

School of Mathematics, Statistics and Computer Science

STAT261
Statistical Inference Notes

Printed at the University of New England, October 4, 2007

Contents

1	Estimation	4
1.1	Statistics	4
1.1.1	Examples of Statistics	5
1.2	Estimation by the Method of Moments	8
1.3	Estimation by the Method of Maximum Likelihood	12
1.4	Properties of Estimators	15
1.5	Examples of Estimators and their Properties	19
1.6	Properties of Maximum Likelihood Estimators	21
1.7	Confidence Intervals	22
1.7.1	Pivotal quantity	25
1.8	Bayesian estimation	28
1.8.1	Bayes' theorem for random variables	28
1.8.2	Post is prior \times likelihood	28
1.8.3	Likelihood	31
1.8.4	Prior	31
1.8.5	Posterior	32
1.9	Normal Prior and Likelihood	34
1.10	Bootstrap Confidence Intervals	36
1.10.1	The empirical cumulative distribution function.	36
2	Hypothesis Testing	40
2.1	Introduction	40
2.2	Terminology and Notation.	41
2.2.1	Hypotheses	41
2.2.2	Tests of Hypotheses	41
2.2.3	Size and Power of Tests	42
2.3	Examples	43
2.4	One-sided and Two-sided Tests	47
2.4.1	Case(a) Alternative is one-sided	48
2.4.2	Case (b) Two-sided Alternative	48
2.4.3	Two Approaches to Hypothesis Testing	50
2.5	Two-Sample Problems	53

2.6	Connection between Hypothesis testing and CI's	55
2.7	Summary	57
2.8	Bayesian Hypothesis Testing	58
2.8.1	Notation	58
2.8.2	Bayesian approach	58
2.9	Non-Parametric Hypothesis testing.	61
2.9.1	Kolmogorov-Smirnov (KS)	61
2.9.2	Asymptotic distribution	62
2.9.3	Bootstrap Hypothesis Tests	65
3	Chi-square Distribution	67
3.1	Distribution of S^2	67
3.2	Chi-Square Distribution	69
3.3	Independence of \bar{X} and S^2	75
3.4	Confidence Intervals for σ^2	75
3.5	Testing Hypotheses about σ^2	77
3.6	χ^2 and Inv- χ^2 distributions in Bayesian inference	79
3.6.1	Non-informative priors	79
3.7	The posterior distribution of the Normal variance	80
3.7.1	Inverse Chi-squared distribution	81
3.8	Relationship between χ^2_ν and Inv- χ^2_ν	82
3.8.1	Gamma and Inverse Gamma	82
3.8.2	Chi-squared and Inverse Chi-squared	82
3.8.3	Simulating Inverse Gamma and Inverse- χ^2 random variables.	82
4	F Distribution	85
4.1	Derivation	85
4.2	Properties of the F distribution	86
4.3	Use of F-Distribution in Hypothesis Testing	89
4.4	Pooling Sample Variances	92
4.5	Confidence Interval for σ_1^2/σ_2^2	94
4.6	Comparing parametric and bootstrap confidence intervals for σ_1^2/σ_2^2	94
5	t-Distribution	96
5.1	Derivation	96
5.2	Properties of the t-Distribution	97
5.3	Use of t-Distribution in Interval Estimation	99
5.4	Use of t-distribution in Hypothesis Testing	104
5.5	Paired-sample t-test	109
5.6	Bootstrap T-intervals	111

6	Analysis of Count Data	115
6.1	Introduction	115
6.2	Goodness-of-Fit Tests	115
6.3	Contingency Tables	125
6.3.1	Method	125
6.4	Special Case: 2×2 Contingency Table	129
6.5	Fisher's Exact Test	131
6.6	Parametric Bootstrap- X^2	134
7	Analysis of Variance	138
7.1	Introduction	138
7.2	The Basic Procedure	138
7.3	Single Factor Analysis of Variance	140
7.4	Estimation of Means and Confidence Intervals	148
7.5	Assumptions Underlying the Analysis of Variance	149
7.5.1	Tests for Equality of Variance	150
7.6	Estimating the Common Mean	152
8	Simple Linear Regression	153
8.1	Introduction	153
8.2	Estimation of α and β	154
8.3	Estimation of σ^2	159
8.4	Inference about $\hat{\alpha}$, β and μ_Y	161
8.5	Correlation	166

The notes

Material for these notes has been drawn and collated from the following sources

- *Mathematical Statistics with Applications* William Mendenhall, Dennis Wakerly, Richard Schaeffer. Duxbury ISBN 0-534-92026-8
- *Bayesian Statistics an introduction, third edition* Peter Lee. Hodder Arnold ISBN 0-340-81405-5.
- *Bayesian Data Analysis* Andrew Gelman, John Carlin, Hal Stern, Donald Rubin. Chapman & Hall. ISBN 1-58488-388-X
- *An Introduction to the Bootstrap* Bradley Effron, Robert Tibshirani. Chapman & Hall. ISBN 0-412-04231-2
- *Introduction to Statistics through Resampling Methods and R/S-PLUS* Phillip Good. Wiley ISBN 0-471-71575-1

There are 3 broad categories of statistical inference

- Parameteric, Frequentist
- Parametric, Bayesian
- Non-parametric

These are not mutually exclusive and both semi-parametric and non-parametric Bayesian models are powerful methods in modern statistics.

Statistical Inference is a vast area which cannot be covered in a 1 semester undergraduate course. This unit shall focus mainly on frequentist parametric statistical inference but it is not intended that this has more weight than the others. Each has their place. However we can introduce the rudimentary concepts about probability, intervals and the mathematics required to derive such. These mathematical techniques apply equally well in the other settings, along with the knowledge of densities, probability etc.

A discussion of statistical inference would be most unbalanced if it were restricted to only one type of inference. Successful statistical modelling requires flexibility and the statistician (with practice) recognises which type of model is suggested by the data and the problem at hand.

The notes are organised to introduce the alternative inference methods in sections. At this stage you may consider the different methods as alternate ways of using the information in the data.

Not all topics include the 3 categories of inference because the Bayesian or Nonparametric counterpart does not always align with the frequentist methods. However, where there exist sufficient alignment, an alternate method of inference is introduced. It is hoped that this will stimulate students to explore the topics of Bayesian and Nonparametric statistics more fully in later units.

Parametric, Frequentist

Both systematic and random components are represented by a mathematical model and the model is a function of parameters which are estimated from the data. For example

$$y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij} \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

is a parametric model where the parameters are

- the coefficients of the systematic model, β_0, β_1
- the variance of the random model, σ^2 .

A rough description of frequentist methods is that population values of the parameters are unknown and based on a sample (x, y) we get estimates of the true, but unknown values. These are denoted as $\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2$ in this case.

Bayesian

Whereas in frequentist inference the data are considered a random sample and the parameters fixed, Bayesian statistics regards the data as fixed and the parameters as random samples. The exercise is that given the data, what are the distributions of the parameters such that the observed sample from those distributions could give rise to the observed data.

Non-parametric

This philosophy does not assume that a mathematical form (with parameters) should be imposed on the data and the model is determined by the data themselves. The techniques include

- permutation tests, bootstrap, Kolmogorov-Smirnov tests etc.
- Kernel density estimation, kernel regression, smoothing splines etc.

This seems a good idea to not impose any predetermined mathematical form on the data. However, the “limitations” are

- the data are not summarized by parameters and so interpretation of the data requires whole curves etc. There is not a ready formula to plug in values to derive estimates.

- Requires sound computing skills and numerical methods.
- The statistical method may be appropriate only when there is sufficient data to reliably indicate associations etc. without the assistance of a parametric model.

Chapter 1

Estimation

The application of the methods of **probability** to the analysis and interpretation of data is known as **statistical inference**. In particular, we wish to make an inference about a **population** based on information contained in a **sample**. Since populations are characterized by numerical descriptive measures called parameters, the objective of many statistical investigations is to make an inference about one or more population parameters. There are two broad areas of inference: **estimation** (the subject of this chapter) and **hypothesis-testing** (the subject of the next chapter).

