MATH2715 - Statistical methods

Dr Sofya Titarenko

first semester, academic year 2022-23

Contents

1 Introduction and Informations **4**
- 1.1 Provisional syllabus . . . 4
- 1.2 Booklist and Resources . . . 7
- 1.3 Assignments . . . 8

2 Introductory Revision **11**
- 2.1 Background Notes: Statistical Modelling Example in R . . . 11
- 2.2 Revision on Distributions . . . 12
- 2.3 Joint Probability Distribution . . . 16
- 2.4 Joint probability function - Discrete rv . . . 16
- 2.5 Joint probability density function - Continuous rv . . . 16

3 Bayesian Statistics **18**
- 3.1 Revision Formulas . . . 18
- 3.2 A first discussion about Bayesian Inference . . . 20
- 3.3 Estimating a probability from binomial data . . . 20
- 3.4 Posterior distribution as compromise between data and prior information . . . 22
- 3.5 Binomial model with uniform prior distributions . . . 23
- 3.6 Binomial model with different prior distributions . . . 23

4 Functions of a random variables **26**
- 4.1 Functions of two random variables . . . 28
- 4.2 How to compute the joint density? . . . 29
- 4.3 Notes about the support . . . 29
- 4.4 Revision of differentiation under the integral sign . . . 30
- 4.5 Examples . . . 32

5 Few observations about the Beta and the Gamma distributions **33**
- 5.1 Beta function . . . 33
- 5.2 Beta distribution . . . 33
- 5.3 Gamma distribution . . . 34

6 Moments and Method of moments **35**
- 6.1 Newton's binomial theorem . . . 35
- 6.2 Parameters estimation with the method of moments . . . 38
- 6.3 One form of the method . . . 38
- 6.4 Alternative form of the method . . . 38
- 6.5 Examples . . . 39

7 Moment Generating Function **40**
- 7.1 Sum of Independent Random Variables . . . 46
- 7.2 MGF properties on Linear Transformations . . . 47
- 7.3 Characteristic function . . . 49
- 7.4 convolutions . . . 52

8 Limit Theorems **54**
- 8.1 Markov's Inequality . . . 54
- 8.2 Chebyshev's Inequality . . . 55
- 8.3 Sequence of Random Variables . . . 55
- 8.4 Weak Law of Large Numbers . . . 57
- 8.5 Weak Law of Large Numbers: EXAMPLE . . . 57
- 8.6 Weak Law of Large Numbers: Interpretation . . . 58

8.7 About Convergence . . . 59
8.7.1 Sequences and Convergence: Recalls . . . 59
8.7.2 Sequence of functions and Convergence: Recalls . . . 59
8.7.3 Convergence in Probability and in Distribution . . . 59
8.7.4 Convergence in Distribution . . . 60
8.8 Central Limit Theorem . . . 61
8.8.1 The Central Limit Theorem, few comments . . . 62
8.8.2 Example of derivation of the CLT for the exponential distribution . . . 65

9 Estimation **68**
9.1 Desirable properties of estimators . . . 68
9.2 Revision and extension of maximum likelihood estimation . . . 69
9.3 Comparing estimators . . . 72
9.3.1 Food for thought . . . 73
9.4 Is the sample mean the "best" estimator of the distribution mean? . . . 73
9.5 Can a biased estimator be "better" than an unbiased estimator? . . . 74
9.6 Are m.l.e.'s the answer? . . . 74
9.7 More about likelihood . . . 75
9.7.1 Invariance property of m.l.e.'s . . . 75
9.7.2 Relative likelihood . . . 76
9.7.3 Likelihood summaries . . . 76
9.7.4 Information . . . 78
9.8 Univariate distributions . . . 78
9.9 Multivariate distributions . . . 79

10 Hypothesis Testing **83**
10.1 Data and questions . . . 84
10.2 Basic ideas . . . 84
10.3 The magic 5% significance level (or p-value of 0.05) . . . 86
10.4 The critical region . . . 87
10.5 Errors in hypothesis testing . . . 88
10.6 Summary of hypothesis testing . . . 90
10.7 The Likelihood Ratio Test . . . 91
10.7.1 The likelihood ratio . . . 91
10.7.2 The likelihood ratio statistic . . . 95
10.7.3 The asymptotic distribution of the likelihood ratio statistic . . . 96
10.7.4 Testing goodness-of-fit for discrete distributions . . . 97
10.7.5 The approximate χ^2 distribution . . . 101

11 Background Notes for the Bivariate normal distribution **103**

1 Introduction and Informations

Lecturer: Sofya Titarenko, Physics Research Deck, 9.308a
email: S.Titarenko@leeds.ac.uk
Regularly updated information about the module is available on **Minerva**

Module objective: To introduce mathematical techniques for analyzing probability distributions and to develop tools for building statistical models .

1.1 Provisional syllabus

- Bayesian Statistics
- Beta Binomial model
- Moments and transformations for univariate probability distributions.
- Conditional and marginal distributions for bivariate distributions; bivariate normal distribution.
- Moment generating functions: law of large numbers; central limit theorem.
- Issues in statistical modelling.
- Estimation: method of moments; maximum likelihood.
- Hypothesis testing. Type 1 and Type 2 errors, power, likelihood ratio test.
- Transformations to normality.
- Bayesian modelling; prior and posterior distributions.

Provisional lecture topics

- Revision. Probability, discrete random variables, continuous random variables, properties of random variables, discrete joint random variables, continuous joint random variables, independence.
- Moments. Sample moments, population moments, skewness and kurtosis, method of moments.
- Gamma distribution. Definition, moments, method of moments estimator, applications.
- Functions of a random variable. Transformation of a random variable, direct and cdf approach
- Functions of two random variables.
- Bivariate transformations II. Beta distribution, mean and variance, convolution.
- Moment generating function I. Properties, differentiation property, power series expansion, probability genetaring function.
- Moment generating function II. Sum of independent random variables, sum of exponential and normal rvs, linear transformations, characteristic function.
- Weak law of large numbers. Convergence in probability, Chebychev's inequality, Markov's inequality, central limit theorem, convergence in distribution.
- Estimators. Optimal estimators. Bias, minimum variance, mean square error.
- Maximum likelihood I. Likelihood function, maximum likelihood principle.
- Maximum likelihood II. Properties of mle, variance of mle, transformation of mle.
- Maximum likelihood III. Several random variables, likelihood ratio test.
- Bayesian statistics. Aims of statistical modelling, likelihood, classical (frequentist) probability, subjective Bayesian degrees of belief, Bayes theorem, Bayesian approach.

- Bayesian priors.
- Bayesian inference.
- Bivariate normal distribution. Linear transformation; matrix representation; orthogonal transformation.
- Bivariate normal distribution. Marginal and conditional pdf; independence and correlation.

1.2 Booklist and Resources

Text Books

- J.A. Rice, Mathematical statistics and data analysis. Duxbury Press, 3rd edition
- Andrew Gelman John B. Carlin Hal S. Stern David B. Dunson Aki Vehtari Donald B. Rubin, Bayesian Data Analysis, Third Edition, CRC press

Suggested R intro

- introductory R document by Johen Voss
- Free R course on DataCamp: https://www.datacamp.com/courses/free-introduction-to-r.

Suggested Readings

- 1st year material in MINERVA
- Probability and Statistics I (MATH1710) by Robert Aykroyd
- Probability and Statistics II (MATH1712) by Johen Voss

1.3 Assignments

Purpose

Completing your coursework responsibly and regularly will help you to understand the material, practice your problem solving skills, and your written presentation skills. As a Maths graduate you will be expected to be able to communicate mathematics clearly in various forms. Writing up your solutions in an organised way will also help your understanding. Do NOT rely on the marker's ability to interpret what you have written.

Marking Scheme

For your coursework, you will need to accomplish five assessments. Four out of five are online tests. The fifth assessment will be done using statistical software R. Each assessment will be marked out of a total of 10. Online tests will provide immediate and automatic feedback. In addition, the samples of online tests will be available on Minerva for you to practice. Please, keep in mind that Minerva keeps only the final grade for online tests. Therefore, attempt the tests only when you are fully prepared. For the fifth assessment, you will need to submit the Jupyter Notebook file with your code and comments. You will submit the file via the Gradescope link available in Minerva. In the submitted file, write in capital letters NAME SURNAME STUDENT-ID at the top of the page. Add plenty of comments to support the results of your code. You will be guided regarding how to work with Jupyter Notebook during some of the workshops. Additional examples will be available on Minerva.

Note on working together

Working on your coursework in a small group is an excellent way of learning. You are encouraged to ask each other questions, explain solutions to each other, but you should always write up the final solutions by yourself.

Timeline

For online tests, feedback and solutions will be provided automatically. The scores of late submissions (either online test or R) will be reduced by 10 % per day late.

- 1st assignment (week 3) will be due on the 21st October, 11.59 pm.
- 2nd assignment (week 5) will be due on the 4th November, 11.59 pm.
- 3rd assignment (week 7) will be due on the 18th November, 11.59 pm.

- 4th assignment (week 9) will be due on the 2nd December, 11.59 pm.
- 5th assignment (week 10) will be due on the 9th December, 11.59 pm.
- In week 5 you will receive an R assignment that will be due on the 9th December, 11.59 pm.

NOTE: THIS IS PROVISIONAL MATERIAL

I apologise for the typos you will encounter reading this draft. This is still a draft and will be updated frequently! All the relevant material is acquired during classes and tutorial sessions. Please let me know any missing part or mistake you might encounter reading this first draft. I would very much appreciate any comment!

Aknowledgments

The materials are largerly prepared by Dr Luisa Cutillo. I would also like to thank Andrew Baczkovski and Robert Aykroyd for their support, guidance and exemplar teaching material.

2 Introductory Revision

2.1 Background Notes: Statistical Modelling Example in R

Example Radiocarbon dates[1].
A laboratory typically quotes a radiocarbon C14 date for an object in the form $y \pm \sigma$ where σ is a standard deviation which depends on the laboratory. For an object with calendar age θ it is usual to model the radiocarbon date y as satisfying $y \sim \mathrm{N}(\mu(\theta), \sigma^2)$ where $\mu(\theta)$ is a function of θ.

Example Earthquake data I[2]
Figure 1 shows the histogram of 236 inter-earthquake times in north-east part of Santa Clara valley during 1992 The R code to produce this figure is:

```
L="https://raw.githubusercontent.com/luisacutillo78/
Public_Math2715/master/Data/quake1.txt"
dd=scan(L)
# Scan data into dd.
hist(dd,breaks=c(0:60),xlab="Time between earthquakes in days",
ylab="Frequency per interval of width one day",main="")
curve(length(dd)*dexp(x,rate=1/mean(dd)),0,60,add=TRUE,lty=1)
# Add exponential pdf.
legend(60,25,c("Fitted exponential pdf"),lty=1,bty="n",cex=0.75,xjust=1)
# Add legend.
```

[1]Source: Christen, J.A. and Buck, C.E. (1998) "Sample selection in radiocarbon dating", *Applied Statistics*, **47**, 543-557.
[2]Source: Northern California Earthquake Catalog
http://quake.geo.berkeley.edu/ncedc/catalog-search.html

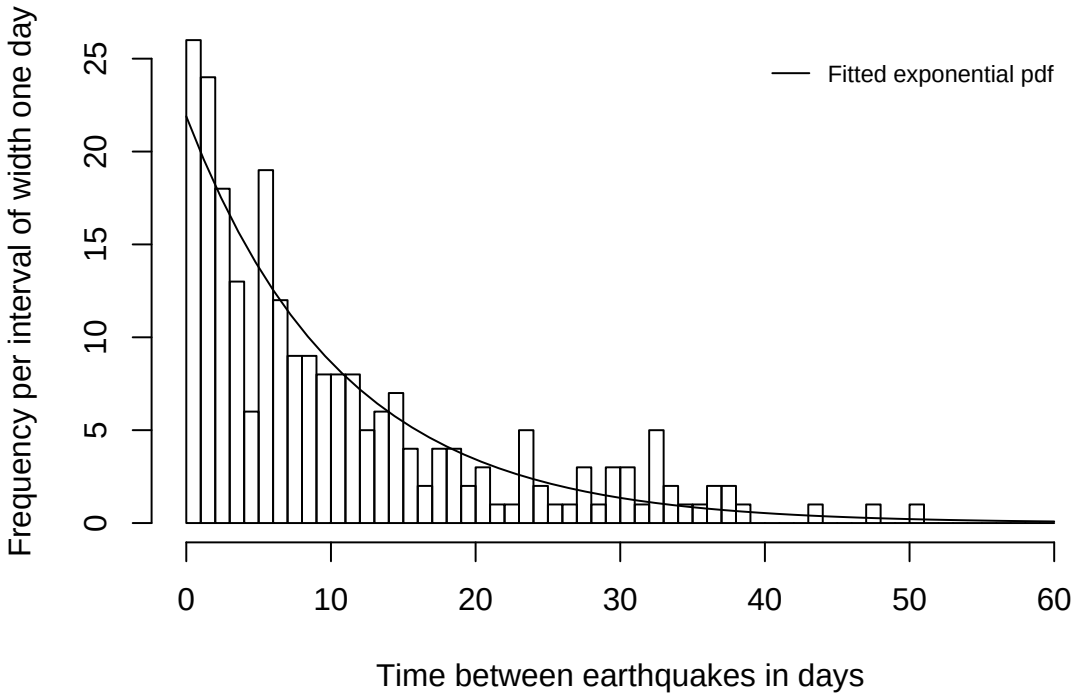


Figure 1: fitted exponential probability density function for earthquake data I.

Abbreviations used in the lecture notes

cdf: cumulative distribution function	iid: independent and identically distributed
mgf: moment generating function	mle: maximum likelihood estimate
pdf: probability density function	rv: random variable

2.2 Revision on Distributions

Discrete random variable

A discrete random variable (rv) X takes probabilities $\text{pr}\{X = x\}$ for $x = 0, 1, 2, \ldots$.

Mean of X is $\text{E}[X] = \sum_x x \, \text{pr}\{X = x\}$, often denoted by μ.

Variance of X is $\text{Var}[X] = \text{E}[(X-\mu)^2] = \text{E}[X^2]-\mu^2$ where $\text{E}[X^2] = \sum_x x^2 \text{pr}\{X = x\}$.

Poisson distribution

If $X \sim \text{Poisson}(\mu)$, then $\text{pr}\{X = x\} = \dfrac{\mu^x e^{-\mu}}{x!}$ for $x = 0, 1, 2, 3, \ldots$. $\text{E}[X] = \mu$, $\text{Var}[X] = \mu$.

Binomial distribution

If $X \sim \text{Bin}(n, \theta)$, then $\text{pr}\{X = x\} = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ for $x = 0, 1, 2, \ldots, n$. $\text{E}[X] = n\theta$, $\text{Var}[X] = n\theta(1-\theta)$. Here X might represent the number of heads in n tosses of a coin.

Geometric distribution[3]

If $X \sim \text{geometric}(\theta)$, then $\text{pr}\{X = x\} = \theta(1-\theta)^x$, $x = 0, 1, 2, 3, \ldots$. $\text{E}[X] = \dfrac{1-\theta}{\theta}$, $\text{Var}[X] = \dfrac{1-\theta}{\theta^2}$. Here X might represent the number of tails observed before the first head occurs in a coin tossing experiment. This is the same as the number of tosses of the coin before the first head!

Bernoulli trials

A random variable X is said to be a Bernoulli random variable with parameter p, shown as $X \sim Bernoulli(p)$, if its PMF is given by

$$P_X(x) = \begin{cases} p & \text{for } x = 1 \\ 1-p & \text{for } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

Where $0 < p < 1$.
with $\text{E}[X] = p$ and $\text{Var}[X] = p(1-p) = pq$.

Suppose $U_1, \ldots, U_n$ are independent Bernoulli random variables taking values 0 or 1 with $\text{pr}\{U_i = 1\} = \theta$, $\text{pr}\{U_i = 0\} = 1 - \theta$. Let $S = U_1 + \cdots + U_n$, be the number of U_i equal to one. It can be shown that $S \sim \text{Bin}(n, \theta)$.

[3]Sometimes the geometric distribution is defined as $\text{pr}\{X = x\} = \theta(1-\theta)^{x-1}$, $x = 1, 2, 3, \ldots$, with mean $1/\theta$ and variance $(1-\theta)/\theta^2$. In this case X might represent the number of tosses of a coin until the first head occurs.

Continuous random variables

A continuous rv X, defined in general for $x \in (-\infty, \infty)$, has probability density function (pdf) $f_X(x)$ and cumulative distribution function (cdf) $F_X(x)$ satisfying:

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}, \qquad \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1,$$

$$\mathrm{pr}\{x < X \le x + \Delta x\} \approx f_X(x)\Delta x,$$

$$F_X(x_0) = \mathrm{pr}\{X \le x_0\} = \int_{-\infty}^{x_0} f(x)\,\mathrm{d}x,$$

$$\mathrm{pr}\{a < X \le b\} = \int_a^b f_X(x)\,\mathrm{d}x = F_X(b) - F_X(a).$$

Mean of X is $\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x$, often denoted μ.

Variance of X is $\mathrm{Var}[X] = \mathrm{E}[(X-\mu)^2] = \mathrm{E}[X^2] - \mu^2$ where $\mathrm{E}[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x$.

Standard normal distribution

If $Z \sim \mathrm{N}(0,1)$, then

$$\text{pdf:} \quad \phi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2} \text{ for } -\infty < z < \infty;$$

$$\text{cdf:} \quad \Phi(z_0) = \mathrm{pr}\{Z \le z_0\} = \int_{-\infty}^{z_0} \phi(z)\,\mathrm{d}z.$$

Normal distribution

If $X \sim \mathrm{N}(\mu, \sigma^2)$, then $\mathrm{E}[X] = \mu$, $\mathrm{Var}[X] = \sigma^2$, and

$$\text{pdf:} \quad f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } -\infty < x < \infty;$$

$$\text{cdf:} \quad F(x_0) = \mathrm{pr}\{X \le x_0\} = \int_{-\infty}^{x_0} f_X(x)\,\mathrm{d}x.$$

Can be shown that $Z = \dfrac{X-\mu}{\sigma} \sim \mathrm{N}(0,1)$ so

$$F_X(x) = \mathrm{pr}\{X \le x\} = \mathrm{pr}\left\{\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right\} = \mathrm{pr}\left\{Z \le \frac{x-\mu}{\sigma}\right\} = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Exponential distribution

If $X \sim \text{exponential}(\lambda)$, then $\mathrm{E}[X] = \dfrac{1}{\lambda}$, $\mathrm{Var}[X] = \dfrac{1}{\lambda^2}$, with

$$\text{pdf:} \quad f_X(x) = \lambda e^{-\lambda x} \text{ for } x > 0;$$

$$\text{cdf:} \quad F_X(x) = \int_0^x \lambda e^{-\lambda x}\,\mathrm{d}x = \left[-e^{-\lambda x}\right]_0^x = 1 - e^{-\lambda x}, \quad x > 0.$$

Uniform distribution

If $X \sim \text{uniform}(0,1)$, then $\mathrm{E}[X] = \frac{1}{2}$, $\mathrm{Var}[X] = \frac{1}{12}$, with

$$\text{pdf:} \quad f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{cdf:} \quad F_X(x) = \int_0^x f_X(x)\,\mathrm{d}x = \begin{cases} 1 & \text{if } x \ge 1, \\ x & \text{if } 0 < x < 1, \\ 0 & \text{if } x \le 0. \end{cases}$$

More generally a uniform distribution might be defined on an interval (a, b).

Gamma function

The gamma function[4] $\Gamma(z)$ is a generalises factorial function that satisfies

- In general $\Gamma(z+1) = z\Gamma(z)$.
- If n is a positive integer, then $\Gamma(n) = (n-1)!$ with $\Gamma(1) = 1$.

[4]Sometimes called the factorial function, the gamma function was first investigated in 1729 by Leonhard Euler (1707-1783), a Swiss mathematician who also introduced the notations e for the exponential function, i for $\sqrt{-1}$ and $\sum$ for summation. Euler gave the definition

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\,\mathrm{d}x.$$

The notation $\Gamma(z)$ was introduced in 1814 by Adrien-Marie Legendre (1752-1833), a French mathematician who was one of the people responsible for the development of the method of least squares. Legendre referred to the above integral definition of $\Gamma(z)$ as the *Eulerian integral of the second kind*.

It can be shown that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Beta Distribution

if $X \sim \text{Beta}(a, b)$, then $\text{E}[X] = \dfrac{a}{a+b}$, $\text{Var}[X] = \dfrac{ab}{(a+b)^2(a+b+1)}$, with

$$\text{pdf:} \quad f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \text{ for } 0 \leq x \leq 1;$$

2.3 Joint Probability Distribution

(SEE TITANIC and INDUSTRIAL PRODUCTION EXAMPLE, LECTURE 3)

- In many situations, there are more than one quantity associated with the experiment
- it is interesting to study the JOINT BEHAVIOUR of these random quantities.

The joint behaviour of two r.v., X and Y, is determined by the **Cumulative Distribution Function**:

$$F(X, Y) = P(X \leq x, Y \leq y)$$

Regardless of whether X and Y are continuous or discrete.

2.4 Joint probability function - Discrete rv

Given X and Y discrete random variables with p.m.f. respectively p_X and p_Y, their joint probability function is $p_{XY}(x, y) = \text{pr}\{X = x \cap Y = y\}$.
The marginal probability function of X is $p_X(x) = \sum_y p_{XY}(x, y) \equiv \text{pr}\{X = x\}$.
If X, Y are independent, then $p_{XY}(x, y) = p_X(x)p_Y(y) \ \ \forall x, y$.[5]
The conditional probability of $Y = y$ given $X = x$, if $p_Y(y_j) > 0$, is defined as: $p_{Y|X}(y|x) = \dfrac{p_{XY}(x, y)}{p_X(x)} \equiv \text{pr}\{Y = y|X = x\}$.

2.5 Joint probability density function - Continuous rv

Suppose we have two continuous random variables X and Y. Their **joint**

[5]The converse is true. Thus if $p_{XY}(x, y) = p_X(x)p_Y(y) \ \ \forall x, y$, then X and Y are independent.

density function is a piecewise continuous function of two variables $f(x, y)$ which has the following properties

- $f_{X,Y}(x, y) \geq 0 \qquad \forall (x, y) \in \mathbb{R}^2$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1.$
- $\forall A \subset \mathbb{R}^2$, $P((X, Y) \in A) = \int \int_A f(x, y) dx dy$

It follows that the **joint probability distribution function** $F_{X,Y}$ is defined as,

$$F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y) dx dy$$

If X and Y are continuous random variables, with joint pdf $f_{X,Y}$ then the individual or **marginal** pdf's f_X and f_Y are given by,

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

If X, Y are independent, then $f_{XY}(x, y) = f_X(x) f_Y(y) \ \ \forall x, y$.[6]
Th conditional probability density function of $Y = y$ given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

[6]The converse is again true. Thus if $f_{XY}(x, y) = f_X(x) f_Y(y) \ \ \forall x, y$, then X and Y are independent.

3 Bayesian Statistics

3.1 Revision Formulas

Conditional Probability

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

For discrete random variables

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$$

For continuous random variables We have conditional densities:

$$f_{y|x}(y|x) = \frac{f_{xy}(x, y)}{f_x(x)}$$

Bayes' Theorem

(The most elementary version)

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \end{aligned}$$

For discrete random variables,

$$\begin{aligned} P(X = x|Y = y) &= \frac{P(X = x, Y = y)}{P(Y = y)} \\ &= \frac{P(Y = y|X = x)P(X = x)}{\sum_t P(Y = y|X = t)P(X = t)} \end{aligned}$$

For continuous random variables

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{XY}(x,y)}{f_y(y)} \\ &= \frac{f_{Y|X}(y|x)f_X(x)}{\int f_{Y|X}(y|t)f_X(t)\,dt} \end{aligned}$$

3.2 A first discussion about Bayesian Inference

Most of the results shown in this chapter are extracted from *Bayesian Data Analysis* by Andrew Gelman, et al.

