) = t_1(y_2)$. Hence the function

$$g(x) = t_1(y)$$

is well-defined and $t_1(y) = g(t_2(y))$. ■

Theorem 2.9.2 *For any $y, z \in \mathbb{Y}$, define the relation on $\mathbb{Y}$ by*

$$y \sim z \Leftrightarrow \frac{f_{Y|\theta}(y;\theta)}{f_{Y|\theta}(z;\theta)} \text{ does not depend on } \theta.$$

Then this relation is an equivalence relation, and its partition is sufficient for θ.

Proof. Easy to show that $\sim$ is reflexive, symmetric and transitive. To show sufficiency, consider the discrete case; let A be a set of the partition defined by $\sim$. Then

$$f_{Y|\theta,T}(y;\theta, y \in A) = \mathrm{P}\left[Y = y|\theta, y \in A\right] = \frac{f_{Y|\theta}(y;\theta)}{\sum\limits_{x\sim y} f_{Y|\theta}(x;\theta)} = \left[\frac{f_{Y|\theta}(y;\theta)}{\sum\limits_{x\sim y} f_{Y|\theta}(x;\theta)}\right]^{-1} \text{for } y \in A$$

which is independent of θ by definition of $\sim$. Hence the partition is sufficient. To show that it is at least as coarse as any other partition, let $v = v(y)$ be any other sufficient statistic. Then, by the Neyman Factorization Theorem, if $v(y) = v(z)$ for $y, z \in \mathbb{Y}$

$$\begin{aligned} f_{Y|\theta}(y;\theta) &= g^*(v(y),\theta)h^*(y) \\ &= g^*(v(z),\theta)\left\{\frac{h^*(y)}{h^*(z)}\right\}h^*(z) \\ &= f_{Y|\theta}(z;\theta)\left\{\frac{h^*(y)}{h^*(z)}\right\} \end{aligned}$$

and the final term is independent of θ. Hence

$$\frac{f_{Y|\theta}(y;\theta)}{f_{Y|\theta}(z;\theta)} = \left\{\frac{h^*(y)}{h^*(z)}\right\}$$

is independent of θ, so $y \sim z$. ■

2.9.4 Ancillarity

Definition 2.9.7 Let T be a minimal sufficient statistic for θ, and suppose $\dim T > \dim \theta$. If $T = (S, A)$ where A has a distribution not depending on θ, then A is an **ancillary** statistic.

Example 2.9.1 Let N be a discrete random variable with $p_n = \mathrm{P}[N = n]$ for $n = 1, 2, ...$ Given $N = n$, let $Y_1, ..., Y_n$ be i.i.d from a one-parameter exponential family

$$f_{Y|N,\theta}(y;n,\theta) = p_n \exp\left\{a(\theta)\sum_{i=1}^{n} b(y_i) + nc(\theta) + \sum_{i=1}^{n} d(y_i)\right\}.$$

Then the Neyman Factorization Theorem implies that

$$\left(N, \sum_{i=1}^{n} b(Y_i)\right)$$

are jointly minimal sufficient for θ, but N has a distribution that does not depend on θ.

There is an argument to support the idea that inferences about θ should be made *conditionally* on the value of any ancillary statistic.

2.9.5 Sufficiency: Examples

Example 2.9.2 If $Y_1, ..., Y_n$ is a random sample from an $Eyp(\lambda)$ distribution. Then

$$f_{Y_1,...,Y_n}(y_1, ..., y_n; \lambda) = \lambda^n eyp\left\{-\lambda\sum_{i=1}^{n} y_i\right\}$$

for $y_i > 0,\ i = 1, ..., n$. Hence the statistic $T = \sum_{i=1}^{n} Y_i$ is sufficient for λ, as

$$f_{Y_1,...,Y_n}(y_1, ..., y_n; \lambda) = l(t;\lambda)c(y_1, ..., y_n)$$

where

$$l(t;\lambda) = \lambda^n eyp\left\{-\lambda\sum_{i=1}^{n} y_i\right\} \qquad c(y_1, ..., y_n) = 1$$

Example 2.9.3 If $Y_1, ..., Y_n$ is a random sample from an $N(\mu, \sigma^2)$ distribution. Then

$$\begin{aligned} f_{Y_1,..,Y_n}(y_1, .., y_n; \mu, \sigma^2) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} eyp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} eyp\left\{-\frac{1}{2\sigma^2}(t_2 - 2\mu t_1 + n\mu^2)\right\} \end{aligned}$$

where $t_1 = \sum_{i=1}^{n} y_i$, $t_2 = \sum_{i=1}^{n} y_i^2$. Hence the statistics

$$\mathbf{T} = \left(\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i^2\right)$$

are jointly sufficient for (μ, σ^2).

Example 2.9.4 If $Y_1, ..., Y_n$ is a random sample from an $Uniform(0, \theta)$ distribution. Then

$$\begin{aligned} f_{Y_1,..,Y_n}(y_1, .., y_n; \theta) &= \frac{1}{\theta^n} \qquad 0 < y_i < \theta \ , i = 1, ..., n \\ &= l(t; \theta)c(y_1, ..., y_n) \end{aligned}$$

where $l(t; \theta) = 1/\theta^n$ if $t < \theta$, $c(y_1, ..., y_n) = 1$, and where $t = \max\{y_1, ..., y_n\}$. Hence the statistic $T = \max\{Y_1, ..., Y_n\}$ (the largest order statistic derived from $Y_1, .., Y_n$) is sufficient for θ.

2.10 Minimum Variance Unbiased Estimation

2.10.1 The Rao-Blackwell Theorem

As identified in previous sections, an objective is to identify MVUEs; the Cramer-Rao lower bound identifies some MVUEs in regular problems, but a more general method is required. Below, it is shown that sufficient statistics can be used to reduce the variance of unbiased estimators.

Lemma 2.10.1 The Law of Iterated Expectation
If X and Y are random variables, then

$$\mathrm{E}_X[X] = \mathrm{E}_Y\left[\mathrm{E}_{X|Y}[X|Y = y]\right].$$

Proof. In the discrete case

$$E_X[X] = \sum_x x f_X(x) = \sum_x x\left\{\sum_y f_{X|Y}(x|y) f_Y(y)\right\} = \sum_y \left\{\sum_x x f_{X|Y}(x|y)\right\} f_Y(y) = E_Y\left[E_{X|Y}[X|Y = y]\right]$$

■

Lemma 2.10.2 If X and Y are random variables, then

$$\mathrm{Var}_X[X] = \mathrm{E}_Y\left[\mathrm{Var}_{X|Y}[X|Y = y]\right] + \mathrm{Var}_Y\left[\mathrm{E}_{X|Y}[X|Y = y]\right]. \tag{2.6}$$

Proof. By definition, the terms on the right-hand side may be expressed

$$\begin{aligned}\mathrm{E}_Y\left[\mathrm{Var}_{X|Y}\left[X|Y=y\right]\right] &= \mathrm{E}_Y\left[\mathrm{E}_{X|Y}\left[X^2|Y=y\right]-\left\{\mathrm{E}_{X|Y}\left[X|Y=y\right]\right\}^2\right]\\ &= \mathrm{E}_X\left[X^2\right]-\mathrm{E}_Y\left[\left\{\mathrm{E}_{X|Y}\left[X|Y=y\right]\right\}^2\right]\\ \mathrm{Var}_Y\left[\mathrm{E}_{X|Y}\left[X|Y=y\right]\right] &= \mathrm{E}_Y\left[\left\{\mathrm{E}_{X|Y}\left[X|Y=y\right]\right\}^2\right]-\left\{\mathrm{E}_Y\left[\mathrm{E}_{X|Y}\left[X|Y=y\right]\right]\right\}^2\\ &= \mathrm{E}_Y\left[\left\{\mathrm{E}_{X|Y}\left[X|Y=y\right]\right\}^2\right]-\left\{E_X\left[X\right]\right\}^2\end{aligned}$$

Hence

$$\mathrm{E}_Y\left[\mathrm{Var}_{X|Y}\left[X|Y=y\right]\right]+\mathrm{Var}_Y\left[\mathrm{E}_{X|Y}\left[X|Y=y\right]\right]=\mathrm{E}_X\left[X^2\right]-\left\{E_X\left[X\right]\right\}^2=\mathrm{Var}_X\left[X\right]$$

■

Theorem 2.10.1 ***The Rao-Blackwell Theorem***
Let $\widetilde{g}(Y)$ be any unbiased estimator of $g(\theta)$. If T is sufficient for θ then

$$\widehat{g}(Y)=E_{Y|\theta}\left[\widetilde{g}(Y)|T=t\right]$$

is also unbiased for θ and

$$Var_{Y|\theta}\left[\widehat{g}\left(T\right)\right]\leq Var_{Y|\theta}\left[\widetilde{g}\left(T\right)\right]$$

Proof. First, T is sufficient for θ so the distribution of Y (and in part the distribution of $\widetilde{g}(Y)$) does not depend on θ (but only on t). Hence $\widehat{g}\left(t\right)$ is purely a function of t and not θ. Thus

$$\mathrm{E}_{Y|\theta}\left[\widehat{g}\left(T\right)\right]=\mathrm{E}_{T|\theta}\left[\mathrm{E}_{Y|T}\left[\widetilde{g}(Y)|T=t\right]\right]=\mathrm{E}_{Y|\theta}\left[\widetilde{g}(Y)\right]=\theta.$$

for all θ, so $\widehat{g}\left(T\right)$ is unbiased for θ. Further

$$\begin{aligned}\mathrm{Var}_{Y|\theta}\left[\widetilde{g}\left(T\right)\right] &= \mathrm{E}_{Y|\theta}\left[\mathrm{Var}_{Y|T}\left[\widetilde{g}(Y)|T=t\right]\right]+\mathrm{Var}_{Y|\theta}\left[\mathrm{E}_{Y|T}\left[\widetilde{g}(Y)|T=t\right]\right]\\ &= \mathrm{E}_{Y|\theta}\left[\mathrm{Var}_{Y|T}\left[\widetilde{g}(Y)|T=t\right]\right]+\mathrm{Var}_{Y|\theta}\left[\widehat{g}(T)\right]\\ &\geq \mathrm{Var}_{Y|\theta}\left[\widehat{g}(T)\right].\end{aligned}$$

■

2.10.2 Completeness

Lemma 2.10.3 If $a(T)$ is a function of a sufficient statistic T, $a(T)$ is unbiased for $g(\theta)$, and $a(T)$ is the only such function of T, then $a(T)$ is MVUE for $g(\theta)$.

Proof. Let $\widetilde{g}(Y)$ be any other unbiased estimator of $g(\theta)$. Let $\widehat{g}(t)=\mathrm{E}_{Y|\theta}\left[\widetilde{g}(Y)|T=t\right]$. Then by the Rao-Blackwell Theorem, $\widehat{g}(T)$ is unbiased for $g(\theta)$ and $\widehat{g}(T)$ is a function of T. Hence let $a(T)=\widehat{g}(T)$ and

$$\mathrm{Var}_{Y|\theta}\left[a\left(T\right)\right]=\mathrm{Var}_{Y|\theta}\left[\widehat{g}\left(T\right)\right]\leq\mathrm{Var}_{Y|\theta}\left[\widetilde{g}\left(T\right)\right]$$

■

Now, consider any statistic $t(Y)$. For each $\theta\in\Theta$ there is probability distribution that is the (sampling) distribution of $T=t(Y)$ that varies as θ varies. This fact motivates the following definition

Definition 2.10.1 The family of distributions of a statistic T is called **complete** if

$$\mathrm{E}_{Y|\theta}\left[h(T)\right]=0,\forall\theta\in\Theta\Rightarrow h(t)=0,\forall t$$

for real function h.
The following result demonstrates a example of when a complete family can be found:

Example 2.10.1 The k-parameter exponential family
Let $\theta = (\theta_1, ..., \theta_q)$, and

$$f_{Y|\theta}(y;\theta) = \exp\left\{\sum_{j=1}^{k} a_j(\theta)b_j(y) + c(\theta) + d(y)\right\}$$

For a random sample $y^T = (y_1, ..., y_n)$

$$f_{Y|\theta}(y;\theta) = \exp\left\{\sum_{j=1}^{k} a_j(\theta)t_j(y) + nc(\theta) + \sum_{i=1}^{n} d(y)\right\} \qquad \text{where } t_j(y) = \sum_{i=1}^{n} b_j(y_i)$$

Consider $T_j = t_j(Y)$; then $T = (T_1, ..., T_k)$ is minimal sufficient for θ, and if $q = \dim\theta$, $k = \dim T$, it can be shown that

$$q = k \Leftrightarrow T \text{ is complete.}$$

Theorem 2.10.2 *Let T be a sufficient statistic whose family of distributions is complete. Then any function of T is an MVUE of its expectation.*

Proof. Let $a(T)$ be a function of T and suppose $\mathrm{E}_{T|\theta}[a(T)] = g(\theta)$. It is required to prove that $a(T)$ is the unique function of T which is unbiased for $g(\theta)$. Let $b(T)$ be any other such function of T. Then

$$E_{T|\theta}[a(T)] - E_{T|\theta}[b(T)] = g(\theta) - g(\theta) = 0$$

for all θ. Then $a(T) - b(T) = 0$ by completeness, and $a(T) = b(T)$. ■

Note. Any MVUE be minimum variance whatever the value of θ, that is, uniformly in θ. Hence the MVUE is sometimes referred to as the Uniform MVUE, or UMVUE. ■

2.11 Methods of Estimation

Recall the **likelihood** construct: let $Y_1, ..., Y_n$ be a random sample from a distribution with mass/density function f_Y that depends on (possibly vector) parameter θ, that is, $f_{Y_1}(y_1) = f_Y(y_1;\theta)$, so that

$$f_{Y_1,...,Y_k}(y_1, ..., y_k) = \prod_{i=1}^{k} f_Y(y_i;\theta)$$

then a statistic $T = t(Y_1, ..., Y_n)$ that is used to represent or **estimate** a function $\tau(\theta)$ of θ based on an **observed** sample of the random variables $y_1, ..., y_n$ is an **estimator**, and $t = t(y_1, ..., y_n)$ is an **estimate**, $\hat{\tau}(\theta)$, of $\tau(\theta)$.

Definition 2.11.1 METHOD OF MOMENTS ESTIMATION: Suppose that $Y_1, ..., Y_n$ is a random sample from a probability distribution with mass/density function f_Y that depends on vector parameter θ of dimension k, and suppose that a sample $y_1, ..., y_n$ has been observed. Let the jth moment of f_Y be denoted $\mu_j^{'}$, and let the jth **sample moment**, denoted $m_j^{'}$ be defined for $j = 1, ..., k$ by

$$m_j^{'} = \frac{1}{n}\sum_{i=1}^{n} y_i^j$$

Then $m_j^{'}$ is an **estimate** of $\mu_j^{'}$, and

$$M_j^{'} = \frac{1}{n}\sum_{i=1}^{n} Y_i^j$$

is an **estimator** of $\mu_j^{'}$.

Interpretation : The method of estimation involves matching the **theoretical** moments to the **sample** moments, giving (in most cases) k equations in the k elements of vector θ which may be solved simultaneously to find the parameter estimates. Intuitively, and recalling the Weak Law of Large Numbers, it is reasonable to suppose that there is a close relationship between the theoretical properties of a probability distribution, and large sample derived estimates; for example, we know that, for large n, the sample mean converges in probability to the theoretical expectation.

2.11.1 Maximum likelihood estimators

Recall the definition of the likelihood function from 2.2.3;

$$l(\theta; y) = f_{Y|\theta}(y; \theta)$$

viewed as a function of θ for constant y.

Definition 2.11.2 The **maximum likelihood estimate** or **M.L.E.** of parameter θ, denoted $\widehat{\theta}_{MLE}$, is defined by

$$\widehat{\theta}_{MLE} = \arg\sup_{\theta\in\Theta} l(\theta; y)$$

that is, the ordinate at which the supremum of the likelihood function is reached. The M.L.E..is typically a function of the data, and the corresponding estimator is termed the **maximum likelihood estimator.**

(i) in general the m.l.e. is not unique

(ii) the m.l.e. need not exist.

Suppose a sample $y_1, ..., y_n$ has been obtained from a probability model specified by mass or density function $f_Y(y; \theta)$ depending on parameter(s) θ lying in parameter space Θ. The maximum likelihood estimate (MLE) is produced as follows;

STEP 1 Write down the **likelihood function**, $L(\theta)$.
STEP 2 Take the natural log of the likelihood, 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. Note that, if parameter space Θ is a bounded interval, then the maximum likelihood estimate may lie on the boundary of Θ. (If the parameter is a k vector, the maximization involves evaluation of partial derivatives).
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. In the single parameter case, if the second derivative of the log -likelihood is negative 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 $\hat{\theta}$).

2.11.2 Properties of M.L.E.s

Some elementary properties of MLEs can be deduced using the results from previous sections:

1. MLEs are not necessarily unbiased

2. MLEs are functionally **invariant**, that is, if $\phi = g(\theta)$ is a 1-1 function of θ,then

$$g\left(\widehat{\theta}_{MLE}\right) \text{ is the MLE of } \phi$$

as

$$l(\phi; y) = f_{Y|\phi}(y; \phi) = f_{Y|\theta}\left(y; g^{-1}(\theta)\right)$$
$$\Longrightarrow \widehat{\phi}_{MLE} = \arg\sup_{\phi\in\Phi} l(\phi; y) = \arg\sup_{\phi\in\Phi} f_{Y|\phi}(y; \phi) = \arg\sup_{\theta\in\Theta} f_{Y|\phi}(y; g^{-1}(\theta)) = g\left(\widehat{\theta}_{MLE}\right)$$

3. In a one parameter regular problem, if there exists an unbiased estimator $t(Y)$ of θ whose variance attains the Cramer-Rao Lower bound, then $t(Y) = \widehat{\theta}_{MLE}$. To see this consider the efficient score U_θ:

$$U_\theta = \frac{\partial}{\partial\theta} \log f_{Y|\theta}(y; \theta) = g(\theta)(t(Y) - \theta).$$

Finding the supremum of the likelihood in a regular problem is equivalent to maximizing the log of the likelihood function, and this is achieved by setting the first partial derivative equal to zero. Hence setting $U_\theta = 0 \Longrightarrow (t(Y) - \theta) = 0 \Longrightarrow \widehat{\theta}_{MLE} = t(Y)$, noting that

$$I(\theta) = -\mathrm{E}_{Y|\theta}\left[\frac{\partial^2}{\partial\theta^2} \log f_{Y|\theta}(Y; \theta)\right] > 0 \Longrightarrow \frac{\partial^2}{\partial\theta^2} \log f_{Y|\theta}(Y; \theta) < 0 \text{ for all } \theta$$

confirms a maximum.

4. An M.L.E. is a function of a minimum sufficient statistic, by the Neyman Factorization

$$l(\theta; y) = f_{Y|\theta}(y; \theta) = g(t(y), \theta)h(y)$$

where t is a minimal sufficient statistic.

2.11.3 Large sample properties of MLEs

The previous section lists some properties of MLEs that hold for finite sample size n. Asymptotic or large sample properties of the estimators can also be derived. Consider the estimation of single parameter θ in a regular problem, from sample of size n. Suppose that the true (but unobserved) value of θ is denoted θ_0, and let $\widehat{\theta}_n$ denote the M.L.E. derived from the sample.