When we say that we have a **random sample** X_1, X_2, \dots, X_n “from a **random variable** X ” or “from a population with **distribution function** $F(x; \theta)$ ”, we mean that X_1, X_2, \dots, X_n are **identically and independently distributed random variables** each with c.d.f. $F(x; \theta)$, that is, depending on some parameter θ . We usually assume that the *form* of the distribution, e.g., binomial, Poisson, Normal, etc. is known but the parameter is unknown. We wish to obtain information from the data (sample) to enable us to make some statement about the parameter. Note that, θ may be a vector, e.g., $\theta = (\mu, \sigma^2)$. See WMS 2.12 for more detailed comments on random samples.

The general problem of estimation is to find out something about θ using the information in the observed values of the sample, x_1, x_2, \dots, x_n . That is, we want to choose a function $H(x_1, x_2, \dots, x_n)$ that will give us a good estimate of the parameter θ in $F(x; \theta)$.

1.1 Statistics

We will introduce the technical meaning of the word **statistic** and look at some **commonly used statistics**.

Definition 1.1 Any function of the elements of a random sample, which does not depend on unknown parameters, is called a **statistic**.

Strictly speaking, $H(X_1, X_2, \dots, X_n)$ is a statistic and $H(x_1, x_2, \dots, x_n)$ is the observed value of the statistic. Note that the former is a random variable, often called an **estimator of θ** , while $H(x_1, x_2, \dots, x_n)$ is called an **estimate** of θ . However, the word *estimate* is sometimes used for both random variable and its observed value.

1.1.1 Examples of Statistics

Suppose that we have a random sample X_1, X_2, \dots, X_n from a distribution with mean μ and variance σ^2 .

1. $\bar{X} = \sum_{i=1}^n X_i/n$ is called the **sample mean**.
2. $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$ is called the **sample variance**.
3. $S = \sqrt{S^2}$ is called the sample standard deviation.
4. $M_r = \sum_{i=1}^n X_i^r/n$ is called the **rth sample moment** about the origin.
5. Suppose that the random variables X_1, \dots, X_n are ordered and re-written as $X_{(1)}, X_{(2)}, \dots, X_{(n)}$. The vector $(X_{(1)}, \dots, X_{(n)})$ is called the **ordered sample**.
 - (a) $X_{(1)}$ is called the **minimum of the sample**, sometimes written X_{\min} or $\min(X_i)$.
 - (b) $X_{(n)}$ is called the **maximum of the sample**, sometimes written X_{\max} or $\max(X_i)$.
 - (c) $X_{\max} - X_{\min} = R$ is called the **sample range**.
 - (d) The **sample median** is $X_{(\frac{n+1}{2})}$ if n is odd, and $\frac{1}{2}(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)})$ if n is even.

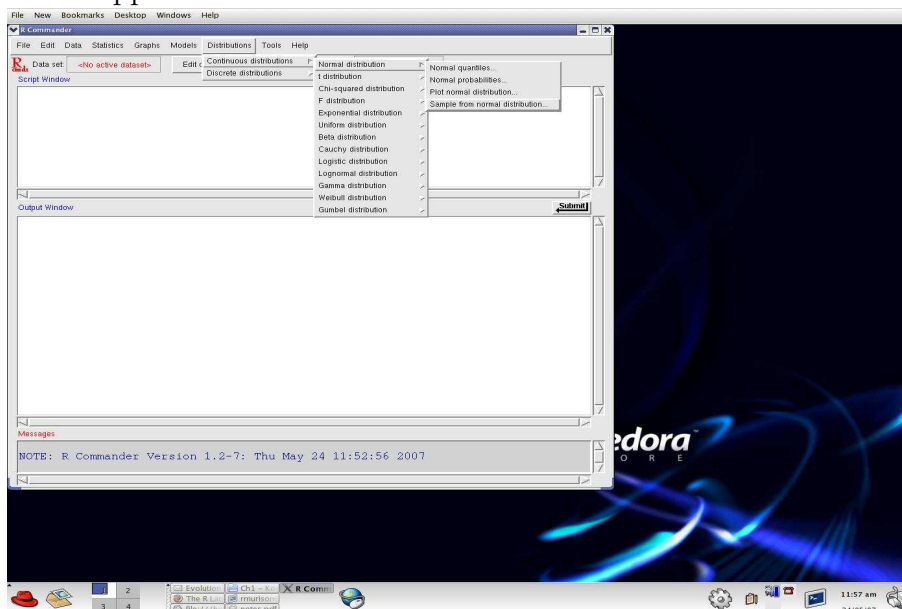
Computer Exercise 1.1

Generate a random sample of size 100 from a **normal distribution** with mean 10 and **standard deviation** 3. Use *R* to find the value of the (sample) mean, variance, standard deviation, minimum, maximum, range, median, and M_2 , the statistics defined above. Repeat for a sample of size 100 from an **exponential distribution** with parameter 1.

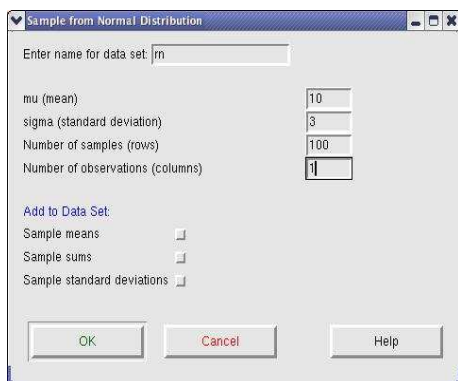
Solution:

<pre>#----- SampleStats.R ----- # Generate the normal random sample rn <- rnorm(n=100,mean=10,sd= 3) print(summary(rn)) cat("mean = ",mean(rn),"\n") cat("var = ",var(rn),"\n") cat("sd = ",sd(rn),"\n") cat("range = ",range(rn),"\n") cat("median = ",median(rn),"\n") cat("Second Moment = ",mean(rn^2),"\n")</pre>	<pre>> source("SampleStats.R") Min. 1st Qu. Median Mean 3rd Qu. Max. 1.8 7.9 9.4 9.8 12.0 18.2 mean = 9.9 var = 9.5 sd = 3.1 range = 1.8 18 median = 9.4 Second Moment = 106</pre>
--	--

If you are using Rcmdr, the menu for generating normal random variables is Distributions → Normal distribution → Sample from a normal distribution which appears like this:-

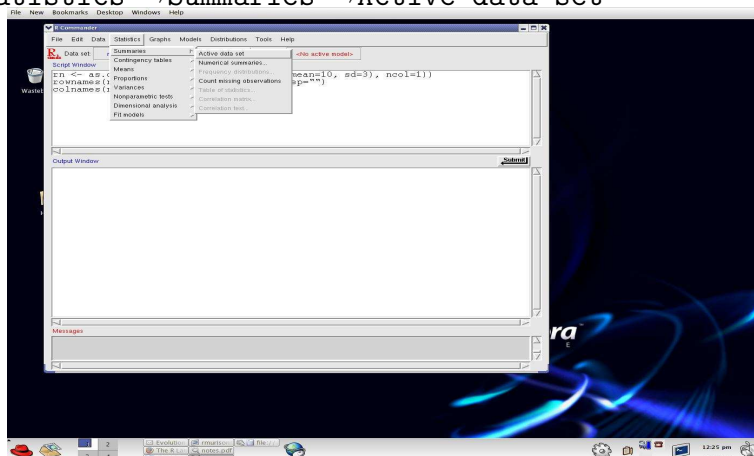


You must then supply a name for the data set (e.g. `rn`) and the parameters $\mu = 10$, $\sigma = 3$. Make the number of rows = 100 and the number of columns = 1.



When you click **OK** a data set containing the numbers is produced. Here the name of the data set is `rn` and this appears in **Data set:** `rn` in the top left of Rcmdr. Observe in the script window that Rcmdr is using the input through the menus (GUI's) to produce a script akin to that above. Rcmdr is an alternative of computing but it is not sufficiently comprehensive to do all our computing and usually requires augmenting with other scripts.