We will start our first detailed topic in connection to the latest statistical models studies last year. In particular we will discuss Bayesian statistical models where only a single scalar parameter, e.g. θ is to be estimated; This means that we will aim to estimate a one-dimensional parameter θ. At the same time we will introduce important concepts and computational methods for Bayesian data analysis. Most of the results shown in this section are extracted from *Bayesian Data Analysis* by Andrew Gelman, et al.

3.3 Estimating a probability from binomial data

We will consider a simple but important starting point for the discussion of Bayesian inference. In the simple binomial model, the aim is to estimate an unknown population proportion from the results of a sequence of 'Bernoulli trials'; that is, data $y_1, \ldots, y_n$, each of which is either 0 or 1.

The binomial distribution provides a natural model for data that arise from a sequence of n exchangeable trials or draws from a large population where each trial gives rise to one of two possible outcomes, conventionally labeled 'success' and 'failure.' Because of the exchangeability, the data can be summarized by the total number of successes in the n trials, which we denote here by y. Converting from a formulation in terms of exchangeable trials to one using independent and identically distributed random variables is achieved quite naturally by letting the parameter θ represent the proportion of successes in the population or, equivalently, the probability of success in each trial. The binomial sampling model states that

$$p(y|\theta) = \text{Bin}(y|n, \theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \tag{1}$$

where on the left side we suppress the dependence on n because it is regarded as part of the experimental design that is considered fixed; all the probabilities discussed for this problem are assumed to be conditional on n.
To perform Bayesian inference in the binomial model, we must specify a prior distribution for θ. We will discuss issues associated with specifying prior distributions many times throughout this book, but for simplicity at this point,

we assume that the prior distribution for θ is uniform on the interval $[0, 1]$.

Elementary application of Bayes' rule, applied to (1), then gives the posterior density for θ as

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}. \tag{2}$$

With fixed n and y, the factor $\binom{n}{y}$ does not depend on the unknown parameter θ, and so it can be treated as a constant when calculating the posterior distribution of θ. As is typical of many examples, the posterior density can be written immediately in closed form, up to a constant of proportionality. In single-parameter problems, this allows immediate graphical presentation of the posterior distribution.

A bit of history

The Reverend Thomas Bayes, an English part-time mathematician whose work was unpublished during his lifetime, and Pierre Simon Laplace, an inventive and productive mathematical scientist whose massive output spanned the Napoleonic era in France, receive independent credit as the first to *invert* the probability statement and obtain probability statements about θ, *given* observed y.

In his famous paper, published in 1763, Bayes sought, in our notation, the probability $\Pr(\theta \in (\theta_1, \theta_2)|y)$; his solution was based on a physical analogy of a probability space to a rectangular table (such as a billiard table):

1. (Prior distribution) A ball is randomly thrown , according to a uniform distribution, on the table. The horizontal position of the ball on the table is θ, expressed as a fraction of the table width.

2. (Likelihood) A ball is randomly thrown n times. The value of y is the number of times lands to the right of B.

Thus, θ is assumed to have a (prior) *uniform distribution* on $[0, 1]$. Using

elementary rules of probability theory, Bayes then obtained

$$\begin{aligned}\Pr(\theta \in (\theta_1, \theta_2)|y) &= \frac{\Pr(\theta \in (\theta_1, \theta_2), y)}{p(y)} \\ &= \frac{\int_{\theta_1}^{\theta_2} p(y|\theta)p(\theta)d\theta}{p(y)} \\ &= \frac{\int_{\theta_1}^{\theta_2} \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta}{p(y)}. \end{aligned} \tag{3}$$

Bayes succeeded in evaluating the denominator, showing that

$$\begin{aligned} p(y) &= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta \\ &= \frac{1}{n+1} \quad \text{for } y = 0, \ldots, n. \end{aligned} \tag{4}$$

This calculation shows that all possible values of y are equally likely *a priori*.

The numerator of (3) is an incomplete beta integral with no closed-form expression for large values of y and $(n-y)$, a fact that apparently presented some difficulties for Bayes.

Laplace, however, independently 'discovered' Bayes' theorem, and developed new analytic tools for computing integrals. For example, he expanded the function $\theta^y(1-\theta)^{n-y}$ around its maximum at $\theta = y/n$ and evaluated the incomplete beta integral using what we now know as the normal approximation.

In analyzing the binomial model, Laplace also used the uniform prior distribution.

3.4 Posterior distribution as compromise between data and prior information

(SEE WORKSHOP 1)

The process of Bayesian inference involves passing from a prior distribution, $p(\theta)$, to a posterior distribution, $p(\theta|y)$, and it is natural to expect that some general relations might hold between these two distributions. For example, we might expect that, because the posterior distribution incorporates the information from the data, it will be less variable than the prior distribution. In can be shown that this notion is formalized in the second of the following expressions:

$$\mathrm{E}(\theta) = \mathrm{E}(\mathrm{E}(\theta|y)) \tag{5}$$

and

$$\mathrm{Var}(\theta) = \mathrm{E}(\mathrm{Var}(\theta|y)) + \mathrm{Var}(\mathrm{E}(\theta|y)), \tag{6}$$

The result expressed by equation (5) is scarcely surprising: the prior mean of θ is the average of all possible posterior means over the distribution of possible data. The variance formula (6) is more interesting because it says that *the posterior variance is on average smaller than the prior variance*, by an amount that depends on the variation in posterior means over the distribution of possible data. The greater the latter variation, the more the potential for reducing our uncertainty with regard to θ, as we shall see in detail for the binomial and normal models in the next chapter.

3.5 Binomial model with uniform prior distributions

(SEE FLIPPED COIN VIDEO LECTURE and Mathematical Statistics and Data Analysis, A.Rice, Chapter 3, Example E, Bayesian Inference)

In the binomial example with the uniform prior distribution, we have shown that starting from:

- a Binomial model for our data $X|\Theta \sim Bin(n, \Theta)$
- a uniform prior on θ, $\Theta \sim U[0, 1]$

we obtain a posterior density that is a beta density with parameters $a = x+1$ and $b = n - x + 1$

- $\Theta|X \sim Beta(a = x + 1, b = n - x + 1)$

In this example, the prior mean is $\frac{1}{2}$, and the prior variance is $\frac{1}{12}$. The posterior mean, $\frac{y+1}{n+2}$, is a compromise between the prior mean and the sample proportion, $\frac{y}{n}$, where clearly the prior mean has a smaller and smaller role as the size of the data sample increases. This is a very general feature of Bayesian inference: the posterior distribution is centered at a point that represents a compromise between the prior information and the data, and the compromise is controlled to a greater extent by the data as the sample size increases.

3.6 Binomial model with different prior distributions

We can generalise the binomial model using a parametric family of prior distributions that includes the uniform as a special case. Considered as a function

of θ, the likelihood (1) is of the form,

$$p(y|\theta) \propto \theta^a(1-\theta)^b.$$

Note that, if the prior density belongs to the same family, then the posterior density will also be of this form. The prior density can be parametrized as:

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1},$$

which is a beta distribution with parameters α and β: $\theta \sim \text{Beta}(\alpha, \beta)$.

Looking closely at the prior $p(\theta)$ and the posterior $p(y|\theta)$ we can note that the prior density is equivalent to $\alpha - 1$ prior successes and $\beta - 1$ prior failures. The parameters of the prior distribution are often referred to as *hyperparameters*.

For example, the beta prior distribution is indexed by two hyperparameters. If you think of the Normal distribution, we could completely specify it by fixing two features of the distribution, for example its mean and variance;

The posterior density for θ is

$$\begin{aligned} p(\theta|y) &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= \text{Beta}(\theta|\alpha+y, \beta+n-y). \end{aligned}$$

The property that the posterior distribution follows the same parametric form as the prior distribution is called *conjugacy*; We say that the beta prior distribution is a *conjugate family* for the binomial likelihood. The main reason to choose a conjugate family is mathematically convenience. Indeed this will ensure that the posterior distribution follows a known parametric form. In practice it could happen that the conjugate prior is not a good fit to the specific example we are analysing, it may be necessary to use a more realistic prior distribution.

In the Beta Binomial model, the posterior mean of θ, which may be interpreted as the posterior probability of success for a future draw from the population, is now

$$\text{E}(\theta|y) = \frac{\alpha+y}{\alpha+\beta+n},$$

which always lies between the sample proportion, y/n, and the prior mean, $\alpha/(\alpha+\beta)$; (see Wokshop 1).

The posterior variance is

$$\text{Var}(\theta|y) = \frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = \frac{E(\theta|y)[1-E(\theta|y)]}{\alpha+\beta+n+1}.$$

As y and $n-y$ become large with fixed α and β, $\text{E}(\theta|y) \approx y/n$ and $\text{Var}(\theta|y) \approx \frac{1}{n}\frac{y}{n}(1-\frac{y}{n})$, which approaches zero at the rate $1/n$.

4 Functions of a random variables

Suppose X is a random variable and $Y = g(X)$, where $g : \Re \to \Re$ is a *monotonic* and *differentiable* function.
Define $g^{-1} : \Re \to \Re$ such that $g^{-1}(g(X)) = X$. Then,

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{\partial g^{-1}(y)}{\partial y} \right|, \text{ if } y = g(x) \text{ for some } x \\ &= 0, \text{ otherwise} \end{aligned}$$

Example Change of Coordinates
Suppose X is a random variable with pdf $f_X(x)$. Suppose $Y = X^n$. Find $f_Y(y)$.

Note that $g^{-1}(y) = y^{1/n}$, hence:

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{\partial g^{-1}(Y)}{\partial y} \right| \\ &= f_X(y^{1/n}) \frac{y^{\frac{1}{n}-1}}{n} \end{aligned}$$

Example Power of wind turbines[7]
For a wind with instantaneous speed V the power is $P = \frac{1}{2}\lambda V^3$ where λ is the density of the air[8]. Typically windspeeds are averaged over a day and the average daily windspeed U is such that $\sqrt{U}$ can be modelled as a normal distribution. The power P available to a wind turbine typically satisfies $P \propto U^{5/2}$.

Example Income in UK[9]
Incomes U are such that $X = \log U$ is well-modelled by a normal distribution; see figure 2.

Reasons for transforming random variables

"A common goal in transforming variables is to induce symmetry and homo-

[7]Source: Haslett, J. and Raftery, A.E. (1989) "Space-time modelling with long-memory dependence: assessing Ireland's wind power", *Applied Statistics*, **38**, 1-50.

[8]Imagine a unit area in space and let λ be the density of the air. For a wind with speed v blowing perpendicular to the area, a volume of air equal to v will be moved in unit time. The mass of air that moves will equal $m = \lambda v$. The kinetic energy of this moving mass equals $\frac{1}{2}mv^2 = \frac{1}{2}\lambda v^3$ and per unit time the power developed will equal $\frac{1}{2}\lambda v^3$.

[9]Source: Inland revenue, `http://www.inlandrevenue.gov.uk/stats/index.htm`

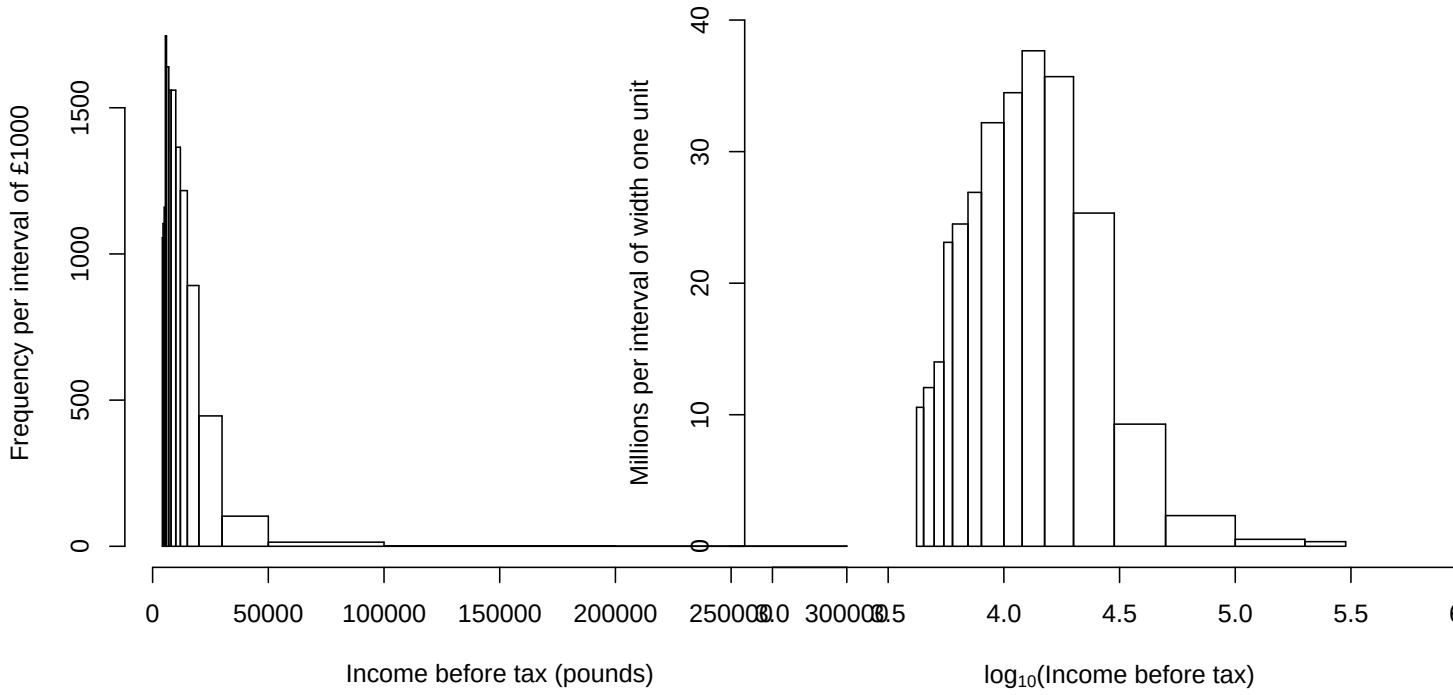


Figure 2: (left) histogram of income before tax 1998–1999, (right) histogram of $\log_{10}$(income).

geneity in the error distribution."[10]

"...a transformation designed to achieve one purpose ...often also helps to achieve another."[11]

[10]Source: Kettl, S. (1991) "Accounting for heteroscedasticity in the transform both sides regression model", *Applied Statistics*, **40**, 261-268.

[11]Source: Kendall, M. and Stuart, A. (1976, p.95) *The Advanced Theory of Statistics, volume III (3rd edition)*, Griffin, London.

4.1 Functions of two random variables

Jacobian of a transformation

Consider the transformation $u = u(x, y)$ and $v = v(x, y)$ which maps R_{xy} in the (x, y) plane into the rectangular region R_{uv} with area $\Delta u \Delta v$ in the (u, v) plane; see figure 3.

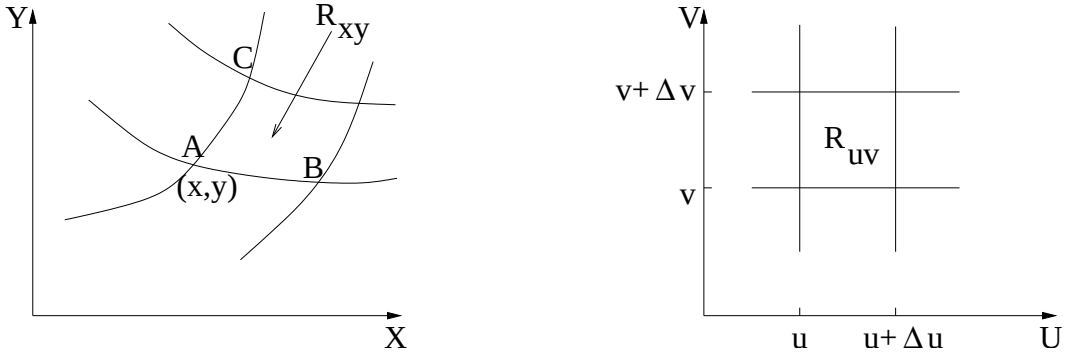


Figure 3: Mapping R_{xy} to R_{uv}.

The determinant

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

is referred to as a Jacobian[12] and the area $|R_{xy}|$ can be shown[13] to satisfy $|R_{xy}| \approx |J| \Delta u \Delta v$.

It can be shown that

$$\frac{\partial(u, v)}{\partial(x, y)} = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} = \left(\frac{\partial(x, y)}{\partial(u, v)} \right)^{-1} = \frac{1}{J}.$$

[12] Carl Jacobi (1804-1851) was a German mathematician who worked extensively with determinants.

[13] In figure 3, let point A map to (u, v) and B map to $(u + \Delta u, v)$ so A has x-coordinate $x(u, v)$ while B has x-coordinate $x(u + \Delta u, v)$ where $x = x(u, v)$ and $y = y(u, v)$ denotes the inverse transformation from the (u, v) plane to the (x, y) plane. Since

$$\frac{\partial x}{\partial u} = \lim_{\Delta u \to 0} \frac{x(u + \Delta u, v) - x(u, v)}{\Delta u}$$

it follows that the x-difference of A and B is $x(u + \Delta u, v) - x(u, v) \approx \frac{\partial x}{\partial u} \Delta u$. Similarly the y-difference of A and B is $y(u + \Delta u, v) - y(u, v) \approx \frac{\partial y}{\partial u} \Delta u$. Hence the vector $\underline{\text{AB}}$ satisfies

$$\underline{\text{AB}} = \left(\frac{\partial x}{\partial u} \Delta u \right) \mathbf{i} + \left(\frac{\partial y}{\partial u} \Delta u \right) \mathbf{j}$$

where $\mathbf{i}$ and $\mathbf{j}$ are unit vectors in the x and y directions respectively. Similarly

$$\underline{\text{AC}} = \left(\frac{\partial x}{\partial v} \Delta v \right) \mathbf{i} + \left(\frac{\partial y}{\partial v} \Delta v \right) \mathbf{j}.$$

For small Δu and Δv, R_{xy} will be approximately a parallelogram with area $|\underline{\text{AB}} \times \underline{\text{AC}}|$ where "$\times$" denotes a vector product. Thus the area of R_{xy} is given by

$$|\underline{\text{AB}} \times \underline{\text{AC}}| \approx \left| \frac{\partial x}{\partial u} \frac{\partial y}{\partial v} - \frac{\partial x}{\partial v} \frac{\partial y}{\partial u} \right| \Delta u \Delta v = \left| \frac{\partial(x, y)}{\partial(u, v)} \right| \Delta u \Delta v.$$

For example, if $u = x/y$ and $v = y$, then $x = uv$ and $y = v$ so that $\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v$ while $\frac{\partial(u,v)}{\partial(x,y)} = \begin{vmatrix} 1/y & -x/y^2 \\ 0 & 1 \end{vmatrix} = 1/y = 1/v$.

4.2 How to compute the joint density?

Suppose X_1 and X_2 has a joint density $f(x_1, x_2)$ and support S_X.
Let $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ with the single-valued inverse $X_1 = v_1(Y_1, Y_2)$ and $X_2 = v_2(Y_1, Y_2)$.

The joint pdf of Y_1 and $Y2$ is:

$$g(y_1, y_2) = |J| f\left[v_1(y_1, y_2), v_2(y_1, y_2)\right]$$

where J is the determinant of the Jacobian Matrix:

$$\begin{pmatrix} \frac{\partial v_1(y_1,y_2)}{\partial y_1} & \frac{\partial v_1(y_1,y_2)}{\partial y_2} \\ \frac{\partial v_2(y_1,y_2)}{\partial y_1} & \frac{\partial v_2(y_1,y_2)}{\partial y_2} \end{pmatrix}$$

Note that S_Y, the support of (Y_1, Y_2), is usually found by considering the image of S_X under the transformation $Y_1, Y2$. Meaning that $\forall (x_1, x_2) \in S_X$ we find $(y_1, y_2) \in S_Y$

$$x_1 = v_1(y_1, y_2), \qquad x_2 = v_2(y_1, y_2)$$

4.3 Notes about the support

The support of a random variable is the set of points where its density is positive. This is a very simple concept, but there are a few issues about supports that are worthwhile stating explicitly.

If a random variable X has support A, then $P(X \in A) = 1$, because if S is the sample space for the distribution of X

$$1 = \int_S f(x)dx = \int_A f(x)dx + \int_{A^c} f(x)dx = \int_A f(x)dx = P(X \in A)$$

because f_X is zero on A^c and the integral of zero is zero. Thus, as long as the only random variables under consideration are X and functions of X it makes no difference whether we consider the sample space to be S (the original sample space) or A (the support of X). In most situations you can use whichever form you prefer. Why would anyone every use thesupport? There are several reasons. One good reason is that there may be many different random

variables, all with different supports, under consideration. If one wants them all to live on the same sample space, which may simplify other parts of the problem, then one needs to spedify their supports. Another reason not so good is mere habit or convention. For example, convention requires that the domain of a c. d. f. be the whole real line. Thus one commonly requires the domain of the matching density to also be the whole real line necessitating something like to express the support if the support is not the whole real line. Think for example iof the Uniform in (a, b), we could consider the sample space to be the interval (a, b).

4.4 Revision of differentiation under the integral sign

Suppose that

$$G(u) = \int_a^b g(x, u)\,\mathrm{d}x$$

where a and b are functions of u and $g(x, u)$ is a function of x and u. Then

$$\frac{\mathrm{d}G(u)}{\mathrm{d}u} = g(b, u)\,\frac{\mathrm{d}b}{\mathrm{d}u} - g(a, u)\,\frac{\mathrm{d}a}{\mathrm{d}u} + \int_a^b \frac{\partial g(x, u)}{\partial u}\,\mathrm{d}x.$$

This is sometimes known as the Liebnitz integral rule.

Example: $G(u) = \displaystyle\int_1^{u^2} e^{xu}\,\mathrm{d}x \quad \Rightarrow \quad \frac{\mathrm{d}G(u)}{\mathrm{d}u} = \left(e^{u^3} \times 2u\right) - (e^u \times 0) + \int_1^{u^2} xe^{xu}\,\mathrm{d}x.$

Example: $G(u) = \displaystyle\int_{-\infty}^{\infty} e^{xu} f(x)\,\mathrm{d}x \quad \Rightarrow \quad \frac{\mathrm{d}G(u)}{\mathrm{d}u} = \int_{-\infty}^{\infty} xe^{xu} f(x)\,\mathrm{d}x$ on writing $g(x, u) = e^{xu} f(x)$ and noting that the integration limits do not depend on u.

Example: $G(u) = \displaystyle\int_{-\infty}^{b} e^{-\frac{1}{2}x^2}\,\mathrm{d}x \quad \Rightarrow \quad \frac{\mathrm{d}G(u)}{\mathrm{d}u} = e^{-\frac{1}{2}b^2} \times \frac{\mathrm{d}b}{\mathrm{d}u}$ with $g(x, u) = e^{-\frac{1}{2}x^2}$ and noting that $\dfrac{\partial g}{\partial u} = 0$.