Recall first that an estimator $\widehat{\theta}_n$ of θ is termed **consistent** for θ_0 if

$$\lim_{n\to\infty} \mathrm{P}\left[\left|\widehat{\theta}_n - \theta_0\right| < \epsilon\right] = 1$$

that is, $\widehat{\theta}_n$ converges in probability (or converges weakly) to θ_0

Theorem 2.11.1 ***The Weak Law of Large Numbers***
Suppose that $Y_1, ..., Y_n$ are i.i.d with finite expectation μ and variance σ^2. Then the sample mean random variable

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

converges in probability to μ.

Proof. By the Chebychev Inequality, for any random variable X with finite expectation μ and variance σ^2, and any $k > 0$,

$$\mathrm{P}[|X - \mu| < k] \geq 1 - \frac{\sigma^2}{k^2}.$$

Setting X equal to $\overline{Y}_n$, $k = \epsilon$, and noting that the expectation of $\overline{Y}_n$ is μ and the variance is σ^2/n,

$$\mathrm{P}[\left|\overline{Y}_n - \mu\right| < \epsilon] \geq 1 - \frac{\sigma^2}{n^2\epsilon^2} \to 1 \text{ as } n \to \infty.$$

■

Definition 2.11.3 A function g is **convex** if

$$\frac{d^2}{dt^2}\left\{g(t)\right\}_{t=x} = g''(x) > 0 \text{ for all } x$$

Lemma 2.11.1 Jensen's Inequality
Suppose that X is a random variable, and function g is convex. Then

$$\mathrm{E}_X[g(X)] \geq g(\mathrm{E}_X[X])$$

Proof.A Taylor expansion of $g(x)$ around $x = \mu$ gives

$$g(x) = g(\mu) + (x - \mu)g'(\mu) + \frac{1}{2}(x - \mu)^2 g''(x_0)$$

for some x_0 such that $x < x_0 < \mu$. Thus, taking expectations,

$$\begin{aligned}\mathrm{E}_{f_X}[g(X)] &= g(\mu) + \mathrm{E}_X[(X - \mu)]g'(\mu) + \frac{1}{2}\mathrm{E}_X[(X - \mu)^2]g''(x_0) \\ &\geq g(\mu) = g(\mathrm{E}_X[X])\end{aligned}$$

as $\mathrm{E}_{f_X}[(X - \mu)] = 0$, and $\mathrm{E}_{f_X}[(X - \mu)^2], g''(x_0) \geq 0$. ■

Theorem 2.11.2 *The maximum likelihood estimate $\widehat{\theta}_n$ is a consistent estimator of θ_0.*

Proof. For finite n

$$f_{Y|\theta}(y;\theta) = \prod_{i=1}^{n} f_{Y_i|\theta}(y_i;\theta) \Longrightarrow \log f_{Y|\theta}(Y;\theta) = \sum_{i=1}^{n} \log f_{Y_i|\theta}(Y_i;\theta).$$

Now the terms $\log f_{Y_i|\theta}(Y_i;\theta)$ are themselves i.i.d random variables, and so, by the Weak Law of Large Numbers,

$$\frac{1}{n}\log f_{Y|\theta}(Y;\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f_{Y_i|\theta}(Y_i;\theta)$$

converges in probability to

$$\mathrm{E}_{Y_0|\theta_0}\left[\log f_{Y_0|\theta}(Y_0;\theta)\right] = z(\theta)$$

say, where Y_0 has the same distribution as $Y_1, ..., Y_n$ and the expectation is with respect to the true conditional density of Y_0 given θ_0. Now

$$\mathrm{E}_{Y_0|\theta_0}\left[-\log\frac{f_{Y_0|\theta}(Y_0;\theta)}{f_{Y_0|\theta_0}(Y_0;\theta_0)}\right] \geq -\log \mathrm{E}_{Y_0|\theta_0}\left[\frac{f_{Y_0|\theta}(Y_0;\theta)}{f_{Y_0|\theta_0}(Y_0;\theta_0)}\right]$$

by Jensen's Inequality. But the right-hand term simplifies

$$\begin{aligned} -\log \mathrm{E}_{Y_0|\theta_0}\left[\frac{f_{Y_0|\theta}(Y_0;\theta)}{f_{Y_0|\theta_0}(Y_0;\theta_0)}\right] &= -\log\left\{\int \frac{f_{Y_0|\theta}(Y_0;\theta)}{f_{Y_0|\theta_0}(Y_0;\theta_0)} f_{Y_0|\theta_0}(Y_0;\theta_0)\, dY_0\right\} \\ &= -\log\left\{\int f_{Y_0|\theta}(Y_0;\theta)\, dY_0\right\} \\ &= -\log 1 = 0 \end{aligned}$$

So therefore

$$z(\theta) \leq z(\theta_0)$$

for all θ, and hence θ_0 maximizes $z(\theta)$ Hence as $\widehat{\theta}_n$ maximizes $\frac{1}{n}\log f_{Y|\theta}(Y;\theta)$ by definition, it follows that $\widehat{\theta}_n$ converges in probability to θ; Let $L(\theta;Y) = \log f_Y(Y;\theta)$. Note that

$$\begin{aligned} z(\theta_0) - z(\widehat{\theta}_n) &= \left[z(\theta_0) - \frac{1}{n}L(\theta_0;Y)\right] + \left[\frac{1}{n}L(\theta_0;Y) - \frac{1}{n}L(\widehat{\theta}_n;Y)\right] + \left[\frac{1}{n}L(\widehat{\theta}_n;Y) - z(\widehat{\theta}_n)\right] \\ &\leq \left[z(\theta_0) - \frac{1}{n}L(\theta_0;Y)\right] + \left[\frac{1}{n}L(\widehat{\theta}_n;Y) - z(\widehat{\theta}_n)\right] \qquad \text{(as second term is negative)} \end{aligned}$$

and as, for **any** θ and $\epsilon_1 > 0$

$$\mathrm{P}\left[\left|\frac{1}{n}L(\theta;Y) - z(\theta)\right| < \epsilon_1\right] \to 1$$

so that

$$\begin{aligned} \mathrm{P}\left[\left|z(\theta_0) - z(\widehat{\theta}_n)\right| < 2\epsilon_1\right] &\geq \mathrm{P}\left[\left|z(\theta_0) - \frac{1}{n}L(\theta_0;Y)\right| + \left|z(\widehat{\theta}_n) - \frac{1}{n}L(\widehat{\theta}_n;Y)\right| < 2\epsilon_1\right] \\ &\geq \mathrm{P}\left[\left|z(\theta_0) - \frac{1}{n}L(\theta_0;Y)\right| < \epsilon_1 \text{ and } \left|z(\widehat{\theta}_n) - \frac{1}{n}L(\widehat{\theta}_n;Y)\right| < \epsilon_1\right] \\ &\to 1 \end{aligned}$$

and the result follows (assuming z is continuous in θ). ■

Theorem 2.11.3 *In a regular problem, $\widehat{\theta}_n$ is asymptotically normal with mean θ_0 and variance attaining the Cramer-Rao lower bound.*

Proof. By definition $\widehat{\theta}_n$ satisfies $L'(\theta, y) = 0$ where

$$L'(\theta, y) = \frac{\partial}{\partial\theta} L(\theta; y) = \frac{\partial}{\partial\theta}\left\{\sum_{i=1}^{n} \log f_{Y|\theta}(y_i;\theta)\right\}. = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\left\{\log f_{Y|\theta}(y_i;\theta)\right\}$$

Using a Taylor expansion about θ_0,for large n,

$$L'(\widehat{\theta}_n, y) \approx L'(\theta_0, y) + \left(\widehat{\theta}_n - \theta_0\right) L''(\theta_0, y) = 0$$

so therefore for large n

$$\widehat{\theta}_n - \theta_0 \approx -\frac{L'(\theta_0, y)}{L''(\theta_0, y)}.$$

Now the numerator is

$$L'(\theta_0, y) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \left\{\log f_{Y|\theta}(y_i;\theta)\right\}_{\theta=\theta_0}$$

that is, a sum of i.i.d variables with finite expectation

$$\mathrm{E}_{Y|\theta_0}\left[L'(\theta_0, Y)\right] = \mathrm{E}_{Y|\theta_0}\left[\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \left\{\log f_{Y|\theta}(Y_i;\theta)\right\}_{\theta=\theta_0}\right] = \sum_{i=1}^{n} \mathrm{E}_{Y|\theta_0}\left[\frac{\partial}{\partial\theta} \left\{\log f_{Y|\theta}(y_i;\theta)\right\}_{\theta=\theta_0}\right] = 0$$

(as each term is the efficient score) and variance

$$\mathrm{Var}_{Y|\theta_0}\left[L'(\theta_0, Y)\right] = \mathrm{E}_{Y|\theta_0}\left[\left\{\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \left\{\log f_{Y|\theta}(Y_i;\theta)\right\}_{\theta=\theta_0}\right\}^2\right] = I(\theta_0)$$

(the Fisher Information, by definition). Hence, by the Central Limit Theorem,

$$L'(\theta_0, y) \sim N\left(0, I\left(\theta_0\right)\right).$$

For the denominator,

$$\mathrm{E}_{Y|\theta_0}\left[L''(\theta_0, Y)\right] = -I(\theta_0)$$

and so, by the Weak Law of Large Numbers, $L''(\theta_0, Y) \approx I(\theta_0)$ as $n \to \infty$. Therefore $\widehat{\theta}_n - \theta_0$ has expectation zero and variance

$$\frac{I(\theta_0)}{\left\{I(\theta_0)\right\}^2} = \frac{1}{I(\theta_0)}$$

which is the Cramer-Rao Lower Bound ■

Definition 2.11.4 If T is an unbiased estimator for θ, then the **efficiency** of T is defined as

$$\frac{\text{CR Bound}}{\mathrm{Var}_{T|\theta}\left[T\right]}$$

If the bound is achieved, then the efficiency is 1 (full efficiency), otherwise it is less than 1. Clearly, the efficiency depends on θ. Note that MLEs are fully efficient.

Summary: Large sample Properties of MLEs

It can be shown that, under certain regularity conditions, maximum likelihood estimators have desirable properties. In particular if $\hat{\theta}_n$ is the maximum likelihood estimator of parameter θ derived from a random sample of size n, then

(i) $\hat{\theta}_n$ exists and is unique

(ii) $\hat{\theta}_n$ is consistent

(iii) $\hat{\theta}_n$ is asymptotically normally distributed, that is,

$$\hat{\theta}_n \xrightarrow{d} N\left(\theta, \frac{1}{n\mathrm{E}_{f_Y}\left[\left\{\partial(Y;\theta)\right\}^2\right]}\right)$$

where

$$\partial(y;\theta) = \frac{\partial}{\partial\theta}\left\{\log f_Y(y;\theta)\right\}$$

that is, $\hat{\theta}_n$ is asymptotically unbiased, and has asymptotic variance equal to the Cramer-Rao lower bound

(iv) $\hat{\theta}_n$ is fully efficient

2.11.4 MLEs and Sufficiency

Theorem 2.11.4 $\mathbf{T} = (T_1, ..., T_k)$ *are sufficient for* θ*, and* $\hat{\theta}$ *is a unique maximum likelihood estimator of* θ*, then* $\hat{\theta}$ *is a function of* $(T_1, ..., T_k)$*.*

Proof. Follows directly from the factorization theorem above ■

2.12 Hypothesis Testing

2.12.1 Introduction

The philosophy of hypothesis testing is as follows; rather than merely producing an estimate of parameter θ, the objective is answer specific questions about the parameter. For example, it may be of interest to enquire

"Is the data consistent with some stated hypothesis ?"

A statistical hypothesis test partitions the sample space $\mathbb{Y}$ into two regions; the region which corresponds to evidence to reject the hypothesis being tested, and its complement. In what follows, the assumption will be made that there are two alternative (complementary) "states of the world" that are encapsulated in two stated hypotheses. The objective is carry out a quantitative statistical analysis to discover the evidence in support of each hypothesis in turn.

2.12.2 Classical Neyman-Pearson Theory

Definition 2.12.1 The **null hypothesis** is denoted H_0; it corresponds to one (usually the simpler or more parsimonious) of the two representations.

Definition 2.12.2 The **alternative hypothesis** is denoted H_1 (or H_A); usually it is the negation of the null hypothesis.

The null and alternative hypotheses define a complementary subsets that partition Θ; say that $\mathcal{W} \subset \Theta$ is the subset corresponding to H_0.

Definition 2.12.3 A **test** of H_0 versus H_1 defines a partition of sample space $\mathbb{Y}$ into two regions. The hypothesis H_0 is **rejected in favour of** H_1 in the test depending on where the data y (or a suitably chosen statistic based on y) fall within $\mathbb{Y}$, that is, in one region or the other.

Definition 2.12.4 A **test statistic** is the function of data y used in a statistical hypothesis test.

Definition 2.12.5 The **critical region,** R**,** is the region within which the data y (or a suitably chosen statistic based on y) must lie in order for hypothesis H_0 to be rejected in favour of H_1. The complement of R will be written $\overline{R}$.

Statistical hypothesis testing centres on finding suitable critical regions. Now, clearly the data themselves are random, and thus the test statistic used in the test is itself a realization of some random variable. The choice of test and critical region will be made by specifying certain properties of the distribution of the test statistic. Generally, the objective is to control the probability that the "correct" hypothesis is incorrectly rejected.

Definition 2.12.6 A **Type I error** occurs when the null hypothesis H_0 is REJECTED when it is in fact TRUE.
A **Type II error** occurs when the null hypothesis H_0 is ACCEPTED when it is in fact FALSE.

Definition 2.12.7 For a particular test, let $R \subset \mathbb{Y}$ be the critical region. If $\theta \in \mathcal{W}$, define the **Type I error probability** $\alpha(\theta)$ by

$$\alpha(\theta) = \mathrm{P}\left[Y \in R | \theta \in \mathcal{W}\right] \tag{2.7}$$

If $\mathcal{W}$ comprises a single value, then $\alpha =$P$[Y \in R | H_0$ is TRUE$]$
For $\theta \in \Theta \backslash \mathcal{W}$, define the **Type II error probability** $\beta(\theta)$ by

$$\beta(\theta) = \mathrm{P}\left[Y \notin R | \theta \in \Theta \backslash \mathcal{W}\right] \tag{2.8}$$

and the **power function** by

$$1 - \beta(\theta) = \mathrm{P}\left[Y \in R | \theta \in \Theta \backslash \mathcal{W}\right]$$

Intuitively, the power function should be high, as it represents the probability of correctly rejecting H_0.

Definition 2.12.8 If $\alpha(\theta) \leq \gamma < 1$ for all $\theta \in \mathcal{W}$. and if $y \in R$, then H_0 is **rejected at significance level** γ.

Note. The value γ is any upper bound of $\alpha(\theta)$, so if H_0 is rejected at level γ then (by definition) it is rejected at level $\gamma + \delta$ for $\delta > 0$ ■

Definition 2.12.9 The **size** of a statistical test is

$$\sup_{\theta \in \mathcal{W}} \alpha(\theta)$$

which is equal to α if $\mathcal{W}$ comprises a single value.

The task remaining is to construct the critical region R; this is achieved by specifying the test statistic and the probabilities described in the previous definitions,, and identifying the subset of $\mathbb{Y}$ for which the target probabilities are met, via (2.7) and (2.8). It would be ideal to be able to specify $\alpha(\theta)$ and $\beta(\theta)$ small, uniformly in θ. These objectives cannot be achieved simultaneously, and hence a compromise is sought. Now H_0 and H_1 are treated asymmetrically; in reality the testing context relates to finding sufficient evidence to reject H_0, rather than deciding between the two hypotheses. Hence rejecting H_0 when it is true (a Type I error) is a more serious error than accepting H_0 when it is false (a Types II error), and this implies that the size of the test, γ, should be fixed at some small level (0.05 or 0.01 say), and subject to this constraint, the test constructed that minimizes $\beta(\theta)$ uniformly in θ. That is, set $\alpha(\theta) \leq \gamma$, for all $\theta \in \mathcal{W}$, and then attempt to find a test to minimize $\beta(\theta)$ for all θ.

Definition 2.12.10 If a test is of size γ and if $\beta(\theta)$ uniformly in θ minimizes the probability of a Type II error among all tests with size $\leq \gamma$, then the test is termed the **uniformly most powerful (UMP)** test of size γ.

Definition 2.12.11 A **simple hypothesis** is one which specifies the distribution of the data completely.

Example 2.12.1 Consider the parameter space $\Theta = \{\theta_0, \theta_1\}$ where $\mathcal{W} = \{\theta_0\}$ and $\Theta \backslash \mathcal{W} = \{\theta_1\}$, and

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

These are two simple hypotheses, and

$$\beta(\theta) = \beta(\theta_1) = \beta, \text{ say}$$

The next result constructs the UMP test for this example as based on a test statistic which is a function of the ratio

$$\frac{f_{Y|\theta_1}(y;\theta_1)}{f_{Y|\theta_0}(y;\theta_0)}$$

Theorem 2.12.1 *The Neyman-Pearson Lemma*
Suppose that $\Theta = \{\theta_0, \theta_1\}$ *and a test of the hypotheses*

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

is required. Suppose that, as a requirement, the size $\leq \gamma$. *Then the most powerful test of* H_0 *against* H_1 *is defined any the critical region*

$$R = \left\{ y : \frac{f_{Y|\theta_1}(y;\theta_1)}{f_{Y|\theta_0}(y;\theta_0)} \geq k \right\}$$

where $k > 0$ *is defined by*

$$\alpha(\theta_0) = P[Y \in R|\theta_0] = \gamma$$

Proof. Let $\widetilde{R}$ be the critical region of any other test of size $\leq \gamma$ and power $1 - \widetilde{\beta}(\theta_1)$, and let $1 - \beta(\theta_1)$ be the power of the test defined by R. Then

$$\begin{aligned}
[1 - \beta(\theta_1)] - \left[1 - \widetilde{\beta}(\theta_1)\right] &= \mathrm{P}\left[Y \in R|\theta_1\right] - \mathrm{P}\left[Y \in \widetilde{R}|\theta_1\right] \\
&= \int_R f_{Y|\theta_1}(y;\theta_1)dy - \int_{\bar{R}} f_{Y|\theta_1}(y;\theta_1)dy \\
&= \int_{R\cap\overline{\widetilde{R}}} f_{Y|\theta_1}(y;\theta_1)dy - \int_{\bar{R}\cap\overline{R}} f_{Y|\theta_1}(y;\theta_1)dy.
\end{aligned}$$

Now,

$$\text{on } R\cap\overline{\widetilde{R}}: \; f_{Y|\theta_1}(y;\theta_1) \geq k f_{Y|\theta_0}(y;\theta_0) \Longrightarrow \int_{R\cap\overline{\widetilde{R}}} f_{Y|\theta_1}(y;\theta_1)dy \geq k \int_{R\cap\overline{\widetilde{R}}} f_{Y|\theta_0}(y;\theta_0)dy$$

and similarly

$$\text{on } \widetilde{R}\cap\overline{R}: \; f_{Y|\theta_1}(y;\theta_1) < k f_{Y|\theta_0}(y;\theta_0) \Longrightarrow \int_{\bar{R}\cap\overline{R}} f_{Y|\theta_1}(y;\theta_1)dy < k \int_{\bar{R}\cap\overline{R}} f_{Y|\theta_0}(y;\theta_0)dy$$

so

$$\begin{aligned}
\mathrm{P}\left[Y \in R|\theta_1\right] - \mathrm{P}\left[Y \in \widetilde{R}|\theta_1\right] &\geq k\left\{\int_{R\cap\overline{\widetilde{R}}} f_{Y|\theta_1}(y;\theta_1)dy - \int_{\bar{R}\cap\overline{R}} f_{Y|\theta_1}(y;\theta_1)dy\right\} \\
&= k\left\{\mathrm{P}\left[Y \in R|\theta_0\right] - \mathrm{P}\left[Y \in \widetilde{R}|\theta_0\right]\right\} \\
&\geq 0
\end{aligned}$$

as $k > 0$,$\mathrm{P}[Y \in R|\theta_0] = \gamma$ and $\mathrm{P}\left[Y \in \widetilde{R}|\theta_0\right] \leq \gamma$. Hence the power of the test with critical region R is not less than the power of the test with critical region $\widetilde{R}$. ■

To evaluate the value of constant k that appears in the Theorem, need to compute $\mathrm{P}[t(Y) \in R|\theta_0]$ for a fixed size γ. To do this, the distribution of $t(Y)$ given that $\theta = \theta_0$ must be used.