If there is an active data set, summary statistics are derived by
Statistics → **Summaries** → **Active data set**



The summary statistics are output in the Output window.

The exponential random sample is generated using: `re <- rexp(n=100, rate=1)`
 or the Rcmdr menus

Distributions → **Exponential distribution** → **Sample from an exponential distribution**

In the probability distributions considered in STAT260 the mean and variance are simple functions of the parameter(s), so when considering statistics, it is helpful to note which ones you'd expect to give information about the mean and which ones give information about the variance. Clearly, \bar{X} and the sample median give information about μ whereas S^2 and R give information about σ^2 .

We have not previously encountered S^2 , M_r , R , etc. but we (should) already know the following facts about \bar{X} .

- (i) It is a random variable with $E(\bar{X}) = \mu$, $\text{Var}(\bar{X}) = \sigma^2/n$, where $\mu = E(X_i)$, $\sigma^2 = \text{Var}(X_i)$.
- (ii) If X_1, X_2, \dots, X_n is from a **normal** distribution, then \bar{X} is also normally distributed.
- (iii) For large n and *any* distribution of the X_i for which a mean (μ) and variance (σ^2) exist, \bar{X} is distributed approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ (by the Central Limit Theorem).

Next we will consider some general methods of estimation. Since different methods may lead to different estimators for the same parameter, we will then need to consider criteria for deciding whether one estimate is better than another.

1.2 Estimation by the Method of Moments

Recall that, for a random variable X , the r th moment about the origin is $\mu'_r = E(X^r)$ and that for a random sample X_1, X_2, \dots, X_n , the r th sample moment about the origin is defined by

$$M_r = \sum_{i=1}^n X_i^r / n, \quad r = 1, 2, 3, \dots$$

and its observed value is denoted by

$$m_r = \sum_{i=1}^n x_i^r / n.$$

Note that the first sample moment is just the sample mean, \bar{X} .

We will first prove a property of sample moments.

Theorem 1.1

Let X_1, X_2, \dots, X_n be a random sample of X . Then

$$E(M_r) = \mu'_r, \quad r = 1, 2, 3, \dots$$

Proof

$$E(M_r) = \frac{1}{n} E\left(\sum_{i=1}^n X_i^r\right) = \frac{1}{n} \sum_{i=1}^n E(X_i^r) = \frac{1}{n} \sum_{i=1}^n \mu'_r = \mu'_r.$$

This theorem provides the motivation for estimation by the method of moments (with the estimator being referred to as the method of moments estimator or MME). The sample

moments, M_1, M_2, \dots , are random variables whose means are μ'_1, μ'_2, \dots . Since the population moments depend on the parameters of the distribution, estimating them by the sample moments leads to estimation of the parameters.

We will consider this method of estimation by means of 2 examples, then state the general procedure

Example 1.1

In this example, the distribution only has one parameter.

Given X_1, X_2, \dots, X_n is a random sample from a $U(0, \theta)$ distribution, find the method of moments estimator (MME) of θ .

Solution: Now, for the uniform distribution ($f(x) = \frac{1}{\theta} I_{[0, \theta]}(x)$),

$$\begin{aligned} \mu = E(X) &= \int_0^\theta x \times \frac{1}{\theta} dx \\ &= \frac{\theta}{2} \end{aligned}$$

Using the Method of Moments we proceed to estimate $\mu = \theta/2$ by m_1 . Thus since $m_1 = \bar{x}$ we have

$$\frac{\tilde{\theta}}{2} = \bar{x}$$

and,

$$\tilde{\theta} = 2\bar{x}.$$

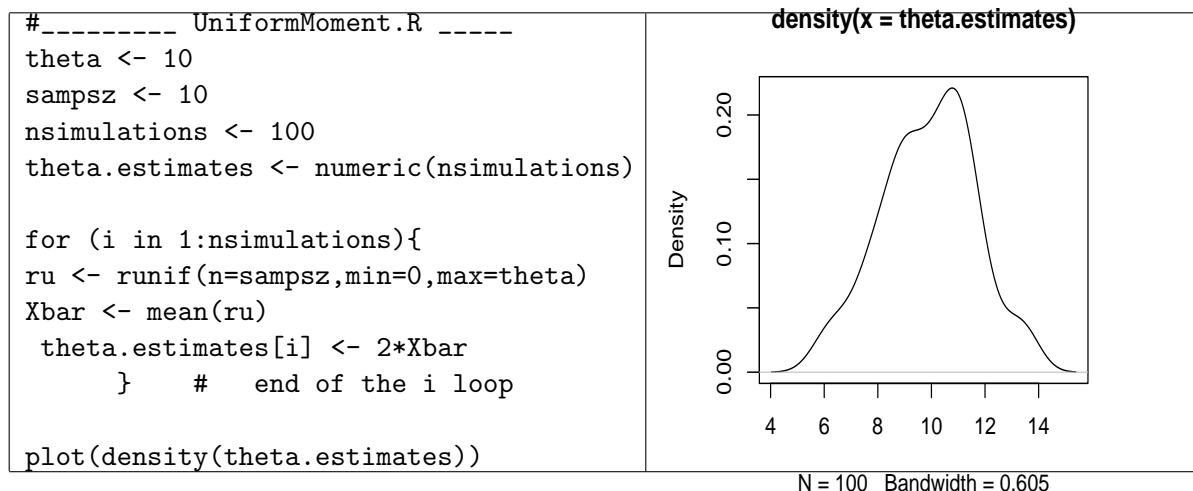
Then, $\tilde{\theta} = 2\bar{x}$ and the MME of θ is $2\bar{X}$.

Computer Exercise 1.2

Generate 100 samples of size 10 from a uniform distribution, $U(0, \theta)$ with $\theta = 10$. Estimate the value of θ from your samples using the method of moments and plot the results. Comment on the results.

In this exercise, we know *a priori* that $\theta = 10$ and have generated random samples. The samples are analysed as if θ unknown and estimated by the method of moments. Then we can compare the estimates with the known value.

Solution:



(You should do the exercise and obtain a plot for yourself).

It should be clear from the plot that about 50% of the estimates are greater than 10 which is outside the parameter space for a $U(0,10)$ distribution. This is undesirable.

Example 1.2

In this example the distribution has two parameters.

Given X_1, \dots, X_n is a random sample from the $N(\mu, \sigma^2)$ distribution, find the method of moments estimates of μ and σ^2 .

Solution:

For the normal distribution, $E(X) = \mu$ and $E(X^2) = \sigma^2 + \mu^2$ (Theorem 2.2 STAT260).

Using the Method of Moments:

Equate $E(X)$ to m_1 and $E(X^2)$ to m_2 so that, $\tilde{\mu} = \bar{x}$ and $\tilde{\sigma}^2 + \tilde{\mu}^2 = m_2$.

That is, estimate μ by \bar{x} and estimate σ^2 by $m_2 - \bar{x}^2$. Then,

$$\tilde{\mu} = \bar{x}, \text{ and } \tilde{\sigma}^2 = \frac{1}{n} \sum x_i^2 - \bar{x}^2.$$

The latter can also be written as $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

Computer Exercise 1.3

Generate 100 samples of size 10 from a normal distribution with $\mu = 14$ and $\sigma = 4$. Estimate μ and σ^2 from your samples using the method of moments. Plot the estimated values of μ and σ^2 . Comment on your results.