To prove the formula for differentiation under the integral suppose $G^\star(x, u) = \displaystyle\int g(x, u)\mathrm{d}x$ so $\dfrac{\partial G^\star(x, u)}{\partial x} = g(x, u)$. Then

$$\frac{\partial^2 G^\star(x, u)}{\partial x \partial u} = \frac{\partial^2 G^\star(x, u)}{\partial u \partial x} \quad \Rightarrow \quad \frac{\partial}{\partial x}\left(\frac{\partial G^\star(x, u)}{\partial u}\right) = \frac{\partial}{\partial u}\left(\frac{\partial G^\star(x, u)}{\partial x}\right) = \frac{\partial g(x, u)}{\partial u}.$$

Integrating with respect to x gives $\dfrac{\partial G^\star(x, u)}{\partial u} = \displaystyle\int \frac{\partial g(x, u)}{\partial u}\mathrm{d}x$. This result implies:

(1) $\dfrac{\partial}{\partial u}\displaystyle\int g(x,u)\mathrm{d}x = \int \frac{\partial g(x,u)}{\partial u}\mathrm{d}x,$ and (2), as a definite integral,
$\dfrac{\partial G^\star(b,u)}{\partial u} - \dfrac{\partial G^\star(a,u)}{\partial u} = \displaystyle\int_a^b \frac{\partial g(x,u)}{\partial u}\mathrm{d}x.$

Considering now the definite integral $G(u) = \displaystyle\int_a^b g(x,u)\,\mathrm{d}x$, it follows that $G(u) = G^\star(b,u) - G^\star(a,u)$. Hence

$$\begin{aligned}\frac{\mathrm{d}G(u)}{\mathrm{d}u} &= \frac{\partial G^\star(b,u)}{\partial b}\cdot\frac{\mathrm{d}b}{\mathrm{d}u} + \frac{\partial G^\star(b,u)}{\partial u} - \frac{\partial G^\star(a,u)}{\partial a}\cdot\frac{\mathrm{d}a}{\mathrm{d}u} - \frac{\partial G^\star(a,u)}{\partial u} \qquad (\dagger)\\ &= g(b,u)\frac{\mathrm{d}b}{\mathrm{d}u} + \frac{\partial G^\star(b,u)}{\partial u} - g(a,u)\frac{\mathrm{d}a}{\mathrm{d}u} - \frac{\partial G^\star(a,u)}{\partial u}\\ &= g(b,u)\frac{\mathrm{d}b}{\mathrm{d}u} - g(a,u)\frac{\mathrm{d}a}{\mathrm{d}u} + \int_a^b \frac{\partial g(x,u)}{\partial u}\mathrm{d}x.\end{aligned}$$

The result (†) follows from obtaining the total derivative of $G^\star(b,u)$ and $G^\star(a,u)$. Recall that for a function $f(x,y)$ with $x = x(u)$ and $y = y(u)$ both functions of u, then $\dfrac{\partial f(x,y)}{\partial u} = \dfrac{\partial f(x,y)}{\partial x}\cdot\dfrac{\partial x}{\partial u} + \dfrac{\partial f(x,y)}{\partial y}\cdot\dfrac{\partial y}{\partial u}$. Writing $f(x,y) = G^\star(b,u)$ gives $\dfrac{\partial G^\star(b,u)}{\partial u} = \dfrac{\partial G^\star(b,u)}{\partial b}\cdot\dfrac{\partial b}{\partial u} + \dfrac{\partial G^\star(b,u)}{\partial u}\cdot\dfrac{\partial u}{\partial u}$
$= \dfrac{\partial G^\star(b,u)}{\partial b}\cdot\dfrac{\partial b}{\partial u} + \dfrac{\partial G^\star(b,u)}{\partial u}.$
Similarly $\dfrac{\partial G^\star(a,u)}{\partial u}$ can be obtained.

4.5 Examples

A collection of solved excercises on the transformation of random variables, can be foun in the file *Examples1_ TransformationsOfRandomVariables.pdf*, in the Worked Examples folder on the module page, under Learning Resources.

5 Few observations about the Beta and the Gamma distributions

5.1 Beta function

The beta function is defined as $B(\alpha, \beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

This was first studied by Euler and Legendre and arose from work on what Legendre referred to as the *Eulerian integral of the first kind*, namely

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x$$

where p and q are positive.

5.2 Beta distribution

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x < 1, \quad \alpha > 0, \quad \beta > 0.$$

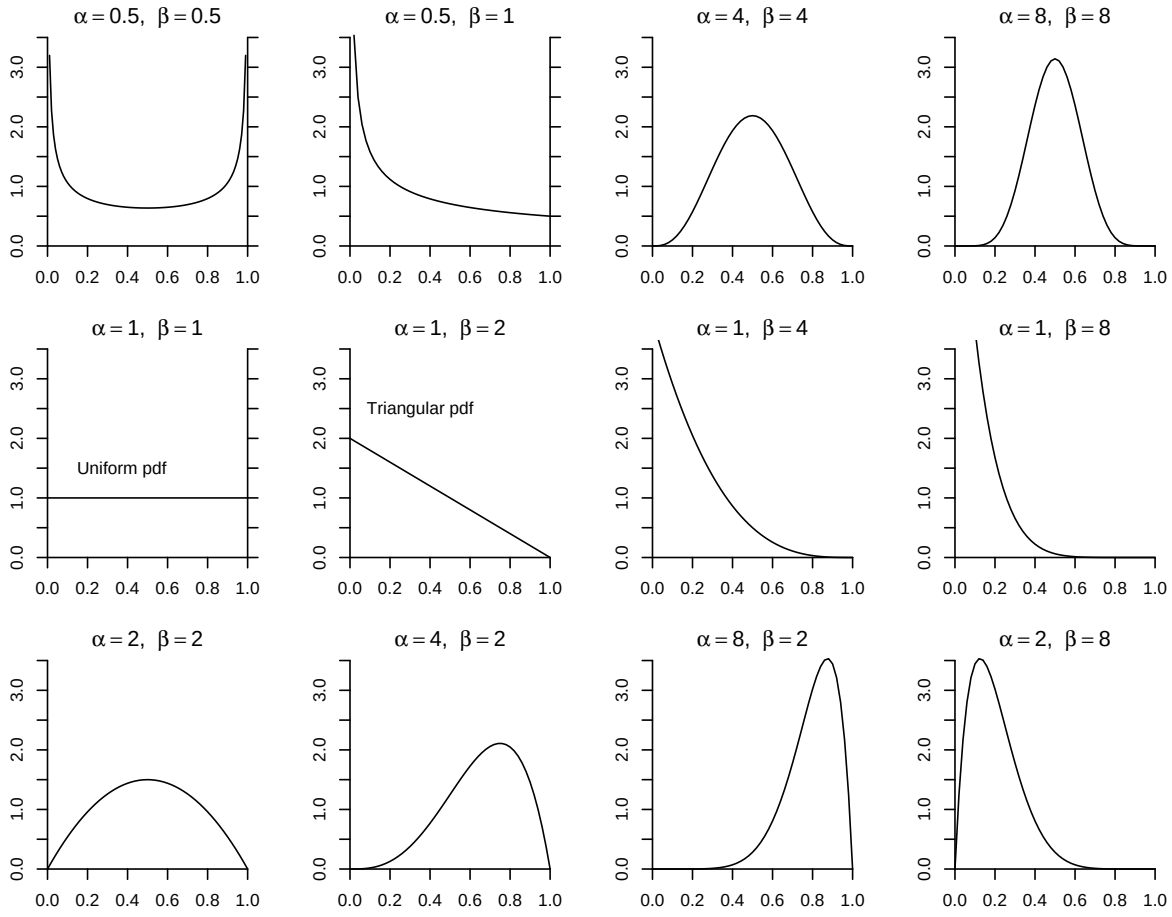


Figure 4: plots show beta probability density function for different values of α and β.

5.3 Gamma distribution

if $X \sim \text{Gamma}(\alpha, \lambda)$, then

$$\text{pdf:} \quad f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \text{ for } x \geq 0; \quad \textit{with } E[X] = \frac{\alpha}{\lambda}; \quad Var[X] = \frac{\alpha}{\lambda^2}$$

$$E[X^2] = Var[X] + E[X]^2 = \frac{\alpha(1+\alpha)}{\lambda^2}$$

Note: When $\alpha = 1$, the gamma reduces to an **exponential distribution**. Another well-known statistical distribution, the **Chi-Square**, is also a special case of the gamma. A Chi-Square distribution with n degrees of freedom is the same as a gamma with $\alpha = n/2$ and $\lambda = 1/2$. The gamma function[14] $\Gamma(z)$ satisfies

- In general $\Gamma(z+1) = z\Gamma(z)$.
- If n is a positive integer, then $\Gamma(n) = (n-1)!$ with $\Gamma(1) = 1$.

[14]Sometimes called the factorial function, the gamma function was first investigated in 1729 by Leonhard Euler (1707-1783), a Swiss mathematician who also introduced the notations e for the exponential function, i for $\sqrt{-1}$ and $\sum$ for summation. Euler gave the definition

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \, \mathrm{d}x.$$

The notation $\Gamma(z)$ was introduced in 1814 by Adrien-Marie Legendre (1752-1833), a French mathematician who was one of the people responsible for the development of the method of least squares. Legendre referred to the above integral definition of $\Gamma(z)$ as the *Eulerian integral of the second kind*.

It can be shown that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

6 Moments and Method of moments

(DEFINITION OF MOMENTS FROM SLIDES)

6.1 Newton's binomial theorem

The binomial theorem (or binomial expansion) describes the algebraic expansion of powers of a binomial.

Isaac Newton (1642-1727) stated the binomial theorem in 1664/5 for arbitrary n although the result for integer values of n was known earlier, namely

$$(a+b)^n = a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{x}a^{n-x}b^x + \cdots + b^n = \sum_{x=0}^{n}\binom{n}{x}a^{n-x}b^x.$$

Examples:
$(X-\mu)^2 = X^2 - 2\mu X + \mu^2,$
$(X-\mu)^3 = X^3 - 3\mu X^2 + 3\mu^2 X - \mu^3,$
$(X-\mu)^5 = X^5 - 5\mu X^4 + 10\mu^2 X^3 - 10\mu^3 X^2 + 5\mu^4 X - \mu^5.$

Example Earthquake data II[15]
Figure 5 shows the distribution of 436 inter-quake times for 437 earthquakes in south-west part of Santa Clara valley for period 1992-1998.

Example Time until AIDS diagnosed[16]
The time between HIV infection and diagnosis of AIDS can be modelled using a gamma(α, λ) distribution.

Example Mean wind power
A good approximate model for the distribution of mean wind power is provided by the gamma distribution. Since $X \sim$ gamma(α, λ) satisfies

$$\mathrm{E}[X] = \frac{\alpha}{\lambda}, \qquad \mathrm{Var}[X] = \frac{\alpha}{\lambda^2}$$

the method of moments estimates for α and λ are found by equating

$$\bar{x} = \frac{\alpha}{\lambda}, \qquad m_2 = \frac{\alpha}{\lambda^2}$$

[15]Source: Northern California Earthquake Catalog
`http://quake.geo.berkeley.edu/ncedc/catalog-search.html`

[16]Source: De Angelis, D., Gilks, W.R., and Day, N.E. (1998) "Bayesian projection of the acquired immune deficiency syndrome epidemic", *Applied Statistics*, **47**, 449-498.

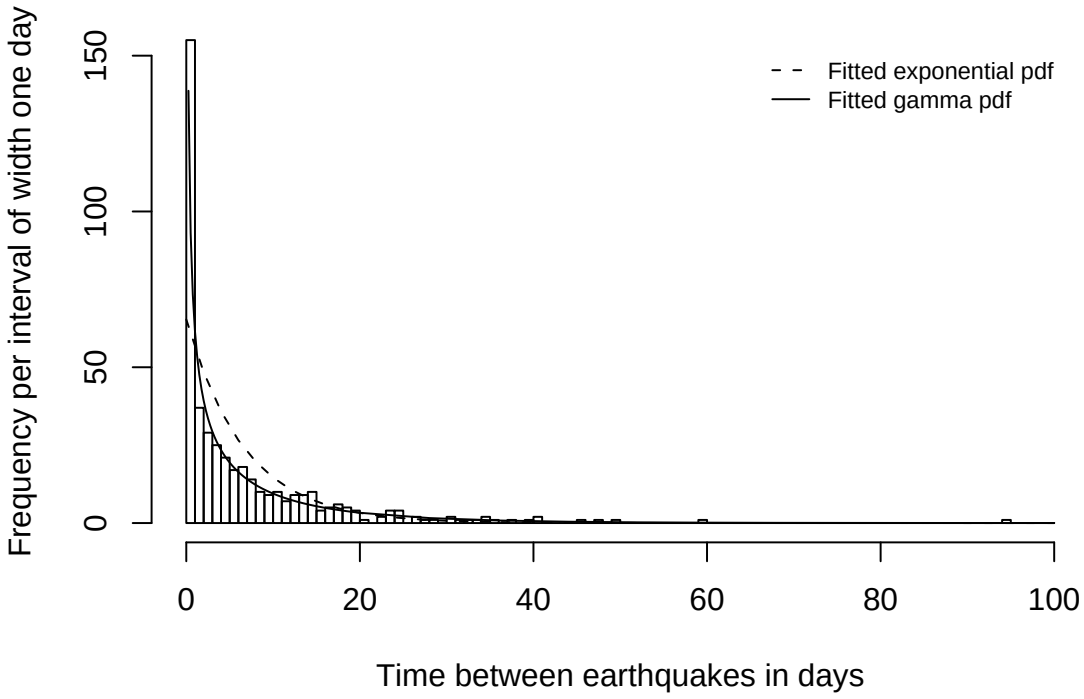


Figure 5: fitted exponential and gamma probability density function for earthquake data II.

giving estimates $\widetilde{\alpha} = \bar{x}^2/m_2$ and $\widetilde{\lambda} = \bar{x}/m_2$. Figure 6 shows the histogram of wind power for 1056 measurements on Cairn Gorm and the probability density function of the fitted gamma distribution based on the method of moments estimates

The R code to produce this figure is:

```
# Read in the data.
L="https://raw.githubusercontent.com/luisacutillo78/
Public_Math2715/master/Data/windmean.txt"

wind=read.table(L,header=FALSE)
pp=0.5*wind[,3]^3 # Set pp as wind-power
mm=mean(pp) # Sample mean of pp.
mm2=mean((pp-mm)^2 # Second sample moment m2 of pp.
alpha=mm^2/mm2 # Method of moments estimates for alpha and lambda.
lambda=mm/mm2

maxpp=ceiling(max(pp)/500) # maxpp = maximum pp value / 500.
```

```
hist(pp,breaks=500*c(0:maxpp),xlab="Mean power (Watts)",
ylab="Frequency per interval of width 500 watts",main="")
legend(15000,50,"Fitted gamma pdf:",bty="n") # Add legends.
legend(15000,44,substitute(list(alpha==t1,lambda==t2),
list(t1=alpha,t2=lambda)),bty="n")
curve(length(pp)*dgamma(x,shape=alpha,rate=lambda)*500,0,50000,
n=501,add=TRUE) # Add pdf.
```

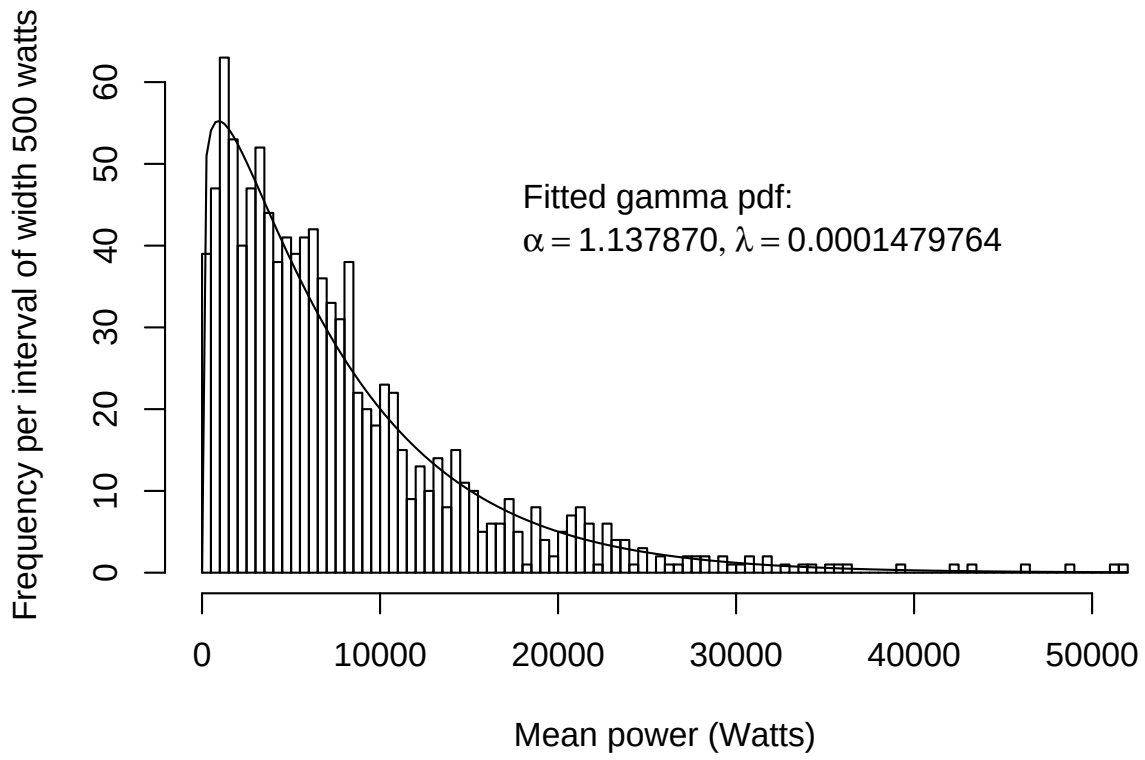


Figure 6: histogram of mean wind power and fitted gamma probability density function.

6.2 Parameters estimation with the method of moments

If $X_1, \ldots, X_n$ are i.i.d. random variables, ($\sim X$), whose the probability distribution function has a few unknown parameters. The methods of moments estimates the unknown parameters by equating sample moments with theoretical moments.

6.3 One form of the method

The basic idea behind this form of the method is to:

- (1) find expressions for the unknown parameters in terms of the lowest possible order moments about the origin $\mu_r = E(X^r)$.
- (2) Equate the first sample moment about the origin $M_1' = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ (sample mean) to the first theoretical moment $\mu = E(X)$. Continue equating sample moments about the origin $M_r' = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ to the corresponding $r - th$ theoretical moment about the origin $\mu_r' = E(X^r)$
- (3) Continue equating sample moments about the origin with the corresponding theoretical moments until you have as many equations as you have parameters
- (4) Solve for the parameters. The resulting values are called *method of moments estimators*.

6.4 Alternative form of the method

In some cases, rather than using the sample moments about the origin, it is easier to use the sample moments about the mean. Doing so, provides us with an alternative form of the method of moments.

- (1) find expressions for the unknown parameters in terms of:
  - the sample mean $\mu = E(X)$
  - and of the lowest possible order moments about the mean $\mu_r = E[(X - \mu)^r]$.
- (2) Equate the first sample moment about the origin $M_1' = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ (sample mean) to the first theoretical moment $\mu_1' = \mu = E(X)$. Continue

equating sample moments about the mean $M_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r$ to the $r-th$ central moment (i.e. about the mean) $\mu_r = E[(X-\mu)^r]$, for $r \geq 2$.

- (3) Continue equating sample moments about the mean with the corresponding central moments until you have as many equations as you have parameters
- (4) Solve for the parameters. Again, the resulting values are called *method of moments estimators.*

6.5 Examples

A collection of solved excercises on moments and on the apllication of the method of moments, can be foun in the file *Examples2_ Moments.pdf*, in the Worked Examples folder on the module page, under Learning Resources.

7 Moment Generating Function

Definition

Given is a random variable X , the function $M(t) = E[e^{tX}]$ is defined **moment-generating function (MGF)** of X, if the expectation is defined. If X is a continuous r.v. with pdf $f(x)$:

$$M(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

If X is a discrete r.v. with probability mass function $p(x)$:

$$\begin{aligned} M(t) &= E[e^{tX}] \\ &= \sum_x e^{tx} p(x) \end{aligned}$$

Proposition

If the MGF exists in an open interval containing zero, then

$$\frac{\partial^n M(t)}{\partial^n t}|_0 = E[X^n]$$

□

Recall the Taylor Expansion of e^{tX} at 0,

$$e^{tX} = 1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \dots$$

$$E[e^{tX}] = 1 + tE[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \dots$$

Differentiate once

$$\begin{aligned} \frac{\partial M(t)}{\partial t} &= 0 + E[X] + \frac{2t}{2!}E[X^2] + \dots \\ M^{'}(0) &= 0 + E[X] + 0 + 0 \dots \end{aligned}$$

Differentiate n times

$$\begin{aligned} \frac{\partial^n M(t)}{\partial^n t} &= 0 + 0 + 0 + \dots + \frac{n \times n-1 \times \dots 2 \times t^0 E[X^n]}{n!} + \frac{n! t E[X^{n+1}]}{(n+1)!} + \dots \\ &= \frac{n! E[X^n]}{n!} + \frac{n! t E[X^{n+1}]}{(n+1)!} + \dots \end{aligned}$$

Evaluated at 0, yields $M^n(0) = E[X^n]$.

Proposition
If two random variables, X and Y have the same moment generating functions, then

$$F_X(x) = F_Y(y)$$

for *almost all* x. ■

Example: Standard Normal distribution
Suppose $Z \sim N(0,1)$, Find its MGF and calculate the moments using it. □

$$\begin{aligned} E[e^{tX}] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx \end{aligned}$$

$tx - \frac{1}{2}x^2 = -\frac{1}{2}\left((x-t)^2 - t^2\right)$

$$\begin{aligned} E[e^{tX}] &= \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} e^{-(x-t)^2/2} dx \\ &= e^{\frac{t^2}{2}} \end{aligned}$$

$$\begin{aligned} M^{'}(0) &= E[X] = e^{t^2/2} t|_0 = 0 \\ M^{''}(0) &= E[X^2] = e^{t^2/2}(t^2+1)|_0 = 1 \\ M^{'''}(0) &= E[X^3] = e^{t^2/2} t(t^2+3)|_0 = 0 \\ M^{''''}(0) &= E[X^4] = e^{t^2/2}(t^4+6t^2+3)|_0 = 3 \\ M^{5}(0) &= E[X^5] = e^{t^2/2} t(t^4+10t^2+15)|_0 = 0 \\ M^{6}(0) &= E[X^6] = e^{t^2/2}(t^6+15t^4+45t^2+15)|_0 = 15 \end{aligned}$$

■

Example: General Normal distribution
Suppose $X \sim N(\mu, \sigma^2)$, Find its MGF. $\square$

$$\begin{aligned}
E[e^{tX}] &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x^2-2\mu x+\mu^2-2tx\sigma^2)}{2\sigma^2}} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu-t\sigma^2)^2}{2\sigma^2}} e^{\frac{t^2\sigma^4+2\mu t\sigma^2}{2\sigma^2}} dx \\
&= e^{\mu t+\frac{\sigma^2}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu-t\sigma^2)^2}{2\sigma^2}} dx \\
&= e^{\mu t+\frac{\sigma^2}{2}t^2}
\end{aligned}$$

■

Example: Poisson Distribution
Suppose $X \sim Poisson(\lambda)$, find its MGF and calculate the moments using it. □

We know that If $X \sim$ Poisson(λ), then

$\text{pr}\{X = k\} = \dfrac{\lambda^k e^{-\lambda}}{k!}$ for $k = 0, 1, 2, 3, \ldots$ and
$\text{E}[X] = \lambda$, $\text{Var}[X] = \lambda$. By definition:

$$\begin{aligned} E[e^{tX}] &= \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} e^{-\lambda} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t - 1)} \\ &= e^{-\lambda(1 - e^t)} \end{aligned} \tag{7}$$

The sum converges for all the t. Differentiating we get:

$$\begin{aligned} M^{'}(t) &= \lambda e^t e^{\lambda(e^t - 1)} \\ M^{''}(t) &= \lambda e^t e^{\lambda(e^t - 1)} + \lambda^2 e^{2t} e^{\lambda(e^t - 1)} \end{aligned}$$

It follows that, when t=0:

$$\begin{aligned} E(X) &= \lambda \\ E(X^2) &= \lambda^2 + \lambda \\ Var(X) &= E(X^2) - E(X)^2 = \lambda \end{aligned}$$

■

Example: Gamma Distribution

Suppose $X \sim gamma(\alpha, \lambda)$, find its MGF and calculate the moments using it. $\square$

By definition:

$$\begin{aligned} E[e^{tX}] &= \int_0^\infty e^{tk} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{x(t-\lambda)} dx \end{aligned}$$

It is easy to see that the previous integral converges for $t < \lambda$. Note that for a $gamma(\alpha, \lambda - t)$ it stands:

$$\int_0^\infty \frac{(\lambda-t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x} dx = 1$$

It follows that:

$$\frac{\Gamma(\alpha)}{(\lambda-t)^\alpha} = \int_0^\infty x^{\alpha-1} e^{(t-\lambda)x} dx$$

and hence:

$$M(t) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\lambda-t)^\alpha} = \left(\frac{\lambda}{\lambda-t}\right)^\alpha$$

If we differentiate $M(t)$ and evaluate the derivatives in $t = 0$ we get:

$$\begin{aligned} M^{'}(0) &= E[X] = \frac{\alpha}{\lambda} \\ M^{''}(0) &= E[X^2] = \frac{\alpha(\alpha+1)}{\lambda^2} \end{aligned}$$

As a consequence we find that:

$$Var(X) = E(X^2) - E(X)^2 = \frac{\alpha}{\lambda^2}$$

Example: Chisquare Distribution

Recall that $X \sim \chi^2_k$ is a special case of gamma distribution $X \sim \Gamma(k/2, 1/2)$.

hence, given the previous example: $M_X(t) = \left(\frac{1/2}{1/2-t}\right)^{\frac{k}{2}} = (1-2t)^{-\frac{k}{2}}$

In particular:

$X \sim \chi^2_1$ is a special case of gamma distribution $X \sim \Gamma(1/2, 1/2)$ and so:
$M_X(t) = (1-2t)^{-\frac{1}{2}}$

Example: use the MGF to prove that if $X \sim N(0,1)$, then a $Y = X^2 \sim \chi^2_1$

$$\begin{aligned} M_Y(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz^2} e^{-z^2/2} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2(1-2t)}{2}} dz \end{aligned}$$

if we let $\sigma^2 = (1-2t)^{-1}$, then

$$\begin{aligned} M_Y(t) &= \frac{\sigma}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-z^2}{2\sigma^2}} dz \\ &= \sigma = (1-2t)^{-1/2} \end{aligned}$$

This is the MGF of a χ^2_1.