2.12.3 Composite Hypotheses.

Often the null and alternative hypotheses do not specify the distribution of the data completely. For example, the specification

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

could be of interest. If, in general, a UMP test of size γ is to be found, then its power must equal the power of the most powerful test of

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

for all $\theta_1 \in \Theta \backslash \theta_0$.

Note. It is possible that, for given alternative hypotheses, no UMP test exists ■

Note. For discrete data, it may not be possible to solve the equation

$$P\left[Y \in R | \theta_0\right] = \alpha$$

for every value of α, and hence only specific values of α may be attained. ■

2.12.4 Local Power

UMP tests do not always exist, so it is convenient to also consider locally most powerful tests. Consider the hypothesis

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

and a partial alternative $\theta_1 = \theta_0 + \delta$ for δ small. Then

$$\log \frac{f_{Y|\theta_1}(y;\theta_1)}{f_{Y|\theta_0}(y;\theta_0)} = \log \frac{f_{Y|\theta_0+\delta}(y;\theta_0+\delta)}{f_{Y|\theta_0}(y;\theta_0)} \approx \delta \frac{\partial}{\partial \theta} \log f_{Y|\theta_0}(y;\theta_0) = \delta U_{\theta_0}$$

For these alternatives H_0 would be rejected if U_{θ_0} is too large; this test is locally most powerful in the sense that it maximizes the slope of the power function at $\theta = \theta_0$. To complete the calculation, the distribution of U_{θ_0} is needed. Using the Central Limit Theorem, it follows that

$$U_{\theta_0} \sim N(0, I(\theta)).$$

2.12.5 Composite Null Hypotheses

To this point all the significance tests have had simple null hypothesis. Now suppose

$$\theta = (\psi, \lambda)$$

where ψ is the parameter of interest, and λ is a nuisance parameter. Consider testing

$$\begin{aligned} H_0 &: \varphi = \varphi_0 \\ H_1 &: \varphi > \varphi_0 \end{aligned}$$

with λ arbitrary.

Definition 2.12.12 Let $0 < \alpha < 1$. Then a **similar test** of size α is one with critical region R such that

$$\mathrm{P}\left[Y \in R|\varphi_0, \lambda\right] = \alpha \text{ for all } \lambda.$$

In this case, R is called a **similar region** of size α.

The objective is to look for a UMP similar test, that is a test which is UMP in a class of similar tests. When H_0 holds, $\varphi = \varphi_0$. Let S_λ be a sufficient statistic for λ when $\varphi = \varphi_0$ (in general S_λ is a function of φ_0) The conditional distribution of Y given $S_\lambda = s$, does not depend on λ when H_0 holds.

Definition 2.12.13 If $R^* \subseteq \mathbb{Y}$, is such that

$$\mathrm{P}\left[Y \in R^*|S_\lambda = s; \varphi = \varphi_0, \lambda\right] = \gamma \text{ for all } \lambda \text{ and } s$$

then R^* is called a **region of Neyman structure**.

Lemma 2.12.1 If S_λ is complete when $\varphi = \varphi_0$, then R^* is a region of Neyman structure if and only if R^* is similar.

Proof. If R^* has Neyman structure then, if $f_{S_\lambda}(s)$ is the density of S_λ under H_0,

$$\mathrm{P}\left[Y \in R^*|\varphi = \varphi_0, \lambda\right] = \int P\left[Y \in R^*|S_\lambda = s; \varphi = \varphi_0, \lambda\right] f_{S_\lambda}(s)ds = \gamma \int f_{S_\lambda}(s)ds == \gamma$$

for all $\lambda$and therefore R^* is similar. Conversely, if R^* is similar, let

$$\phi_{R^*}(y) = \begin{cases} 1 & y \in R^* \\ 0 & \text{otherwise} \end{cases}$$

Now for all γ

$$\gamma = \mathrm{P}\left[Y \in R^*|\varphi = \varphi_0, \lambda\right]$$

that is, when H_0 holds,

$$\gamma = \mathrm{E}_{Y|\psi_0}\left[\phi_{R^*}(Y)\right] \text{ for all } \lambda \Longrightarrow \mathrm{E}_{S_\lambda}\left[\mathrm{E}_{Y|\psi_0}\phi_{R^*}(Y)|S_\lambda = s\right] - \gamma = 0 \text{ for all } \lambda$$

and by completeness of S_λ when $\varphi = \varphi_0$

$$\mathrm{E}_{Y|\psi_0}\left[\phi_{R^*}(Y)|S_\lambda = s\right] - \gamma = 0$$

for all s and λ when H_0 holds. Therefore

$$\mathrm{P}\left[Y \in R^*|S_\lambda = s; \varphi = \varphi_0, \lambda\right] = \gamma \text{ for all } s \text{ and } \lambda$$

and R^* has Neyman structure. ■

If S_λ is complete (under H_0) the only way to find similar regions is through regions of Neyman structure, that is by conditioning on S_λ under H_0.

2.12.6 The Likelihood Ratio Test

As an alternative to the search for most powerful tests, consider a general approach to test construction, and assess the qualities of the types of test constructed.

Definition 2.12.14 The **Likelihood Ratio Test** statistic for testing H_0 against H_1 is

$$t(y) = \frac{\sup\limits_{H_1} f_{Y|\theta}(y;\theta)}{\sup\limits_{H_0} f_{Y|\theta}(y;\theta)}$$

where H_0 is rejected if $t(y)$ is too large, that is, if

$$\mathrm{P}\left[t(Y) \geq k | H_0\right] = \gamma$$

Theorem 2.12.2 *If H_0 imposes s independent constraints then as $n \to \infty$*

$$2\log t(Y) = 2\log \frac{\sup\limits_{H_1} f_{Y|\theta}(Y;\theta)}{\sup\limits_{H_0} f_{Y|\theta}(Y;\theta)} \simeq \chi^2_s$$

that is, $2\log t(Y)$ has an approximate Chi-squared distribution with s degrees of freedom.

Proof.(For $s = 1$,

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

and data i.i.d). Then, if $\widehat{\theta}_n$ is the M.L.E. of θ using a Taylor expansion about θ_0

$$2\log t(Y) = 2\left[\log f_{Y|\widehat{\theta}_n}(Y;\widehat{\theta}_n) - \log f_{Y|\theta_0}(Y;\theta_0)\right]$$
$$\approx -2\left[\left(\theta_0 - \widehat{\theta}_n\right)\frac{\partial}{\partial\theta}\left\{\log f_{Y|\theta}(Y;\theta)\right\}_{\theta=\widehat{\theta}_n} - \frac{1}{2}\left(\theta_n - \widehat{\theta}_0\right)^2 \frac{\partial^2}{\partial\theta^2}\left\{\log f_{Y|\theta}(Y;\theta)\right\}_{\theta=\widehat{\theta}_n}\right]$$

But, by definition,

$$\frac{\partial}{\partial\theta}\left\{\log f_{Y|\theta}(Y;\theta)\right\}_{\theta=\widehat{\theta}_n} = 0$$

and thus

$$2\log t(Y) \approx \left(\theta_0 - \widehat{\theta}_n\right)^2 \frac{\partial^2}{\partial\theta^2}\left\{\log f_{Y|\theta}(Y;\theta)\right\}_{\theta=\widehat{\theta}_n} = \left(\theta_0 - \widehat{\theta}_n\right)^2 I\left(\theta_0\right).$$

Now the Weak Law of Large Numbers and the Central Limit Theorem gives (approximately)

$$\widehat{\theta}_n \sim N\left(\theta_0, \frac{1}{I\left(\theta_0\right)}\right) \Rightarrow \left(\theta_n - \widehat{\theta}_0\right)\sqrt{I(\theta_0)} \sim N(0,1)$$

and thus

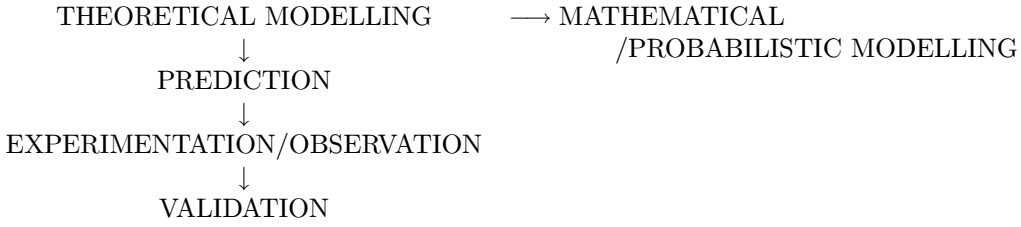
$$2\log t(Y) = \left(\theta_0 - \widehat{\theta}_n\right)^2 I\left(\theta_0\right) \sim \chi^2_1$$

■

3. Practical Classical Statistical Inference

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.

Broadly, the "*Scientific Process*" involves several different stages:



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

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.

3.1 General Framework

Suppose that an experiment or **trial** is to be repeated n times under identical conditions. Let Y_i be the random variable corresponding to the outcome of the ith trial, and suppose that each of the n random variables $Y_1, ..., Y_n$ takes values in sample space $\mathbb{Y}$.

Often, assumptions can reasonably be made about the experimental conditions that lead to simplifications of the joint probability model for the random variables. Essentially, the assumption of identical experimental conditions for each of the n trials implies that the random variables corresponding to the trial outcomes are **identically distributed**, that is, in the usual notation, the (marginal) mass/density function of Y_i is denoted $f(y)$ dropping the subscript on the function f. Another common assumption is that the random variables $Y_1, ..., Y_n$ are **independent**. Thus $Y_1, ..., Y_n$ are usually treated as **i.i.d.** random variables.

In practice, it is commonly assumed that f takes one of the familiar forms (*Binomial*, *Poisson*, *Eyponential*, *Normal* etc.). Thus f depends on one or more parameters (θ, λ, (μ, σ) etc.). As before, the role of these parameters could be indicated by re-writing the function $f(y)$ as

$$f(y) \equiv f(y; \theta) \qquad y \in \mathbb{Y}$$

where θ here is a generic parameter, which may possibly be vector-valued.

It is important here to specify precisely the range of values which this parameter can take; in a Poisson model, we have parameter $\lambda > 0$, and in a Normal model, we have parameters $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$. In the general case, we have parameter $\theta \in \Theta$ where Θ is some subset of $\mathbb{R}^d$ and $d = 1, 2$, say, is the number of parameters; Θ is the **parameter space**. In practice, of course, parameter θ is **unknown** during the experiment.

3.1.1 Objectives of a statistical analysis

After the experiment has been carried out, a sample of **observed data** will have been obtained. Suppose that we have observed outcomes $y_1, ..., y_n$ on the n trials (that is, we have observed $Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n$), 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 $\{y_1, ..., y_n\}$ 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.

3.1.2 Exploratory data analysis

Objectives : To summarize and describe the sample data $\{y_1, ..., y_n\}$, in order to propose a plausible probability model.
The aim of this stage of analysis is 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 (Poisson, Exponential, Normal) 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 the "average value" in the sample.
(2) The **mode**, or "most likely" value or interval observed 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.

Numerical Summaries

Sample mean	y	$= \frac{1}{n}\sum_{i=1}^{n} y_i$
Sample variance	S^2	$= \frac{1}{n}\sum_{i=1}^{n}(y_i - y)^2$
(S^2 or s^2 may be used)	s^2	$= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - y)^2$
Sample quantiles		Suppose that the sample has been sorted into ascending order and re-labelled $y_{(1)} < ... < y_{(n)}$ Then the pth quantile, $0 < p < 100$, is given by
	$y^{(p)}$	$= y_{(k)}$ where k is the nearest integer to $pn/100$.
Median	m	$= y^{(50)}$, the 50th quantile
Lower quartile	q_{25}	$= y^{(25)}$, the 25th quantile
Upper quartile	q_{75}	$= y^{(75)}$, the 75th quantile
Inter-quartile range	IQR	$= q_{75} - q_{25}$
Sample minimum	y_{min}	$= y_{(1)}$
Sample maximum	y_{may}	$= y_{(n)}$
Sample range	R	$= y_{(n)} - y_{(1)}$
Sample skewness	κ	$= \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - y)^3}{S^2}$

Each of these summary statistics can be routinely computed from sample data using a calculator or statistical computer package.

Example 3.1.1 Connection between sample statistics and probability models.
Consider the discrete probability distribution defined on the set of observed sample outcomes $\{y_1, ..., y_n\}$, by placing equal probability $1/n$ on each value, that is, the probability distribution specified by mass function denoted $f_{(n)}$

$$f_{(n)}(y) = \frac{1}{n} \qquad y \in \{y_1, ..., y_n\}\,.$$

Then the expectation of this probability distribution is given by

$$\mathrm{E}_{f(n)}[Y] = \sum_{i=1}^{n} y_i f_{(n)}(y_i) = \sum_{i=1}^{n} y_i \left\{\frac{1}{n}\right\} = \frac{1}{n}\sum_{i=1}^{n} y_i = y$$

that is, the sample mean. Similarly, the variance of this probability distribution is equal to sample variance,

$$\mathrm{Var}_{f(n)}[Y] = \frac{1}{n}\sum_{i=1}^{n}(y_i - y)^2 = S^2.$$

In fact, each of the summary statistics listed above can be viewed as a feature of the probability distribution described by mass function $f_{(n)}$.
Now, consider this probability distribution as n increases to infinity. Then the sample mass function $f_{(n)}$ tends to a function f which can be regarded as the "true" mass/density function, and the sample mean, variance, percentiles etc. tend to the true mean, variance, percentiles of the distribution from which the

data are generated. In practice, of course, n is always finite, and thus the true distribution, true mean etc., cannot be known exactly. Therefore, we approximate the true distribution by an appropriately chosen distribution (Poisson, Exponential, Normal etc.) with parameters chosen to correspond to the observed sample properties.

3.1.3 Graphical Summaries

The most common graphical summary technique is the **histogram**. Typically, the sample space $\mathbb{Y}$ is divided into a number of subsets $\mathbb{Y}_1, ..., \mathbb{Y}_H$, 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 a set of bars, where the hth bar has height n_h and occupies the region of corresponding to $\mathbb{Y}_h$.

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

3.1.4 Outliers

Each of the summaries described above is used on the presumption that each of the observed data values $y_1, ..., y_n$ is essentially an observation from the same probability distribution, and, because of this, each y_i is treated as equally important. Sometimes, however, 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 variance are extremely sensitive to the presence of outliers in the sample. Other summary statistics, for example those based on sample percentiles, are less sensitive to outliers. Outliers can usually be identified by inspection of the raw data, or from careful plotting of histograms.

3.2 Practical Estimation

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.

Example 3.2.1 Suppose that 20 product lifetime measurements are to be made; beginning at time $y = 0$, each product survives for some random time until failure. This yields a sample of observed failure times $y_1, y_2, ..., y_{20}$. Typically, we would assume a probability model, such as the Exponential, for such data; this distribution has a single parameter λ, which is unknown in practice, and thus requires estimation. Subsequently, we may be interested in prediction for a future product. For example, we may be interested in assessing the probability that an identical product would survive for longer than y^*; this probability is given under the Exponential model by $\mathrm{P}[\ Y > y^*\ ] = e^{-\lambda y^*}$, which we may only evaluate if we know, or have an estimate of λ.

It is therefore important to find a simple and yet general technique for parameter estimation. Maximum Likelihood is the most commonly used technique:

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 3.2.2 A sample $y_1, ..., y_n$ is modelled by a Poisson distribution with parameter denoted λ; hence

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

for some $\lambda > 0$.

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

$$L(\lambda) = \prod_{i=1}^{n} f(y_i;\lambda) = \prod_{i=1}^{n} \left\{ \frac{\lambda^{y_i}}{y_i!}e^{-\lambda} \right\} = \frac{\lambda^{y_1+...+y_n}}{y_1!....y_n!}e^{-n\lambda}$$

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

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

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

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

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

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

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

Example 3.2.3 The following data are the numbers of household accidents recorded across a number of households ($n = 647$) :

Number of accidents	Frequency
0	447
1	132
2	42
3	21
4	3
5	2

so that the estimate of λ if a Poisson model is assumed is

$$\hat{\lambda}_{ML} = y = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{(447 \times 0) + (132 \times 1) + (42 \times 2) + (21 \times 3) + (3 \times 4) + (2 \times 5)}{647} = 0.465.$$

The log-likelihood function and MLE are depicted on the following figure:

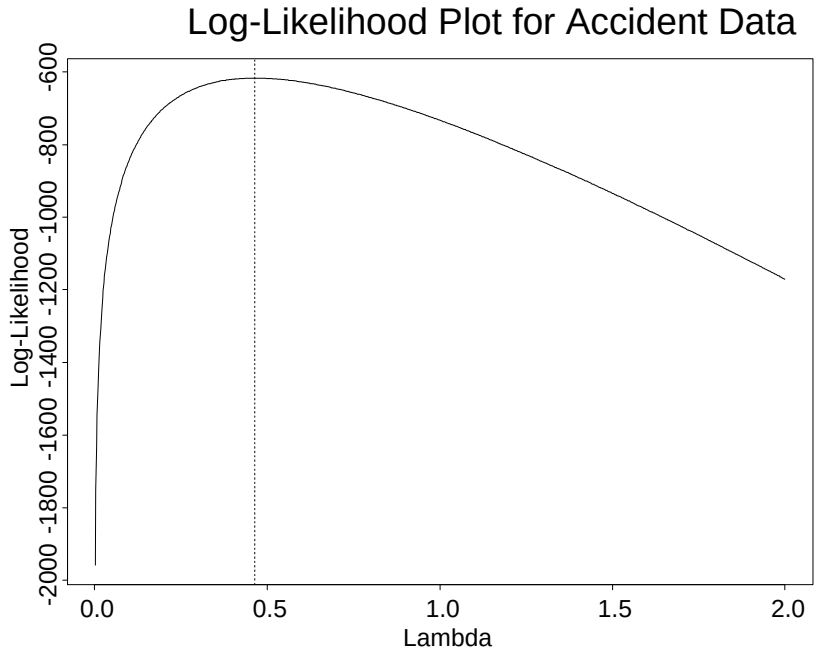


Figure 3.1: Log-likelihood function and maximum likelihood estimate for the household accident data.