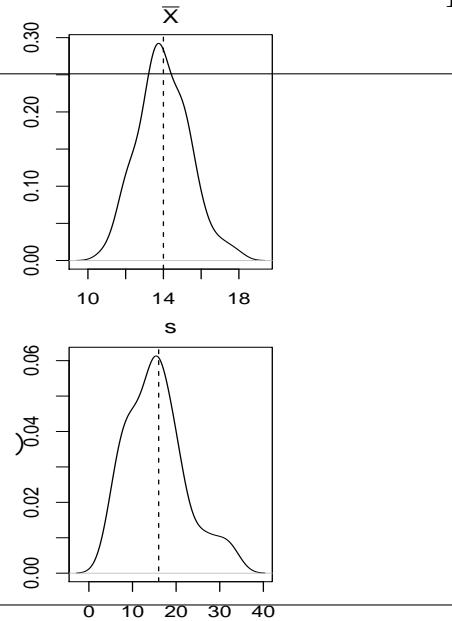
Solution:

```

#-----NormalMoments.R-----
mu <- 14
sigma <- 4
sampsz <- 10
nsimulations <- 100
mu.estimates <- numeric(nsimulations)
var.estimates <- numeric(nsimulations)
for (i in 1:nsimulations){
  rn <- rnorm(mean=mu,sd=sigma,n=sampsz)
  mu.estimates[i] <- mean(rn)
  var.estimates[i] <- mean( (rn -mean(rn))^2
                          ) # end of i loop

plot(density(mu.estimates))
plot(density(var.estimates))

```



The plot you obtain for the row means should be centred around the true mean of 14. However, you will notice that the plot of the variances is not centred about the true variance of 16 as you would like. Rather it will appear to be centred about a value less than 16. The reason for this will become evident when we study the properties of estimators in section 1.5.

General Procedure

Let X_1, X_2, \dots, X_n be a random sample from $F(x : \theta_1, \dots, \theta_k)$. That is, suppose that there are k parameters to be estimated. Let μ'_r, m_r ($r = 1, 2, \dots, k$) denote the first k population and sample moments respectively, and suppose that each of these population moments are certain known functions of the parameters. That is,

$$\begin{aligned}
 \mu'_1 &= g_1(\theta_1, \dots, \theta_k) \\
 \mu'_2 &= g_2(\theta_1, \dots, \theta_k) \\
 &\vdots \\
 \mu'_k &= g_k(\theta_1, \dots, \theta_k) .
 \end{aligned}$$

Solving simultaneously the set of equations,

$$\mu'_r = g_r(\tilde{\theta}_1, \dots, \tilde{\theta}_k) = m_r, \quad r = 1, 2, \dots, k$$

gives the required estimates, $\tilde{\theta}_1, \dots, \tilde{\theta}_k$.

1.3 Estimation by the Method of Maximum Likelihood

First the term **likelihood of the sample** must be defined. This has to be done separately for discrete and continuous distributions.

Definition 1.2

Let x_1, x_2, \dots, x_n be sample observations taken on the random variables X_1, X_2, \dots, X_n . Then the likelihood of the sample, $L(\theta|x_1, x_2, \dots, x_n)$, is defined as:

- (i) the joint probability of x_1, x_2, \dots, x_n if X_1, X_2, \dots, X_n are discrete, and
- (ii) the joint probability density function of X_1, \dots, X_n evaluated at x_1, x_2, \dots, x_n if the random variables are continuous.

In general the value of the likelihood depends not only on the (fixed) sample x_1, x_2, \dots, x_n but on the value of the (unknown) parameter θ . and can be thought of as a function of θ . The **likelihood function** for a set of n identically and independently distributed (iid) random variables, X_1, X_2, \dots, X_n , can thus be written as:

$$L(\theta; x_1, \dots, x_n) = \begin{cases} P(X_1 = x_1).P(X_2 = x_2)...P(X_n = x_n) & \text{for X discrete} \\ f(x_1; \theta).f(x_2; \theta)...f(x_n; \theta) & \text{for X continuous.} \end{cases} \quad (1.1)$$

For the discrete case, $L(\theta; x_1, \dots, x_n)$ is the probability (or likelihood) of observing $(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$. It would then seem that a sensible approach to selecting an estimate of θ would be to find the value of θ which maximizes the probability of observing $(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$, (the event which occurred).

The **maximum likelihood estimate** (MLE) of θ is defined as that value of θ which maximizes the likelihood. To state it more mathematically, the MLE of θ is that value of θ , say $\hat{\theta}$ such that

$$L(\hat{\theta}; x_1, \dots, x_n) > L(\theta'; x_1, \dots, x_n).$$

where θ' is any other value of θ .

Before we consider particular examples of MLE's, some comments about notation and technique are needed.

Comments

1. It is customary to use $\hat{\theta}$ to denote both estimator (random variable) and estimate (its observed value). Recall that we used $\tilde{\theta}$ for the MME.

2. Since $L(\theta; x_1, x_2, \dots, x_n)$ is a product, and sums are usually more convenient to deal with than products, it is customary to maximize $\log L(\theta; x_1, \dots, x_n)$ which we usually abbreviate to $l(\theta)$. This has the same effect. Since $\log L$ is a strictly increasing function of L , it will take on its maximum at the same point.
3. In some problems, θ will be a vector in which case $L(\theta)$ has to be maximized by differentiating with respect to 2 (or more) variables and solving simultaneously 2 (or more) equations.
4. The method of differentiation to find a maximum only works if the function concerned actually has a turning point.

Example 1.3

Given X is distributed $\text{bin}(1, p)$ where $p \in (0, 1)$, and a random sample x_1, x_2, \dots, x_n , find the maximum likelihood estimate of p .

Solution: The likelihood is,

$$\begin{aligned}
 L(p; x_1, x_2, \dots, x_n) &= P(X_1 = x_1)P(X_2 = x_2) \dots P(X_n = x_n) \\
 &= \prod_{i=1}^n \binom{1}{x_i} p^{x_i} (1-p)^{1-x_i} \\
 &= p^{x_1+x_2+\dots+x_n} (1-p)^{n-x_1-x_2-\dots-x_n} \\
 &= p^{\sum x_i} (1-p)^{n-\sum x_i}
 \end{aligned}$$

So

$$\log L(p) = \sum x_i \log p + (n - \sum x_i) \log(1-p)$$

Differentiating with respect to p , we have

$$\frac{d \log L(p)}{dp} = \frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p}$$

This is equal to zero when $\sum x_i(1-p) = p(n - \sum x_i)$, that is, when $p = \sum x_i/n$.

This estimate is denoted by \hat{p} .

Thus, if the random variable X is distributed $\text{bin}(1, p)$, the MLE of p derived from a sample of size n is

$$\hat{p} = \bar{X}. \quad (1.2)$$

Example 1.4

Given x_1, x_2, \dots, x_n is a random sample from a $N(\mu, \sigma^2)$ distribution, where both μ and σ^2 are unknown, find the maximum likelihood estimates of μ and σ^2 .

Solution: Write the likelihood as:

$$\begin{aligned} L(\mu, \sigma^2; x_1, \dots, x_n) &= \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i - \mu)^2/2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^n (x_i - \mu)^2/2\sigma^2} \end{aligned}$$

So

$$\log L(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \sum_{i=1}^n (x_i - \mu)^2/2\sigma^2$$

To maximize this w.r.t. μ and σ^2 we must solve simultaneously the two equations

$$\partial \log L(\mu, \sigma^2) / \partial \mu = 0 \quad (1.3)$$

$$\partial \log L(\mu, \sigma^2) / \partial \sigma^2 = 0. \quad (1.4)$$

These equations become, respectively,

$$-\frac{1}{2} \cdot \left(-\frac{2}{\sigma^2}\right) \sum_{i=1}^n (x_i - \mu) = 0 \quad (1.5)$$

$$\frac{-n}{2\sigma^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^4} = 0 \quad (1.6)$$

From (1.5) we obtain $\sum_{i=1}^n x_i = n\mu$, so that $\hat{\mu} = \bar{x}$. Using this in equation (1.6), we obtain

$$\hat{\sigma}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n.$$

Thus, if X is distributed $N(\mu, \sigma^2)$, the MLE's of μ and σ^2 derived from a sample of size n are

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / n. \quad (1.7)$$

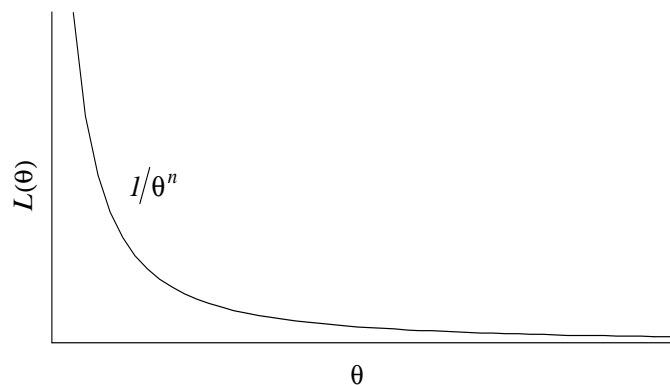
Note that these are the same estimators as obtained by the method of moments.