Example: Exponential Distribution

Suppose $X \sim exponential(\lambda)$, find its MGF and calculate the moments using it. □

By definition:

$$E[e^{tX}] = \int_0^{\infty} e^{tk} \lambda e^{-\lambda x} dx = \int_0^{\infty} \lambda e^{-(\lambda - t)x} dx$$

the latter integral converges for $t < \lambda$. It follows that:

$$E[e^{tX}] = \left[-\frac{\lambda}{\lambda - t} e^{-(\lambda - t)x} \right]_0^{\infty} \tag{8}$$

hence:

$$M(t) = \frac{\lambda}{\lambda - t}$$

and

$$\begin{aligned} M^{'}(0) &= E[X] = \frac{1}{\lambda} \\ M^{''}(0) &= E[X^2] = \frac{2}{\lambda^2} \\ Var(X) &= E(X^2) - E(X)^2 = \frac{1}{\lambda} \end{aligned}$$

Note that this result could have been indirectly obtained observing that an exponential distribution of parameter λ is a special case of a $gamma(\alpha, \lambda)$, with the parameter $\alpha = 1$.

7.1 Sum of Independent Random Variables

Suppose X_i are a sequence of independent random variables. Define

$$Y = \sum_{i=1}^{N} X_i$$

Then

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

□

$$\begin{aligned} M_Y(t) &= E[e^{tY}] \\ &= E[e^{t\sum_{i=1}^{N} X_i}] \\ &= E[e^{tX_1+tX_2+\ldots tX_N}] \\ &= E[e^{tX_1}]E[e^{tX_2}]\ldots E[e^{tX_N}] \text{ (by independence)} \\ &= \prod_{i=1}^{N} E[e^{tX_i}] \end{aligned}$$

■

7.2 MGF properties on Linear Transformations

Proposition
If X has MGF $M_X(t)$ and $Y = a + bX$, then Y has the MGF $M_Y(t) = e^{at}M_X(bt)$.
□

By definition of MGF we have:

$$\begin{aligned} M_Y(t) &= E[e^{tY}] \\ &= E[e^{at+btX}] \\ &= E[e^{at}e^{btX}] \\ &= e^{at}E[e^{btX}] \\ &= e^{at}M_X(bt) \end{aligned}$$

■

Example: Sum of i.i.d. exponential(λ) random variables

Let $X_1, \ldots, X_n$ be independent and identically distributed exponential(λ) random variables.
If $X_i \sim$ exponential(λ) and $S_n = X_1 + \cdots + X_n$, then

$\mathrm{E}[X_i] = 1/\lambda$, $\mathrm{Var}[X_i] = 1/\lambda^2$
sand so:

$$\mathrm{E}[S_n] = \frac{n}{\lambda}, \ \mathrm{Var}[S_n] = \frac{n}{\lambda^2}.$$

How to find the distribution of the sum?
The moment generating function of X_i is $m_{X_i}(t) = \dfrac{\lambda}{\lambda - t}$ so that the moment generating function of $S_n = X_1 + \cdots + X_n$ is $m_{S_n}(t) = \left(\dfrac{\lambda}{\lambda - t}\right)^n$.

We recognize that this is the MGF of a Gamma random variable and hence $S_n \sim Gamma(n, lambda)$.

Example: Sum of i.i.d. Poisson(μ) random variables

Suppose $X_1, \ldots, X_n$ are independent and identically distributed Poisson(μ) random variables.

If $X_i \sim$ Poisson(μ), and $S_n = X_1 + \cdots + X_n$
then

$$\mathrm{E}[X_i] = \mu \; and \; \mathrm{Var}[X_i] = \mu$$

so

$$\mathrm{E}[S_n] = n\mu$$

and $\mathrm{Var}[S_n] = n\mu$.

How to find the distribution of the sum?

The moment generating function of X_i is $m_{X_i}(t) = \exp(-\mu(1 - e^t))$ so that the moment generating function of $S_n = X_1 + \cdots + X_n$ is

$$m_{S_n}(t) = \big(\exp(-\mu(1 - e^t))\big)^n = \exp(-n\mu(1 - e^t)).$$

and hence $S_n \sim Poisson(n\mu)$.

7.3 Characteristic function

One of the shortcomings of the MGF is that given a random variable X, $M_X(t)$ does not always exist.

For example, assume $X \sim Cauchy$ with $f_X(x) = \frac{1}{\pi(1+x^2)}$.
The MGF of X is

$$M_X(t) = E\left[e^{tX}\right] = \int_{-\infty}^{+\infty} \frac{e^{tX}}{\pi(1+x^2)} dx.$$

This is only finite for t=0, not defined otherwise!

Definition The characteristic function of a rv X is

$$\phi_X(t) = E\left[e^{itX}\right]$$

if X has MFG $M_X(t)$ then:

$$\phi_X(t) = E\left[e^{itX}\right] = M_X(it)$$

continuous case Assume X is a continuous r.v. with pdf $f_X(x)$. The characteristic function of X is

$$\phi_X(t) = E\left[e^{itX}\right] = \int_{-\infty}^{+\infty} e^{itX} f_X(x) dx.$$

note that this is the Fourier transform of $f_X(x)$

The Fourier inversion theorem relates a density function $f_U(u)$ and the characteristic function $\phi_U(t)$ by

$$f_U(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu} \phi_U(t)\, \mathrm{d}t.$$

For a free access source of explanation and examples about the characteristic function, I suggest this wikipage:

`https://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)`

Example

if $X \sim exponential(\lambda)$:

$$M_X(t) = \frac{\lambda}{\lambda - t}$$

and:

$$\phi_X(t) = \frac{\lambda}{\lambda - it} = M_X(it)$$

Example

Random variables X and Y are independent and each has a uniform$(0,1)$ distribution. Let $U = X - Y$. Obtain the characteristic function and probability density function $f_U(u)$ of U. Use the Fourier inversion theorem to deduce the characteristic function of the random variable Z with probability density function $f_Z(z) = \dfrac{(1-\cos z)}{\pi z^2}$, where $-\infty < z < +\infty$.

If $X \sim uniform(0,1)$, then $f_X(x) = 1$ for $0 < x < 1$. Hence X (and Y) has moment generating function

$$m_X(t) = E[e^{tX}] = \int_{x=-\infty}^{\infty} e^{tx} f_X(x)\,\mathrm{d}x =$$

$$= \int_{x=0}^{1} e^{tx}\,\mathrm{d}x = \left[\frac{e^{tx}}{t}\right]_{x=0}^{1} = \frac{e^t - 1}{t}, \quad -\infty < t < \infty.$$

The mgf of U is $m_U(t) = E[e^{tU}] = E[e^{t(X-Y)}] = E[e^{tX}e^{-tY}] = E[e^{tX}] \cdot E[e^{(-t)Y}] = m_X(t)m_Y(-t)$ as X, Y are independent. Hence

$$m_U(t) = \left(\frac{e^t - 1}{t}\right)\left(\frac{e^{-t} - 1}{-t}\right) = \frac{2 - e^t - e^{-t}}{-t^2}.$$

Thus the characteristic function of U is

$$\phi_U(t) = E[e^{itU}] = m_U(it) = \frac{2 - e^{it} - e^{-it}}{-(it)^2} =$$

$$= \frac{2 - (\cos t + i\sin t) - (\cos t - i\sin t)}{t^2} = \frac{2(1-\cos t)}{t^2}.$$

Joint pdf of (U,V): As X and Y are independent, $f_{XY}(x,y) = f_X(x)f_Y(y) = 1, \quad 0 < x, y < 1.$

Put $u = x - y$, $v = y$ so $x = u + v$ and $y = v$ with Jacobian

$$J = \left|\frac{\partial(x,y)}{\partial(u,v)}\right| = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix} = 1.$$

Hence

$$f_{UV}(u,v) = f_{XY}(x,y)|J| = 1.$$

Range of U and V: $0 < x < 1,\ 0 < y < 1 \implies -1 < x - y < 1$ so $-1 < u < 1$ while $0 < v < 1$.

But also $0 < x < 1 \implies -y < x - y < 1 - y$ so $-v < u < 1 - v$

giving $\quad -u < v < 1-u.$
Clearly $\max(0,-u) < v < \min(1,1-u)$.

Figure **??**(left) shows the (x,y) region on the left. The elemental strip B[17] with $u > 0$ is defined for $y \in (0, 1-u)$. The elemental strip A[18] with $u < 0$ is defined for $y \in (-u, 1)$. Probability density function of U:

$$f_U(u) = \int_{v=-\infty}^{\infty} f_{UV}(u,v)\,\mathrm{d}v = \int_{v=\max(0,-u)}^{\min(1,1-u)} \mathrm{d}v = \Big[v\Big]_{v=\max(0,-u)}^{\min(1,1-u)}.$$

There are thus two cases to consider.

$$f_U(u) = \left\{ \begin{array}{ll} \Big[v\Big]_{v=-u}^{v=1} & = 1+u \quad \text{if } u < 0, \\ \\ \Big[v\Big]_{v=0}^{v=1-u} & = 1-u \quad \text{if } u \geq 0. \end{array} \right\} = 1 - |u| \quad \text{for } -1 < u < 1.$$

Fourier inversion theorem:

The Fourier inversion theorem gives $f_U(u) = \dfrac{1}{2\pi}\displaystyle\int_{t=-\infty}^{\infty} e^{-itu}\phi_U(t)\,\mathrm{d}t$. Thus

$$1 - |u| = \frac{1}{2\pi}\int_{t=-\infty}^{\infty} e^{-itu} \cdot \frac{2(1-\cos t)}{t^2}\,\mathrm{d}t.$$

Putting $t = -z$ and $\mathrm{d}t = -\mathrm{d}z$ shows that

$$\int_{z=-\infty}^{\infty} e^{izu}\frac{(1-\cos z)}{\pi z^2}\,\mathrm{d}z = 1 - |u|$$

so that the characteristic function of a random variable Z having probability density function $f_Z(z) = \dfrac{(1-\cos z)}{\pi z^2}$ is $\phi_Z(u) = 1 - |u|$. Clearly $f_Z(z)$ is a probability density function since $f_Z(z) \geq 0$ for all z and $\phi_Z(0) = 1$. [19]

[17] The line $y = x - u$ for $0 < x < 1$ with $u > 0$.
[18] The line $y = x - u$ for $0 < x < 1$ with $u < 0$.
[19] The probability density function $f_Z(z)$ can be plotted using the following R command:
`curve((1-cos(x))/(pi*x*x),-30,30,xlab="z",ylab="Pdf")`

7.4 convolutions

convolution: discrete case Assume X and Y are discrete rv taking values on integers with joint pmf $p(x, y)$. Let $Z = X + Y$. Note that $Z = z$ whenever $X = x$ and $Y = z - x$. It follows that:

$$P(Z = z) = p_Z(z) = \sum_{-\infty}^{\infty} p(x, z - x)$$

if X and Y are independent then:

$$P(Z = z) = p_Z(z) = \sum_{-\infty}^{\infty} p_X(x) p_Y(z - x)$$

This sum is called the **convolution** of the sequences p_X and p_Y.

convolution: continuous case Assume X and Y are continuous rv. Let $Z = X + Y$. Let's start computing the cdf $F(Z)$:

$$F_Z(z) = \int \int_{\{(x,y):x+y\leq z\}} f(x, y)dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, y)dxdy$$

Make a change of variable from y to $v = x + y$ and then reverse the order of integration: if X and Y are independent then:

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f(x, v - x)dvdy = \int_{-\infty}^{z-x} \int_{-\infty}^{\infty} f(x, v - x)dvdy$$

Finally, differentiate to find the density:
$f_Z(z) = \int_{-\infty}^{\infty} f(x, z - x)dx$

Again, if X and Y are independent, the result is a **convolution**: $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)dx$

Example

If X and Y are independent random variables satisfying $X \sim Poisson(\lambda)$ and $Y \sim Poisson(\mu)$, what is the distribution of $U = X + Y$?

Example

If X and Y are independent and identically distributed geometric(θ) random variables, what is the distribution of $U = X + Y$?

Proof. Here $X \sim$ geometric(θ) and $Y \sim$ geometric(θ) with probability functions

$$p_X(x) = \theta(1 - \theta)^x, \quad x = 0, 1, 2, \ldots,$$

and

$$p_Y(y) = \theta(1-\theta)^y, \quad y = 0, 1, 2, \ldots.$$

As X and Y are discrete independent random variables, then the convolution formula becomes

$$P(U = u) = \sum_v p_X(u-v)p_Y(v) = \sum_{v=0}^{u} \theta(1-\theta)^{u-v}\theta(1-\theta)^v$$

$$= \sum_{v=0}^{u} \theta^2(1-\theta)^u = (u+1)\theta^2(1-\theta)^u.$$

Hence $U = X + Y$ has a negative binomial distribution with parameters $r = 2$ and θ. The limits of the summation follow because $p_Y(v) = 0$ for $v < 0$ whilst $p_X(u-v) = 0$ for $v > u$ (so $u - v < 0$). More generally if $X_1, X_2, \ldots, X_r$ are independent geometric random variables with common probability function $p_X(x) = \theta(1-\theta)^x$ for $x = 0, 1, 2, \ldots$, then $U = X_1 + X_2 + \cdots + X_r$ has a negative binomial distribution with probability function

$$p_U(u) = \binom{r+u-1}{u} \theta^r(1-\theta)^u, \quad u = 0, 1, 2, \ldots.$$

□

8 Limit Theorems

In this chapter we will mainly deal with the behaviour of the sum of random variables, as the numer of summands becomes large. Many commonly used statistics, such as tne averages, are a special case of sums. Hence it i useful and inetresting to investigate the results presented in this chapter.

8.1 Markov's Inequality

Proposition Suppose X is a random variable that takes on non-negative values. Then, for all $a > 0$,

$$P(X \geq a) \ \leq \ \frac{E[X]}{a}$$

Proof. For $a > 0$,

$$\begin{aligned} E[X] &= \int_0^\infty x f(x) dx \\ &= \int_0^a x f(x) dx + \int_a^\infty x f(x) dx \end{aligned}$$

Because $X \geq 0$,

$$E[X] \geq \int_a^\infty x f(x) dx \geq \int_a^\infty a f(x) dx = a P(X \geq a)$$

$$\frac{E[X]}{a} \geq P(X \geq a)$$

□

8.2 Chebyshev's Inequality

Proposition If X is a random variable with mean μ and variance σ^2, then, for any value $k > 0$,

$$P(|X - \mu| \geq k) \ \leq \ \frac{\sigma^2}{k^2}$$

Proof. Define the non negative r.v. random variable $Y = (X - \mu)^2$. Set $a = k^2$ and apply Markov's inequality:

$$P(Y \geq k^2) \leq \frac{E[Y]}{k^2}$$
$$P((X - \mu)^2 \geq k^2) \leq \frac{E[(X - \mu)^2]}{k^2}$$
$$P((X - \mu)^2 \geq k^2) \leq \frac{\sigma^2}{k^2}$$

Further we know that, if $(X - \mu)^2 \geq k^2$, we get $|X - \mu| \geq k$. Thus:

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

□

Pafnuti Chebyshev (1821-1894) was a Russian mathematician. Many variants of the spelling of his name exist, Pafnut-i/y (T/Ts)-cheb-i/y-sh/ch/sch-ev/ff/w.

Uses of Chebyshev's (1867) inequality include constructing confidence intervals, so that

$$\text{pr}\{|X - \mu| < 4.472\sigma_X\} > 0.95.$$

If X is known to be a normal random variable, we have the tighter bound

$$\text{pr}\{|X - \mu| < 1.96\sigma_X\} = 0.95.$$

The general application of Chebyshev's inequality has lead it to be used in many fields including digital communication, gene matching, and detecting computer security attacks. lease see the R code file 03-ChebyshevsInequality.R for an application of Chebyshev's inequality.

8.3 Sequence of Random Variables

In the following we would like to clarify the concept of a sequence of independent and identically, distributed random variables.

A single variable X

Observe that, in any probability model, we have a sample space S and a probability measure P. For sake of simplicity, assume that our sample space consists of a finite number of elements:

$$S = \{s_1, s_2, \cdots, s_k\}. \tag{9}$$

We can interpret a r.v. X as a mapping that assigns a real number to any of the possible outcomes $s_i, i = 1, 2, \ldots, k$:

$$X(s_i) = x_i, \qquad \text{for } i = 1, 2, \cdots, k. \tag{10}$$

A sequence of variables $X_1, \ldots, X_n, ...$

Also in the case of a sequence of random variables $X_1, \ldots, X_n, ...$ it is useful to consider that each X_j is a function from S to real numbers. Hence:

$$X_j(s_i) = x_{ji}, \qquad \text{for } i = 1, 2, \cdots, k. \tag{11}$$

It is worth to point out in conclusion that a sequence of random variables is in fact, is a sequence of functions $X_i : S \to \mathbb{R}$. Also, given a sequence of r.v. $X_1, X_2, \ldots, X_n$ and a sequence as sampled **data** $x_1, x_2, \ldots, x_n$ (i.e. n observations), we can say that each observation will be a random variable, say X_i with realization x_i.

8.4 Weak Law of Large Numbers

Proposition Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with mean μ and $Var(X_i) = \sigma^2$. Then, for all $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0 \; as \; n \to \infty$$

Proof. We will prove that $P\left\{\left|\overline{X} - \mu\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}$. It is easy to show that:

$$\frac{E[X_1 + X_2 + \cdots + X_n]}{n} = \frac{\sum_{i=1}^n E[X_i]}{n} = \mu$$

Further,

$$\begin{aligned} E[\left(\frac{\sum_{i=1}^n X_i - \mu}{n}\right)^2] &= \frac{Var(X_1 + X_2 + \cdots + X_n)}{n^2} \\ &= \frac{\sum_{i=1}^n Var(X_i)}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Apply Chebyshev's Inequality:

$$P\left\{\left|\frac{X_1 + X_2 + \ldots + X_n}{n} - \mu\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}$$

□

8.5 Weak Law of Large Numbers: EXAMPLE

Suppose $X_1, X_2, \ldots$ are iid normal distributions,

$$X_i \sim N(0, 10)$$

Suppose we want to guarantee that we have at most a 0.01 probability of being more than 0.1 away from the true μ. How big do we need n?

For the Week Law of Large Numbers we have:

$$P\left\{\left|\overline{X} - \mu\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}$$

Using $\epsilon = 0.1$ and $\sigma^2 = 10$

$$\begin{aligned} 0.01 &= \frac{10}{n(0.1^2)} \\ n &= \frac{1000}{0.01} \\ n &= 100,000 \end{aligned}$$

8.6 Weak Law of Large Numbers: Interpretation

It is a common belief that if we toss a coin *many* times, the proportion of heads will be close to $\frac{1}{2}$.
The law of large numbers is a mathematical interpretation of this believe! Indeed for the coin example we can consider that:

- Successive tosses of a coin can be modelled as independent random trials X_i.
- Each X_i takes on 0 (if the $i-th$ result is tail) or 1 (if the $i-th$ result is head).
- The proportion of heads in n trials is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Hence, the law of large numbers says that $\bar{X}$ *approaches* μ as the number of trials grows.

Example

Let

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Figure 7 shows the cumulative distribution function of Z_1 and Z_4 together with the standard normal cumulative distribution function in the two cases $X_i \overset{\text{ind}}{\sim}$ exponential($\lambda = 1$) and $X_i \overset{\text{ind}}{\sim}$ uniform$(0, 1)$. As $n \to \infty$ it can be seen that in both cases the cumulative distribution function of Z_n tends to that of the standard normal cumulative distribution function.

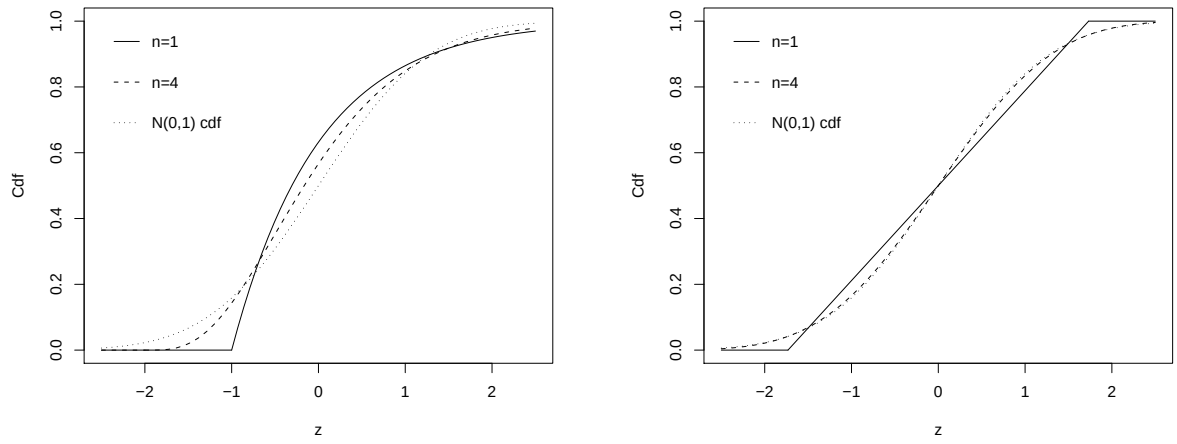


Figure 7: cumulative distribution function of Z_1 and Z_4 for (left) $X_i \overset{\text{ind}}{\sim}$ exponential($\lambda = 1$), (right) $X_i \overset{\text{ind}}{\sim}$ uniform$(0, 1)$.