3.3 Practical Hypothesis Testing

Given a sample $y_1, ..., y_n$ from a probability model $f(y;\theta)$ 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	$Y_1, ..., Y_n \sim N(\mu, \sigma^2)$	Possible Models	$\mu = c_1$
	sample observations	$y_1, ...y_n$		$\sigma = c_2$
TWO SAMPLE :	Random variables	$X_1, ..., X_n \sim N(\mu_X, \sigma_X^2)$	Possible Models	$\mu_X = \mu_Y$
	sample one observations	$x_1, ...x_n$		$\sigma_X = \sigma_Y$
	Random variables	$Y_1, ..., Y_n \sim N(\mu_Y, \sigma_Y^2)$		
	sample two observations	$y_1, ...y_n$		

3.3.1 Hypothesis Testing for Normal data I - The Z-test (σ Known)

Recall that, if $Y_1, ..., Y_n \sim N(\mu, \sigma^2)$ are the i.i.d. outcome random variables of n experimental trials, then

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

, with Y and S^2 statistically independent. Suppose we want to test the **hypothesis** that $\mu = c$, for some specified constant c, (where, for example, $c = 20.0$) is a plausible model; more specifically, we want to test

$$\begin{array}{ll} H_0 & : \mu = c \qquad \text{the } \mathbf{NULL} \text{ hypothesis} \\ H_1 & : \mu \neq c \qquad \text{the } \mathbf{ALTERNATIVE} \text{ hypothesis} \end{array}$$

[i.e. we want to test whether H_0 is true, or whether H_1 is true]. Now, we know that, in the case of a Normal sample, the distribution of the estimator Y is Normal, and

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

where Z is a **random variable**. Now, when we have observed the data sample, we can calculate y, and therefore we have a way of testing whether $\mu = c$ is a plausible model; we calculate y from $y_1, ..., y_n$, and then calculate

$$z = \frac{y - c}{\sigma/\sqrt{n}}$$

If H_0 is true, and $\mu = c$, 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

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

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

3.3.2 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 3.3.1 A sample of size 10 has sample mean $y = 19.7$. 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.

3.3.3 Hypothesis Testing 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 $\overline{Y}$ 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 $\overline{Y}$ and s^2, we see that

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

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

$$t = \frac{(\overline{y} - \mu)}{s/\sqrt{n}}$$

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

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

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.

3.3.4 Hypothesis Testing for Normal data II - 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

$$\begin{array}{ll} H_0 & : \sigma^2 = c \\ H_1 & : \sigma^2 \neq c \end{array}$$

We construct a test based on the estimate of variance, s_2. In particular, we saw from the Theorem on p.32 that 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

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

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

$$\begin{array}{ll} H_0 & : \sigma^2 = 1.0 \\ H_1 & : \sigma^2 \neq 1.0 \end{array}$$

we compute test statistic

$$q = \frac{(n-1)s^2}{c} = \frac{90.78^2}{1.0} = 5.43.75$$

and compare with

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

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

3.3.5 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 $N(\mu_X, \sigma^2_X)$ distribution, and sample two, $y_1, ..., y_{n_Y}$, independently from a $N(\mu_Y, \sigma^2_Y)$ 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 $(\bar{x}, s^2_X)$ and $(\bar{y}, s^2_Y)$ respectively.

First, consider testing the hypothesis

$$\begin{aligned} H_0 &: \mu_X = \mu_Y \\ H_1 &: \mu_X \neq \mu_Y \end{aligned}$$

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

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

and hence

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

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 $\sigma_Y = \sigma_Y = \sigma$ 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^2_P, defined by

$$s^2_P = \frac{(n_Y - 1)s^2_Y + (n_Y - 1)s^2_Y}{n_X + n_Y - 2}$$

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

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

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.

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 $(n_Y - 1, n_Y - 1)$ degrees of freedom; this distribution can be denoted $F(n_Y - 1, n_Y - 1)$, and its quantiles are tabulated. Hence we can look up the 0.025 and 0.975 quantiles of this distribution (the F distribution is not symmetric), and hence define the critical region; informally, if the test statistic q is very small or very large, then it is a surprising observation from the F distribution and hence we reject the hypothesis of equal variances.

3.3.6 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

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

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

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

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

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

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 $[c_1, c_2]$.

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.

Test	Pivotal Quantity T_{PQ}	Distribution	Parameter	Confidence Interval
ONE SAMPLE Z	$Z = \dfrac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$	$N(0,1)$	μ	$\overline{y} \pm C_R\ \sigma/\sqrt{n}$
ONE SAMPLE T	$T = \dfrac{\overline{Y} - \mu}{s/\sqrt{n}}$	$St(n-1)$	μ	$\overline{y} \pm C_R\ s/\sqrt{n}$
ONE SAMPLE σ	$Q = \dfrac{(n-1)s^2}{\sigma^2}$	χ^2_{n-1}	σ^2	$\left[\dfrac{(n-1)s^2}{C_{R_2}} : \dfrac{(n-1)s^2}{C_{R_1}}\right]$
$\sigma_Y = \sigma_Y = \sigma$ known TWO SAMPLE Z	$Z = \dfrac{(\overline{X} - \mu_X) - (\overline{Y} - \mu_Y)}{\sigma\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$	$N(0,1)$	$\mu_X - \mu_Y$	$(\bar{x} - \bar{y}) \pm C_R\ \sigma\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}$
$\sigma_Y = \sigma_Y = \sigma$ unknown TWO SAMPLE T	$T = \dfrac{(\overline{X} - \mu_X) - (\overline{Y} - \mu_Y)}{s_P\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}}$	$St(n_X + n_Y - 2)$	$\mu_X - \mu_Y$	$(\bar{x} - \bar{y}) \pm C_R\ s_P\sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}$
$\sigma_Y \neq \sigma_Y$ known TWO SAMPLE Z	$Z = \dfrac{(\overline{X} - \mu_X) - (\overline{Y} - \mu_Y)}{\sqrt{\dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}}}$	$N(0,1)$	$\mu_X - \mu_Y$	$(\bar{x} - \bar{y}) \pm C_R\sqrt{\dfrac{\sigma_X^2}{n_X} + \dfrac{\sigma_Y^2}{n_Y}}$
TWO SAMPLE σ	$Q = \dfrac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$	$F(n_X - 1, n_Y - 1)$	$\dfrac{\sigma_X^2}{\sigma_Y^2}$	$\left[\dfrac{s_X^2}{C_{R_2}s_Y^2} : \dfrac{s_X^2}{C_{R_1}s_Y^2}\right]$

3.4 Linear Regression Analysis

Suppose that we have n measurements of two variables X and Y, that is, a sample of pairs of observations

$$\{(x_i, y_i) : i = 1, ..., n\}$$

and it is believed that there is a **linear** relationship between X and Y. Suppose that we regard X as a **controlled** variable, that is, we can control the values of Y at which X is measured. Our aim is to try and **predict** Y for a given value of X, and thus we have to build a probability model for Y conditional on $X = x$ that incorporates the linear dependence.

3.4.1 Terminology

X is the **response** or **dependent** variable

Y is the **covariate** or **independent** variable

A simple relationship between X and Y is the **linear regression model**, where

$$\mathrm{E}[Y|X = x] = \alpha + \beta x,$$

that is, conditional on $X = x$, the expected or "predicted" value of X is given by $\alpha + \beta x$, where α and β are unknown parameters; in other words, we model the relationship between X and Y as a straight line with **intercept** α and **slope** β.

For data $\{(x_i, y_i) : i = 1, ..., n\}$, the objective is to estimate the unknown parameters α and β. A simple estimation technique, is **least-squares estimation**.

3.4.2 Least-Squares Estimation

Suppose that a sample, $\{(x_i, y_i) : i = 1, ..., n\}$, is believed to follow a linear regression model, $\mathrm{E}[Y|X = x] = \alpha + \beta x$. For fixed values of α and β, let $y_i^{(P)}$ denote the expected value of Y conditional on $X = x_i$, that is

$$y_i^{(P)} = \alpha + \beta x_i$$

Now define error terms e_i, $i = 1, ..., n$ by

$$e_i = y_i - y_i^{(P)} = y_i - \alpha - \beta x_i$$

that is, e_i is the vertical discrepancy between the **observed** and **expected** values of Y.

The objective in least-squares estimation is find a "line of best fit", and this is achieved by inspecting the squares of the error terms e_i, and choosing α and β such that the sum of the squared errors is **minimized**; we aim to find the straight line model for which the total error is smallest.

Let $S(\alpha, \beta)$ denote the total error in fitting a linear regression model with parameters α and β. Then

$$S(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i^{(P)})^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

To calculate the least-squares estimates, we have to minimize $S(\alpha, \beta)$ as a function of α and β. This can be achieved in the usual way by taking partial derivatives with respect to the two parameters, and equating the partial derivatives to zero simultaneously.

$$\begin{aligned} (1) \quad \frac{\partial}{\partial \alpha}\{S(\alpha,\beta)\} &= -2\sum_{i=1}^{n}(y_i - \alpha - \beta x_i) = 0 \\ (2) \quad \frac{\partial}{\partial \beta}\{S(\alpha,\beta)\} &= -2\sum_{i=1}^{n}x_i(y_i - \alpha - \beta x_i) = 0 \end{aligned}$$

Solving (1), we obtain an equation for the least-squares estimates $\hat{\alpha}$ and $\hat{\beta}$

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}\,\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \hat{\beta}\bar{x}.$$

Solving (2) in the same way, and combining the last two equations, and solving for $\hat{\beta}$ gives

$$\begin{aligned} \hat{\alpha} &= \frac{\sum_{i=1}^{n} x_i y_i - \hat{\beta}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i} \\ \hat{\beta} &= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left\{\sum_{i=1}^{n} x_i\right\}^2} \end{aligned}$$

Thus the least-squares estimates of α and β are given by

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad \hat{\beta} = \frac{nS_{xy} - S_x S_y}{nS_{xx} - \{S_x\}^2}$$

where

$$S_x = \sum_{i=1}^{n} x_i \qquad S_y = \sum_{i=1}^{n} y_i \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i$$

Therefore it is possible to produce estimates of parameters in a linear regression model using least-squares, without any specific reference to probability models. In fact, the least-squares approach is very closely related to maximum likelihood estimation for a specific probability model.

The **correlation coefficient**, r, measures the degree of association between X and Y variables and is given by

$$r = \frac{nS_{xy} - S_x S_y}{\sqrt{(nS_{xx} - S_x^2)(nS_{yy} - S_y^2)}}$$

and therefore is quite closely related to $\hat{\beta}$.

3.4.3 Relationship between least-squares and maximum likelihood

Suppose that X and Y follow a linear regression model,

$$\mathrm{E}[Y|X = x] = \alpha + \beta x,$$

and recall that the error terms e_i were defined

$$e_i = y_i - \alpha - \beta x_i.$$

Now, e_i is the vertical discrepancy between observed and expected behaviour, and thus e_i could be interpreted as the observed version of a **random variable**, say ϵ_i, which represents the random uncertainty involved in measuring Y for a given Y. A plausible probability model might therefore be that the random variables ϵ_i, $i = 1, ...n$, were independent and identically distributed, and

$$\epsilon_i \sim N(0, \sigma^2),$$

for some error variance parameter σ^2. Implicit in this assumption is that the distribution of the random error in measuring Y does not depend on the value of Y at which the measurement is made. This distributional assumption about the error terms leads to a probability model for the variable Y. As we can write

$$Y = \alpha + \beta X + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$, then given on $X = x_i$, we have the conditional distribution Y_i as

$$Y_i|X = x_i \sim N(\alpha + \beta x_i, \sigma^2),$$

where random variables Y_i and Y_j are **independent** (as ϵ_i and ϵ_j are independent). On the basis of this probability model, we can derive a likelihood function, and hence derive maximum likelihood estimates. For example, we have the likelihood $L(\theta) = L(\alpha, \beta, \sigma^2)$ defined as the product of the n conditional density terms derived as the conditional density of the observed y_i given x_i,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f(y_i; x_i, \theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right\} \end{aligned}$$

The maximum likelihood estimates of α and β, and error variance σ^2, are obtained as the values at which $L(\alpha, \beta, \sigma^2)$ is maximized. But, $L(\alpha, \beta, \sigma^2)$ is maximized when the term in the exponent, that is,

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

is minimized. But this is **precisely** the least-squares criterion described above, and thus the m.l.e s of α and β assuming a Normal error model are **exactly equivalent** to the least-squares estimates.

3.4.4 Estimates of Error Variance and Residuals

In addition to the estimates of α and β, we can also obtain the maximum likelihood estimate of σ^2,

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = S^2$$

Often, a **corrected** estimate, s^2, of the error variance is used, defined by

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \ = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

where $\hat{y_i} = \hat{\alpha} + \hat{\beta}x_i$ is the **fitted value** of Y at $X = x_i$. Note also that, having fitted a model with parameters $\hat{\alpha}$ and $\hat{\beta}$, we can calculate the error in fit at each data point, or **residual**, denoted $e_i, i = 1, ..., n$, where $e_i = y_i - \hat{y_i} = y_i - \hat{\alpha} - \hat{\beta}x_i$.

3.4.5 Prediction for a new covariate value

Suppose that, having fitted a model, and obtained estimates $\hat{\alpha}$ and $\hat{\beta}$ using maximum likelihood or least-squares, we want to predict the Y value for a new value x^* of covariate Y. By considering the nature of the regression model, we obtain the predicted value y^* as

$$y^* = \hat{\alpha} + \hat{\beta}x^*$$

3.4.6 Standard Errors of Estimators and t-statistics

We need to be able to understand how the estimators corresponding to $\hat{\alpha}$ and $\hat{\beta}$ behave, and by how much the estimate is likely to vary. This can be partially achieved by inspection of the **standard errors** of estimates, that is, the square-root of the variance in the sampling distribution of the corresponding estimator. It can be shown that

$$s.e.(\hat{\alpha}) = s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} \qquad s.e.(\hat{\beta}) = s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}}$$

where s is the square-root of the corrected estimate of the error variance. It is good statistical practice to report standard errors whenever estimates are reported. The standard error of a parameter also allows a test of the hypothesis "parameter is equal to zero". The test is carried out by calculation of the **t-statistic**, that is, the ratio of a parameter estimate to its standard error. The t-statistic must be compared with the 0.025 and 0.975 percentiles of a Student-t distribution with $n - 2$ degrees of freedom as described below.

3.4.7 Hypothesis Tests and Confidence Intervals for Parameters

We may carry out hypothesis tests for the parameters in a linear regression model; as usual we need to be able to understand the sampling distributions of the corresponding estimators. In the linear regression model, the sampling distributions of the estimators of α and β have **Student-t distributions** with $n - 2$

degrees of freedom, hence we use the test statistics

$$t_\alpha = \frac{\hat{\alpha} - c}{s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}}} = \frac{\hat{\alpha} - c}{s.e.(\hat{\alpha})}$$

$$t_\beta = \frac{\hat{\beta} - c}{s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}}} = \frac{\hat{\beta} - c}{s.e.(\hat{\beta})}$$

to test the null hypothesis that the parameter is equal to c. Typically, we use a test at the 5 % significance level, so the appropriate critical values are the 0.025 and 0.975 quantiles of a $St(n-2)$ distribution.

It is also useful to report, for each parameter, a confidence interval in which we think the **true** parameter value (that we have estimated by $\hat{\alpha}$ or $\hat{\beta}$) lies with high probability. It can be shown that the 95% confidence intervals are given by

$$\alpha \ : \ \hat{\alpha} \ \pm \ t_{n-2}(0.975)\ s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} \qquad \beta \ : \ \hat{\beta} \ \pm \ t_{n-2}(0.975)\ s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}}$$

where $t_{n-2}(0.975)$ is the 97.5th percentile of a Student-t distribution with $n-2$ degrees of freedom.

The confidence intervals are useful because they provide an alternative method for carrying out hypothesis tests. For example, if we want to test the hypothesis that $\alpha = c$, say, we simply note whether the 95% confidence interval contains c. If it does, the hypothesis can be accepted; if not the hypothesis should be rejected, as the confidence interval provides evidence that $\alpha \neq c$.

We may carry out a hypothesis test to carry out whether there is significant correlation between two variables. We denote by ρ the true correlation; then to test the hypothesis

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

we use the test statistic

$$t_r = r\sqrt{\frac{n-2}{1-r^2}}$$

which we compare with the null distribution which is Student-t with $n-2$ degrees of freedom. If $|t_r| > t_{n-2}(0.975)$, then we can conclude that the true correlation ρ is significantly different from zero.

3.4.8 Multiple Linear Regression

In everything that is described above, we have used a model in which we predicted a response Y from a single covariate X. This simple model can be extended to the case where Y is modelled as a function of p covariates $X_1, ..., X_p$, that is, we have the conditional expectation of Y given by

$$\mathrm{E}[Y|X_1 = x_1, ..., X_p = x_p] = \alpha + \beta_1 x_1 + ... + \beta_p x_p,$$

so that the observation model is given by

$$Y_i|X_1 = x_{i1}, ..., X_p = x_{ip} \sim N(\alpha + \beta_1 x_{i1} + ... + \beta_p x_{ip}, \sigma^2).$$

Again, we can use maximum likelihood estimation to obtain estimates of the parameters in the model, that is, parameter vector $(\alpha, \beta_1, ..., \beta_p, \sigma^2)$, but the details are slightly more complex, as we have to solve $p+1$ equations simultaneously. The procedure is simplified if we write the parameters as a single vector, and perform matrix manipulation and calculus to obtain the estimates.

Example 3.4.1 The following data are believed to follow a linear regression model;

x	0.54	2.03	3.15	3.96	6.25	8.17	11.08	12.44	14.04	14.34	18.71	19.90
y	11.37	11.21	11.61	8.26	14.08	16.25	11.00	14.94	16.91	15.78	21.26	20.25

We want to calculate estimates of α and β from these data. First, we calculate the summary statistics;

$$S_x = \sum_{i=1}^{n} x_i = 118.63 \qquad S_y = \sum_{i=1}^{n} y_i = 172.92$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 = 1598.6 \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i = 1930.9$$

with $n = 12$ which leads to parameter estimates

$$\hat{\beta} = \frac{nS_{xy} - S_x S_y}{nS_{xx} - \{S_x\}^2} = \frac{12 \times 1930.9 - 118.63 \times 172.92}{12 \times 1598.6 - (118.63)^2} = 0.5201$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 14.410 - 0.5201 \times 9.8842 = 9.269$$

This fit leads to the following fitted values and residuals;

x	0.54	2.03	3.15	3.96	6.25	8.17	11.08	12.44	14.04	14.34	18.71	19.90
y	11.37	11.21	11.61	8.26	14.08	16.25	11.00	14.94	16.91	15.78	21.26	20.25
$\hat{y}$	9.55	10.33	11.95	12.37	12.52	13.52	15.03	15.73	16.57	16.73	19.00	19.62
e	1.82	0.88	−0.34	−4.11	1.56	2.73	−4.03	−0.80	0.34	−0.95	2.26	0.63

The **corrected variance estimate**, s^2, is given by

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

$$= \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ = 3.438 \qquad \Longrightarrow \qquad s = 2.332$$

The **standard errors** for the two parameters are given by

$$s.e.(\hat{\alpha}) = s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} = 1.304 \qquad s.e.(\hat{\beta}) = s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}} = 0.113$$

The **t-statistics** for the two parameters are given by

$$t_\alpha = \frac{\hat{\alpha}}{s.e.(\hat{\alpha})} = \frac{9.269}{1.304} = 7.109 \qquad t_\beta = \frac{\hat{\beta}}{s.e.(\hat{\beta})} = \frac{0.520}{0.113} = 4.604.$$

The 0.975 percentile of a Student-t distribution with $n - 2 = 10$ degrees of freedom is found from tables to be 2.228. Both t-statistics are more extreme than this critical value, and hence it can be concluded that both parameters are significantly different from zero.