Example 1.5

Given random variable X is distributed uniformly on $[0, \theta]$, find the MLE of θ based on a sample of size n .

Solution: Now $f(x_i; \theta) = 1/\theta$, $x_i \in [0, \theta]$, $i = 1, 2, \dots, n$. So the likelihood is

$$L(\theta; x_1, x_2, \dots, x_n) = \prod_{i=1}^n (1/\theta) = 1/\theta^n.$$

Figure 1.1: $L(\theta) = 1/\theta^n$ 

When we come to find the maximum of this function we note that the slope is not zero anywhere, so there is no use finding $\frac{dL(\theta)}{d\theta}$ or $\frac{d \log L(\theta)}{d\theta}$.

Note however that $L(\theta)$ increases as $\theta \rightarrow 0$. So $L(\theta)$ is maximized by setting θ equal to the *smallest* value it can take. If the observed values are x_1, \dots, x_n then θ can be no smaller than the *largest* of these. This is because $x_i \in [0, \theta]$ for $i = 1, \dots, n$. That is, each $x_i \leq \theta$ or $\theta \geq$ each x_i .

Thus, if X is distributed $U(0, \theta)$, the MLE of θ is

$$\hat{\theta} = \max(X_i). \quad (1.8)$$

Comment

The Method of Moments was first proposed near the turn of the century by the British statistician Karl Pearson. The Method of Maximum Likelihood goes back much further. Both Gauss and Daniel Bernoulli made use of the technique, the latter as early as 1777. Fisher though, in the early years of the twentieth century, was the first to make a thorough study of the method's properties and the procedure is often credited to him.

1.4 Properties of Estimators

Using different methods of estimation can lead to different estimators. Criteria for deciding which are *good* estimators are required. Before listing the qualities of a good estimator, it is important to understand that they are random variables. For example, suppose that we take a sample of size 5 from a uniform distribution and calculate \bar{x} . Each time we repeat

the experiment we will probably get a different sample of 5 and therefore a different \bar{x} . The behaviour of an estimator for different random samples will be described by a probability distribution. The actual distribution of the estimator is not a concern here and only its mean and variance will be considered. As a first condition it seems reasonable to ask that the distribution of the estimator be *centered* around the parameter it is estimating. If not it will tend to overestimate or underestimate θ . A second property an estimator should possess is *precision*. An estimator is precise if the dispersion of its distribution is small. These two concepts are incorporated in the definitions of *unbiasedness* and *efficiency* below.

In the following, X_1, X_2, \dots, X_n is a random sample from the distribution $F(x; \theta)$ and $H(X_1, \dots, X_n) = \hat{\theta}$ will denote an estimator of θ (not necessarily the MLE).

Definition 1.3 Unbiasedness

An estimator $\hat{\theta}$ of θ is **unbiased** if

$$E(\hat{\theta}) = \theta \text{ for all } \theta. \quad (1.9)$$

If an estimator $\hat{\theta}$ is biased, the **bias** is given by

$$b = E(\hat{\theta}) - \theta. \quad (1.10)$$

There may be large number of unbiased estimators of a parameter for any given distribution and a further criterion for choosing between all the unbiased estimators is needed.

Definition 1.4 Efficiency

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be 2 unbiased estimators of θ with variances $\text{Var}(\hat{\theta}_1)$, $\text{Var}(\hat{\theta}_2)$ respectively, We say that $\hat{\theta}_1$ is **more efficient** than $\hat{\theta}_2$ if

$$\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2).$$

That is, $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if it has a smaller variance.

Definition 1.5 Relative Efficiency

The **relative efficiency** of $\hat{\theta}_2$ with respect to $\hat{\theta}_1$ is defined as

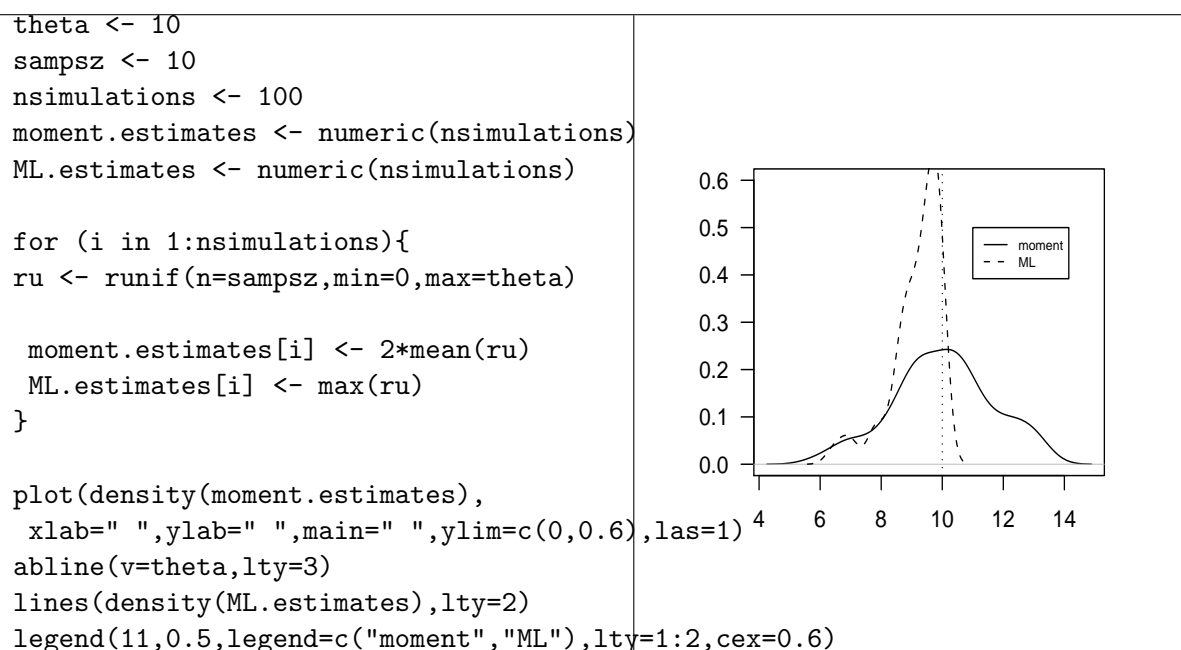
$$\text{efficiency} = \text{Var}(\hat{\theta}_1)/\text{Var}(\hat{\theta}_2). \quad (1.11)$$

Computer Exercise 1.4

Generate 100 random samples of size 10 from a $U(0,10)$ distribution. For each of the 100 samples generated calculate the MME and MLE for μ and graph the results.

- From the graphs does it appear that the estimators are biased or unbiased? Explain.
- Estimate the variance of the two estimators by finding the sample variance of the 100 estimates (for each estimator). Which estimator appears more efficient?

Solution:



You should see that the Method of Moments gives unbiased estimates of which many are not in the range space as noted in Computer Example 1.3. The maximum likelihood estimates are all less than 10 and so are biased.

It will now be useful to indicate that the estimator is based on a sample of size n by denoting it by $\hat{\theta}_n$.

Definition 1.6 Consistency $\hat{\theta}_n$ is a **consistent** estimator of θ if

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0 \text{ for all } \epsilon > 0. \quad (1.12)$$

We then say that $\hat{\theta}_n$ **converges in probability** to θ as $n \rightarrow \infty$. Equivalently,

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1.$$

This is a large-sample or *asymptotic* property. Consistency has to do only with the limiting behaviour of an estimator as the sample size increases without limit and does not imply that the observed value of $\hat{\theta}$ is necessarily close to θ for any specific size of sample n . If only a relatively small sample is available, it would seem immaterial whether a consistent estimator is used or not.