8.7 About Convergence

8.7.1 Sequences and Convergence: Recalls

Sequence of real numbers:

$$\{a_i\}_{i=1}^{\infty} = \{a_1, a_2, a_3, \ldots, a_n, \ldots,\}$$

Definition. We say that the sequence $\{a_i\}_{i=1}^{\infty}$ converges to real number A if for each $\epsilon > 0$ there is a positive integer n_ϵ such that for $n \geq n_{epsilon}$, $|a_n - A| < \epsilon$

8.7.2 Sequence of functions and Convergence: Recalls

$$\{f_i\}_{i=1}^{\infty} = \{f_1, f_2, f_3, \ldots, f_n, \ldots,\}$$

Definition. Suppose $f_i : X \to \Re$ for all i. Then $\{f_i\}_{i=1}^{\infty}$ converges *pointwise* to f if, for all $x \in X$ and $\epsilon > 0$, there is an n_ϵ such that for all $n \geq n_\epsilon$,

$$|f_n(x) - f(x)| < \epsilon$$

This is a strong statement!

8.7.3 Convergence in Probability and in Distribution

Let $\widehat{\theta}_i$ be an estimator for θ based on i observations. Observe that Increasing the sample size we get a Sequence of estimators:

$$\left\{\widehat{\theta}_i\right\}_{i=1}^{n} = \left\{\widehat{\theta}_1, \widehat{\theta}_2, \widehat{\theta}_3, \ldots, \widehat{\theta}_n\right\}$$

We aim to address the questions:

- What can we say about $\left\{\widehat{\theta}_i\right\}_{i=1}^{n}$ as $n \to \infty$?
- What is the probability that $\widehat{\theta}_n$ differs from θ?
- What is the probability $\left\{\widehat{\theta}_i\right\}_{i=1}^{n}$ converges to θ?

Convergence in Probability. We will say the sequence $\widehat{\theta}_n$ converges in probability to θ (perhaps a non-degenerate RV) if,

$$\lim_{n\to\infty} Prob(|\widehat{\theta}_n - \theta| > \epsilon) = 0$$

For any $\epsilon > 0$ In particular:

- ϵ is a *tolerance* parameter. Broadly speaking it expressed how much error around θ we are goinf to *tolerate*.
- In the limit, convergence in probability implies the $\widehat{\theta}_n$ distribution collapses on a spike at θ.
- $\left\{\widehat{\theta}_i\right\}$ does not need actually converge to θ, only P($|\theta_n$ - $\theta| > \epsilon) = 0$

In symbols we say:

$$\widehat{\theta}_n \rightarrow^p \theta$$

0r

$$p\ lim_{n\to\infty}\widehat{\theta}_n = \theta$$

8.7.4 Convergence in Distribution

$\widehat{\theta}_n$, with cdf $F_n(x)$, converges in distribution to random variable X with cdf $F(x)$ if

$$\lim_{n\to\infty} |F_n(x) - F(x)| = 0$$

For all $x \in \Re$ where $F(x)$ is continuous.
Note that the previous definition provide a result for the cdfs, but says nothing about convergence of underlying RV The convergence in distribution will help us sometimes for justifying the use of some sampling distributions

In symbols we say:

$$\widehat{\theta}_n \rightarrow^D X$$

or

$$\widehat{\theta}_n \rightarrow^D pdf_X$$

where pdf_X is the probability distribution of X.

8.8 Central Limit Theorem

Proposition. Let $X_1, X_2, \ldots$ be a sequence of independent random variables with mean μ and variance σ^2. Let X_i have a cdf $P(X_i \leq x) = F(x)$ and moment generating function $M(t) = E[e^{tX_i}]$. Let $S_n = \sum_{i=1}^{n} X_i$. Then

$$\lim_{n\to\infty} P\left(\frac{S_n - \mu n}{\sigma\sqrt{n}} \leq x\right) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{x^2}{2}\right)$$

Proof plan:

1) Rely on Fact that convergence of MGFs$\rightsquigarrow$ convergence in CDFs

2) Show that MGFs, in limit, converge on normal MGF

In our proof, we will make use of the following results:

Proposition. Let F_n be a sequence of cumulative distribution functions with the corresponding moment generating functions M_n. F be a cdf with the moment generating functions M. If $\lim_{n\to\infty} M_n(t) \to M(t)$ for all t in some interval, then $F_n(x) \rightsquigarrow F(x)$ for all x (when F is continuous).

Proposition. Suppose $\lim_{n\to\infty} a_n = a$, then

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n \;=\; e^a$$

Proposition. Suppose $M(t)$ is a moment generating function some random variable X. Then $M(0) = 1$.

Proof. Suppose $X_1, \ldots, X_n$ are iid variables with $E[X] = 0$, variance σ_x^2, Moment Generating Function (MGF) $M_x(t)$.

Let $S_n = \sum_{i=1}^{n} X_i$ and $Z_n = \frac{S_n}{\sigma_x\sqrt{n}}$.
$M_{S_n} = (M_x(t))^n$ and $M_{Z_n}(t) = \left(M_x\left(\frac{t}{\sigma_x\sqrt{n}}\right)\right)^n$
Using Taylor's Theorem we can write

$$M_x(s) \;=\; M_x(0) + sM_x^{'}(0) + \frac{1}{2}s^2 M_x^{''}(0) + e_s$$

$e_s/s^2 \to 0$ as $s \to 0$.

$$M_x(s) = M_x(0) + sM_x^{'}(0) + \frac{1}{2}s^2M_x^{''}(0) + e_s$$

Filling in the values we have

$$M_x(s) = 1 + 0 + \frac{\sigma_x^2}{2}s^2 + \underbrace{e_s}_{\text{Goes to zero}}$$

Set $s = \frac{t}{\sigma_x\sqrt{n}}$ $\lim_{n\to\infty} s \to 0$. Then

$$\begin{aligned} M_{Z_n}(t) &= \left(1 + \frac{\sigma_x^2}{2}\left(\frac{t}{\sigma_x\sqrt{n}}\right)^2\right)^n \\ &= \left(1 + \frac{t^2/2}{n}\right)^n \\ \lim_{n\to\infty} M_{Z_n}(t) &= e^{\frac{t^2}{2}} \end{aligned}$$

□

8.8.1 The Central Limit Theorem, few comments

The Central Limit Theorem (CLT) deals with the long-run behaviour of the sample mean as n grows. In general, the colloquial result is that

everything becomes Normal eventually.

During the lecture we formalized this concept and proved the theorem. In the following I'll try to give you more insight into this incredibly powerful result.

Consider i.i.d. random variables $X_1, X_2, \ldots, X_n$ each with meanμ and variance σ^2 (NOTE WELL: it doesn't necessarily have to be a Normal distribution! For example, if we had that each X was distributed Unif(0,1), then μ=1/2 and σ^2=1/12).

Again, define $\bar{X}_n$ as the *sample mean* of X. We can write this out as:

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}$$

As discussed above, we know that the sample mean is itself a random variable, and we know that it approaches the true mean in the long run by the Law of Large Numbers (LLN). However, are we able to nail down a specific distribution for this random variable as we approach the long run, not just the value

that it converges to? A good place to start is to find the mean and variance; these parameters won't tell us what the distribution is, but they will be useful once we determine the distribution. To find the expectation and variance, we can just *brute force* our calculations. First, for the expectation, we take the expectation of both sides:

$$E(\bar{X}_n) = E(\frac{X_1 + \ldots + X_n}{n})$$

By linearity:

$$= E(\frac{X_1}{n}) + \ldots + E(\frac{X_n}{n})$$

Since n is a known constant, we can factor it out of the expectation:

$$= \frac{1}{n}E(X1) + \ldots + \frac{1}{n}E(Xn)$$

Now, we are left with the expectation, or mean, of each X. Do we know these values? Well, recall above that, by the set-up of the problem, each X is a random variable with mean μ. That is, the expectation of each X is μ. We get:

$$= \mu_n + ... + \mu_n$$

We have n of these terms, so they sum to:

$$= \mu$$

Hence, we get that $E(\bar{X}_n) = \mu$. Think about this result: it says that the average of the sample mean is equal to μ, where μ is the average of each of the random variables that make up the sample mean. This is intuitive; the sample mean should have an average of μ (you could say that $\bar{X}_n$ is unbiased for μ, since it has expectation μ; this is a concept that you will explore more in detail). Let's now turn to the Variance:

$$Var(\bar{X}_n) = Var\big(\frac{X_1 + ... + X_n}{n}\big)$$

We know that the X terms are independent, so the variance of the sum is the sum of the variances:

$$= Var\big(\frac{X_1}{n}\big) + ... + Var\big(\frac{X_n}{n}\big)$$

Since n is a constant, we factor it out (remembering to square it):

$$= \frac{1}{n^2}Var(X_1) + ... + \frac{1}{n^2}Var(X_n)$$

Do we know the variance of each X term? In the set-up of the problem, it was defined as ?2, so we can simply plug in ?2 for each variance:

$$= \frac{\sigma^2}{n^2} + ... + \frac{\sigma^2}{n^2}$$

We have n of these terms (since we have n random variables and thus n variances) so this simplifies to:

$$= \frac{\sigma^2}{n}$$

Consider this result for a moment. First, consider when n=1. In this case, the sample mean is just X_1 (since, by definition, we would have $\frac{X_1}{1} = X_1$). The variance that we calculated above, $\frac{\sigma^2}{n}$, comes out to σ^2 when $n = 1$, which makes sense, since this is just the variance of X_1. Next, consider what happens to this variance as n grows; it gets smaller and smaller, since n is in the denominator. Does this make sense? As n grows, we are essentially adding up more and more random variables in our sample mean calculation. It makes sense, then, that the overall sample mean will have less variance; among other things, adding up more random variables means that the effect of *outlier* random variables is lessened (i.e., if we observe an extremely large value for X_1, it is mediated by the sheer number of random variables).

So, we found that the sample mean has mean μ and variance $\frac{\sigma^2}{n}$, where μ is the mean of each underlying random variable, σ^2 is the variance of each underlying random variable, and n is the total number of random variables. Now that we have the parameters, we are ready for the main result of the CLT.

The CLT states that, for large n, the distribution of the sample mean approaches a Normal distribution. This is an extremely powerful result, because it holds no matter what the distribution of the underlying random variables (i.e., the X's) is. We know that Normal random variables are governed by the mean and variance (i.e., these are the two parameters), and we already found the mean and variance of $\bar{X}_n$, so we can say:

$$\bar{X}_n \rightarrow^D N(\mu, \frac{\sigma^2}{n})$$

Where $\rightarrow^D$ means *converges in distribution*; it's implied here that this convergence takes place as n, or the number of underlying random variables, grows.

Think about this distribution as n gets extremely large. The mean, μ, will be unaffected, but the variance will be close to 0, so the distribution will essentially be a constant (specifically, the constant μ with no variance). This makes sense:

if we take an extremely high number of draws from a distribution, we should get that this sample mean is at the true mean, with very little variance. It?s also the result we saw from the LLN, which said that the sample mean approaches a constant: as n grows here, we approach a variance of 0, which essentially means we have a constant (since constants have variance 0). The CLT just describes the distribution *on the way* to the LLN convergence.

Hopefully this brings some clarity to the statement "everything becomes Normal": taking the sum of i.i.d. random variables (we worked with the sample mean here, but the sample mean is just the sum divided by a constant n), regardless of the underlying distribution of the random variables, yields a Normal distribution.
Please check out the Rcode file, 04-CentralLimitTheorem.R, in the R examples material, for a graphical interpretation of the CLT.
Further Reference: `https://bookdown.org/probability/beta/`.

8.8.2 Example of derivation of the CLT for the exponential distribution

Few observations.

- The distribution of an average tends to be Normal, even when the distribution from which the average is computed is decidedly non-Normal!
- This normal distribution will have the same mean as the parent distribution, **AND**, variance equal to the variance of the parent divided by the sample size.
- The distribution of the phenomenon under study does not have to be Normal, but its average will be.

Given these premises, let's try to **derive the CLT for the exponential distribution**.
Assume $X_1, \ldots, X_n \sim exponential(\lambda)$ and independent. We know that:

$$\begin{aligned} \mu &= 1/\lambda \\ \sigma^2 &= 1/\lambda^2 \\ M_X(t) &= \frac{\lambda}{\lambda - t} \end{aligned}$$

and also: $M_{\sum_i X_i}(t) = \left(\frac{\lambda}{\lambda - t}\right)^n$
and $Z_n = \frac{\frac{\sum_i X_i}{n} - \frac{1}{\lambda}}{\frac{1}{\lambda\sqrt{n}}} = \frac{\lambda \sum_i X_i}{\sqrt{n}} - \sqrt{n} = a \sum_i X_i + b$

It follows that:

$$\begin{aligned} M_{Z_n}(t) &= E[e^{tZ_n}] = e^{bt} M_{\sum_i X_i(at)} = e^{-\sqrt{n}t} \left(\frac{\lambda}{\lambda - \lambda t/\sqrt{n}} \right)^n \\ &= e^{-\sqrt{n}t}(1 - \frac{t}{\sqrt{n}})^{-n}. \end{aligned}$$

Let's consider the *log* and remember that: $log(1+x) \simeq x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + ...$ we get

$$\begin{aligned} log(M_{Z_n}(t)) &= -\sqrt{n}t - nlog(1 - \frac{t}{\sqrt{n}}) \simeq -\sqrt{n}t - n(-\frac{t}{\sqrt{n}} - \frac{t^2}{2n} - \frac{t^3}{3n\sqrt{n}}) \\ &\simeq \frac{1}{2}t^2 + \frac{t^3}{3n\sqrt{n}} \end{aligned}$$

It is easy to see that as $n \to \infty$, then $log(M_{Z_n}(t)) \to \frac{1}{2}t^2$ and $(M_{Z_n}(t)) \to e^{\frac{1}{2}t^2}$hence:

$$Z_n \to^D N(0,1)$$

Counter example CLT

The distribution of an average will tend to be Normal as the sample size increases, regardless of the distribution from which the average is taken except when the moments of the parent distribution do not exist.

if $X \sim Cauchy$ with $f_X(x) = \frac{1}{\pi(1+x^2)}$ where $-\infty < x < \infty$.

The Characteristic function of X is:

$$\phi_X(t) = E\left[e^{itX}\right] = \int_{-\infty}^{\infty} \frac{e^{itx}}{\pi(1+x^2)} dx = e^{-|t|}$$

If $X_1, \ldots, X_n \sim^{iid} Cauchy$ and $S_n = X_1 + \ldots X_n$ then:

$$\begin{aligned} \phi_{S_n}(t) &= E\left[e^{it(X_1 + \ldots X_n}\right] = E\left[e^{it(X_1)} e^{it(X_2)} \ldots e^{it(X_n)}\right] \\ &= E[e^{it(X_1)}]E[e^{it(X_2)}] \ldots E[e^{it(X_n)}] \\ &= (e^{-|t|})^n = e^{-n|t|} \end{aligned}$$

It follows that for $\bar{X}_n = \frac{S_n}{n}$ we have:

$$\begin{aligned} \phi_{\bar{X}_n}(t) &= E[e^{it\bar{X}}] = E[e^{it(S_n/n)}] = E[e^{i(t/n)S_n}] = \phi_{S_n}(t/n) \\ &= e^{-n|t/n|} = e^{-|t|} \end{aligned}$$

and so:

$$\bar{X}_n \sim Cauchy, \forall n$$

9 Estimation

The main content of this chapter relies on professor Simon Myers' teaching material.
Url: `http://www.stats.ox.ac.uk/~myers/stats_materials.html`

9.1 Desirable properties of estimators

Suppose we have $X = X_1, X_2, \ldots, X_n$ drawn from a distribution with some parameter θ. Often $X_1, X_2, \ldots, X_n$ form a *sample*.

Definition *Estimators*

An *estimator* $\widehat{\theta}_n$ of θ is just a function of the observed random variables which (we hope) forms a useful approximation to the parameter, once the data are observed:

$$\widehat{\theta}_n = g(X_1, X_2, \ldots, X_n).$$

Note that $\widehat{\theta}_n$ can depend only on the observed random variables, and not on any unknown parameters. You have already met a number of estimators in this and past statistic's modules , such as the sample mean and sample variance. The estimator is a function of random variables, so is itself a random variable, with a distribution, mean, and variance, etc.

Here is a familiar property which you have already met.

Definition *Unbiasedness*

$\widehat{\theta}_n$ is said to be *unbiased* for θ if

$$E\left(\widehat{\theta}_n\right) = \theta, \quad \forall \theta \in \Theta.$$

□

Next a new property, but one which appeals to common-sense. The idea is that, the larger the amount of data, the closer the estimate should be to the parameter to be estimated. This is expressed in terms of the estimator converging in probability to the parameter value.

Convergence in probability (borrowed from probability course)

A sequence of random variables $Z_1, Z_2, \ldots, Z_n, \ldots$ is said to *converge in probability* to a constant z if for any $\epsilon > 0$, as $n \to \infty$

$$P(|Z_n - z| > \epsilon) \to 0.$$

Definition *Consistency*

$\widehat{\theta}_n$ is said to be *consistent* for θ if

$$\widehat{\theta}_n \xrightarrow{P} \theta.$$

□

Definition *Efficiency*

$\widehat{\theta}_A$ is said to be more *efficient* than $\widehat{\theta}_B$ if

$$Var\left(\widehat{\theta}_A\right) < Var\left(\widehat{\theta}_B\right), \quad \forall \theta \in \Theta.$$

□

Again, this appeals to common-sense.

9.2 Revision and extension of maximum likelihood estimation

The basic idea starts with the joint distribution of $X = X_1, X_2, \ldots, X_n$ depending upon a parameter θ,

$$f(\mathbf{x}; \theta) = f(x_1, x_2, \ldots, x_n; \theta).$$

For fixed θ, probability statements can be made about X. If we have observations, x, but θ is unknown, we regard information about θ as being contained in the likelihood

$$L(\theta; \mathbf{x}) = f(\mathbf{x}; \theta),$$

where L is regarded as a function of θ with $\mathbf{x}$ fixed.

Example

Suppose $X = X_1, X_2, \ldots, X_n$ are independent Bernoulli random variables with parameter $\theta \in [0, 1]$.

$$i.e. \quad P(X_i = 1) = \theta, \; P(X_i = 0) = 1 - \theta.$$

Observations are $x = (1, 0, 0, 1, 0, 1, 1)$ and

$$\begin{aligned} L(\theta; \mathbf{x}) &= \prod_{i=1}^{7} f(x_i; \theta) \\ &= \prod_{i=1}^{7} \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^4 (1-\theta)^3 . \end{aligned}$$

In general, for a sample size n,

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum x_i} (1-\theta)^{n-\sum x_i} .$$

■

Maximum likelihood estimation

The value of θ which maximises $L(\theta)$ is called the *maximum likelihood estimate* of θ.

Example (continued)

In the previous example, we can find this by differentiating $L(\theta)$, or equivalently by differentiating $l(\theta) = \log L(\theta)$.

$$\begin{aligned} l(\theta) &= \textstyle\sum_i x_i \log\theta + (n - \sum_i x_i) \log(1-\theta) \\ \frac{\partial l(\theta)}{\partial \theta} &= \textstyle\sum_i x_i / \theta - (n - \sum_i x_i) / (1-\theta) \end{aligned}$$

Putting $\dfrac{\partial l(\theta)}{\partial \theta} = 0$ we obtain

$$\theta = \textstyle\sum_i x_i \Big/ n.$$

This leads us to $\widehat{\theta}_n$, the *maximum likelihood estimator* based on a sample of size n

$$\widehat{\theta}_n = \textstyle\sum_i X_i \Big/ n.$$

■

Example

Suppose $X = X_1, X_2, \ldots, X_n$ are independent normal random variables, $N\left(\mu, \sigma^2\right)$.

$$\begin{aligned} L\left(\mu, \sigma^2\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \end{aligned}$$

and

$$\begin{aligned} l\left(\mu, \sigma^2\right) &= \log L\left(\mu, \sigma^2\right) \\ &= -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

Differentiating,

$$\begin{aligned} \frac{\partial l\left(\mu, \sigma^2\right)}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu), \\ \frac{\partial l\left(\mu, \sigma^2\right)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

Equating these derivatives to zero results in

$$\widehat{\mu} = \overline{X}, \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

■

9.3 Comparing estimators

Remember that we began with three desirable properties of estimators. These were

Unbiasedness	$E\left(\widehat{\theta}_n\right) = \theta, \quad \forall\theta \in \Theta.$
Consistency	$\widehat{\theta}_n \xrightarrow{P} \theta.$
Efficiency	$\widehat{\theta}_A$ is more *efficient* than $\widehat{\theta}_B$ if $Var\left(\widehat{\theta}_A\right) < Var\left(\widehat{\theta}_B\right), \quad \forall\theta \in \Theta.$

Even if we assume unbiasedness and consistency to be desirable, it is possible to have more than one such estimator.

Example

Consider the linear estimator

$$\widehat{\theta}_n = \sum_{i=1}^{n} a_i X_i$$

where $E(X_i) = \theta,\ Var(X_i) = \sigma^2$ for $1 \leq i \leq n$.

$$E\left(\widehat{\theta}_n\right) = \sum_{i=1}^{n} a_i E(X_i) = \theta \sum_{i=1}^{n} a_i,$$

so the estimator is unbiased provided

$$\sum_{i=1}^{n} a_i = 1.$$

For i.i.d. random variables,

$$Var\left(\widehat{\theta}_n\right) = \sum_{i=1}^{n} a_i^2 \sigma^2.$$

Now

$$\begin{aligned} \sum_{i=1}^{n} \left(a_i - \frac{1}{n}\right)^2 &\geq 0 \\ \Rightarrow \sum_{i=1}^{n} a_i^2 - \frac{2}{n}\sum_{i=1}^{n} a_i + \frac{1}{n} &\geq 0 \end{aligned}$$

and, if $\sum_{i=1}^{n} a_i = 1$,

$$\Rightarrow \quad \sum_{i=1}^{n} a_i^2 \geq \frac{1}{n}.$$

Equality occurs iff $a_i = \dfrac{1}{n},\ 1 \leq i \leq n.$

The conclusion then is that, if $\widehat{\theta}_n$ is a linear unbiased estimator of the form $\sum_{i=1}^{n} a_i X_i$ and if $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$, then

$$Var\left(\overline{X}\right) \leq Var\left(\widehat{\theta}_n\right).$$

Of all linear unbiased estimators, $\overline{X}$ is most efficient.

■

9.3.1 Food for thought

1. Is the sample mean the "best" estimator of the distribution mean?

2. If $\overline{X}$ is unbiased for θ, is $g\left(\overline{X}\right)$ unbiased for $g\left(\theta\right)$?

3. Can a biased estimator be "better" (whatever that means!) than an unbiased estimator?

9.4 Is the sample mean the "best" estimator of the distribution mean?

It is not true that sample mean is the 'best' choice of estimator of the population mean for any underlying parent distribution. The only thing true regardless of the population distribution is that the sample mean is an unbiased estimator of the population mean, i.e. $E(\overline{X}) = \mu$.

Now unbiasedness is often not the only criteria considered for choosing an estimator of your unknown quantity of interest. We usually prefer estimators that have smaller variance or smaller mean squared error (MSE) in general, because it is a desirable property to have in an estimator. And it might be the case that $\overline{X}$ does not attain the minimum variance/MSE among all possible estimators.

Consider a sample $(X_1, X_2, \ldots, X_n)$ drawn from a uniform distribution on $(0, \theta)$. Now $T_1 = \overline{X}$ is an unbiased estimator of the population mean $\theta/2$, but it does not attain the minimum variance among all unbiased estimators of $\theta/2$. It can be shown that the uniformly minimum variance unbiased estimator

(UMVUE) of the population mean is instead $T_2 = \frac{n+1}{2n}\max(X_1, \ldots, X_n)$. So T_2 is the best estimator within the unbiased class where 'best' means 'having the smallest variance'.

9.5 Can a biased estimator be "better" than an unbiased estimator?

Look at the distributions of the following two estimators of θ.

Figure 2.8

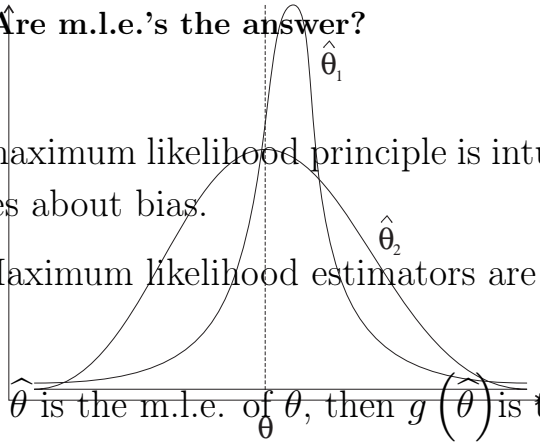
Which do you prefer?