To calculate the **confidence intervals** for the two parameters. we need to use the 0.975 percentile of a

$St(10)$ distribution. From above, we have that $St(10)(0.975) = 2.228$, and so the confidence intervals are given by

$$\alpha \;:\; \hat{\alpha} \pm t_{n-2}(0.975)\, s\sqrt{\frac{S_{xx}}{nS_{xx} - \{S_x\}^2}} \;=\; 9.269 \pm 2.228 \times 1.304 \;=\; (6.364 : 12.174)$$

$$\beta \;:\; \hat{\beta} \pm t_{n-2}(0.975)\, s\sqrt{\frac{n}{nS_{xx} - \{S_x\}^2}} \;=\; 0.5201 \pm 2.228 \times 0.113 \;=\; (0.268 : 0.772)$$

so that, informally, we are 95% certain that the true value of α lies in the interval $(6.724 : 12.174)$, and that the true value of β lies in the interval $(0.268 : 0.772)$. This amounts to evidence that, for example, $\alpha \neq 0$ (as the confidence interval for α does not contain 0), and evidence that $\beta \neq 1$ (as the confidence interval for β does not contain 1).

3.5 Model Validation

Techniques used for estimation and hypothesis testing allow specific and quantitative questions about the parameters in a probability model to be posed and resolved on the basis of a collection of sample data $x_1, ..., x_n$. However, the question as to the validity of the assumed probability model (for example, Binomial, Poisson, Exponential, Normal etc.) has yet to be addressed.

3.5.1 Probability Plots

The probability plotting technique involves comparing *predicted* and *observed* behaviour by comparing quantiles of the proposed probability distribution with sample quantiles. Suppose that a sample of data of size n are to be modelled using a proposed probability model with cdf F_Y which possibly depends on unknown parameter(s) θ. The sample data are first sorted into ascending order, and then the ith datum, y_i, corresponds to the $100i/(n+1)$th quantile of the sample. Now, the equivalent *hypothetical* quantile of the distribution, q_i is found as the solution of

$$F_Y(q_i) = \frac{i}{n+1} \qquad i = 1, ..., n.$$

If the model encapsulated in F_Y is an acceptable model for the sample data, then for large n, $y_i \approx q_i$, so a plot of $\{(q_i, y_i) : i = 1, ..., n\}$ should be a straight line through the origin with slope 1. Hence the validity of F_Y as a model for the sample data can be assessed through such a plot.

Example 3.5.1 For the *Eyponential*(1) model, F_Y is given by

$$F_Y(y) = 1 - e^{-y} \qquad y \geq 0$$

so the probability plot consists of examining $\{(q_i, y_i) : i = 1, ..., n\}$ where

$$1 - e^{-q_i} = \frac{i}{n+1} \Longrightarrow q_i = -\log\left\{1 - \frac{i}{n+1}\right\}$$

Example 3.5.2 For the $N(0,1)$ model, $F_Y \equiv \Phi$ is only available numerically (for example via statistical tables). Here the probability plot consists of examining $\{(q_i, y_i) : i = 1, ..., n\}$ where

$$\Phi(q_i) = \frac{i}{n+1} \Longrightarrow q_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$$

Example 3.5.3 For the *Eyponential*(λ) model, we plot $\{(q_i, y_i) : i = 1, ..., n\}$ where

$$F_Y(q_i) = 1 - e^{-\lambda q_i} = \frac{i}{n+1} \Longrightarrow q_i = -\frac{1}{\lambda} \log \left\{ 1 - \frac{i}{n+1} \right\}.$$

Hence, if we define q_i^* by

$$q_i^* = -\log \left\{ 1 - \frac{i}{n+1} \right\}$$

then if the model is correct, a plot of $\{(q_i^*, y_i) : i = 1, ..., n\}$ should be approximately a straight line through the origin with slope $1/\lambda$; hence λ can be estimated from this plot by using linear regression.

Example 3.5.4 For the $N(\mu, \sigma^2)$ model, is again only available numerically (for example via statistical tables). Here the probability plot consists of examining $\{(q_i, y_i) : i = 1, ..., n\}$ where

$$F_Y(q_i) = \Phi\left(\frac{q_i - \mu}{\sigma} \right) = \frac{i}{n+1} \Longrightarrow q_i = \mu + \sigma \Phi^{-1} \left(\frac{i}{n+1} \right).$$

Hence, if we define q_i^* by

$$q_i^* = \Phi^{-1} \left(\frac{i}{n+1} \right)$$

then if the model is correct, a plot of $\{(q_i^*, y_i) : i = 1, ..., n\}$ should be approximately a straight line with intercept μ and slope σ; hence μ, σ can again be estimated from this plot by using linear regression.

3.5.2 The Chi-Squared Goodness-of-Fit Test

The problem of testing a hypothesis as to whether a data sample $y_1, ..., y_n$ is well-modelled by a specified probability distribution can be approached from a "goodness-of-fit" perspective.

Suppose that the data are recorded as the number of observations, O_i, say in a sample of size n that fall into each of k categories or "bins". Suppose that under the hypothesized model with mass/density function f_Y or cdf F_Y, the data follow a specific probability distribution specified by probabilities $\{p_i : i = 1, ..., k\}$. These probabilities can be calculated directly from f_Y or F_Y, possibly after parameters in the model have been estimated using maximum likelihood. Then, if the hypothesized model is correct, $E_i = np_i$ observations would be expected to fall into category i. An intuitively sensible measure of the **goodness-of-fit** of the data to the hypothesized distribution is given by the **chi-squared statistic**

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

A formal hypothesis test of model adequacy can be carried out in the usual framework; here the chi-squared statistic is the test statistic, and the null distribution (the distribution of the test statistic if the hypothesis is TRUE) is approximately a **chi-squared distribution** with $k - d - 1$ degrees of freedom, where d is the number of parameters in f_Y or F_Y that were estimated in order calculate the probabilities $p_1, ..., p_k$.

Example 3.5.5 Testing the fit of a Poisson distribution
An early experiment into the properties of radioactive materials involved counting the number of alpha particles emitted from a radioactive source in 2612 consecutive 7.5 second intervals. A total of 10126 particles were counted, and the observed frequencies for each of the numbers of counts (per 7.5s) from 0 to 12 were recorded. To test the hypothesis that the data follow a Poisson distribution, a chi-squared test can

Count	O_i	p_i	E_i	$(O_i - E_i)^2/E_i$
0	57	0.021	54	0.167
1	204	0.080	210	0.171
2	383	0.156	407	1.415
3	525	0.201	525	0.000
4	532	0.195	510	0.949
5	408	0.151	395	0.428
6	273	0.098	255	1.271
7	139	0.054	141	0.028
8	49	0.026	68	5.309
9	27	0.011	30	0.300
10	10	0.004	11	0.091
11	4	0.002	4	0.000
12	2	0.000	1	1.000
>12	0	0.001	1	1.000
Total	2612	1.000	2612	12.129

be performed. First, we estimate Poisson parameter λ by its m.l.e., which is $\hat{\lambda} = \bar{x} = 10126/2612 = 3.877$. Secondly, we calculate probabilities p_i using the Poisson formula. Thirdly, we calculate the theoretical (expected) frequencies $E_i = np_i$ for each category. Finally, we calculate the χ^2 statistic as the sum of the (standardized) squared differences between observed and expected frequencies.

In this case, $\chi^2 = 12.129$. To complete the test we find that the 95th percentile of a Chi-squared distribution with $k - 1 - 1 = 12$ degrees of freedom is 21.03. This implies that the χ^2 statistic would only be surprising at a significance level of 0.05 if it was larger than 21.03. Here, as $\chi^2 = 12.129$, and therefore not surprising. Hence there is no evidence to indicate that the data are not from a Poisson distribution. Clearly, the categorization is arbitrary, and several of the categories in example 1 could be combined. As a general rule, the categories should be chosen so that there is at least **five** observed counts in each.

Example 3.5.6 Contingency Tables
One common use of the Chi-Squared Goodness-of-Fit test is in the context of **contingency tables**; 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

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

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. This test is readily carried out using a Chi-Squared Goodness of Fit test; it is easy to show that, if the independence model is correct, the expected cell frequencies E_{ij} can be calculated as

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

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

Summary 3.1 *If a given hypothesis is true, it can be shown that the chi-squared statistic χ^2 for a sample of data has a particular Chi-squared distribution. If χ^2 takes a value that is surprising or unlikely under that probability distribution (for example if its value lies in the extreme right-hand tail and is larger, say, than the 95th percentile of the distribution) it is very likely that the hypothesis is false and should be be rejected. In general, to test a hypothesis, consider a statistic calculated from the sample data. Derive mathematically the probability distribution of the statistic when the hypothesis is true, and compare the actual value of the statistic with the hypothetical probability distribution. Ask the question "Is the value a likely observation from this probability distribution ?". If the answer is "No", then reject the hypothesis.*

3.5.3 Model Validation in Linear Regression

The adequacy of the fit of a linear regression model can be assessed on a **global** (overall) level using the R^2 (R-squared) statistic, which is calculated as the square of the correlation coefficient, r; R^2 quantifies the proportion of the total variation observed in the data that is explained by the regression (i.e. systematic) component of the model, as compared to the random variation. If the covariate Y is a good predictor, then the majority of the total variation Y will be explained by variation in Y, and so R^2 will be near 1.

The adequacy of the fit of a linear regression model can be assessed on a **local** (point-by-point) level by inspection of the *residuals*; recall that, if $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ is the **fitted value** of Y at $X = x_i$ for parameter estimates $\hat{\alpha}$ and $\hat{\beta}$, the residual e_i is the error in fit at each data point, so $e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}x_i$. If the regression model is adequate, then because of the assumptions underlying the model, the residuals $e_i, i = 1, ..., n$ should form an i.i.d. sample from a $N(0, \sigma^2)$ distribution, whose magnitude should not vary systematically with x_i, y_i or the fitted value $\hat{y}_i$; this could be checked using simple scatterplots. The normality of the residuals can be checked using, for example, a probability plot. Clearly, in practice, the random variation in the model, as encapsulated in variance σ^2, is unknown, but can be estimated using the corrected error variance estimate

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \ = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Using this estimate, it is possible to compute **standardized residuals**

$$e_i^* = \frac{y_i - \hat{y}_i}{s}$$

which can be shown to follow a Student-t distribution with $n-2$ degrees of freedom. The advantage of using standardized residuals is that they have, approximately, variance equal to 1.

4. Bayesian Inference and Decision Making

4.1 Introduction and Terminology

In elementary probability theory, Bayes Theorem gave a relationship between conditional probabilities. For events E and F such that $\mathrm{P}(E)$, $\mathrm{P}(F) > 0$,

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

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

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

For random variables, the same result can be seen to hold: for random variables X and Y,

$$f_{Y|Y}(y|x) = \frac{f_{XY}(x|y)f_Y(y)}{f_Y(x)} = \frac{f_{X|Y}(x|y)f_Y(y)}{\int f_{X|Y}(x|y)f_Y(y)dy}. \tag{4.1}$$

Bayes Theorem can be used as the basis for inferences about a parameter θ in a probability model.

In Bayesian analysis, θ is treated as a random variable (instead of a fixed but unknown constant), with a **prior** density encapsulating the beliefs about θ before the data are collected. Denoting by $\pi(\theta)$ the prior density, by (4.1),

$$\pi_{\theta|Y}(\theta|y) = \frac{f_{Y|\theta}(y;\theta)\pi(\theta)}{f_Y(y)} = \frac{f_{Y|\theta}(y;\theta)\pi(\theta)}{\int f_{Y|\theta}(y;\theta)\pi(\theta)d\theta} \tag{4.2}$$

Definition 4.1.1 The denominator in (4.2)

$$f_Y(y) = \int f_{Y|\theta}(y;\theta)\pi(\theta)d\theta$$

is termed the **marginal likelihood** of the data y, and does not depend on θ. The term $\pi_{\theta|Y}(\theta|y)$ in (4.2) is the **posterior distribution** of θ given y, which encapsulates the beliefs about θ in light of the data.

Lemma 4.1.1 ***Lemma 4.1.2*** The variance of the posterior distribution, $\mathrm{Var}_{\theta|Y}[\theta|Y = y]$ satisfies

$$\mathrm{E}_Y\left[Var_{\theta|Y}[\theta|Y = y]\right] < Var_\theta[\theta]$$

, that is, the prior variance is larger than the expected posterior variance.

Proof. Using (2.6) for the joint distribution of θ and Y,

$$\begin{aligned} \mathrm{Var}_\theta[\theta] &= \mathrm{E}_Y\left[\mathrm{Var}_{\theta|Y}[\theta|Y=y]\right] + \mathrm{Var}_Y\left[\mathrm{E}_{\theta|Y}[\theta|Y=y]\right] \\ &\geq \mathrm{E}_Y\left[\mathrm{Var}_{\theta|Y}[\theta|Y=y]\right] \end{aligned}$$

as the second term is non-negative. ■

Note. The Bayesian calculation in (4.2) acts sequentially, that is, for data y_1

$$\pi_{\theta|Y}(\theta|y_1) = \frac{f_{Y|\theta}(y_1;\theta)\pi(\theta)}{f_Y(y_1)} = \frac{f_{Y|\theta}(y_1;\theta)\pi(\theta)}{\int f_{Y|\theta}(y_1;\theta)\pi(\theta)d\theta} \tag{4.3}$$

contains the information about θ in light of the data y_1 and prior assumptions. If new (independent and identically distributed to y_1) data y_2 becomes available, then the posterior for θ in light of the combined data (y_1, y_2) is

$$\pi_{\theta|Y}(\theta|y_1,y_2) = \frac{f_{Y|\theta}(y_1,,y_2;\theta)\pi(\theta)}{f_Y(y_1,y_2)} = \frac{f_{Y|\theta}(y_1,y_2;\theta)\pi(\theta)}{\int f_{Y|\theta}(y_1,y_2;\theta)\pi(\theta)d\theta}.$$

But note also that

$$\pi_{\theta|Y}(\theta|y_1,y_2) = \frac{f_{Y|\theta}(y_2;\theta)\pi_{\theta|Y}(\theta|y_1)}{f_Y(y_2;y_1)}$$

where $\pi_{\theta|Y}(\theta|y_1)$ is the posterior for θ from (4.3), and

$$f_Y(y_2;y_1) = \frac{f_Y(y_1,y_2)}{f_Y(y_1)} = \frac{\int f_{Y|\theta}(y_1;\theta)\pi(\theta)d\theta}{\int f_{Y|\theta}(y_1,y_2;\theta)\pi(\theta)d\theta} \tag{4.4}$$

■

Note. Note that for the posterior calculation in (4.2)

$$\pi_{\theta|Y}(\theta|y) \propto f_{Y|\theta}(y;\theta)\pi(\theta)$$

ignoring terms that do not involve θ. Hence

$$f_Y(y) = \int f_{Y|\theta}(y;\theta)\pi(\theta)d\theta$$

acts as a **normalizing constant** in the posterior density. ■

4.2 The Role of Sufficiency

Direct from 4.2), of $t(Y)$ is a sufficient statistic for θ, it follows that

$$\pi_{\theta|Y}(\theta|y) = \frac{f_{Y|\theta}(y;\theta)\pi(\theta)}{f_Y(y)} = \frac{g(t(y),\theta)h(y)\pi(\theta)}{f_Y(y)} = \left[\frac{h(y)}{f_Y(y)}\right] g(t(y),\theta)\pi(\theta)$$

where $f_{Y|\theta}(y;\theta) = g(t(y),\theta)h(y)$ by the Neyman factorization result. Thus the posterior distribution of θ only depends on the data through $t(y)$.

Lemma 4.2.1 If $t(Y)$ is a sufficient statistic for θ (in the classical sense) then

$$\pi_{\theta|Y}(\theta|y) = \pi_{\theta|Y}(\theta|t(y))$$

for all prior specifications $\pi(\theta)$.

Proof. By definition

$$f_{Y|\theta}(y;\theta) = f_{Y,T|\theta}(y,t;\theta)$$

if $t = t(y)$, and zero otherwise. Thus, by sufficiency,

$$f_{Y|\theta}(y;\theta) = f_{Y|T}(y;t)f_{T|\theta}(t;\theta)$$

and hence

$$\begin{aligned}\pi_{\theta|Y}(\theta|y) &\propto f_{Y|\theta}(y;\theta)\pi(\theta)\\ &= f_{Y|T}(y;t)f_{T|\theta}(t;\theta)\pi(\theta)\\ &\propto f_{T|\theta}(t;\theta)\pi(\theta)\\ &\propto \pi_{\theta|T}(\theta|t)\end{aligned}$$

with the constant of proportionality equal to one, as both sides must integrate to one. ■

Definition 4.2.1 Bayesian Sufficiency: statistic $t(Y)$ is sufficient for θ in the Bayesian sense if

$$\pi_{\theta|Y}(\theta|y) \propto f_{T|\theta}(t;\theta)\pi(\theta)$$

Lemma 4.2.2 For statistic $t(Y)$, $t(Y)$ is sufficient in the Bayesian sense if and only if it is sufficient in the Classical sense.

Proof. For the if, see the previous Lemma. For the only if, suppose

$$\pi_{\theta|Y}(\theta|y) = f_{T|\theta}(t;\theta)\pi(\theta)h(y).$$

By (4.2),

$$\pi_{\theta|Y}(\theta|y) = \frac{f_{Y|\theta}(y;\theta)\pi(\theta)}{f_Y(y)} \Longrightarrow \frac{f_{Y|\theta}(y;\theta)}{f_Y(y)} = \frac{\pi_{\theta|Y}(\theta|y)}{\pi(\theta)} = f_{T|\theta}(t;\theta)h(y).$$

Hence

$$f_{Y|\theta}(y;\theta) = f_{T|\theta}(t;\theta)h(y)f_Y(y) = g(t,\theta)h^*(y)$$

where $g(t,\theta) = f_{T|\theta}(t;\theta)$ and $h^*(y) = h(y)f_Y(y)$. Thus t is sufficient in the classical sense. ■

4.3 Construction of Prior Distributions

In the Bayesian formulation, the prior density $\pi(\theta)$ plays an important role. There are several methods via which the prior can be specified quantitatively.

1. From historical or **training** data
2. By subjective assessment, similar to the subjective assessment of probabilities in elementary probability theory.

3. Matching to a functional form
4. In a **non-informative** or **vague** specification, where the prior probability is spread evenly across the parameter space.

Definition 4.3.1 A **non-informative** prior expresses prior ignorance about the parameter of interest.

Example 4.3.1 Suppose that $\Theta = \{\theta_1, ..., \theta_k\}$ (that is, θ is known to take one of a finite number of possible values). Then a non-informative prior places equal probability on each value, that is,

$$\pi(\theta) = \frac{1}{k}$$

If the parameter space Θ is unbounded, however, a non-informative prior specification is more difficult to construct. A naive prior specification would be

$$\pi(\theta) = \text{constant}$$

Note that this prior specification does not give a valid probability measure, as π does not integrate to 1 on Θ.

Definition 4.3.2 A prior distribution $\pi(\theta)$ for parameter θ is termed **improper** if it does not integrate to 1.