The following theorem (which will not be proved) gives a method of testing for consistency.

Theorem 1.2

If, $\lim_{n \rightarrow \infty} E(\hat{\theta}_n) = \theta$ and $\lim_{n \rightarrow \infty} \text{Var}(\hat{\theta}_n) = 0$, then $\hat{\theta}_n$ is a consistent estimator of θ .

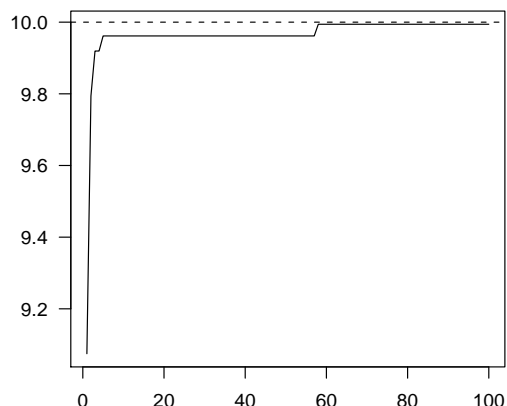
Computer Exercise 1.5

Demonstrate that the MLE is consistent for estimating θ for a $U(0, \theta)$ distribution.

Method: Generate the random variables one at a time. After each is generated calculate the MLE of θ for the sample of size n generated to that point and so obtain a sequence of estimators, $\{\hat{\theta}_n\}$. Plot the sequence.

Solution: Uniform random variables are generated one at a time and $\hat{\theta}_n$ is found as the maximum of $\hat{\theta}_{n-1}$ and n th uniform rv generated. The estimates are plotted in order.

```
#----- UniformConsistency.R -----
theta <- 10
sampsz <- 10
nsimulations <- 100
ML.est <- numeric(nsimulations)
for (i in 1:nsimulations){
  ru <- runif(n=sampsz,min=0,max=theta)
  if(i==1) ML.est[i] <- max(ru)
  else ML.est[i] <- max(ML.est[i-1],max(ru) )
}
plot(ML.est,type='l')
abline(h=theta,lty=2)
```



As n increases $\hat{\theta} \rightarrow \theta$.

The final concept of **sufficiency** requires some explanation before a formal definition is given. The random sample X_1, X_2, \dots, X_n drawn from the distribution with $F(x; \theta)$ contains information about the parameter θ . To estimate θ , this sample is first condensed

to a single random variable by use of a statistic $\theta^* = H(X_1, X_2, \dots, X_n)$. The question of interest is whether any information about θ has been lost by this condensing process. For example, a possible choice of θ^* is $H(X_1, \dots, X_n) = X_1$ in which case it seems that some of the information in the sample has been lost since the observations X_2, \dots, X_n have been ignored. In many cases, the statistic θ^* does contain all the relevant information about the parameter θ that the sample contains. This is the concept of **sufficiency**.

Definition 1.7 Sufficiency

Let X_1, X_2, \dots, X_n be a random sample from $F(x; \theta)$ and let $\theta^* = H(X_1, X_2, \dots, X_n)$ be a statistic (a function of the X_i only). Let $\theta' = H'(X_1, X_2, \dots, X_n)$ be any other statistic which is not a function of θ^* . If for each of the statistics θ' , the conditional density of θ' given θ^* does not involve θ , then θ^* is called a **sufficient statistic** for θ . That is, if $f(\theta'|\theta^*)$ does not contain θ , then θ^* is sufficient for θ .

Note: Application of this definition will not be required, but you should think of sufficiency in the sense of using all the relevant information in the sample. For example, to say that \bar{x} is sufficient for μ in a particular distribution means that knowledge of the actual observations x_1, x_2, \dots, x_n gives us no more information about μ than does only knowing the average of the n observations.

1.5 Examples of Estimators and their Properties

In this section we will consider the sample mean \bar{X} and the sample variance S^2 and examine which of the above properties they have.

Theorem 1.3

Let X be a random variable with mean μ and variance σ^2 . Let \bar{X} be the sample mean based on a random sample of size n . Then \bar{X} is an **unbiased** and **consistent** estimator of μ .

Proof Now $E(\bar{X}) = \mu$, no matter what the sample size is, and $\text{Var}(\bar{X}) = \sigma^2/n$. The latter approaches 0 as $n \rightarrow \infty$, satisfying Theorem 1.2.

It can also be shown that of all linear functions of X_1, X_2, \dots, X_n , \bar{X} has minimum variance. Note that the above theorem is true no matter what distribution is sampled. Some applications are given below.

For a random sample X_1, X_2, \dots, X_n , \bar{X} is an unbiased and consistent estimator of:

- (i) μ when the X_i are distributed $N(\mu, \sigma^2)$;

- (ii) p when the X_i are distributed $\text{bin}(1, p)$;
- (iii) λ when the X_i are distributed $\text{Poisson}(\lambda)$;
- (iv) $1/\alpha$ when the X_i have p.d.f. $f(x) = \alpha e^{-\alpha x}, x > 0$.

Sample Variance

Recall that the sample variance is defined by

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

Theorem 1.4

Given X_1, X_2, \dots, X_n is a random sample from a distribution with mean μ and variance σ^2 , then S^2 is an unbiased estimator of σ^2 .

Proof

$$\begin{aligned}
 (n-1)E(S^2) &= E \sum_{i=1}^n (X_i - \bar{X})^2 \\
 &= E \sum_{i=1}^n [X_i - \mu - (\bar{X} - \mu)]^2 \\
 &= E \left[\sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + n(\bar{X} - \mu)^2 \right] \\
 &= E \left[\sum_{i=1}^n (X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2 \right] \\
 &= E \sum_{i=1}^n (X_i - \mu)^2 - nE(\bar{X} - \mu)^2 \\
 &= \sum_{i=1}^n \text{Var}(X_i) - n\text{Var}(\bar{X}) \\
 &= n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2 \\
 \text{So } E(S^2) &= \sigma^2 .
 \end{aligned} \tag{1.13}$$

We make the following comments.

- (i) In the special case of Theorem 1.4 where the X_i are distributed $N(\mu, \sigma^2)$ with both μ and σ^2 unknown, the MLE of σ^2 is $\sum_{i=1}^n (X_i - \bar{X})^2 / n$ which is $(n-1)S^2/n$. So in this case the MLE is biased.
- (ii) The number in the denominator of S^2 , that is, $n-1$, is called the **number of degrees of freedom**. The numerator is the sum of n deviations (from the mean) squared but the deviations are not independent. There is one constraint on them, namely the fact that $\sum (X_i - \bar{X}) = 0$. As soon as $n-1$ of the $X_i - \bar{X}$ are known, the n th one is determined.
- (iii) In calculating the observed value of S^2 , s^2 , the following form is usually convenient.

$$s^2 = \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right] / (n-1) \quad (1.14)$$

or, equivalently,

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

The equivalence of the two forms is easily seen:

$$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2$$

where the RHS can readily be seen to be $\sum x_i^2 - \frac{(\sum x_i)^2}{n}$.

- (iv) For any distribution that has a fourth moment,

$$\text{Var}(S^2) = \frac{\mu_4 - 3\mu_2^2}{n} - \frac{2\mu_2^2}{n-1} \quad (1.15)$$

Clearly $\lim_{n \rightarrow \infty} \text{Var}(S^2) = 0$, so from Theorem 1.2, S^2 is a consistent estimator of σ^2 .

1.6 Properties of Maximum Likelihood Estimators

The following four properties are the main reasons for recommending the use of Maximum Likelihood Estimators.

- (i) The MLE is consistent.
- (ii) The MLE has a distribution that tends to normality as $n \rightarrow \infty$.
- (iii) If a sufficient statistic for θ exists, then the MLE is sufficient.