Even though $\widehat{\theta}_1$ is biased,

$$E\left[\left(\widehat{\theta}_1 - \theta\right)^2\right] < Var\left(\widehat{\theta}_2\right).$$

$\widehat{\theta}_1$ has smaller *mean squared error.*

9.6 Are m.l.e.'s the answer?



The maximum likelihood principle is intuitively appealing and does not involve worries about bias.

(i) Maximum likelihood estimators are asymptotically unbiased.

(ii) If $\widehat{\theta}$ is the m.l.e. of θ, then $g\left(\widehat{\theta}\right)$ is the m.l.e. of $g(\theta)$.

This is the *invariance property* of maximum likelihood estimators.

9.7 More about likelihood

9.7.1 Invariance property of m.l.e.'s

Lemma 2.4 If $\widehat{\theta}$ is an m.l.e. of θ and if g is a function, then $g\left(\widehat{\theta}\right)$ is an m.l.e. of $g(\theta)$.
□

Proof If g is one-to-one, then

$$L(\theta) = L\left(g^{-1}\left(g(\theta)\right)\right)$$

are both maximised by $\widehat{\theta}$, so

$$\widehat{\theta} = g^{-1}\left(\widehat{g(\theta)}\right)$$

or

$$g\left(\widehat{\theta}\right) = \widehat{g(\theta)}.$$

If g is many-to-one, then $\widehat{\theta}$ which maximises $L(\theta)$ still corresponds to $g\left(\widehat{\theta}\right)$, so $g\left(\widehat{\theta}\right)$ still corresponds to the maximum of $L(\theta)$
■

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution $B(1, \theta)$. Consider m.l.e.'s of the mean, θ, and variance, $\theta(1-\theta)$.

Note, by the way, that $\theta(1-\theta)$ is not a 1-1 function of θ.

The log-likelihood is

$$l(\theta) = \sum_i x_i \log\theta + (n - \sum_i x_i)\log(1-\theta)$$

and

$$\frac{dl(\theta)}{d\theta} = \sum_i x_i / \theta - (n - \sum_i x_i)/(1-\theta)$$

so it is easily shown that the m.l.e. of θ is $\widehat{\theta} = \overline{X}$.

Putting $\nu = \theta(1-\theta)$,

$$\frac{dl(\nu)}{d\nu} = \frac{dl\left(\nu(\theta)\right)}{d\theta} \cdot \frac{d\theta}{d\nu}$$

so it is easily seen that, since $\dfrac{d\theta}{d\nu}$ is not, in general, equal to zero,

$$\widehat{\nu} = \nu\left(\widehat{\theta}\right) = \overline{X}\left(1 - \overline{X}\right).$$

■

9.7.2 Relative likelihood

If $\sup_\theta L(\theta) < \infty$, the *relative likelihood* is

$$RL(\theta) = \frac{L(\theta)}{\sup_\theta L(\theta)}; \quad 0 \leq RL(\theta) \leq 1.$$

Relative likelihood is invariant to known 1-1 transformations of x, for if y is a 1-1 function of x,

$$f_Y(y;\theta) = f_X\left(x(y);\theta\right)\left|\frac{dx}{dy}\right|.$$

$\left|\frac{dx}{dy}\right|$ is independent of θ, so

$$RL_X(\theta) = RL_Y(\theta).$$

9.7.3 Likelihood summaries

Realistic statistical problems often have many parameters. These cause problems because it can be hard to visualise $L(\theta)$, and it becomes necessary to use summaries.

Key idea

In large samples, log-likelihoods are often approximately quadratic near the maximum.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ.
i.e.

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Then

$$l(\lambda) = n\log\lambda - \lambda\textstyle\sum_i x_i, \qquad \frac{dl(\lambda)}{d\lambda} = n/\lambda - \textstyle\sum_i x_i,$$

$$\frac{d^2l(\lambda)}{d\lambda^2} = -n/\lambda^2, \qquad \frac{d^3l(\lambda)}{d\lambda^3} = 2n/\lambda^3.$$

The log-likelihood has a maximum at $\widehat{\lambda} = n \,/\sum_i x_i$, so

$$\begin{aligned} RL(\lambda) = & \left(\frac{\lambda}{\widehat{\lambda}}\right)^n e^{n-\lambda\sum_i x_i} \\ = & \left(\frac{\lambda}{\widehat{\lambda}} e^{1-\lambda/\widehat{\lambda}}\right)^n, \quad \lambda > 0. \\ \to & \ 1 \quad \text{as} \quad \lambda \to \widehat{\lambda}. \end{aligned}$$

Now, what does the likelihood look like in the neighbourhood of $\widehat{\lambda}$, as $n \to \infty$?

$$\begin{aligned} \log RL(\lambda) &= l(\lambda) - l(\widehat{\lambda}) \\ &= l(\widehat{\lambda}) + l'(\widehat{\lambda})\left(\lambda - \widehat{\lambda}\right) + \frac{1}{2} l''(\widehat{\lambda})\left(\lambda - \widehat{\lambda}\right)^2 - l(\widehat{\lambda}) \\ &+ O(\lambda - \widehat{\lambda})^3 \end{aligned}$$

using Taylor series.

Now $l'(\widehat{\lambda}) = 0$ and $l''(\widehat{\lambda}) = -n \Big/ \widehat{\lambda}^2$, so

$$\log RL(\lambda) \simeq -\frac{n\left(\lambda - \widehat{\lambda}\right)^2}{2\widehat{\lambda}^2} \to -\infty \quad \text{as} \quad n \to \infty$$

unless $\lambda = \widehat{\lambda}$.

Thus, as $n \to \infty$,

$$RL(\lambda) \to \begin{cases} 1, & \lambda = \widehat{\lambda}, \\ 0, & \text{otherwise.} \end{cases}$$

Conclusion

Likelihood becomes more concentrated about the maximum as $n \to \infty$, and values far from the maximum become less and less plausible.

■

In general

We call the value $\widehat{\theta}$ which maximises $L(\theta)$ or, equivalently, $l(\theta) = \log L(\theta)$ the *maximum likelihood estimate*, and

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2}$$

is called the *observed information.*

Usually $J(\theta) > 0$ and $J(\widehat{\theta})$ measures the concentration of $l(\theta)$ at $\widehat{\theta}$. Close to $\widehat{\theta}$, we summarise

$$l(\theta) \simeq l(\widehat{\theta}) - \frac{1}{2}\left(\theta - \widehat{\theta}\right)^2 J(\widehat{\theta}).$$

9.7.4 Information

In a model with log-likelihood $l(\theta)$, the *observed information* is

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2}.$$

When observations are independent, $L(\theta)$ is a product of densities so

$$l(\theta) = \sum_i \log f\left(x_i; \theta\right)$$

and

$$J(\theta) = -\sum_i \frac{\partial^2}{\partial \theta^2} \log f\left(x_i; \theta\right).$$

Since

$$l(\theta) \simeq l(\widehat{\theta}) - \frac{1}{2}\left(\theta - \widehat{\theta}\right)^2 J(\widehat{\theta}),$$

for θ near to $\widehat{\theta}$, we see that large $J(\widehat{\theta})$ implies that $l(\theta)$ is more concentrated about $\widehat{\theta}$.

This means that the data are less ambiguous about possible values of θ, *i.e.* we have more information about θ.

9.8 Univariate distributions

Before an experiment is conducted, we have no data so that we cannot evaluate $J(\theta)$.

But we can find its expected value

$$I(\theta) = E\left(-\frac{\partial^2 l(\theta)}{\partial \theta^2}\right).$$

This is called the *expected information* or *Fisher's information.*

If the observations are a random sample, then the whole sample expected information is

$$I(\theta) = ni(\theta)$$

where

$$i(\theta) = E\left(-\frac{\partial^2}{\partial\theta^2}\log f\left(X_i;\theta\right)\right),$$

the single observation Fisher information.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Poisson distribution with parameter θ.

$$L(\theta) = \prod_{i=1}^{n}\frac{\theta^{x_i}e^{-\theta}}{x_i!},$$

giving

$$l(\theta) = \log L(\theta) = \sum_i x_i \log\theta - n\theta - \sum_i \log x_i!$$

Thus

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial\theta^2} = \sum\nolimits_i x_i \Big/ \theta^2.$$

To find $I(\theta)$, we need $E\left(X_i\right) = \theta$ and

$$I(\theta) = \frac{1}{\theta^2}\sum_i E\left(X_i\right) = \frac{n}{\theta}.$$

■

9.9 Multivariate distributions

If θ is a $(p \times 1)$ vector of parameters, then $\mathbf{I}(\theta)$ and $\mathbf{J}(\theta)$ are $(p \times p)$ matrices.

$$\left\{\mathbf{J}(\theta)\right\}_{rs} = -\frac{\partial^2 l(\theta)}{\partial\theta_r\partial\theta_s} \quad \text{and} \quad \left\{\mathbf{I}(\theta)\right\}_{rs} = E\left(-\frac{\partial^2 l(\theta)}{\partial\theta_r\partial\theta_s}\right).$$

These matrices are obviously symmetric.

We can also write the above as

$$\mathbf{J}(\theta) = -\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T} \quad \text{and} \quad \mathbf{I}(\theta) = E\left(-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}\right).$$

Example

$X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with parameters μ and σ^2. We have already seen that

$$L\left(\mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum\nolimits_i \left(x_i - \mu\right)^2\right],$$

so

$$l\left(\mu, \sigma^2\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum\nolimits_i \left(x_i - \mu\right)^2.$$

and

$$\begin{aligned}
\frac{\partial l}{\partial \mu} &= \frac{1}{\sigma^2}\sum\nolimits_i \left(x_i - \mu\right), \\
\frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum\nolimits_i \left(x_i - \mu\right)^2, \\
\frac{\partial^2 l}{\partial \mu^2} &= -\frac{n}{\sigma^2}, \\
\frac{\partial^2 l}{\partial \mu \partial \sigma^2} &= -\frac{1}{\sigma^4}\sum\nolimits_i \left(x_i - \mu\right). \\
\frac{\partial^2 l}{\partial \left(\sigma^2\right)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum\nolimits_i \left(x_i - \mu\right)^2.
\end{aligned}$$

$$\mathbf{J}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_i \left(x_i - \mu\right) \\ \frac{1}{\sigma^4}\sum_i \left(x_i - \mu\right) & \frac{1}{\sigma^6}\sum_i \left(x_i - \mu\right)^2 - \frac{n}{2\sigma^4} \end{pmatrix}.$$

To find $\mathbf{I}(\mu, \sigma^2)$, use

$$\begin{aligned}
E\left(X_i\right) &= \mu, \\
Var\left(X_i\right) &= E\left[\left(X_i - \mu\right)^2\right] = \sigma^2,
\end{aligned}$$

so that

$$\mathbf{I}(\mu, \sigma^2) = E\left(\mathbf{J}(\mu, \sigma^2)\right) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Example *Censored exponential data*

Lifetimes of n components, safety devices, etc. are observed for a time c, when r have failed and $(n-r)$ are still OK.

We have two kinds of observation:

(i) Exact failure times x_i observed if $x_i \leq c$, so that

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0;$$

(ii) x_i unobserved if $x_i > c$,

$$P(X > c) = e^{-\lambda c}.$$

Data are therefore $x_1, \ldots, x_r, \underbrace{c, \ldots, c}_{n-r \text{ times}}$

The $(n-r)$ components, safety devices, etc. which have not failed are said to be *censored*.

The likelihood is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{r} \lambda e^{-\lambda x_i} \prod_{i=r+1}^{n} e^{-\lambda c} \\ &= \lambda^r \exp\left[-\lambda\left(\sum_{i=1}^{r} x_i + (n-r)c\right)\right]. \end{aligned}$$

$$l(\lambda) = r\log\lambda - \lambda\left(\textstyle\sum_{i=1}^{r} x_i + (n-r)c\right)$$

$$l'(\lambda) = r/\lambda - \left(\textstyle\sum_{i=1}^{r} x_i + (n-r)c\right)$$

$$l''(\lambda) = -r/\lambda^2 .$$

Thus $J(\lambda) = r/\lambda^2 > 0$ if $r > 0$ so we must observe *at least one* exact failure time.

$$I(\lambda) = E\left(r/\lambda^2\right) = \frac{1}{\lambda^2} E\left(\#X_i \text{ observed exactly.}\right)$$

Now $P(X_i \text{ observed exactly}) = P(X_i \leq c) = 1 - e^{-\lambda c}$, so

$$I_c(\lambda) = \frac{n\left(1 - e^{-\lambda c}\right)}{\lambda^2}.$$

No censoring if $c \to \infty$, giving

$$I_\infty(\lambda) = \frac{n}{\lambda^2} > I_c(\lambda)$$

as one might expect.

The asymptotic efficiency when there is censoring at c relative to no censoring is

$$I_c(\lambda) / I_\infty(\lambda) = 1 - e^{-\lambda c}.$$

■

Revision Example *Events in a Poisson process*

Events are observed for period $(0, T)$.

n events occur at times $0 < t_1 < t_2 < \ldots < t_n < T$

Two observers A and B. A records exact times, B uses an automatic counter and goes to the pub (*i.e.* B merely records how many events there are).

A knows exact times, and times between events are independent and exponentially distributed, so

$$\begin{aligned} L_A(\lambda) &= \lambda e^{-\lambda t_1} \times \lambda e^{-\lambda(t_2 - t_1)} \times \cdots \times \lambda e^{-\lambda(t_n - t_{n-1})} \times e^{-\lambda(T - t_n)} \\ &= \lambda^n e^{-\lambda T}. \end{aligned}$$

B merely observes the event $[N = n]$, where $N \sim Poi(\lambda T)$, so

$$L_B(\lambda) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}.$$

Log-likelihoods are

$$\begin{aligned} l_A(\lambda) &= n \log \lambda - \lambda T, \\ l_B(\lambda) &= n \log \lambda + n \log T - \lambda T - \log n! \end{aligned}$$

and

$$J_A(\lambda) = J_B(\lambda) = n / \lambda^2 .$$

$E(N) = \lambda T$, so $I_A(\lambda) = I_B(\lambda) = T / \lambda$, and both observers get the same information. As usual, the one who went to the pub did the right thing.

■

10 Hypothesis Testing

The main content of this chapter relies on professor Simon Myers' teaching material.
Url: `http://www.stats.ox.ac.uk/~myers/stats_materials.html`

A question associated with the data migth be formulated in terms of a hypothesis. In particular, we have a so-called null hypothesis which refers to some basic premise which to we will adhere unless evidence from the data causes us to abandon it.

Example In a clinical treatment data may be collected to compare two treatments (old v. new).

The *null hypothesis* is likely to be

no difference between treatments

The *alternative hypothesis* might be:-

a) treatments are different (*2-sided*),
b) new treatment is better (*1-sided*),
c) old treatment is better (*1-sided*).

■

In general we are often in a position to specify the form of the p.m.f., $p(x;\theta)$, say, or the p.d.f., $f(x;\theta)$, but there is doubt about the value of the parameter θ. All that is known is that θ is some element of a specified parameter space Θ. We assume that the null hypothesis of interest specifies that θ is an element of some subset Θ_0 of Θ, and so is true if $\theta \in \Theta_0$ but false if $\theta \notin \Theta_0$.

Example

A coin is tossed and we hypothesise that it is fair. Hence Θ_0 is the set $\left\{\frac{1}{2}\right\}$ containing just one element of the parameter space $\Theta = [0, 1]$.

■

As a convention we shall denote the complement of Θ_0 in Θ by Θ_1. We call the original hypothesis that $\theta \in \Theta_0$ the *null hypothesis* and denote it by H_0. The hypothesis that $\theta \in \Theta_1$ is referred to as the *alternative hypothesis* and denoted by H_1.

10.1 Data and questions

□

Data set: *Patients with glaucoma in one eye.*

The following data (Ehlers, N., *Acta Opthalmologica*, **48**) give corneal thicknesses in microns for patients with one glaucomatous eye and one normal eye.

Table 4.1 Glaucoma in one eye

Corneal thickness		
Glaucoma	Normal	Difference
488	484	4
478	478	0
480	492	−12
426	444	−18
440	436	4
410	398	12
458	464	−6
460	476	−16

Is there a difference in corneal thickness between the eyes? To answer this we take the differences $Glaucoma - Normal$ for each patient and test for the mean of those differences being zero.

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta \neq 0.$$

10.2 Basic ideas

The first, and main idea, is that we need to use statistics which contain all of the relevant information about the parameter (or parameters) we are going to test: in other words we will be looking towards using *sufficient statistics*. Therefore it is hardly surprising that we usually use the same statistics as we would in calculating confidence intervals. Let us try to work out how we might do this by applying common-sense to an example.■

Example *Patients with glaucoma in one eye*

Here is Table 4.1 again, and we ask "Is there a difference in corneal thickness between the eyes?"

Table 4.1 Glaucoma in one eye

Corneal thickness		
Glaucoma	Normal	Difference
488	484	4
478	478	0
480	492	−12
426	444	−18
440	436	4
410	398	12
458	464	−6
460	476	−16

Formally we are testing the difference θ between the corneal thicknesses.

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta \neq 0.$$

Assuming the data to be normally distributed (for the sake of this example), the mean difference is $\overline{x} = -4$ and the estimated standard deviation is $s = 10.744$. Under H_0 we obtain a t-statistic of

$$t = \sqrt{n}\frac{\overline{x}}{s} = \frac{-4\sqrt{8}}{10.744} = -1.053.$$

The t-statistic has 7 degrees of freedom for a p-value of 0.327.
We cannot reject the null hypothesis of no difference in corneal thickness.

Graph of $t(7)$ p.d.f.

■

The p-value is different under different alternative hypotheses.

In Example above the natural alternative hypothesis is $H_1 : \theta \neq 0$. However it is possible that we may believe that glaucoma can only reduce corneal thickness, and that no other outcome is possible. In such a case the alternative hypothesis would be $H_1 : \theta < 0$. Does this affect the p-value?

In such a case the tail of interest in the t-distribution would be the *lower tail*. For the one-sided alternative the upper tail no longer provides evidence against the null-hypothesis, so the p-value becomes

$$p = P(T < -1.053) = 0.1635.$$

In this example even in the case of the alternative hypothesis $H_1 : \theta < 0$ there is no strong evidence to suggest that the null hypothesis is false. Whatever the alternative, we have no grounds to reject the hypothesis of no difference in the corneal thickness.

Definition *Hypothesis test*

A *hypothesis test* is conducted using a test statistic whose distribution is known under the null hypothesis H_0 , and is used to consider the likely truth of the null hypothesis as opposed to a stated alternative hypothesis H_1.

□

Definition *p-value*

The *p-value* (or *significance level* or *size*) is the probability of the test statistic taking a value, in the light of the alternative hypothesis, at least as extreme as its observed value. It is calculated under the assumption that the test statistic has the distribution which it would have if the null hypothesis were true.

If the alternative hypothesis is two-sided it will usually be the case that extreme values occur in two disjoint regions, referring to two tails of the distribution under the null hypothesis.

□

10.3 The magic 5% significance level (or p-value of 0.05)

The question arises in each example considered so far: *what is the critical level for the p-value? Is there some generally accepted level at which null hypotheses are automatically rejected?* Alas, the literature is filled with what purports to be the definitive answer to this question, which is so misleading and ridiculous

that it needs special mention.

A significance level of $p < 0.05$ is often taken to be of interest, because it is below the "magic" level of 0.05**.** For example suppose that we had tested a new drug (new drug versus standard drug), which under the null hypothesis of no difference between the two drugs, gave $p = 0.04$. This says that the apparent difference between the two drugs being due to chance is less than 1 in 20. The p-value of 0.05 is the watershed used by the American control board (the FDA, which stands for Food and Drugs Administration) which licences new drugs from pharmaceutical companies. As a result it has been almost universally accepted right across the board in all walks of life.

However this level can be, to say the least, inappropriate and possibly even catastrophic. Suppose, for example, we were considering test data for safety critical software for a nuclear power station, N representing the number of faults detected in the first 10 years. Would we be happy with a p-value on trials which suggests that

$$P(N \geq 1) = 0.05?$$

We might be more comfortable if $p = 0.0001$, but even then, given the number of power stations (over 1000 in Europe alone) we would be justified in worrying. The significance level which should be used in deciding whether or not to reject a null hypothesis ought to depend entirely on the question being asked; it quite properly should depend upon the consequences of being wrong. At the very least we should qualify our rejection with something like the following.

$0.05 < p \leq 0.06$	"Weak evidence for rejection"
$0.03 < p \leq 0.05$	"Reasonable evidence for rejection"
$0.01 < p \leq 0.03$	"Good evidence for rejection"
$0.005 < p \leq 0.01$	"Strong evidence for rejection"
$0.001 < p \leq 0.005$	"Very strong evidence for rejection"
$0.0005 < p \leq 0.001$	"Extremely strong evidence for rejection"
$p \leq 0.0005$	"Overwhelming evidence for rejection"

10.4 The critical region

Suppose we have data $\mathbf{x} = x_1, x_2, \ldots, x_n,\ x \in \mathbb{R}_x$, which constitute evidence about the truth or falsehood of a null hypothesis H_0. Suppose further that we have decided to formulate our test as a decision rule by electing a p-value in

advance, say α, and rejecting the null hypothesis in situations where the data lead to a p-value less than or equal to α. In such circumstances we can decide, in advance, on a region $C_1 \subset \mathbb{R}_x$ such that H_0 is rejected if $\mathbf{x} \in C_1$. Should $\mathbf{x} \in C_0$, the complement of C_1, H_0 is not rejected.

C_1 is called the *critical region* of the test and α is called the *significance level*.

Note that α is the probability of rejection of H_0 given that it is true. In other words

$$P(\mathbf{x} \in C_1 \mid H_0) = \alpha.$$

Example(revisited): *Patients with glaucoma in one eye*

The significance level is set at 0.05 and H_0 is rejected if $t \leq -2.365$ or $t \geq 2.365$. Here $t = -1.053$ and the null hypothesis is not rejected.

Figure: The critical region

■

10.5 Errors in hypothesis testing

There are two types of possible error.
A *Type I error* is the error of rejecting the null hypothesis when it is, in fact, true.
A *Type II error* is the error of not rejecting the null hypothesis when it is, in fact, false.

	H_0 not rejected	H_0 rejected
H_0 true	no error	Type I error
H_0 false	Type II error	no error

Thus

$$\begin{aligned} P(\text{Type I error}) &= P(x \in C_1 \mid H_0) = \alpha \\ P(\text{Type II error}) &= P(x \in C_0 \mid H_1) = \beta. \end{aligned}$$

The probability that H_0 is correctly rejected, $P(x \in C_1 \mid H_1) = 1 - \beta$ is called the *power* of the test.

Example: *Do air bags save lives?*

Suppose that deaths in crashes involving a particular make of car have been at an average rate of 6 per week and that the company has introduced air bags. They want to use the figures over the next year (*i.e.* 52 weeks) to test their effectiveness. Assume the data are from a Poisson distribution with mean μ. The company plans to test

$$H_0 : \mu = 6 \quad \text{against} \quad H_1 : \mu < 6,$$

using a significance level of 0.05.

$$Y = \sum_{i=1}^{52} X_i \sim Poisson(52\mu)$$

and we use a critical region of the form

$$C_1 = \{y : y \leq k\}.$$

Now the distribution of Y may be approximated by $N(52\mu, 52\mu)$ or, under H_0, $N(312, 312)$.

$$\begin{aligned} 0.05 &= P(Y \leq k) \\ &\simeq P\left(Z \leq \frac{k - 312}{\sqrt{312}}\right) \end{aligned}$$

where $Z \sim N(0, 1)$, so that

$$\frac{k - 312}{\sqrt{312}} \simeq -1.645$$

giving $k = 283$ to the nearest integer.
The power of the test is $P(Y \leq 283)$ where $Y \sim Poisson(52\mu)$. Thus

$$\text{Power} \simeq P\left(Z \leq \frac{283 - 52\mu}{\sqrt{52\mu}}\right) = \Phi\left(\frac{283 - 52\mu}{\sqrt{52\mu}}\right).$$

Note that at $\mu = 6$ the power has value 0.05 and the power increases as μ decreases, approaching 1 as μ approaches 0.