Even for improper priors, (4.2) can be used to compute the posterior density, which itself will often not be improper (that is, the posterior is **proper**) However, if $\phi = g(\theta)$ is a transformation of θ that may of inferential interest, then by elementary transformation results,

$$\pi(\theta) = \text{constant} \Rightarrow \pi(\phi) = \text{constant} \times \left| \frac{d}{d\phi} g^{-1}(\phi) \right|$$

which may **not** be constant, and hence a **non-uniform** prior on $\pi(\phi)$ results. This is perhaps unsatisfactory, and so the following procedure may be preferable.

Definition 4.3.3 Jeffrey's Non-Informative Prior
Consider the prior $\pi(\theta)$ determined by

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

where $I(\theta)$ is Fisher's Information.

Lemma 4.3.1 Jeffrey's prior is invariant under 1-1 transformations

Proof. Let $\phi = \phi(\theta)$ be a 1-1 transformation. Then, by a previous result,

$$I(\phi) = \frac{I(\theta)}{\left\{\phi'(\theta)\right\}^2}$$

and hence by the transformation result

$$\pi(\phi) \propto \sqrt{I(\theta)} \times \frac{1}{\left|\phi'(\theta)\right|} = \sqrt{\frac{I(\theta)}{\left\{\phi'(\theta)\right\}^2}} = \sqrt{I(\phi)}.$$

■

4.4 Scale and Location Parameters

Definition 4.4.1 θ is a **location parameter** if

$$f_{Y|\theta}(y;\theta) = f_{Y|\theta}(y-\theta).$$

θ is a **scale parameter** if

$$f_{Y|\theta}(y;\theta) = \frac{1}{\theta} f_{Y|\theta}\left(\frac{y}{\theta}\right)$$

For a location parameter, it is required to have, for set $A \subset \Theta$

$$\int_A \pi(\theta)d\theta = \int_{A_c} \pi(\theta)d\theta$$

where $A_c = \{\theta : \theta - c \in A\}$ for scalar c. But, therefore, for all c,

$$\int_{A_c} \pi(\theta)d\theta = \int_A \pi(\theta - c)d\theta \Longrightarrow \pi(\theta) = \pi(\theta - c) \Longrightarrow \pi(\theta) = \text{constant}$$

For a scale parameter, it is required to have, for set $A \subset \Theta$

$$\int_A \pi(\theta)d\theta = \int_{A_c} \pi(\theta)d\theta$$

where now $A_c = \{\theta : c\theta \in A\}$ for scalar c. But, therefore, for all c,

$$\int_{A_c} \pi(\theta)d\theta = \int_A \pi(c\theta)d\theta \Longrightarrow \pi(\theta) = c\pi(c\theta) \Longrightarrow \pi(\theta) \propto \frac{1}{\theta}$$

4.5 Bayesian Estimation

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**

$$\widehat{\theta}_B = \mathrm{E}_{\theta|Y}\left[\theta|y\right] = \int \theta \pi_{\theta|Y}(\theta|y)d\theta$$

whereas another is the **posterior mode**

$$\widehat{\theta}_B = \arg\max_{\theta} \pi_{\theta|Y}(\theta|y)$$

and finally the **posterior median** that satisfies

$$\int_{-\infty}^{\widehat{\theta}_B} \pi_{\theta|Y}(\theta|y)d\theta = \frac{1}{2}$$

Definition 4.5.1 A $\mathbf{100(1-\alpha)\%}$**Bayesian Credible Interval** for θ is a subset C of Θ such that

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

Definition 4.5.2 The $\mathbf{100(1-\alpha)\%}$**Highest Posterior Density Bayesian Credible Interval** for θ, subject to

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

is a subset C of Θ such that

$$C = \left\{\theta \in \Theta : \pi_{\theta|Y}(\theta|y) \geq k\right\}$$

where k is the largest constant such that $\mathrm{P}[\theta \in C] \geq 1 - \alpha$.

4.6 Bayesian Hypothesis Testing

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

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

For example,

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

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

$$\frac{\mathrm{P}\left[H_0|y\right]}{\mathrm{P}\left[H_1|y\right]}$$

(the **posterior odds** on H_0)

Example 4.6.1 To test two simple hypothesis

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

define the prior probabilities of H_0 and H_1as π_0 and $\pi_1$respectively. Then, by Bayes Theorem

$$\frac{\mathrm{P}\left[H_0|y\right]}{\mathrm{P}\left[H_1|y\right]} = \frac{\dfrac{f_{Y|\theta}(y;\theta_0)\pi_0}{f_{Y|\theta}(y;\theta_0)\pi_0 + f_{Y|\theta}(y;\theta_1)\pi_1}}{\dfrac{f_{Y|\theta}(y;\theta_1)\pi_1}{f_{Y|\theta}(y;\theta_0)\pi_0 + f_{Y|\theta}(y;\theta_1)\pi_1}} = \frac{f_{Y|\theta}(y;\theta_0)\pi_0}{f_{Y|\theta}(y;\theta_1)\pi_1}$$

4.7 Bayesian Inference and Decision Making

Suppose that, in an inference setting, a decision is to be made, and the decision is selected from some set $\mathcal{D}$ of alternatives. Regarding the parameter space Θ as a set of potential "states of nature", within which the "true" state θ lies.

Definition 4.7.1 Define the **loss function** for decision d and state θ as the loss (or penalty) incurred when the true state of nature is θ and the decision made is d. Denote this loss as

$$L(d, \theta)$$

Definition 4.7.2 With prior $\pi(\theta)$ and no data, the **expected loss** (or the **Bayes loss**) is defined as

$$\mathrm{E}_\theta\left[L(d,\theta)\right] = \int L(d,\theta)\pi(\theta)d\theta$$

The optimal Bayesian decision is

$$d_B = \arg\min_{d\in\mathcal{D}} \mathrm{E}_\theta\left[L(d,\theta)\right]$$

that is, the decision that minimizes the Bayes loss.
If data are available, the optimal decision will intuitively become a function of the data. Suppose now that the decision in light of the data is denoted $\delta(y)$ (a function from $\mathbb{Y}$ to $\mathcal{D}$, and the associated loss is $L(\delta(y),\theta)$

Definition 4.7.3 The **risk** associated with decision $\delta(Y)$ is the expected loss associated with $\delta(Y)$, with the expectation taken over the distribution of Y given θ

$$R_\theta(\delta) = \mathrm{E}_{Y|\theta}\left[L(\delta(Y),\theta)\right] = \int L(\delta(Y),\theta) f_{Y|\theta}(Y;\theta)dY$$

Definition 4.7.4 The **Bayes risk** expected risk $R_\theta(\delta)$ associated with $\delta(Y)$, with the expectation taken over the prior distribution of θ

$$\begin{aligned} R(\delta) &= \mathrm{E}_\theta\left[R_\theta(\delta)\right] = \mathrm{E}_\theta\left[\mathrm{E}_{Y|\theta}\left[L(\delta(Y),\theta)\right]\right] \\ &= \int\left\{\int L(\delta(y),\theta) f_{Y|\theta}(y;\theta)dy\right\}\pi(\theta)d\theta \\ &= \int\int L(\delta(y),\theta) f_Y(y)\pi_{\theta|Y}(\theta|y)dyd\theta \\ &= \int\left\{\int L(\delta(y),\theta)\pi_{\theta|Y}(\theta|y)d\theta\right\} f_Y(y)dy \end{aligned}$$

where

$$f_Y(y) = \int f_{Y|\theta}(y;\theta)\pi(\theta)d\theta$$

Definition 4.7.5 With prior $\pi(\theta)$ and fixed data y, the optimal Bayesian decision, termed the **Bayes rule** is

$$\begin{aligned} d_B &= \arg\min_{\delta\in\mathcal{D}} R(\delta) \\ &= \arg\min_{\delta\in\mathcal{D}} \int\left\{\int L(\delta(y),\theta)\pi_{\theta|Y}(\theta|y)d\theta\right\} f_Y(y)dy \\ &= \arg\min_{\delta\in\mathcal{D}} \int L(\delta(y),\theta)\pi_{\theta|Y}(\theta|y)d\theta \end{aligned}$$

that is, the decision that minimizes the Bayes risk, or equivalently **(posterior) expected loss** in making decision δ, with expectation taken with respect to the posterior distribution $\pi_{\theta|Y}(\theta|y)$.

4.8 Applications of Decision Theory to Estimation

Theorem 4.8.1 *Under squared error loss*

$$L(\delta(y),\theta) = (\delta(y)-\theta)^2$$

the Bayes rule for estimating θ is

$$\delta(y) = \widehat{\theta}_B = E_{\theta|Y}\left[\theta|y\right] = \int \theta\pi_{\theta|Y}(\theta|y)d\theta$$

that is, the posterior expectation.

Proof. The expected posterior loss is

$$\int L(\delta(y),\theta)\pi_{\theta|Y}(\theta|y)d\theta = \int (\delta(y)-\theta)^2\pi_{\theta|Y}(\theta|y)d\theta$$

which needs to be minimized with respect to $\delta(y)$. Set $\widehat{\theta} = \delta(y)$. Then

$$\begin{aligned} \frac{d}{d\widehat{\theta}}\left\{\int\left(\widehat{\theta}-\theta\right)^2\pi_{\theta|Y}(\theta|y)d\theta\right\} &= \int\frac{d}{d\widehat{\theta}}\left\{\left(\widehat{\theta}-\theta\right)^2\right\}\pi_{\theta|Y}(\theta|y)d\theta \\ &= \int 2\left(\widehat{\theta}-\theta\right)\pi_{\theta|Y}(\theta|y)d\theta \end{aligned}$$

and equating this to zero gives

$$\int\left(\widehat{\theta}-\theta\right)\pi_{\theta|Y}(\theta|y)d\theta = 0 \Longrightarrow \widehat{\theta} = \int\theta\pi_{\theta|Y}(\theta|y)d\theta = \mathrm{E}_{\theta|Y}\left[\theta|y\right]$$

■

Theorem 4.8.2 *Under absolute error loss*

$$L(\delta(y), \theta) = |\delta(y) - \theta|$$

the Bayes rule for estimating θ is the solution of

$$\int_{-\infty}^{\delta(y)} \pi_{\theta|Y}(\theta|y)d\theta = \frac{1}{2}$$

that is, the posterior median.

Proof. The expected posterior loss is

$$\int L(\delta(y), \theta)\pi_{\theta|Y}(\theta|y)d\theta = \int |\delta(y) - \theta| \, \pi_{\theta|Y}(\theta|y)d\theta$$

which needs to be minimized with respect to $\delta(y)$. Set $\widehat{\theta} = \delta(y)$. Then

$$\int \left|\widehat{\theta} - \theta\right| \pi_{\theta|Y}(\theta|y)d\theta = \int_{-\infty}^{\widehat{\theta}} \left(\widehat{\theta} - \theta\right) \pi_{\theta|Y}(\theta|y)d\theta + \int_{\widehat{\theta}}^{\infty} \left(\theta - \widehat{\theta}\right) \pi_{\theta|Y}(\theta|y)d\theta$$

Differentiating with respect to $\widehat{\theta}$

$$\int_{-\infty}^{\widehat{\theta}} \pi_{\theta|Y}(\theta|y)d\theta - \int_{\widehat{\theta}}^{\infty} \pi_{\theta|Y}(\theta|y)d\theta$$

and equating this to zero gives

$$\int_{-\infty}^{\widehat{\theta}} \pi_{\theta|Y}(\theta|y)d\theta = \int_{\widehat{\theta}}^{\infty} \pi_{\theta|Y}(\theta|y)d\theta = \frac{1}{2}$$

and hence $\widehat{\theta}$ is the median. ■

4. Appendix A: Probability Distributions

Appendix A Probability Distributions

A.1 DISCRETE PROBABILITY DISTRIBUTIONS

A.1.1 DISCRETE UNIFORM DISTRIBUTION

NOTATION $Y \sim Uniform(n)$ **RANGE** $\mathbb{Y} = \{1, 2, ..., n\}$
MASS FUNCTION

$$f_Y(x) = \frac{1}{n} \quad \text{for} \quad x \in \{1, 2, ..., n\}$$

CDF

$$F_Y(x) = \frac{x}{n} \quad \text{for} \quad x \in \{1, 2, ..., n\}$$

MGF

$$\begin{aligned} M_Y(t) &= \sum_{x=1}^{n} e^{tx} \frac{1}{n} \\ &= \frac{e^t}{n} \left[1 + e^t + ... + e^{(n-1)t}\right] \\ &= \frac{e^t}{n} \frac{[1 - e^{nt}]}{[1 - e^t]} \end{aligned}$$

rth MOMENT

$$M_Y^{(r)}(t) = \sum_{x=1}^{n} x^r \; e^{tx} \frac{1}{n} \qquad \Longrightarrow M_Y^{(r)}(0) = \frac{1}{n} \sum_{x=1}^{n} x^r$$

$$\Longrightarrow \mathrm{E}_{f_Y}[Y] = M_Y^{(1)}(0) = \frac{1}{n} \sum_{x=1}^{n} x = \frac{(n+1)}{2}$$

$$\mathrm{E}_{f_Y}[Y^2] = M_Y^{(2)}(0) = \frac{1}{n} \sum_{x=1}^{n} x^2 = \frac{(n+1)(2n+1)}{6}$$

$$\Longrightarrow \mathrm{Var}_{f_Y}[Y] = \mathrm{E}_{f_Y}[Y^2] - \{\mathrm{E}_{f_Y}[Y]\}^2 = \frac{(n+1)(n-1)}{12}$$

Note

We can define a discrete uniform distribution over any finite set of values rather that merely the integers $\{1, 2, ..., n\}$; in this case, the moments of the distribution will depend on the nature of the range $\mathbb{Y}$, but in the same techniques for calculation of moments, mgf etc. can be used.

A.1.2 BERNOULLI DISTRIBUTION

NOTATION $Y \sim Bernoulli(\theta)$
RANGE $\mathbb{Y} = \{0, 1\}$
MASS FUNCTION

$$f_Y(y) = \theta^y(1-\theta)^{1-y} \qquad for \qquad y \in \{0,1\} \qquad \text{where} 0 \leq \theta \leq 1.$$

CDF

$$F_Y(x) = \begin{cases} 0 & x < 0 \\ 1-\theta & for \quad x = 0 \\ 1 & x \geq 1 \end{cases}$$

MGF

$$M_Y(t) \ = \sum_{y=0}^{1} e^{ty}\theta^y(1-\theta)^{1-y} \ = 1 - \theta + \theta e^t$$

rth MOMENT

$$\begin{aligned} M_Y^{(r)}(t) \quad &= \quad \theta e^t \Longrightarrow M_Y^{(r)}(0) = \theta \qquad r = 1, 2, ... \\ &\Longrightarrow \quad \mathrm{E}_{f_Y}[Y] = \theta \qquad \mathrm{E}_{f_Y}[Y^2] = \theta \\ &\Longrightarrow \quad \mathrm{Var}_{f_Y}[Y] = \ \theta - \theta^2 \ = \ \theta(1-\theta) \end{aligned}$$

Note:
(1) The Bernoulli distribution is used for modelling when the outcome of an experiment is either a "success" or a 'failure", where the probability of getting a success is equal to θ.
(2) The mass function of Y can be re-written

$$\mathrm{P}[Y = y] = \exp \qquad \{t(y)q(\theta) + \log h(y) + \log c(\theta)\}$$

where $t(y) = y$, $q(\theta) = \log(\theta/(1-\theta))$, $h(y) = 1$, $c(\theta) = 1 - \theta$.

A.1.3 BINOMIAL DISTRIBUTION

NOTATION $Y \sim Bin(n, \theta)$
RANGE $\mathbb{Y} = \{0, 1, 2, ..., n\}$
MASS FUNCTION

$$f_Y(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad for \qquad x \in \{0,1,2,...,n\} \qquad \text{where} n \geq 0, \ 0 \leq \theta \leq 1.$$

MGF

$$\begin{aligned} M_Y(t) &= \sum_{y=0}^{n} e^{ty}\binom{n}{y}\theta^y(1-\theta)^{n-y} \\ &= \sum_{y=0}^{n}\binom{n}{y}\left(\theta e^t\right)^y(1-\theta)^{n-y} \\ &= \left(1-\theta+\theta e^t\right)^n \end{aligned}$$

***r*th MOMENT**

No simple general expression for $M_Y^{(r)}(t)$, but

$$\begin{aligned} M_Y^{(1)}(t) &= n\theta e^t\left(1-\theta+\theta e^t\right)^{n-1} \\ M_Y^{(2)}(t) &= n(n-1)\left\{\theta e^t\right\}^2\left(1-\theta+\theta e^t\right)^{n-2} + n\theta e^t\left(1-\theta+\theta e^t\right)^{n-1} \end{aligned}$$

so that $M_Y^{(1)}(0) = n\theta$ and $M_Y^{(2)}(0) = n(n-1)\theta^2 + n\theta$, and thus

$$\mathrm{E}_{f_Y}[Y] = n\theta \qquad \mathrm{Var}_{f_Y}[Y] = n(n-1)\theta^2 + n\theta - n^2\theta^2 = n\theta(1-\theta)$$

Notes

(1) If $Y_1, ..., Y_k$ are independent and identically distributed (i.i.d.) $Bernoulli(\theta)$ random variables, and $Y = Y_1 + ..., Y_k$, then by the standard result for mgfs,

$$M_Y(t) = \{M_Y(t)\}^k = \left(1-\theta+\theta e^t\right)^k$$

so therefore $Y \sim Bin(k, \theta)$ because of the uniqueness of mgfs. Thus the binomial distribution is used to model the **total** number of successes in a series of independent and identical experiments.
(2) Alternatively, consider **sampling without replacement** from **infinite** collection, or **sampling with replacement** from a **finite** collection of objects, a proportion θ of which are of Type I, and the remainder are of Type II. If Y is the number of Type I objects in a sample of n, $Y \sim Bin(n, \theta)$.
(3) The mass function of Y can be re-written

$$\mathrm{P}[Y = y] = \exp \qquad \{t(y)q(\theta) + logh(y) + logc(\theta)\}$$

where

$$t(y) = y, q(\theta) = \log\left(\frac{\theta}{1-\theta}\right), h(y) = \binom{n}{y}, c(\theta) = (1-\theta)^n.$$

A.1.4 POISSON DISTRIBUTION

NOTATION $Y \sim Poisson(\lambda)$
RANGE $\mathbb{Y} = \{0, 1, 2, ..\}$
MASS FUNCTION

$$f_Y(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad for \qquad x \in \{0, 1, 2, ...\} \qquad \text{where} \lambda > 0.$$

MGF

$$M_Y(t) = \sum_{x=0}^{\infty} e\frac{e^{-\lambda}\lambda^x}{x!} \;=\; e^{-\lambda}\sum_{x=0}^{\infty}\frac{\left(\lambda e^t\right)^x}{x!} = e^{-\lambda}e^{\lambda e^t} \;=\; \exp\left\{\lambda\left(e^t-1\right)\right\}$$

rth MOMENT

No simple general expression for $M_Y^{(r)}(t)$, but

$$M_Y^{(1)}(t) = \;\lambda e^t \exp\left\{\lambda\left(e^t-1\right)\right\} \qquad M_Y^{(2)}(t) = \left(\lambda e^t\right)^2 \exp\left\{\lambda\left(e^t-1\right)\right\} + \lambda e^t \exp\left\{\lambda\left(e^t-1\right)\right\}$$

so that $M_Y^{(1)}(0) = \lambda$ and $M_Y^{(2)}(0) = \lambda^2 + \lambda$, and thus

$$\mathrm{E}_{f_Y}[Y] = \;\lambda \qquad \mathrm{Var}_{f_Y}[Y] = \;\lambda^2 + \lambda - \lambda^2 = \lambda$$

Notes

(1) If $Y \sim Bin(n, \theta)$, let $\lambda = n\theta$. Then

$$M_Y(t) = \left(1-\theta+\theta e^t\right)^n = \left(1+\lambda(e^t-1))/n\right)^n \longrightarrow \exp \qquad \left\{\lambda\left(e^t-1\right)\right\}$$

as $n \longrightarrow \infty$, which is the mgf of a Poisson random variable. Therefore, the Poisson distribution arises as the limiting case of the binomial distribution, when $n \longrightarrow \infty$,$\theta \longrightarrow 0$ with $n\theta = \lambda$ constant (that is, for "large" n and "small" θ).