- (iv) The MLE is **invariant** under functional transformations. That is, if $\hat{\theta} = H(X_1, X_2, \dots, X_n)$ is the MLE of θ and if $u(\theta)$ is a continuous monotone function of θ , then $u(\hat{\theta})$ is the MLE of $u(\theta)$. This is known as the **invariance property** of maximum likelihood estimators.
- For example, in the normal distribution where the mean is μ and the variance is σ^2 , $(n-1)S^2/n$ is the MLE of σ^2 , so the MLE of σ is $\sqrt{(n-1)S^2/n}$.

1.7 Confidence Intervals

In the earlier part of this chapter we have been considering **point estimators** of a parameter. By *point estimator* we are referring to the fact that, after the sampling has been done and the observed value of the estimator computed, our end-product is the single number which is hopefully a good approximation for the unknown true value of the parameter. If the estimator is good according to some criteria, then the estimate should be reasonably close to the unknown true value. But the single number itself does not include any indication of how high the probability might be that the estimator has taken on a value close to the true unknown value. The method of **confidence intervals** gives both an idea of the actual numerical value of the parameter, by giving it a *range* of possible values, and a measure of how confident we are that the true value of the parameter is in that range. To pursue this idea further consider the following example.

Example 1.6

Consider a random sample of size n for a normal distribution with mean μ (unknown) and known variance σ^2 . Find a 95% confidence interval for the unknown mean, μ .

Solution: We know that the best estimator of μ is \bar{X} and the sampling distribution of \bar{X} is $N(\mu, \frac{\sigma^2}{n})$. Then from the standard normal,

$$P\left(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < 1.96\right) = .95 .$$

The event $\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < 1.96$ is equivalent to the event

$$\mu - \frac{1.96\sigma}{\sqrt{n}} < \bar{X} < \mu + \frac{1.96\sigma}{\sqrt{n}} ,$$

which is equivalent to the event

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} .$$

Hence

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95 \quad (1.16)$$

The two statistics $\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}$, $\bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$ are the endpoints of a 95% confidence interval for μ . This is reported as

The 95% CI for μ is $\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$

Computer Exercise 1.6

Generate 100 samples of size 9 from a $N(0,1)$ distribution. Find the 95% CI for μ for each of these samples and count the number that do (don't) contain zero. (You could repeat this say 10 times to build up the total number of CI's generated to 1000.) You should observe that about 5% of the intervals don't contain the true value of $\mu(=0)$.

Solution: Use the commands:

```
#----- ConfInt.R -----
sampsz <- 9
nsimulations <- 100
non.covered <- 0
for (i in 1:nsimulations){
  rn <- rnorm(mean=0,sd=1,n=sampsz)

  Xbar <- mean(rn)
  s <- sd(rn)
  CI <- qnorm(mean=Xbar,sd=s/sqrt(sampsz),p=c(0.025,0.975) )

  non.covered <- non.covered + (CI[1] > 0) + (CI[2] < 0)
}
cat("Rate of non covering CI's",100*non.covered/nsimulations," % \n")

> source("ConfInt.R")
Rate of non covering CI's 8 %
```

This implies that 8 of the CI's don't contain 0. With a larger sample size we would expect that about 5% of the CI's would not contain zero.

We make the following definition:

Definition 1.8

An interval, at least one of whose endpoints is a random variable is called a **random interval**.

In (1.16), we are saying that the probability is 0.95 that the random interval $(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}})$ contains μ . A confidence interval (CI) has to be interpreted carefully. For a particular sample, where \bar{x} is the observed value of \bar{X} , a 95% CI for μ is

$$\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right), \quad (1.17)$$

but the statement

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

is either true or false. The parameter μ is a constant and either the interval contains it in which case the statement is true, or it does not contain it, in which case the statement is false. How then is the probability 0.95 to be interpreted? It must be considered in terms of the relative frequency with which the indicated event will occur “in the long run” of similar sampling experiments.

Each time we take a sample of size n , a different \bar{x} , and hence a different interval (1.17) would be obtained. Some of these intervals will contain μ as claimed, and some will not. In fact, if we did this many times, we’d expect that 95 times out of 100 the interval obtained would contain μ . The measure of our confidence is then 0.95 because **before a sample is drawn** there is a probability of 0.95 that the confidence interval to be constructed will cover the true mean.

A statement such as $P(3.5 < \mu < 4.9) = 0.95$ is incorrect and should be replaced by :

A 95% confidence interval for μ is (3.5, 4.9).

We can generalize the above as follows: Let $z_{\alpha/2}$ be defined by

$$\Phi(z_{\alpha/2}) = 1 - (\alpha/2). \quad (1.18)$$

That is, the area under the normal curve **above** $z_{\alpha/2}$ is $\alpha/2$. Then

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha.$$

So a $100(1 - \alpha)\%$ CI for μ is

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right). \quad (1.19)$$

Commonly used values of α are 0.1, 0.05, 0.01.

Confidence intervals for a given parameter are not unique. For example, we have considered a **symmetric, two-sided** interval, but

$$\left(\bar{x} - z_{2\alpha/3} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/3} \frac{\sigma}{\sqrt{n}} \right)$$

is also a $100(1 - \alpha)\%$ CI for μ . Likewise, we could have one-sided CI's for μ . For example,

$$\left(-\infty, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}}\right) \text{ or } \left(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty\right).$$

[The second of these arises from considering $P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_\alpha\right) = 1 - \alpha$.]

We could also have a CI based on say, the sample median instead of the sample mean. Methods of obtaining confidence intervals must be judged by their various statistical properties. For example, one desirable property is to have the length (or expected length) of a $100(1 - \alpha)\%$ CI as short as possible. Note that for the CI in (1.19), the length is constant for given n .

1.7.1 Pivotal quantity

We will describe a general method of finding a confidence interval for θ from a random sample of size n . It is known as the **pivotal method** as it depends on finding a pivotal quantity that has 2 characteristics:

- (i) It is a function of the sample observations and the unknown parameter θ , say $H(X_1, X_2, \dots, X_n; \theta)$ where θ is the only unknown quantity,
- (ii) It has a probability distribution that does not depend on θ .

Any probability statement of the form

$$P(a < H(X_1, X_2, \dots, X_n; \theta) < b) = 1 - \alpha$$

will give rise to a probability statement about θ .

Example 1.7

Given X_1, X_2, \dots, X_{n_1} from $N(\mu_1, \sigma_1^2)$ and Y_1, Y_2, \dots, Y_{n_2} from $N(\mu_2, \sigma_2^2)$ where σ_1^2, σ_2^2 are known, find a symmetric 95% CI for $\mu_1 - \mu_2$.

Solution: Consider $\mu_1 - \mu_2$ ($= \theta$, say) as a single parameter. Then \bar{X} is distributed $N(\mu_1, \sigma_1^2/n_1)$ and \bar{Y} is distributed $N(\mu_2, \sigma_2^2/n_2)$ and further, \bar{X} and \bar{Y} are independent. It follows that $\bar{X} - \bar{Y}$ is normally distributed, and writing it in standardized form,

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \text{ is distributed as } N(0, 1).$$

So we have found the pivotal quantity which is a function of $\mu_1 - \mu_2$ but whose distribution does not depend on $\mu_1 - \mu_2$. A 95% CI for $\theta = \mu_1 - \mu_2$ is found by considering

$$P\left(-1.96 < \frac{\bar{X} - \bar{Y} - \theta}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} < 1.96\right) = .95,$$

which, on rearrangement, gives the appropriate CI for $\mu_1 - \mu_2$. That is,

$$\left(\bar{x} - \bar{y} - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x} - \bar{y} + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right). \quad (1.20)$$

Example 1.8

In many problems where we need to estimate proportions, it is reasonable to assume that sampling is from a binomial population, and hence that the problem is to estimate p in the $\text{bin}(n, p)$ distribution, where p is unknown. Find a $100(1 - \alpha)\%$ CI for p , making use of the fact that for large sample sizes, the binomial distribution can be approximated by the normal.