Figure: The power function

■

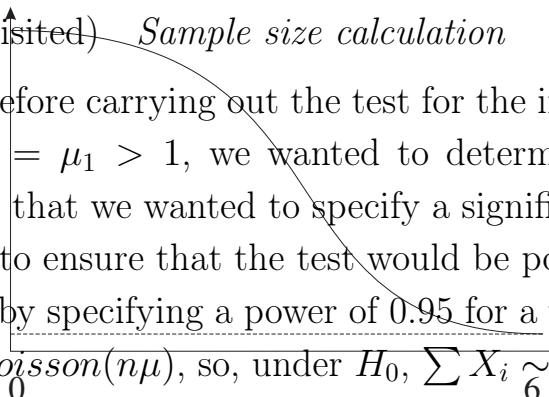
10.6 Summary of hypothesis testing

There are four main ingredients to a test.

- The critical region C_1.
- The sample size n.
- The significance level (or size) $\alpha = P(x \in C_1 | H_0)$
- The power $Q = P(x \in C_1 | H_1)$.

If any two of these are known, the other two may be determined.

Example: (revisited) *Sample size calculation*



Suppose that, before carrying out the test for the insect traps $H_0 : \mu = \mu_0 = 1$ against $H_1 : \mu = \mu_1 > 1$, we wanted to determine a suitable sample size. Suppose further that we wanted to specify a significance level of $\alpha = 0.01$ and that we wished to ensure that the test would be powerful enough to reject the null hypothesis by specifying a power of 0.95 for a value of $\mu_1 = 1.5$. We know that $\sum X_i \sim Poisson(n\mu)$, so, under H_0, $\sum X_i \sim Poisson(n)$, and under H_1 with $\mu_1 = 1.5$ we know that $\sum X_i \sim Poisson(1.25n)$.

A normal approximation would give us, under H_0, $\sum X_i \dot{\sim} N(n, n)$ so that

$$\frac{\sum X_i - n}{\sqrt{n}} \dot{\sim} N(0,1) \quad \Rightarrow \quad P\left(\frac{\sum X_i - n}{\sqrt{n}} \geq 2.326\right) \simeq 0.01$$

and the critical region is

$$\sum x_i \geq n + 2.326\sqrt{n}.$$

For a power of 0.95, we require

$$P\left(\sum X_i \geq n + 2.326\sqrt{n} \mid \mu_1 = 1.5\right) = 0.95,$$

which may be re-written

$$P\left(\frac{\sum X_i - 1.5n}{\sqrt{1.5n}} \geq \frac{-0.5n + 2.326\sqrt{n}}{\sqrt{1.5n}}\right) = 0.95.$$

For a standard normal distribution, $P(Z \geq -1.645) = 0.95$, so the approximate sample size can be calculated from

$$\frac{-0.5n + 2.326\sqrt{n}}{\sqrt{1.5n}} \simeq -1.645$$

giving

$$\sqrt{n} = 8.681, \quad n = 75.367,$$

so the recommended sample size would be 76.

■

10.7 The Likelihood Ratio Test

10.7.1 The likelihood ratio

We often want to test in situations where the adopted probability model involves several unknown parameters. Thus we may denote an element of the parameter space by

$$\theta = (\theta_1, \theta_2, \ldots \theta_k)$$

Some of these parameters may be *nuisance* parameters, (*e.g.* testing hypotheses on the unknown mean of a normal distribution with unknown variance, where the variance is regarded as a nuisance parameter).

We use the *likelihood ratio*, $\lambda(\mathbf{x})$, defined as

$$\lambda(\mathbf{x}) = \frac{\sup\{L(\theta; \mathbf{x}) : \theta \in \mathbf{\Theta}_0\}}{\sup\{L(\theta; \mathbf{x}) : \theta \in \mathbf{\Theta}\}}, \quad \mathbf{x} \in \mathbb{R}_X^n.$$

The informal argument for this is as follows.

For a realisation x, determine its best chance of occurrence under H_0 and also its best chance overall. The ratio of these two chances can never exceed unity, but, if small, would constitute evidence for rejection of the null hypothesis.

A *likelihood ratio test* for testing $H_0 : \theta \in \mathbf{\Theta}_0$ against $H_1 : \theta \in \mathbf{\Theta}_1$ is a test with critical region of the form

$$C_1 = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\},$$

where k is a real number between 0 and 1.

Clearly the test will be at significance level α if k can be chosen to satisfy

$$\sup\{P(\lambda(\mathbf{X}) \leq k;\ \theta \in \boldsymbol{\Theta}_0)\} = \alpha.$$

If H_0 is a simple hypothesis with $\boldsymbol{\Theta}_0 = \{\theta_0\}$, we have the simpler form

$$P(\lambda(\mathbf{X}) \leq k;\ \theta_0) = \alpha.$$

To determine k, we must look at the c.d.f. of the random variable $\lambda(\mathbf{X})$, where the random sample $\mathbf{X}$ has joint p.d.f. $f_{\mathbf{X}}(\mathbf{x}; \theta_0)$.

Example: *Exponential distribution*

Test $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$.

Here $\boldsymbol{\Theta}_0 = \{\theta_0\}$, $\boldsymbol{\Theta}_1 = [\theta_0, \infty)$.

The likelihood function is

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta) = \theta^n e^{-\theta \sum x_i}.$$

The numerator of the likelihood ratio is

$$L(\theta_0; \mathbf{x}) = \theta_0^n e^{-n\theta_0 \overline{x}}.$$

We need to find the supremum as θ ranges over the interval $[\theta_0, \infty)$. Now

$$l(\theta; \mathbf{x}) = n \log \theta - n\theta\overline{x}$$

so that

$$\frac{\partial l(\theta; \mathbf{x})}{\partial \theta} = \frac{n}{\theta} - n\overline{x}$$

which is zero only when $\theta = 1/\overline{x}$. Since $L(\theta; \mathbf{x})$ is an increasing function for $\theta < 1/\overline{x}$ and decreasing for $\theta > 1/\overline{x}$,

$$\sup\{L(\theta; \mathbf{x}) : \theta \in \Theta\} = \begin{cases} \overline{x}^{-n} e^{-n}, & \text{if } 1/\overline{x} \geq \theta_0 \\ \theta_0^n e^{-n\theta_0 \overline{x}} & \text{if } 1/\overline{x} < \theta_0 \end{cases}.$$

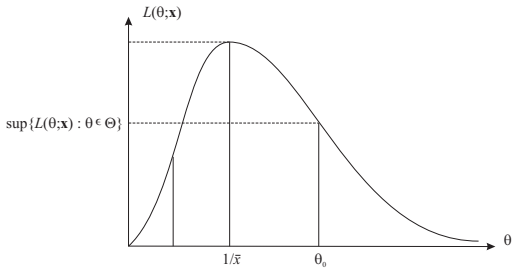
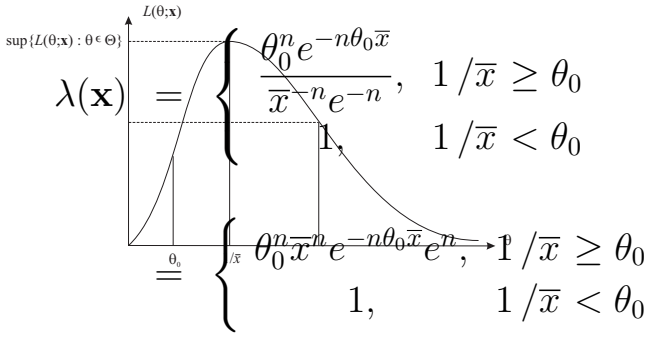


Figure 4.10 Likelihood function



$$\lambda(\mathbf{x}) = \begin{cases} \dfrac{\theta_0^n e^{-n\theta_0\overline{x}}}{\overline{x}^n e^{-n}}, & 1/\overline{x} \geq \theta_0 \\ 1, & 1/\overline{x} < \theta_0 \end{cases}$$

$$= \begin{cases} \theta_0^n \overline{x}^n e^{-n\theta_0\overline{x}} e^n, & 1/\overline{x} \geq \theta_0 \\ 1, & 1/\overline{x} < \theta_0 \end{cases}$$

Since

$$\frac{d}{d\overline{x}}\left(\overline{x}^n e^{-n\theta_0\overline{x}}\right) = n\overline{x}^{n-1}e^{-n\theta_0\overline{x}}\left(1-\theta_0\overline{x}\right)$$

is positive for values of $\overline{x}$ between 0 and $1/\theta_0$ where $\theta_0 > 0$, it follows that $\lambda(\mathbf{x})$ is a non-decreasing function of $\overline{x}$. Therefore the critical region of the likelihood ratio test is of the form

$$C_1 = \left\{\mathbf{x} : \sum_{i=1}^{n} x_i \leq c\right\}.$$

■

Example : *The one-sample t-test*

The null hypothesis is $H_0 : \theta = \theta_0$ for the mean of a normal distribution with unknown variance σ^2.

We have

$$\begin{aligned} \boldsymbol{\Theta} &= \left\{(\theta, \sigma^2) : \theta \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\right\} \\ \boldsymbol{\Theta}_0 &= \left\{(\theta, \sigma^2) : \theta = \theta_0,\ \sigma^2 \in \mathbb{R}^+\right\} \end{aligned}$$

and

$$f(x;\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right), \quad x \in \mathbb{R}.$$

The likelihood function is

$$L(\theta,\sigma^2;\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)$$

Since

$$l(\theta_0,\sigma^2;\mathbf{x}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta_0)^2$$

and

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\theta_0)^2,$$

which is zero when

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\theta_0)^2$$

we conclude that

$$\sup\left\{L(\theta_0,\sigma^2;\mathbf{x})\right\} = \left(\frac{2\pi}{n}\sum_{i=1}^{n}(x_i-\theta_0)^2\right)^{-n/2} e^{-n/2}.$$

For the denominator, we already know from previous examples that the m.l.e. of θ is $\overline{x}$, so

$$\sup\left\{L(\theta,\sigma^2;\mathbf{x})\right\} = \left(\frac{2\pi}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{-n/2} e^{-n/2}$$

and

$$\lambda(\mathbf{x}) = \left(\frac{\sum_{i=1}^{n}(x_i-\theta_0)^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)^{-n/2}.$$

This may be written in a more convenient form. Note that

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\theta_0)^2 &= \sum_{i=1}^{n}\left((x_i-\overline{x})+(\overline{x}-\theta_0)\right)^2 \\ &= \sum_{i=1}^{n}(x_i-\overline{x})^2 + n(\overline{x}-\theta_0)^2\end{aligned}$$

so that

$$\lambda(\mathbf{x}) = \left(1+\frac{n(\overline{x}-\theta_0)^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)^{-n/2}.$$

The critical region is

$$C_1 = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\}$$

so it follows that H_0 is to be rejected when the value of

$$\frac{|\overline{x} - \theta_0|}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

exceeds some constant.

Now we have already seen that

$$\frac{\overline{X} - \theta}{S/\sqrt{n}} \sim t(n-1)$$

where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Therefore it makes sense to write the critical region in the form

$$C_1 = \left\{\mathbf{x} : \frac{|\overline{x} - \theta_0|}{s/\sqrt{n}} \geq c\right\}$$

which is the standard form of the two-sided t-test for a single sample.

■

10.7.2 The likelihood ratio statistic

Since the function $-2\log\lambda(\mathbf{x})$ is a decreasing function, it follows that the critical region of the likelihood ratio test can also be expressed in the form

$$C_1 = \{\mathbf{x} : -2\log\lambda(x) \geq c\}.$$

Writing

$$\Lambda(\mathbf{x}) = -2\log\lambda(\mathbf{x}) = 2\left[l(\widehat{\theta} : \mathbf{x}) - l(\theta_0 : \mathbf{x})\right]$$

the critical region may be written as

$$C_1 = \{\mathbf{x} : \Lambda(\mathbf{x}) \geq c\}$$

and $\Lambda(\mathbf{X})$ is called the *likelihood ratio statistic*.

We have been using the idea that values of θ close to $\widehat{\theta}$ are well supported by the data so, if θ_0 is a possible value of θ, then it turns out that, for large samples,

$$\Lambda(\mathbf{X}) \xrightarrow{D} \chi^2_p$$

where $p = \dim(\theta)$.

Let us see why.

10.7.3 The asymptotic distribution of the likelihood ratio statistic

Write

$$l(\theta_0) = l(\widehat{\theta}) + (\widehat{\theta} - \theta_0)l'(\widehat{\theta}) + \frac{1}{2}(\widehat{\theta} - \theta_0)^2 l''(\widehat{\theta}) + \dots$$

and, remembering that $l'(\widehat{\theta}) = 0$, we have

$$\begin{aligned} \Lambda &\simeq (\widehat{\theta} - \theta_0)^2 \left[-l''(\widehat{\theta})\right] \\ &= (\widehat{\theta} - \theta_0)^2 J(\widehat{\theta}) \\ &= (\widehat{\theta} - \theta_0)^2 I(\theta_0)\frac{J(\widehat{\theta})}{I(\theta_0)}. \end{aligned}$$

But

$$(\widehat{\theta} - \theta_0)I(\theta_0)^{1/2} \xrightarrow{D} N(0,1) \quad \text{and} \quad \frac{J(\widehat{\theta})}{I(\theta_0)} \xrightarrow{P} 1$$

so

$$(\widehat{\theta} - \theta_0)^2 I(\theta_0) \xrightarrow{D} \chi^2_1$$

or

$$\Lambda \xrightarrow{D} \chi^2_1$$

provided θ_0 is the true value of θ.

Example: *Poisson distribution*

Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random sample from a Poisson distribution with parameter θ, and test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at significance level 0.05.

The p.m.f. is

$$p(x;\theta) = \frac{e^{-\theta}\theta^x}{x!}, \quad x = 0, 1, \dots$$

so that

$$l(\theta : \mathbf{x}) = -n\theta + \sum_{i=1}^{n} x_i \log\theta - \log\prod_{i=1}^{n} x_i!$$

and

$$\frac{\partial l(\theta : \mathbf{x})}{\partial\theta} = -n + \frac{1}{\theta}\sum_{i=1}^{n} x_i$$

giving $\widehat{\theta} = \overline{x}$.

Therefore

$$\Lambda = 2n\left[\theta_0 - \overline{x} + \overline{x}\log\left(\frac{\overline{x}}{\theta_0}\right)\right].$$

The distribution of Λ under H_0 is approximately χ^2_1 and $\chi^2_1(0.95) = 3.84$, so the critical region of the test is

$$C_1 = \left\{\mathbf{x} : 2n\left[\theta_0 - \overline{x} + \overline{x}\log\left(\frac{\overline{x}}{\theta_0}\right)\right] \geq 3.84\right\}.$$

■

10.7.4 Testing goodness-of-fit for discrete distributions

Example: *Pielou's data on Armillaria root rot in Douglas fir trees*

They were collected by the ecologist E.C. Pielou, who was interested in the pattern of healthy and diseased trees. The subject of her research was *Armillaria* root rot in a plantation of Douglas firs. She recorded the lengths of 109 runs of diseased trees.

Table 4.4 Run lengths of diseased trees

Run length	1	2	3	4	5	6
Number of runs	71	28	5	2	2	1

On biological grounds, Pielou proposed a geometric distribution as a probability model. Is this plausible?

□

Let's try to answer this by first looking at the general case.

Suppose we have k groups with n_i in the i^{th} group. Thus

Group	1	2	3	4	$\cdots$	k
Number	n_1	n_2	n_3	n_4	$\cdots$	n_k

where $\sum_i n_i = n$.

Suppose further that we have a probability model such that $\pi_i(\theta),\ i = 1, 2, \ldots, k$, is the probability of being in the i^{th} group. Clearly $\sum_i \pi_i(\theta) = 1$.

The likelihood is

$$L(\theta) = n!\prod_{i=1}^{k}\frac{\pi_i(\theta)^{n_i}}{n_i!}$$

and the log-likelihood is

$$l(\theta) = \sum_{i=1}^{k} n_i \log \pi_i(\theta) + \log n! - \sum_{i=1}^{k}\log n_i!$$

Suppose $\widehat{\theta}$ maximises $l(\theta)$, being the solution of $l'(\widehat{\theta}) = 0$.

The general alternative is to take π_i as unrestricted by the model and subject only to $\sum_i \pi_i = 1$. Thus we maximise

$$l(\pi) = \sum_{i=1}^{k} n_i \log \pi_i + \log n! - \sum_{i=1}^{k} \log n_i! \quad \text{with} \quad g(\pi) = \sum_i \pi_i = 1.$$

Using Lagrange multiplier γ we obtain the set of k equations

$$\frac{\partial l}{\partial \pi_i} - \gamma \frac{\partial g}{\partial \pi_i} = 0, \quad 1 \leq i \leq k,$$

or

$$\frac{n_i}{\pi_i} - \gamma = 0, \quad 1 \leq i \leq k.$$

Writing this as

$$n_i - \gamma \pi_i = 0, \quad 1 \leq i \leq k$$

and summing over i we find $\gamma = n$ and

$$\widehat{\pi}_i = \frac{n_i}{n}.$$

The likelihood ratio statistic is

$$\begin{aligned} \Lambda &= 2\left[\sum_{i=1}^{k} n_i \log \frac{n_i}{n} - \textstyle\sum_{i=1}^{k} n_i \log \pi_i(\widehat{\theta})\right] \\ &= 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right). \end{aligned}$$

General statement of asymptotic result for the likelihood ratio statistic

Testing $H_0 : \theta \in \Theta_0 \subset \Theta$ against $H_1 : \theta \in \Theta$, the likelihood ratio statistic

$$\Lambda = 2\left[\sup_{\theta \in \Theta} l(\theta) - \sup_{\theta \in \Theta_0} l(\theta)\right] \xrightarrow{D} \chi^2_p,$$

where

$$p = \dim \Theta - \dim \Theta_0$$

In the case above where we are looking at the fit of a one-parameter distribution

$$\Lambda = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right),$$

the restriction $\sum_{i=1}^{k}\pi_i = 1$ means that $\dim\Theta = k-1$. Clearly $\dim\Theta_0 = 1$ so $p = k-2$ and

$$\Lambda \xrightarrow{D} \chi^2_{k-2}.$$

Example: (revisited) *Pielou's data on Armillaria root rot in Douglas fir trees*

The data are

Run length	1	2	3	4	5	6
Number of runs	71	28	5	2	2	1

and Pielou proposed a geometric model with p.m.f.

$$p(x) = (1-\theta)^{x-1}\theta, \quad x = 1, 2, \ldots$$

where x is the observed run length. Thus, if $x_j,\ 1 \le j \le n$, are the observed run lengths, the log-likelihood for Pielou's model is

$$l(\theta) = \sum_{j=1}^{n}(x_j - 1)\log(1-\theta) + n\log\theta$$

and, maximising,

$$\frac{\partial l(\theta)}{\partial\theta} = -\frac{\sum_{j=1}^{n} x_j - n}{(1-\theta)} + \frac{n}{\theta}$$

which gives

$$\widehat{\theta} = \frac{1}{\overline{x}}.$$

By the invariance property of m.l.e.'s

$$\pi_i(\widehat{\theta}) = (1-\widehat{\theta})^{i-1}\widehat{\theta} = \frac{(\overline{x}-1)^{i-1}}{\overline{x}^i}.$$

The data give $\overline{x} = 1.523$. We can therefore use the expression for $\pi_i(\widehat{\theta})$ to calculate

$$\Lambda = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right) = 3.547.$$

There are six groups, so $p = 6 - 1 - 1 = 4$.

The approximate distribution of Λ is therefore χ^2_4 and

$$P(\Lambda \geq 3.547) = 0.471.$$

There is no evidence against Pielou's conjecture that a geometric distribution is an appropriate model.

■

Example: *Flying bomb hits on London*

Data set 4.5 gave the number of flying bomb hits recorded in each of 576 small areas of $\frac{1}{4}km^2$ in the south of London during World War II.

Table 4.5 Flying bomb hits on London

Number of hits in an area	0	1	2	3	4	5	≥ 6
Frequency	229	211	93	35	7	1	0

Propaganda broadcasts claimed that the weapon could be aimed accurately. If, however, this was not the case, the hits should be randomly distributed over the area and should therefore be fitted by a Poisson distribution. Is this the case?

□

The first thing to do is to calculate the m.l.e. of the Poisson parameter. The likelihood function for a sample of size n is

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!},$$

so that the log-likelihood is

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

$$\frac{\mathrm{d}l}{\mathrm{d}\theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - n = 0$$

and

$$\widehat{\theta} = \overline{x} = \frac{535}{576} = 0.928.$$

Using the Poisson probability mass function with $\theta = 0.929$ we therefore obtain

i	0	1	2	3	4	5	≥ 6
$\pi_i\left(\widehat{\theta}\right)$	0.3949	0.3669	0.1704	0.0528	0.0123	0.0023	0.0004

and hence

$$\Lambda = 2\sum_{i=1}^{k} n_i \log\left(\frac{n_i}{n\pi_i(\widehat{\theta})}\right) = 1.4995.$$

This is tested against $\chi^2(\nu)$ where $\nu = k - 2 = 7 - 2 = 5$. This gives $P(\Lambda \geq 1.4995) = 0.913$. Clearly there is not a shred of evidence in favour of rejection.

■

10.7.5 The approximate χ^2 distribution

The tests carried out in the previous 2 Examples are not, strictly speaking, correct. The reason for this is that the χ^2-distribution we have used to calculate the p-values is an approximation, and the quality of that approximation depends upon the sample size. Happily there is a general rule of thumb you can use.

Rule of thumb for the χ^2 approximation

An approximate χ^2-distribution may be used for testing count data provided that the expected value of each cell in the table is at least 5. If the expected value of a cell is less than 5, it should be pooled with an adjacent cell or cells to obtain a suitable value.

□

Example:(revisited) *Pielou's data on Armillaria root rot in Douglas fir trees*

Look at the table with the expected values written in.

Run length	1	2	3	4	5	6	≥7
Number of runs n_i	71	28	5	2	2	1	0
Expected number of runs $n\pi_i(\widehat{\theta})$	71.569	24.577	8.440	2.898	0.995	0.342	0.218

Clearly we need to pool cells to obtain

Run length	1	2	≥ 3
Number of runs n_i	71	28	10
Expected number of runs $n\pi_i(\widehat{\theta})$	71.569	24.577	12.893

The test statistic is now re-calculated to obtain $\Lambda = 1.087$, which is tested as $\chi^2(1)$ to give a p-value of 0.297. The conclusion that there is no evidence against Pielou's conjecture that the underlying distribution is geometric is unaltered.

■

Example: (revisited) *Flying bomb hits on London*

Including the expected values in the table for flying bomb hits, we obtain the table below.

Number of hits in an area	0	1	2	3	4	5	≥ 6
Frequency	229	211	93	35	7	1	0
Expected frequency	227.462	211.334	98.150	30.413	7.027	1.325	0.230

After pooling, we obtain

Number of hits in an area	0	1	2	3	≥ 4
Frequency	229	211	93	35	8
Expected frequency	227.462	211.334	98.150	30.413	8.582

The test statistic is now re-calculated to obtain $\Lambda = 1.101$, which is tested as $\chi^2(3)$ to give a p-value of 0.777. Again we find no evidence for rejection of the null hypothesis. We have therefore found no evidence that V1 flying bomb could be aimed with any degree of precision.

■

11 Background Notes for the Bivariate normal distribution

In this section we will revise few background knowledge you need to study the Bivariate normal distribution.