(2) Suppose that Y_1 and Y_2 are independent, with $Y_1 \sim Poisson(\lambda_1)$, $Y_2 \sim Poisson(\lambda_2)$, then if $Y = Y_1 + Y_2$, using the general mgf result for independent random variables,

$$M_Y(t) = \;M_{Y_1}(t)M_{Y_2}(t) = \exp \qquad \left\{\lambda_1\left(e^t-1\right)\right\}\exp \qquad \left\{\lambda_2\left(e^t-1\right)\right\} = \exp \qquad \left\{\left(\lambda_1+\lambda_2\right)\left(e^t-1\right)\right\}$$

so that $Y \sim Poisson(\lambda_1 + \lambda_2)$. Therefore, the sum of two independent Poisson random variables also has a Poisson distribution.

By induction, this result can be extended to give the distribution of the sum of k independent random Poisson random variables, that is, if $Y_1, ..., Y_k$ are independent random variables with $Y_i \sim Poisson(\lambda_i)$ for $i = 1, ..., k$, then

$$Y = \sum_{i=1}^{k} Y_i \Longrightarrow Y \sim Poisson\left(\sum_{i=1}^{k}\lambda_i\right).$$

(3) **The Poisson Process**

Consider an experiment involving events that occur repeatedly in time. Let $Y(t)$ be the random variable representing the number of events that occur in the interval $(0, t]$, so that $Y(t)$ takes values $0, 1, 2, ...$ Suppose that

1. $Y(0) = 0$
2. For all $0 < s \leq t$, $h > 0$, and non-negative integers n and m,

$$\mathrm{P}[Y(t+h) - Y(t) = n | Y(s) = m] = \mathrm{P}[Y(t+h) - Y(t) = n]$$

that is, the numbers of events occurring in disjoint intervals are probabilistically **independent**.

3. For $\delta t > 0$ small,

$$\mathrm{P}[Y(t+\delta t) - Y(t) = 1] = \lambda \delta t + O(\delta t)$$

for some $\lambda > 0$, where

$$\lim_{\delta t \longrightarrow 0} \frac{O(\delta t)}{\delta t} = 0$$

that is, the probability of **exactly one** event occurring in the small interval $(t, t+\delta t]$ is, for small δt, proportional to the length of the interval, δt.

4. For $\delta t > 0$ small,

$$\mathrm{P}[Y(t+\delta t) - Y(t) \geq 2] = O(\delta t)$$

for some $\lambda > 0$, that is, the probability of **more than one** event occurring in a small interval $(t, t+\delta t]$ is essentially zero.
Then, if

$$P_n(t) = \mathrm{P}[n \text{ events occur in } (0, t]]$$

it can be shown that

$$P_n(t) = P[Y(t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$$

(that is, the random variable corresponding to the number of events that occurs in the interval $(0, t]$ has a Poisson distribution with parameter λt.)

Examples : Failures/breakdowns of mechanical components, occurrence of accidents, emission of particles from radioactive sources etc.

(4) The mass function of Y can be re-written

$$\mathrm{P}[Y = y] = \exp \qquad \{t(y)q(\lambda) + \log h(y) + \log c(\lambda)\}$$

where

$$t(y) = y, q(\lambda) = \log \lambda, h(y) = \frac{1}{y!}, c(\lambda) = e^{-\lambda}$$

.

A.1.5 GEOMETRIC DISTRIBUTION

NOTATION $Y \sim Geometric(\theta)$ **RANGE** $\mathbb{Y} = \{1, 2, ...\}$
MASS FUNCTION

$$f_Y(x) = (1-\theta)^{x-1}\theta \qquad for \qquad x \in \{1, 2, ...\} \qquad \text{where} 0 \leq \theta \leq 1.$$

CDF

$$F_Y(x) = 1 - (1-\theta)^x \qquad for \qquad x = 1, 2, ...$$

MGF

$$M_Y(t) = \sum_{x=1}^{\infty} e^{tx}(1-\theta)^{x-1}\theta = \theta e^t \sum_{x=1}^{\infty} e^{t(x-1)}(1-\theta)^{x-1} = \theta e \sum_{x=0}^{\infty} \left(e^t(1-\theta)\right)^x = \frac{\theta e^t}{1 - e^t(1-\theta)}$$

rth MOMENT
No simple general expression for $M_Y^{(r)}(t)$, but

$$M_Y^{(1)}(t) = \frac{\theta e^t}{\left[1 - e^t(1-\theta)\right]^2} \qquad M_Y^{(2)}(t) = \frac{\theta e^t \left[1 - e^t(1-\theta)\right]\left[1 + e^t(1-\theta)\right]}{\left[1 - e^t(1-\theta)\right]^4}$$

so that $M_Y^{(1)}(0) = \frac{1}{\theta}$ and $M_Y^{(2)}(0) = \frac{2-\theta}{\theta^2}$, and thus

$$\mathrm{E}_{f_Y}[Y] = \frac{1}{\theta} \qquad \mathrm{Var}_{f_Y}[Y] = \frac{2-\theta}{\theta^2} - \frac{1}{\theta^2} = \frac{1-\theta}{\theta^2}$$

Notes
(1) If $Y \sim Geometric(\theta)$, then for $x, j \geq 1$,

$$\begin{aligned} \mathrm{P}[Y = y + j | Y > j] &= \frac{\mathrm{P}[Y = y + j, Y > j]}{\mathrm{P}[Y > j]} \\ &= \frac{\mathrm{P}[Y = y + j]}{\mathrm{P}[Y > j]} \\ &= \frac{(1-\theta)^{y+j-1}\theta}{(1-\theta)^j} \\ &= (1-\theta)^{y-1}\theta \\ &= \mathrm{P}[Y = y] \end{aligned}$$

So $\mathrm{P}[Y = y + j | Y > j] = \mathrm{P}[Y = y]$. This property is unique (among **discrete** distributions) to the geometric distribution, and is called the **lack of memory** property.
(2) The geometric distribution is used to model the number, Y, of independent, identical Bernoulli trials until the first success is obtained. It is a discrete **waiting time** distribution.
(3) The mass function of Y can be re-written

$$\mathrm{P}[Y = y] = \exp \qquad \{t(y)q(\theta) + logh(y) + logc(\theta)\}$$

A.1.6 NEGATIVE BINOMIAL DISTRIBUTION

NOTATION $Y \sim NeBi(n, \theta)$
RANGE $\mathbb{Y} = \{n, n+1, n+2, ...\}$
MASS FUNCTION

$$f_Y(y) = \binom{y-1}{n-1}\theta^n(1-\theta)^{y-n} \qquad \text{where } n \in \{1, 2, 3, ...\},\ 0 \leq \theta \leq 1.$$

MGF

$$\begin{aligned} M_Y(t) &= \sum_{y=n}^{\infty} e^{ty}\binom{y-1}{n-1}\theta^n(1-\theta)^{y-n} \\ &= \left(\theta e^t\right)^n \sum_{y=n}^{\infty}\binom{y-1}{n-1}\left(e^t(1-\theta)\right)^{y-n} \\ &= \left\{\frac{\theta e^t}{1-e^t(1-\theta)}\right\}^n \end{aligned}$$

rth MOMENT No simple general expression for $M_Y^{(r)}(t)$, but

$$\begin{aligned} M_Y^{(1)}(t) &= \frac{n(\theta e^t)^n}{[1-e^t(1-\theta)]^{n+1}} \\ M_Y^{(2)}(t) &= \frac{n(\theta e^t)^n\left[n+e^t(1-\theta)\right]}{[1-e^t(1-\theta)]^{n+2}} \end{aligned}$$

so that

$$M_Y^{(1)}(0) = \frac{n}{\theta} \qquad M_Y^{(2)}(0) = \frac{n(n+(1-\theta))}{\theta^2}$$

and thus

$$\mathrm{E}_{f_Y}[Y] = \frac{n}{\theta} \qquad \mathrm{Var}_{f_Y}[Y] = \frac{n(n+(1-\theta))}{\theta^2} - \frac{n^2}{\theta^2} = \frac{n(1-\theta)}{\theta^2}$$

Notes
(1) If $Y \sim Bin(n, \theta)$, $Y \sim NeBi(r, \theta)$, then for $r \leq n$, $\mathrm{P}[Y \geq r] = \mathrm{P}[Y \leq n]$.
(2) The Negative Binomial distribution is used to model the number, Y, of independent, identical Bernoulli trials needed to obtain **exactly** n successes.
(3) Alternative representation: let Y be the number of **failures** in a sequence of independent, identical Bernoulli trials that contains **exactly** n successes. Then $Y = Y - n$, and hence

$$f_Y(y) = \binom{n+y-1}{n-1}\theta^n(1-\theta)^y \qquad y \in \{0, 1, ...\}$$

(4) If $Y_i \sim Geometric(\theta)$, for $i = 1, ...n$, are i.i.d. random variables, and $Y = Y_1 + ... + Y_n$, then $Y \sim NeBi(n, \theta)$ (result immediately follows using mgfs).
(5) If $Y \sim NeBi(n, \theta)$, let $n(1-\theta) = \lambda$ and $Y = Y - n$. Then

$$M_Y(t) = e^{-nt}M_Y(t) = \left\{\frac{\theta}{1-e^t(1-\theta)}\right\}^n = \left\{1 + \lambda\frac{(e^t-1)}{n-\lambda e^t}\right\}^n \longrightarrow \exp\left\{\lambda(e^t-1)\right\}$$

as $n \longrightarrow \infty$, hence the alternate form of the negative binomial distribution tends to the Poisson distribution as $n \longrightarrow \infty$ with $n(1-\theta) = \lambda$ constant.

(6) The mass function of Y can be re-written

$$\mathrm{P}[X = y] = \exp \qquad \{t(y)q(\theta) + \log h(y) + \log c(\theta)\}$$

A.2 CONTINUOUS PROBABILITY DISTRIBUTIONS

A.2.1 CONTINUOUS UNIFORM DISTRIBUTION

NOTATION $Y \sim Uniform(a, b)$
RANGE $\mathbb{Y} = [a, b]$ or (a, b), for $a \leq b$
PDF

$$f_Y(x) = \frac{1}{b-a} \qquad a \leq x \leq b$$

CDF

$$F_Y(x) = \frac{x-a}{b-a} \qquad a \leq x \leq b$$

MGF

$$M_Y(t) \; = \int_a^b e^{ty} \frac{1}{b-a} \; dy \; = \; \frac{(e^{tb} - e^{ta})}{t(b-a)} \qquad if \; t \neq 0$$

and $M_Y(0) = 1$.
rth MOMENT

$$\mathrm{E}_{f_Y}[Y^r] \; = \int_a^b y^r \frac{1}{b-a} \; dy \; = \left[\; \frac{b^{r+1}}{r+1} - \frac{a^{r+1}}{r+1} \right]$$

so therefore

$$\begin{aligned} \mathrm{E}_{f_Y}[Y] &= \frac{1}{b-a} \left[\frac{b^2 - a^2}{2} \right] = \frac{(a+b)}{2} \\ \mathrm{E}_{f_Y}[Y^2] &= \frac{1}{b-a} \left[\frac{b^3 - a^3}{3} \right] = \frac{(a^2 + ab + b^2)}{3} \\ \Longrightarrow \mathrm{Var}_{f_Y}[Y] &= \frac{(a^2 + ab + b^2)}{3} - \left\{ \frac{(a+b)}{2} \right\}^2 = \frac{(a-b)^2}{12} \end{aligned}$$

A.2.2 EXPONENTIAL DISTRIBUTION

NOTATION $Y \sim Exp(\lambda)$
RANGE $\mathbb{Y} = \mathbb{R}^+$
PDF

$$f_Y(y) = \lambda e^{-\lambda y} \qquad y > 0 \qquad \lambda > 0.$$

CDF

$$F_Y(y) = 1 - e^{-\lambda y} \qquad y > 0$$

MGF

$$\begin{aligned} M_Y(t) &= \int_0^\infty e^{tx}\lambda e^{-\lambda x}\,dx \\ &= \lambda\int_0^\infty e^{-(\lambda-t)x}\,dx \\ &= \frac{\lambda}{\lambda-t} \qquad \text{for } t<\lambda \end{aligned}$$

rth MOMENT

$$M_Y^{(r)}(t) = \frac{r!\lambda}{(\lambda-t)^{r+1}} \Longrightarrow M_Y^{(r)}(0) = \frac{r!}{\lambda^r}, \qquad \mathrm{E}_{f_Y}[Y] = \frac{1}{\lambda}, \mathrm{E}_{f_Y}[Y^2] = \frac{2}{\lambda^2} \Longrightarrow \mathrm{Var}_{f_Y}[Y] = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Notes
(1) Alternative representation uses $\theta = 1/\lambda$ as the parameter of the distribution.
(2) If $Y \sim Exp(\lambda)$, then, for all $x, t > 0$,

$$\begin{aligned} \mathrm{P}[Y > y+t | Y > t] &= \frac{\mathrm{P}[Y > y+t, Y > t]}{\mathrm{P}[Y > t]} \\ &= \frac{\mathrm{P}[Y > y+j]}{\mathrm{P}[Y > t]} \\ &= \frac{e^{-\lambda(y+t)}}{e^{-\lambda t}} = e^{-\lambda y} \\ &= \mathrm{P}[\ Y > y\ ] \end{aligned}$$

Thus, for all $x, t > 0$, $\mathrm{P}[\ Y > x+t \mid Y > t\ ] = \mathrm{P}[\ Y > x\ ]$ - this is known as the **Lack of Memory Property**, and is unique to the exponential distribution amongst continuous distributions.
(3) Suppose that $Y(t)$ is a Poisson process with rate parameter $\lambda > 0$, so that

$$\mathrm{P}[Y(t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$$

Let $Y_1, ..., Y_n$ be random variables defined by Y_1 = "time that first event occurs", and, for $i = 2, ..., n$, Y_i = "time interval between occurrence of $(i-1)$st and ith events". Then $Y_1, ..., Y_n$ are i.i.d. $Exp(\lambda)$.

Proof : $Y_1, ..., Y_n$ are i.i.d. because of the assumption 2. underlying the Poisson process. So consider the distribution of Y_1; in particular, consider the probability $\mathrm{P}[Y_1 > y]$ for $y > 0$. The event $[Y_1 > y]$ is equivalent to the event "No events occur in the interval (0,y]", which has probability $e^{-\lambda y}$. But

$$F_{Y_1}(y) = \mathrm{P}[\ Y_1 \le y\ ] = 1 - \mathrm{P}[\ Y_1 > y\ ] = 1 - e^{-\lambda y} \Longrightarrow Y_1 \sim Eyp(\lambda).$$

(4) The exponential distribution is used to model failure times in continuous time. It is a continuous **waiting time** distribution, the continuous analogue of the geometric distribution.
(5) If $Y \sim Uniform(0,1)$, and $Y = -\log(1-Y)/\lambda$, then $Y \sim Eyp(\lambda)$.
(6) If $Y \sim Exp(\lambda)$, then $X = Y^{1/\alpha}$ for $\alpha > 0$ has a (two-parameter) **Weibull** distribution, and

$$f_X(x) = \alpha\lambda x^{\alpha-1}e^{-\lambda x^\alpha} \qquad x > 0$$

(7) The density function of Y can be re-written

$$f_Y(y) = \exp \qquad \{t(y)q(\lambda) + logh(y) + logc(\lambda)\}$$

where $t(y) = -y$, $q(\lambda) = \lambda$, $h(y) = 1$, $c(\lambda) = \lambda$.

A.2.3 GAMMA DISTRIBUTION

NOTATION $Y \sim Ga(\alpha, \beta)$ **RANGE** $\mathbb{Y} = \mathbb{R}^+$
PDF

$$f_Y(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad x > 0 \qquad \text{where} \alpha, \beta > 0.$$

and where, for any real number $\alpha > 0$,

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$$

is the **gamma function**, $\Gamma(.)$.
MGF

$$\begin{aligned} M_Y(t) &= \int_0^\infty e^{ty} \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}\, dy \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty y^{\alpha-1}\, e^{-(\beta-t)y}\, dy \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\beta-t)^\alpha} = \left(\frac{\beta}{\beta-t}\right)^\alpha \end{aligned}$$

rth MOMENT
No simple general expression for $M_Y^{(r)}(t)$, but

$$M_Y^{(1)}(t) = \frac{\alpha\beta^\alpha}{(\beta-t)^{\alpha+1}} \qquad M_Y^{(2)}(t) = \frac{\alpha(\alpha+1)\beta^\alpha}{(\beta-t)^{\alpha+2}}$$

so that

$$M_Y^{(1)}(0) = \frac{\alpha}{\beta} \qquad M_Y^{(2)}(0) = \frac{\alpha(\alpha+1)}{\beta^2}$$

and thus

$$\mathrm{E}_{f_Y}[Y] = \frac{\alpha}{\beta} \qquad \mathrm{Var}_{f_Y}[Y] = \frac{\alpha(\alpha+1)}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}$$

Notes
(1) If $Y_1 \sim Ga(\alpha_1, \beta), Y_2 \sim Ga(\alpha_2, \beta)$ are independent random variables, and $Y = Y_1 + Y_2$, then $Y \sim Ga(\alpha_1 + \alpha_2, \beta)$ (directly from properties of mgfs).

(2) $Ga(1,\beta) \equiv Exp(\beta)$.
(3) If $Y_1, ..., Y_n \sim Exp(\lambda)$ are independent random variables, and $Y = Y_1 + ... + Y_n$, then $Y \sim Ga(n, \lambda)$ (directly from (1) and (2)).
(4) For $\alpha > 0$,

$$\begin{aligned}\Gamma(\alpha) &= \int_0^\infty t^{\alpha-1}e^{-t}\,dt \\ &= \left[-t^{\alpha-1}e^{-t}\right]_0^\infty + \int_0^\infty (\alpha-1)t^{\alpha-2}e^{-t}\,dt \\ &= (\alpha-1)\int_0^\infty t^{\alpha-2}e^{-t}\,dt \;=\; (\alpha-1)\Gamma(\alpha-1)\end{aligned}$$

so $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$. Thus if $\alpha = 1, 2, ...$, then $\Gamma(\alpha) = (\alpha-1)!$.
(5) **Special Case :** If $\alpha = 1, 2, ...$ the $Ga(\alpha/2, 1/2)$ distribution is also known as the **Chi-squared distribution with α degrees of freedom**
(6) If $Y_1 \sim \chi^2_{n_1}$ and $Y_2 \sim \chi^2_{n_2}$ are independent Chi-squared random variables with n_1 and n_2 degrees of freedom respectively, then random variable defined as the ratio $(Y_1/n_1)/(Y_2/n_2)$ has an **F-distribution with (n_1, n_2) degrees of freedom**.
(7) The density function of Y can be re-written

$$f_Y(y) = \exp \qquad \left\{\mathbf{t}(\mathbf{y})^T\mathbf{q}(\boldsymbol{\psi}) + logh(y) + logc(\boldsymbol{\psi})\right\}$$

where

$$\mathbf{t}(\mathbf{y}) = \begin{pmatrix}\log y \\ y\end{pmatrix}, \mathbf{q}(\boldsymbol{\psi}) = \begin{pmatrix}\alpha-1 \\ -\beta\end{pmatrix}, h(y) = 1, c(\boldsymbol{\psi}) = \frac{\beta^\alpha}{\Gamma(\alpha)}$$

.