Revision of matrix properties – much NOT necessary to learn

Matrix

Let $\boldsymbol{A} = (a_{ij})$, a matrix with element a_{ij} in ith row and jth column.

Example $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ of order 2 (rows) $\times$ 2 (columns).

Transpose

If $\boldsymbol{A}$ is of order $m \times n$ and has elements a_{ij}, then $\boldsymbol{A}^\top$ is of order $n \times m$ and has elements a_{ji}.

Properties: $(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$, $(\boldsymbol{A}+\boldsymbol{B})^\top = \boldsymbol{A}^\top + \boldsymbol{B}^\top$, $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$.

Example If $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, then $\boldsymbol{A}^\top = \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix}$.

Symmetric matrix

$\boldsymbol{S} = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}$ is symmetric if $s_{12} = s_{21}$. More generally $s_{ij} = s_{ji},\ \forall i \neq j$.

If $\boldsymbol{S}$ is symmetric, then $\boldsymbol{S}^\top = \boldsymbol{S}$.

Matrix addition (not needed)

The matrices must have the same order. $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$ has elements $c_{ij} = a_{ij} + b_{ij}\ \ \forall i, j$.

Example

For $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ and $\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$, $\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} \\ a_{21}+b_{21} & a_{22}+b_{22} \end{pmatrix}$.

Matrix multiplication

Note that number of columns in first matrix must equal number of rows in second, and $\boldsymbol{AB} \neq \boldsymbol{BA}$ necessarily. If $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$, then $\boldsymbol{C} = \boldsymbol{AB} = (c_{ik})$, where $c_{ik} = \sum_j a_{ij} b_{jk}$.

Example

For $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ and $\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$,

$$\boldsymbol{AB} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}.$$

Vectors

Let $\boldsymbol{a} = (a_i)$, a column vector with element a_i in ith row, and $\boldsymbol{a}^\top$ be the corresponding row vector.

Example If $\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$, then $\boldsymbol{a}^\top = \begin{pmatrix} a_1 & a_2 \end{pmatrix}$.

Row vector multiplied by a matrix

Let $\boldsymbol{a}^\top = (a_1, a_2, \ldots, a_n)$ be a row vector and $\boldsymbol{B} = (b_{ij})$ be a $(n \times n)$ matrix. Then

$$\boldsymbol{c}^\top = \boldsymbol{a}^\top \boldsymbol{B} = (c_1, c_2, \ldots, c_n), \quad \boldsymbol{d} = \boldsymbol{Ba} = \begin{pmatrix} d_1 \\ \vdots \\ d_n \end{pmatrix}, \quad \text{where} \quad c_j = \sum_i a_i b_{ij}$$

and $d_i = \sum_j b_{ij} a_j$.

Example Let $\boldsymbol{a}^\top = (a_1, a_2)$ and $\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$. Then,

$$\boldsymbol{a}^\top \boldsymbol{B} = \begin{pmatrix} a_1 & a_2 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = (a_1 b_{11} + a_2 b_{21} \quad a_1 b_{12} + a_2 b_{22}),$$

and

$$\boldsymbol{Ba} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} b_{11}a_1 + b_{12}a_2 \\ b_{21}a_1 + b_{22}a_2 \end{pmatrix}.$$

Example

If $\mathbf{1} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$, then $\boldsymbol{B1} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} b_{11} + b_{12} \\ b_{21} + b_{22} \end{pmatrix}$.

Example If $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, then $\boldsymbol{y}^\top \boldsymbol{y} = \begin{pmatrix} y_1 & \cdots & y_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \sum_i y_i^2$,

a scalar quantity.

Example $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{y} = \sum_i \sum_j a_{ij} x_i y_j$, a scalar, called a quadratic form.

Identity matrix

For example $\boldsymbol{I} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Properties: for any matrix $\boldsymbol{A}$, $\boldsymbol{A}\boldsymbol{I} = \boldsymbol{I}\boldsymbol{A} = \boldsymbol{A}$.

Diagonal matrix

Off-diagonal elements equal zero. For example $\boldsymbol{D} = \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix}$.

Determinant of a matrix

If $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ then $|\boldsymbol{A}| = a_{11}a_{22} - a_{12}a_{21}$. The matrix $\boldsymbol{A}$ is said to be singular if $|\boldsymbol{A}| = 0$.

Matrix inverse

If $\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$, then $\boldsymbol{A}^{-1} = \frac{1}{|\boldsymbol{A}|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$.
Properties: $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.

Differentiation (not needed)
Let $\boldsymbol{x}$, $\boldsymbol{y}$ be vectors, $\boldsymbol{c}^\top$ = row vector of constants, and $\boldsymbol{A}$ = matrix of constants. Then

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\boldsymbol{c}^\top\boldsymbol{x} = \boldsymbol{c}, \qquad \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}, \qquad \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}, \qquad \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\boldsymbol{x}^\top\boldsymbol{x} = 2\boldsymbol{x}.$$

Example Let $\boldsymbol{c}^\top = (c_1, c_2)$ and $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ so $\boldsymbol{c}^\top\boldsymbol{x} = \sum_i c_i x_i = c_1x_1 + c_2x_2$.
Then

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\boldsymbol{c}^\top\boldsymbol{x} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}x_1}\boldsymbol{c}^\top\boldsymbol{x} \\ \frac{\mathrm{d}}{\mathrm{d}x_2}\boldsymbol{c}^\top\boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}x_1}(c_1x_1 + c_2x_2) \\ \frac{\mathrm{d}}{\mathrm{d}x_2}(c_1x_1 + c_2x_2) \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \boldsymbol{c}.$$

Example If $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, then $\boldsymbol{x}^\top\boldsymbol{x} = \sum_i x_i^2$ and

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{x}}\boldsymbol{x}^\top\boldsymbol{x} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}x_1}\boldsymbol{x}^\top\boldsymbol{x} \\ \frac{\mathrm{d}}{\mathrm{d}x_2}\boldsymbol{x}^\top\boldsymbol{x} \end{pmatrix} = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}x_1}(x_1^2 + x_2^2) \\ \frac{\mathrm{d}}{\mathrm{d}x_2}(x_1^2 + x_2^2) \end{pmatrix} = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} = 2\boldsymbol{x}.$$

Mean and variance

Let $\boldsymbol{Y} = (Y_i)$ be a vector of variables with $\mathrm{E}[Y_i] = \mu_i$ and $\mathrm{cov}(Y_i, Y_j) = \sigma_{ij}$. Then

$$\mathrm{E}[\boldsymbol{Y}] = \begin{pmatrix} \mathrm{E}[Y_1] \\ \vdots \\ \mathrm{E}[Y_n] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu},$$

where the ith element of $\boldsymbol{\mu}$ is μ_i. Similarly,

$$\mathrm{Var}[\boldsymbol{Y}] = \begin{pmatrix} \mathrm{Var}[Y_1] & \mathrm{cov}(Y_1, Y_2) & \cdots & \mathrm{cov}(Y_1, Y_n) \\ \mathrm{cov}(Y_2, Y_1) & \mathrm{Var}[Y_2] & \cdots & \mathrm{cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(Y_n, Y_1) & \mathrm{cov}(Y_n, Y_2) & \cdots & \mathrm{Var}[Y_n] \end{pmatrix} =$$

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix} = \boldsymbol{\Sigma}$$

where the matrix $\boldsymbol{\Sigma}$ has elements σ_{ij}.

subsection*Mean and variance properties (not needed)

Suppose that $\boldsymbol{C}$ is a matrix of constants and $\boldsymbol{b}$ is a vector of constants. Then

$$\mathrm{E}[\boldsymbol{CY} + \boldsymbol{b}] = \boldsymbol{C\mu} + \boldsymbol{b} \quad \text{and} \quad \mathrm{Var}[\boldsymbol{CY} + \boldsymbol{b}] = \boldsymbol{C\Sigma C}^\top.$$

Quesestion for you: Can you prove these results for the mean and variance?

Hint: If $\boldsymbol{C} = (c_{ij})$, $\boldsymbol{b} = (b_i)$, and $\boldsymbol{U} = \boldsymbol{CY} + \boldsymbol{b}$, then verify that

$$\begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} \equiv \boldsymbol{U} = \begin{pmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{n1} & \cdots & c_{nn} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} + \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \sum_j c_{1j} Y_j + b_1 \\ \vdots \\ \sum_j c_{nj} Y_j + b_n \end{pmatrix}.$$

Can you now determine the mean of $U_i = \sum_j c_{ij} Y_j + b_i$? Is it the same as the ith row of the vector $\boldsymbol{C\mu} + \boldsymbol{b}$?

Show that the (i, j)th element of $\boldsymbol{C\Sigma C}^\top$ is:

$$\mathrm{cov}(U_i, U_j) = \mathrm{cov}\left(\sum_k c_{ik} Y_k, \sum_l c_{jl} Y_l \right) = \sum_k \sum_l c_{ik} \mathrm{cov}(Y_k, Y_l) c_{jl}.$$

Notice that $\boldsymbol{C}^\top = (c_{jl})$.

Bivariate normal distribution: Introduction

If X_1, X_2 are independent $\mathrm{N}(0, \sigma^2 = 4)$ random variables, the joint probability density function of (X_1, X_2) is

$$f_{X_1X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{x_1^2}{2\sigma^2}\right\} \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{x_2^2}{2\sigma^2}\right\}$$
$$= \frac{1}{2\pi\sigma^2} \exp\left\{-\frac{(x_1^2 + x_2^2)}{2\sigma^2}\right\}.$$

This is shown below.

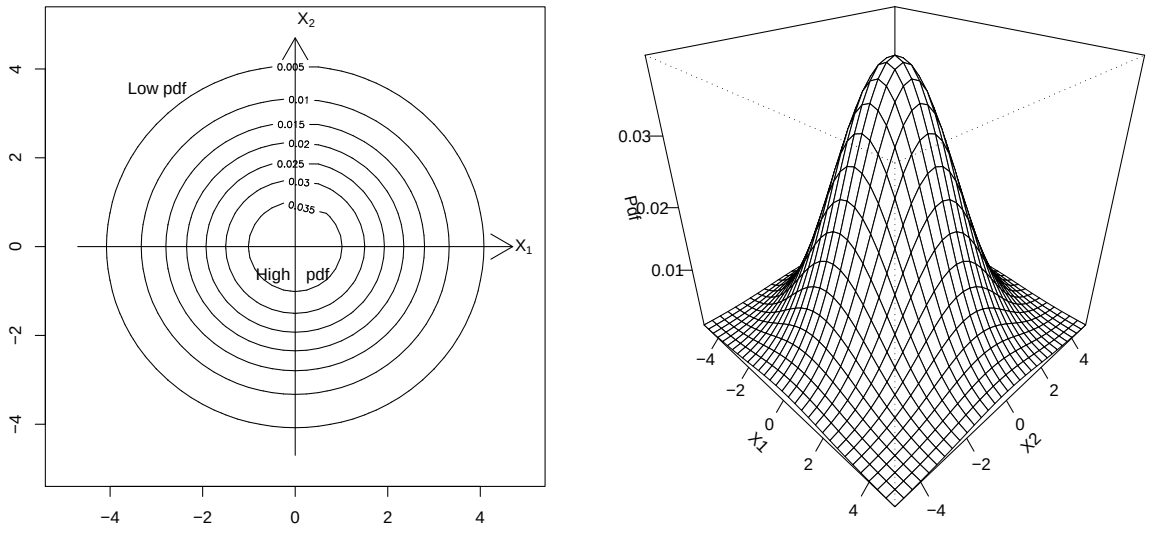


Figure 8: bivariate normal distribution with $X_1, X_2 \overset{\text{ind}}{\sim} \mathrm{N}(0, \sigma^2 = 4)$; (left) contours of joint probability density function; (right) joint probability density function surface.

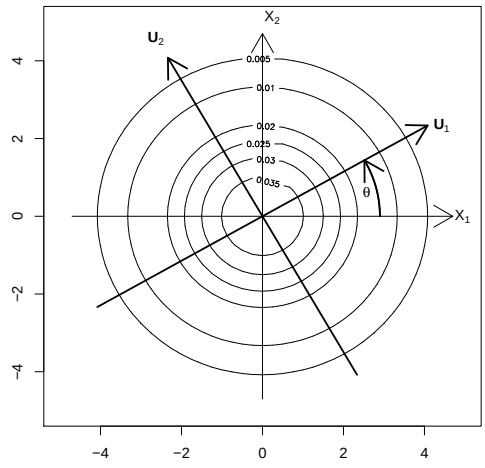


Figure 9: rotation of (X_1, X_2) through $\theta = 30°$ to give new axes (U_1, U_2) and showing the probability density function contours of $f_{X_1X_2}(x_1, x_2)$.

Bivariate normal distribution: more details

Suppose $\boldsymbol{X} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}$, then the probability density function of $\boldsymbol{X}$ is

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) &= \frac{1}{2\pi|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right) \\ &= \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left(-\frac{(x_1^2 + x_2^2 - 2\rho x_1 x_2)}{2\sigma^2(1-\rho^2)}\right), \qquad -\infty < x_1, x_2 < \infty. \end{aligned}$$

More generally $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

and probability density function

$$f_{\boldsymbol{X}}(\boldsymbol{x}) =$$

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{\frac{-1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right\}.$$

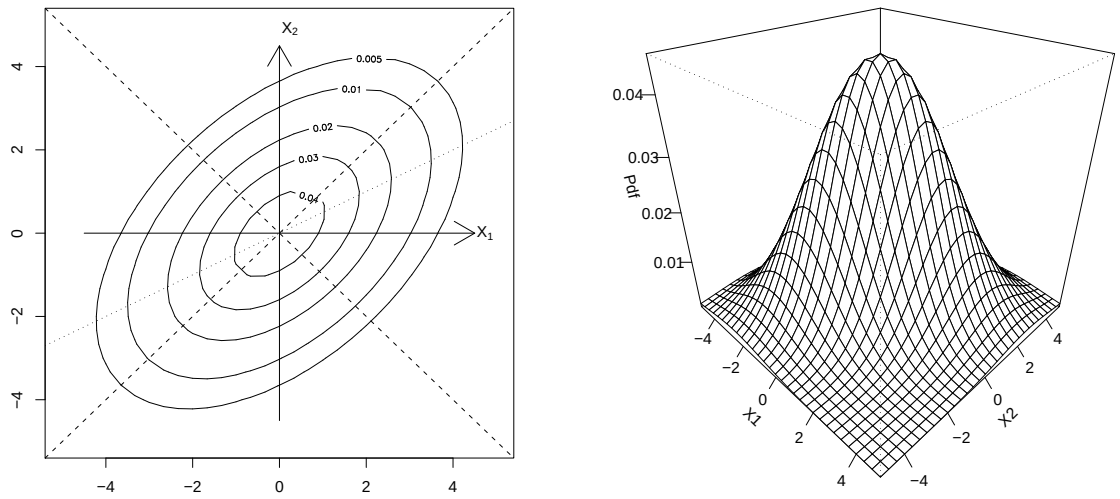


Figure 10: bivariate normal distribution with $\sigma = 2$, $\rho = 0.5$; (left) contours of joint probability density function, (dashed line) ellipse principal axes, (dotted line) $X_2 = \rho X_1$; (right) probability density function surface.

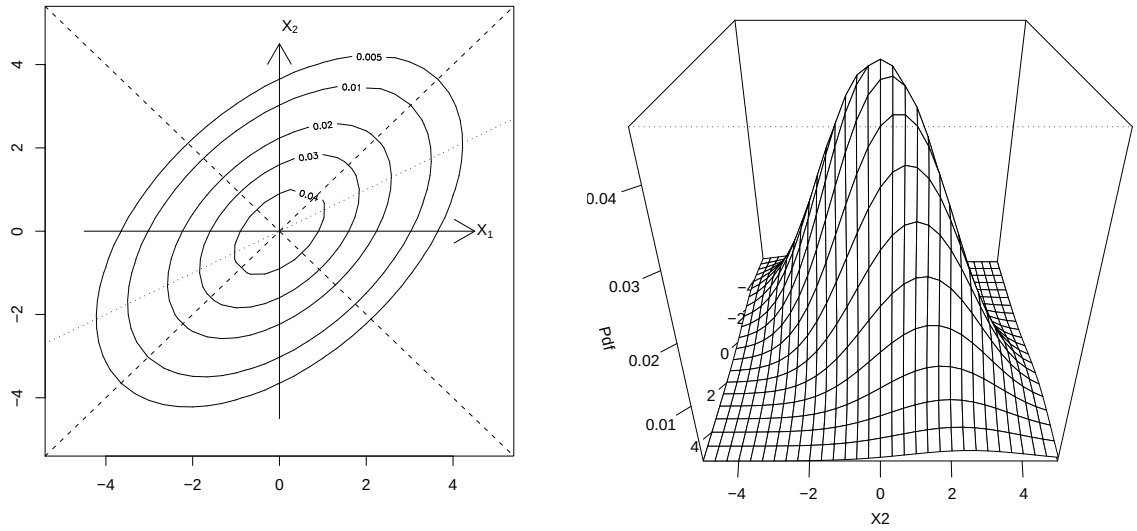


Figure 11: bivariate normal distribution with $\sigma = 2$, $\rho = 0.5$; (left) contours of joint probability density function, (dashed line) ellipse principal axes, (dotted line) regression line $X_2 = \rho X_1$; (right) showing conditional probability density function for X_2 given $X_1 = x_1$ is a normal density.

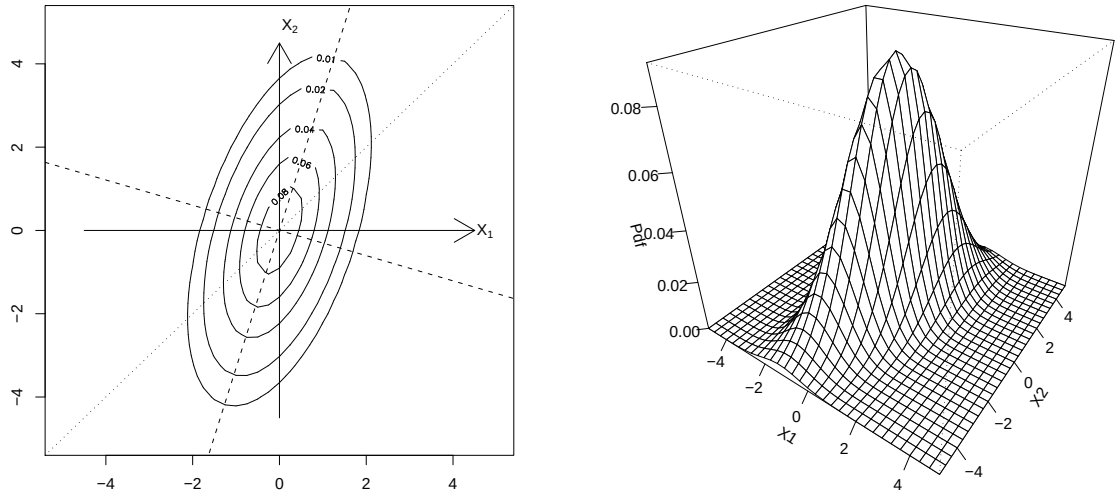


Figure 12: bivariate normal distribution with $\mu_1 = \mu_2 = 0$, $\sigma_1 = 1$, $\sigma_2 = 2$, $\rho = 0.5$; (left) contours of joint probability density function, (dashed line) ellipse principal axes, (dotted line) regression line $X_2 = \rho\sigma_2 X_1/\sigma_1$; (right) probability density function surface.

Example Heights and weights[20]

The data below and shown in figure 13 gives the heights and weights of 58703 National servicemen, born in 1933 and recruited in 1951.

				Weight (lbs) Y						
			Class	80–	110–	140–	170–	200–	230–	Marginal
			Mid-point y_j	95	125	155	185	215	245	total for X
	53–		54.5		5					5
X	57–		58.5	78	97	12				187
	61–	x_i	62.5	1000	6022	481	28	1		7532
Height	65–		66.5	419	21706	9271	396	36	1	31829
(inches)	69–		70.5	12	5382	10951	1189	105	11	17650
	73–		74.5		110	989	339	31	6	1475
	77–		77.5		2	12	9	2		25
	Marginal total for Y			1509	33324	21716	1961	175	18	Total 58703

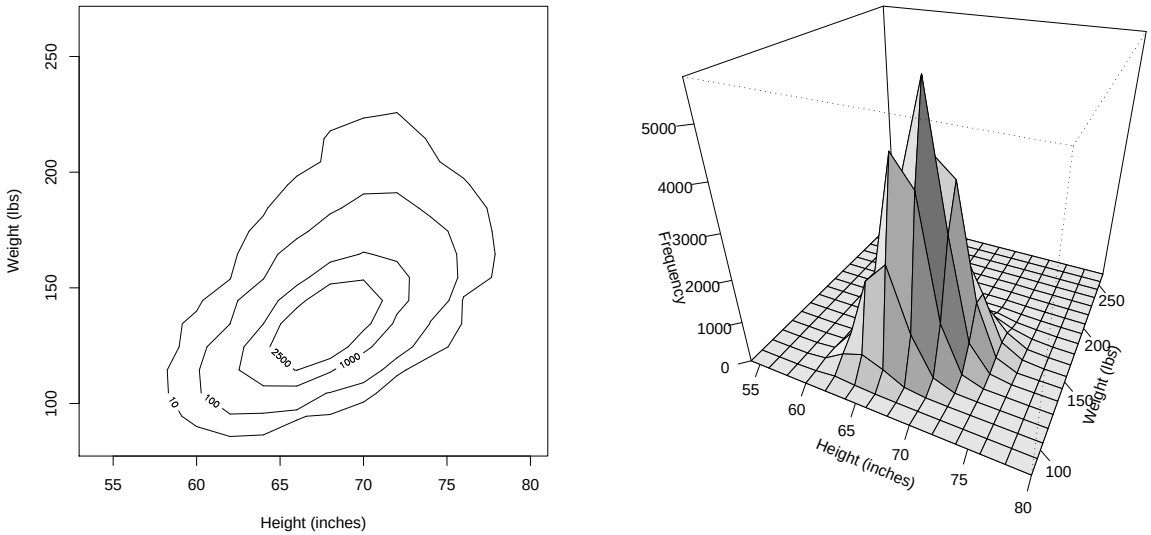


Figure 13: (left) frequency contours of 10, 100, 1000, 2500 men, (right) frequency as a surface.

You can rotate the perspective plot in figure 13 by typing, within R,

```
source("http://www.maths.leeds.ac.uk/~sta6ajb/math2715/army1951.rr")
```

Example CD4 cell counts in AIDS patients[21]

The CD4 cell count of a patient is used as a marker of the progress of HIV infection. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_T)$

[20]Source: Rosenbaum, S. (1954) "Heights and weights of the army intake, 1951", *Journal of Royal Statistical Society, series A*, **54**, 331–347.

[21]Source: Lipsitz, S.R., Ibrahim, J., and Molenberghs, G. (2000) "Using a Box-Cox transformation in the analysis of longitudinal data with incomplete responses", *Applied Statistics*, **49**, 287-296.

denote the CD4 cell count at times 1, 2, ..., T. For some transformation $\boldsymbol{U} = \boldsymbol{u}(\boldsymbol{Y})$ a suitable model is $\boldsymbol{U} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu}$ depends on the age of the patient, the treatment used and the stage of disease reached. The distribution of $\boldsymbol{Y}$ can then be derived knowing the Jacobian of the transformation.

Further reading related to this section

Rice, J.A. (1995) *Mathematical Statistics and Data Analysis (2nd edition)*, sections 3.3, 3.5, 4.4.
Hogg, R.V., McKean, J.W. and Craig, A.T. (2005) *Introduction to Mathematical Statistics (6th edition)*, section 3.5.
Larsen, R.J. and Marx, M.L. (2010) *An Introduction to Mathematical Statistics and its Applications (5th edition)*, sections 11.5.
Miller, I. and Miller, M. (2004) *John E. Freund's Mathematical Statistics with Applications*, sections 6.7.