A.2.4 BETA DISTRIBUTION

NOTATION $Y \sim Be(\alpha, \beta)$
RANGE $\mathbb{Y} = (0, 1)$
PDF

$$f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, y^{\alpha-1}(1-y)^{\beta-1} \qquad 0 < y < 1 \text{ where} \alpha, \beta > 0.$$

rth MOMENT
For $r = 1, 2, ...$,

$$\begin{aligned}\mathrm{E}_{f_Y}[Y^r] &= \int_0^1 y^r \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 y^{r+\alpha-1}\,(1-y)^{\beta-1}\,dy \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \times \frac{\Gamma(r+\alpha)\Gamma(\beta)}{\Gamma(r+\alpha+\beta)} \\ \Longrightarrow \mathrm{E}_{f_Y}[Y] &= \frac{\alpha}{\alpha+\beta} \qquad \mathrm{E}_{f_Y}[Y^2] = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \\ \Longrightarrow \mathrm{Var}_{f_Y}[Y] &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\end{aligned}$$

Notes
(1) The beta distribution arises naturally in the context of order statistics; if $Y_1, ..., Y_k$ are i.i.d. random variables with CDF F_Y, say, consider first the random variables $U_1, ..., U_k$ defined by $U_i = F_Y(Y_i)$ for $i = 1, ..., k$. It can be shown that $U_1, ..., U_k$ are i.i.d. $Uniform(0, 1)$ random variables. Now, consider the order statistics $Y_1, ..., Y_k$ derived from $U_1, ..., U_k$; using previous results, it can be shown that the marginal distribution of the jth order statistic is $Be(j, k - j + 1)$, for $j = 1, ..., k$.
(2) If $Y_1 \sim Ga(\alpha_1, \beta)$, $Y_2 \sim Ga(\alpha_2, \beta)$ are independent random variables, and $Y = Y_1/(Y_1 + Y_2)$, then $Y \sim Be(\alpha_1, \alpha_2)$ (using standard multivariate transformation techniques).
(3) Suppose that random variables X and Y have a joint probability distribution such that the **conditional** distribution of Y, given $Y = y$ for $0 < y < 1$, is binomial, $Bin(n, y)$, and the **marginal** distribution of Y is beta, $Be(\alpha, \beta)$, so that

$$f_{X|Y}(x|y) = y^x(1-y)^{n-x} \qquad x = 0, 1, ..., n \qquad f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \qquad 0 < y < 1.$$

Then the **marginal** distribution of Y is given by

$$f_X(x) = \int_0^1 f_{Y|Y}(x|y) f_Y(y)\, dy = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+\alpha+\beta)} \quad x = 0, 1, 2...n$$

(3) If $\alpha = \beta = 1$, $Be(\alpha, \beta) \equiv Uniform(0, 1)$.
(4) The density function of Y can be re-written

$$f_Y(y) = \exp \qquad \left\{ \mathbf{t}(x)^T \mathbf{q}(\psi) + \log h(y) + \log c(\psi) \right\}$$

where

$$\mathbf{t}(y) = \begin{pmatrix} \log y \\ \log(1-y) \end{pmatrix}, \mathbf{q}(\boldsymbol{\psi}) = \begin{pmatrix} \alpha - 1 \\ \beta - 1 \end{pmatrix}, h(y) = 1, c(\boldsymbol{\psi}) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

.

A.2.5 NORMAL DISTRIBUTION

NOTATION $Y \sim N(\mu, \sigma^2)$
RANGE $\mathbb{Y} = \mathbb{R}$
PDF

$$f_Y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \qquad \left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \qquad x \in \mathbb{R} \qquad \text{where } \mu \in \mathbb{R}, \sigma > 0.$$

MGF

$$\begin{aligned} M_Y(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} dx \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 + tx \right\} dx \\ &= \exp\left\{ \mu t + \frac{t^2\sigma^2}{2} \right\} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}\left(x - (\mu + t\sigma^2)\right)^2 \right\} dx \\ &= \exp\left\{ \mu t + \frac{t^2\sigma^2}{2} \right\} \end{aligned}$$

because the integrand is a PDF , and thus the integral is equal to one.

rth MOMENT

No simple general expression for $M_Y^{(r)}(t)$, but

$$M_Y^{(1)}(t) = (\mu + t\sigma^2)\exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\} \qquad M_Y^{(2)}(t) = (\mu^2 + 2t\sigma^2\mu + t^2\sigma^4 + \sigma^2)\exp\left\{\mu t + \frac{t^2\sigma^2}{2}\right\}$$

so that $M_Y^{(1)}(0) = \mu$ and $M_Y^{(2)}(0) = \mu^2 + \sigma^2$, and thus

$$\mathrm{E}_{f_Y}[Y] = \mu \qquad \mathrm{Var}_{f_Y}[Y] = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$$

Notes

(1) Special Case : If $\mu = 0$, $\sigma^2 = 1$, then Y has a **Standard** or **Unit** normal distribution. Usually, the PDF of the unit normal is written $\phi(x)$, and the CDF is written $\Phi(x)$.

(2) If $Y \sim N(0,1)$, and $Y = \sigma Y + \mu$, then $Y \sim N(\mu, \sigma^2)$. Re-expressing this result, if $Y \sim N(\mu, \sigma^2)$, and $Y = (Y - \mu)/\sigma$, then $Y \sim N(0,1)$. (using transformation or mgf techniques)

(3) **The Central Limit Theorem:** Suppose $Y_1, ..., Y_n$ are i.i.d. random variables with mgf M_Y, with $\mathrm{E}_{f_Y}[Y_i] = \mu$ and $\mathrm{Var}_{f_Y}[Y_i] = \sigma^2$ that is, the mgf and the expectation and variance of the Y_is are specified, but the PDF is not. Let the random variable Z_n be defined by

$$Z_n = \frac{\sum_{i=1}^{n} Y_i - n\mu}{\sqrt{n\sigma^2}}$$

and let Z_n have mgf M_{Z_n}. Then, as $n \longrightarrow \infty$, $M_{Z_n}(t) \longrightarrow \exp\left\{t^2/2\right\}$ **irrespective of the distribution of the Y_is**, that is, the distribution of Z_n tends to a **unit normal** distribution as n tends to infinity.

(4) If $Y \sim N(0,1)$, and $Y = Y^2$, then $Y \sim \chi_1^2$, so that the square of a unit normal random variable has a chi-square distribution with 1 degree of freedom.

(5) If $Y \sim N(0,1)$, and $Y \sim N(0,1)$ are independent random variables, and Z is defined by $Z = Y/Y$, the Z has a **Cauchy** distribution

$$f_Z(z) = \frac{1}{\pi\,(1 + z^2)} \qquad z \in \mathbb{R}$$

(6) If $Y \sim N(0,1)$, and $Y \sim Ga(n/2, 1/2)$ for $n = 1, 2, ...$ (so that $Y \sim \chi_n^2$), are independent random variables, and T is defined by

$$T = \frac{Y}{\sqrt{Y/n}}$$

then T has a **Student-t** distribution with n degrees of freedom, $T \sim St(n)$,

$$f_T(t) = \frac{\Gamma\left(\dfrac{n+1}{2}\right)}{\Gamma\left(\dfrac{n}{2}\right)}\left(\frac{1}{n\pi}\right)^{1/2}\left\{\frac{1}{1 + t^2/n}\right\}^{(n+1)/2} \qquad t \in \mathbb{R}$$

Taking limiting cases of the Student-t distribution

$$\begin{array}{ll} n \longrightarrow \infty & St(n) \longrightarrow N(0,1) \\ n \longrightarrow 1 & St(n) \longrightarrow Cauchy \end{array}$$

(7) The density function of Y can be re-written

$$f_Y(x) = \exp\left\{\mathbf{t}(\mathbf{x})^T\mathbf{q}(\boldsymbol{\psi}) + \log h(x) + \log c(\boldsymbol{\psi})\right\}$$

where $\mathbf{t}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$, $\mathbf{q}(\boldsymbol{\psi}) = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}$, $h(x) = 1$, and

$$c(\boldsymbol{\psi}) = (2\pi\sigma^2)^{1/2} \exp\left\{\frac{-\mu^2}{2\sigma^2}\right\}$$

.

A.3 MULTIVARIATE PROBABILITY DISTRIBUTIONS

A.3.1 MULTINOMIAL DISTRIBUTION

The **multinomial** distribution is a multivariate generalization of the **binomial** distribution. Recall that the binomial distribution arose from an infinite Urn model with two types of objects being sampled without replacement. Suppose that the proportion of "Type 1" objects in the urn is θ (so $0 \leq \theta \leq 1$) and hence the proportion of "Type 2" objects in the urn is $1-\theta$. Suppose that n objects are sampled, and Y is the random variable corresponding to the number of "Type 1" objects in the sample. Then $Y \sim \text{Bin}(n, \theta)$, and

$$f_Y(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad for \qquad x \in \{0, 1, 2, ..., n\}$$

Now consider a generalization; suppose that the Urn contains $k+1$ types of objects ($k = 1, 2, ...$), with

$$\text{"Proportion of type } i\text{" objects } = \theta_i$$

for $i = 1, ..., k+1$. Let Y_i be the random variable corresponding to the number of type i objects in a sample of size n, for $i = 1, ..., k$. Then the joint distribution of vector $Y = (Y_1, ..., Y_k)$ is given by

$$f_{Y_1,...,Y_k}(x_1, ..., x_k) = \frac{n!}{x_1!...x_k!x_{k+1}!}\theta_1^{x_1}....\theta_k^{x_k}\theta_{k+1}{}^{x_{k+1}} = \frac{n!}{\left\{\prod_{i=1}^{k+1} x_i!\right\}}\prod_{i=1}^{k+1}\theta_i^{x_i}$$

where $0 \leq \theta_i \leq 1$ for all i, and $\theta_1 + ... + \theta_k + \theta_{k+1} = 1$, and where x_{k+1} is defined by $x_{k+1} = n - (x_1 + ... + x_k)$. This is the mass function for the **multinomial distribution** which reduces to the binomial if $k = 1$. It can also be shown that the marginal distribution of Y_i is $\text{Bin}(n, \theta_i)$.

A.3.2 DIRICHLET DISTRIBUTION

The **Dirichlet** distribution is a multivariate generalization of the **beta** distribution. Recall that the beta distribution arose a s follows; suppose that V_1 and V_2 are independent Gamma random variables with $V_1 \sim Ga(\alpha_1, \beta), V_2 \sim Ga(\alpha_2, \beta)$. Then if Y is defined by

$$Y = \frac{V_1}{V_1 + V_2}$$

we have that $Y \sim Be(\alpha_1, \alpha_2)$, and

$$f_Y(x) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1} \quad 0 < x < 1$$

Now consider a generalization; suppose that $V_1, ..., V_{k+1}$ are independent Gamma random variables with $V_i \sim Ga(\alpha_i, \beta)$, for $i = 1, ..., k+1$. Define

$$Y_i = \frac{V_i}{V_1 + ... + V_{k+1}}$$

for $i = 1, ..., k$. Then the joint distribution of vector $Y = (Y_1, ..., Y_k)$ is given by

$$f_{Y_1,...,Y_k}(x_1, ..., x_k) = \frac{\Gamma(\alpha)}{\Gamma(\alpha_1)...\Gamma(\alpha_k)\Gamma(\alpha_{k+1})} x_1^{\alpha_1 - 1} ... x_k^{\alpha_k - 1} x_{k+1}{}^{\alpha_{k+1} - 1}$$

for $0 \leq x_i \leq 1$ for all i such that $x_1 + ... + x_k + x_{k+1} = 1$, where $\alpha = \alpha_1 + ... + \alpha_{k+1}$ and where x_{k+1} is defined by $x_{k+1} = 1 - (x_1 + ... + x_k)$. This is the density function which reduces to the beta distribution if $k = 1$. It can also be shown that the marginal distribution of Y_i is $Beta(\alpha_i, \alpha)$.

A.3.3 THE MULTIVARIATE NORMAL DISTRIBUTION

The **multivariate normal** distribution is a multivariate generalization of the **normal** distribution which can be generated in the following way. Suppose that $Y_1, ..., Y_k$ are i.i.d. $N(0, \sigma^2)$ random variables. Using vector notation, we can write the joint density function of $Y_1, ..., Y_k$ as

$$f_{Y_1,...,Y_k}(x_1, ..., x_k) = \left(\frac{1}{2\pi\sigma^2}\right)^{k/2} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{x}\right\}$$

where $\mathbf{x} = (x_1, ..., x_k)$. Now consider the multivariate transformation from $(Y_1, ..., Y_k)$ to $(Y_1, ..., Y_k)$ (that is, from Y to $\mathbf{Y}$) defined by

$$\mathbf{Y} = A^T\mathbf{Y} + \boldsymbol{\mu}$$

where A is a $k \times k$ invertible matrix of real numbers, and $\boldsymbol{\mu}$ is a $k \times 1$ vector. This is a 1-1 transformation, so using the usual multivariate transformation formula, we can obtain the joint density function of $(Y_1, ..., Y_k)$ as

$$f_{Y_1,...,Y_k}(y_1, ..., y_k) = \left(\frac{1}{2\pi}\right)^{k/2} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})\right\}$$

where $\Sigma = \sigma^2 A^T A$. This is the pdf of the **multivariate normal distribution**. It can be shown that any marginal, joint marginal, or conditional distribution of a subset of $Y_1, ..., Y_k$ is normal or multivariate normal.

A.4 EXPECTATION AND VARIANCE RESULTS FOR STANDARD DISTRIBUTIONS

The expectations and variances for the special distributions described in previous sections are as follows:

		Parameters	EXPECTATION	VARIANCE
DISCRETE	$Bernoulli(\theta)$	θ	θ	$\theta(1-\theta)$
	$Binomial(n,\theta)$	n,θ	$n\theta$	$n\theta(1-\theta)$
	$Poisson(\lambda)$	λ	λ	λ
	$Geometric(\theta)$	θ	$\frac{1}{\theta}$	$\frac{(1-\theta)}{\theta^2}$
	$NegBinomial(n,\theta)$	n,θ	$\frac{n}{\theta}$	$\frac{n(1-\theta)}{\theta^2}$
CONTINUOUS	$Uniform(a,b)$	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	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
	$Beta(\alpha,\beta)$	α,β	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
	$Normal(\mu,\sigma^2)$	μ,σ^2	μ	σ^2

Index

alternative hypothesis, 59
ancillarity, 50

Bayes Theorem, 87
Bayesian credible interval, 91
Bayesian credible interval, HPD, 91
Bayesian estimate, posterior expectation, 91
Bayesian estimate, posterior median, 91
Bayesian estimate, posterior mode, 91
Bayesian Inference, 87
bias, 38
binomial, limiting form, 28

canonical form exponential family, 43
Central Limit Theorem, 31, 35, 107
chain rule, 3
chain rule, random variables, 19
Chebychev Inequality, 32
completeness, 52
composite hypothesis, 62
composite null hypothesis, 62
conditional mass/density function, 16
conditional probability, 3
consistency, 55
convergence in distribution, 33
convergence in probability, 34
correlation, 22
covariance, 22
Cramer-Rao Lower Bound, 47
critical region, 59

density function, 8
distribution - Bernoulli, 27, 96
distribution - beta, 31, 105
distribution - binomial, 28, 96
distribution - Cauchy, 107
distribution - chi-squared, 105
distribution - continuous Uniform, 29, 102
distribution - Discrete Uniform, 26, 95
distribution - exponential, 29, 102
distribution - F, 105
distribution - gamma, 30, 104
distribution - Geometric, 27
distribution - geometric, 100
distribution - negative Binomial, 28
distribution - negative binomial, 101
distribution - normal, 31
distribution - normal (Gaussian), 106
distribution - Poisson, 97
distribution - Student-t, 32, 107
distribution - Weibull, 103
distribution function, continuous, 8
distribution function, discrete, 7
distribution, chi-squared, 31
distributions, continuous, 29
distributions, discrete, 26

efficiency, 58
efficient score, 41
equivalence, 48
estimate, 38
estimation, method of moments, 53
estimator, 38
estimators, consistency, 44
estimators, invariance, 44
estimators, unbiasedness, 44
event, 1
event, certain, 1
event, impossible, 1
expectation, 11
expectation, conditional, 20
expectations, multivariate, 22
exponential family, 43

Factorization Theorem, 48
Fisher Information, 42

generating functions, 24

Hypothesis Testing, 59

independence, 3
independence, random variables, 19
invariance, 55

Jensen's Inequality, 33, 56
joint cdf, 15
joint cdf, continuous, 15
joint mass function, 14
joint pdf, 15

Law of Iterated Expectation, 51
Likelihood function, 37
likelihood ratio test, 64
limiting distribution, 33
limiting distribution, degenerate, 33
local power, 62
location parameter, 91

loss function, 92
loss, absolute error, 94
loss, expected, 92
loss, squared error, 93

marginal likelihood, 87
marginal mass function, 15
marginal pdf, 15
mass function, 6
maximum likelihood estimate, 54
Minimal sufficient partition, 49
Minimal sufficient statistic, 49
MLES, asymptotic properties, 57
moment generatig functions, properties, 24
moment generating function, 24
moments, 12
moments, central, 12
mse, 38
mutual exclusive events, 2
MVUE, 38

Neyman Pearson Lemma, 61
Neyman-Pearson, 59
null hypothesis, 59

parameter space, 37
partition, 2
posterior distribution, 87
posterior odds, 92
power function, 60
prior density, 87
prior specification, 89
prior, improper, 90
prior, non-informative, 90
probability, 2
probability generating function
    generating function, probability, 26
probability tree, 3

random sample, 37
random variable, 6
random variable, continuous, 8
random variable, discrete, 6
Rao-Blackwell Theorem, 52
regular (estimation), 38
risk, 92
risk, Bayes, 93
rule, Bayes, 93

sample outcome, 1
sample space, 1, 37
sampling distribution, 38, 39
scale parameter, 91
significance level, 60
simple hypothesis, 60
size of a test, 60
standard deviation, 12
statistic, 37, 48
statistical inference, practical considerations, 65
statistical test, 59
sufficiency, Bayesian
    Bayesian sufficiency, 89
sufficient partition, 48
sufficient statistic, 48

Taylor expansion, 56
test statistic, 59
Theorem of Total Probability, 4
transformations, 13
transformations, Jacobian
    Jacobian, 21
transformations, multivariate, 20
Type I error, 60
Type II error, 60

UMVUE, 53
unbiasedness, 38
uniformly most powerful test, 60

variance, 12

Weak Law of Large Numbers, 34, 55