Solution: Given X is distributed as $\text{bin}(n, p)$, an unbiased estimate of p is $\hat{p} = X/n$. For n large, X/n is approximately normally distributed. Then,

$$E(\hat{p}) = E(X)/n = p,$$

and

$$\text{Var}(\hat{p}) = \frac{1}{n^2} \text{Var}(X) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n},$$

so that

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \text{ is distributed approximately } N(0, 1).$$

[Note that we have found the required pivotal quantity whose distribution does not depend on p .]

An approximate $100(1 - \alpha)\%$ CI for p is obtained by considering

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) = 1 - \alpha. \quad (1.21)$$

where $z_{\alpha/2}$ is defined in (1.18).

Rearranging (1.21), the confidence limits for p are obtained as

$$\frac{2n\hat{p} + z_{\alpha/2}^2 \pm z_{\alpha/2} \sqrt{4n\hat{p}(1-\hat{p}) + z_{\alpha/2}^2}}{2(n + z_{\alpha/2}^2)}. \quad (1.22)$$

A simpler expression can be found by dividing both numerator and denominator of (1.22) by $2n$ and neglecting terms of order $1/n$. That is, a 95% CI for p is

$$\left(\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/n} \right). \quad (1.23)$$

Note that this is just the expression we would have used if we replaced $\text{Var}(\hat{p}) = p(1-p)/n$ in (1.21) by $\widehat{\text{Var}}(\hat{p}) = \hat{p}(1-\hat{p})/n$. In practice, confidence limits for p are generally obtained by means of specially constructed tables which makes it possible to find confidence intervals when n is small.

Example 1.9

Construct an appropriate 90% confidence interval for λ in the Poisson distribution. Evaluate this if a sample of size 30 yields $\sum x_i = 240$.

Solution: Now \bar{X} is an unbiased estimator of λ for this problem, so λ can be estimated by $\hat{\lambda} = \bar{x}$ with $E(\hat{\lambda}) = \lambda$ and $\text{Var}(\hat{\lambda}) = \text{Var}(\bar{X}) = \sigma^2/n = \lambda/n$. By the Central Limit Theorem, for large n , the distribution of \bar{X} is approximately normal, so

$$\frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \text{ is distributed approximately } N(0, 1).$$

An approximate 90% CI for λ can be obtained from considering

$$P\left(-1.645 < \frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} < 1.645\right) = .90. \quad (1.24)$$

Rearrangement of the inequality in (1.24) to give an inequality for λ , is similar to that in Example 1.8 where it was necessary to solve a quadratic. But, noting the comment following (1.23), replace the variance of \bar{X} by its estimate $\hat{\lambda}/n = \bar{X}/n$, giving for the 90% CI for λ ,

$$(\bar{x} - 1.645\sqrt{\bar{x}/n} \quad \bar{x} + 1.645\sqrt{\bar{x}/n})$$

which on substitution of the observed value $240/30 = 8$ for \bar{x} gives (7.15, 8.85).

1.8 Bayesian estimation

Fundamental results from probability

Some results from STAT260 are presented without elaboration, intended as a revision to provide the framework for Bayesian data analyses.

$$\begin{aligned}
 P(E|FH)P(F|H) &= P(EF|H) = P(EFH|H) = P(EH|H) = P(E|H) \\
 P(E) &= \sum_n P(E|H_n)P(H_n) \\
 P(H_n|E)P(E) &= P(EH_n) = P(H_n)P(E|H_n) \\
 P(H_n|E) &\propto P(H_n)P(E|H_n)
 \end{aligned}
 \tag{1.25}$$

$$\tag{1.26}$$

The results at (1.25) is Bayes' theorem and in this form shows how we can “invert” probabilities, getting $P(H_n|E)$ from $P(E|H_n)$.

When H_n consists of exclusive and exhaustive events,

$$P(H_n|E) = \frac{P(H_n)P(E|H_n)}{\sum_m P(H_m)P(E|H_m)} \tag{1.27}$$

1.8.1 Bayes' theorem for random variables

$$p(y|x) \propto p(y)p(x|y) \tag{1.28}$$

The constant of proportionality is

$$\begin{aligned}
 \text{continuous: } \frac{1}{p(x)} &= \frac{1}{\int p(x|y)p(y)dy} \\
 \text{discrete: } \frac{1}{p(x)} &= \frac{1}{\sum_y p(x|y)p(y)dy}
 \end{aligned}$$

1.8.2 Post is prior \times likelihood

Suppose we are interested in the values of k unknown quantities,

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$$

and *a priori* beliefs about their values can be expressed in terms of the pdf $p(\boldsymbol{\theta})$.

Then we collect data,

$$\mathbf{X} = (X_1, X_2, \dots, X_n)$$

which have a probability distribution that depends on $\boldsymbol{\theta}$, expressed as

$$p(\mathbf{X}|\boldsymbol{\theta})$$

From (1.28),

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$$

The term, $p(\mathbf{X}|\boldsymbol{\theta})$ may be considered as a function of \mathbf{X} for fixed $\boldsymbol{\theta}$, i.e. a density of \mathbf{X} which is parameterized by $\boldsymbol{\theta}$.

We can also consider the same term as a function of $\boldsymbol{\theta}$ for fixed \mathbf{X} and then it is termed the *likelihood function*,

$$\ell(\boldsymbol{\theta}|\mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta})$$

These are the names given to the terms of (1.28),

- $p(\boldsymbol{\theta})$ is the prior
- $\ell(\boldsymbol{\theta}|\mathbf{X})$ is the likelihood
- $p(\boldsymbol{\theta}|\mathbf{X})$ is the posterior

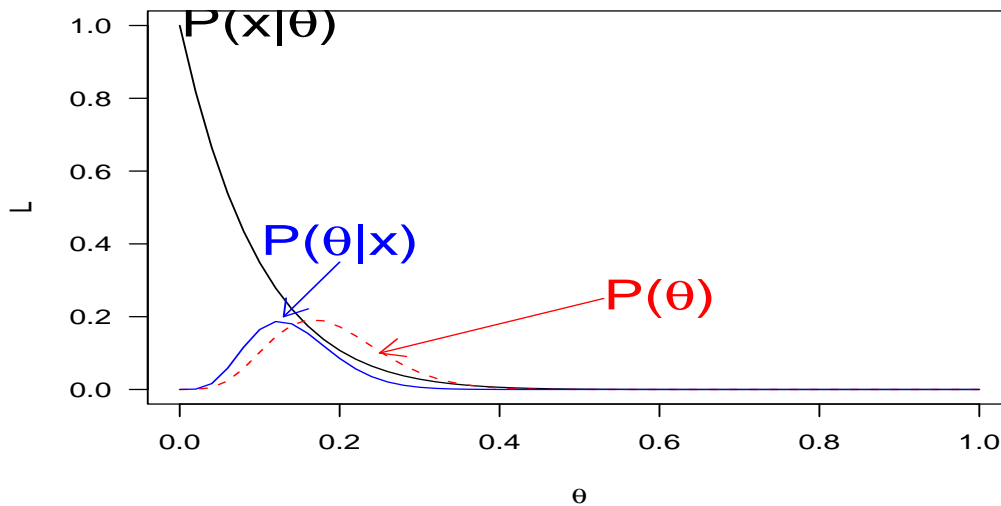
and Bayes' theorem is

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

The function $p(\cdot)$ is not the same in each instance but is a generic symbol to represent the density appropriate for prior, density of the data given the parameters, and the posterior. The form of p is understood by considering its arguments, i.e. $p(\theta)$, $p(x|\theta)$ or $p(\theta|x)$.

A diagram depicting the relationships amongst the different densities is shown in Figure 1.2

Figure 1.2: Posterior distribution



The posterior is a combination of the likelihood, where information about $\boldsymbol{\theta}$ comes from the data \mathbf{X} and the prior $p(\boldsymbol{\theta})$ where the information is the knowledge of $\boldsymbol{\theta}$ independent of

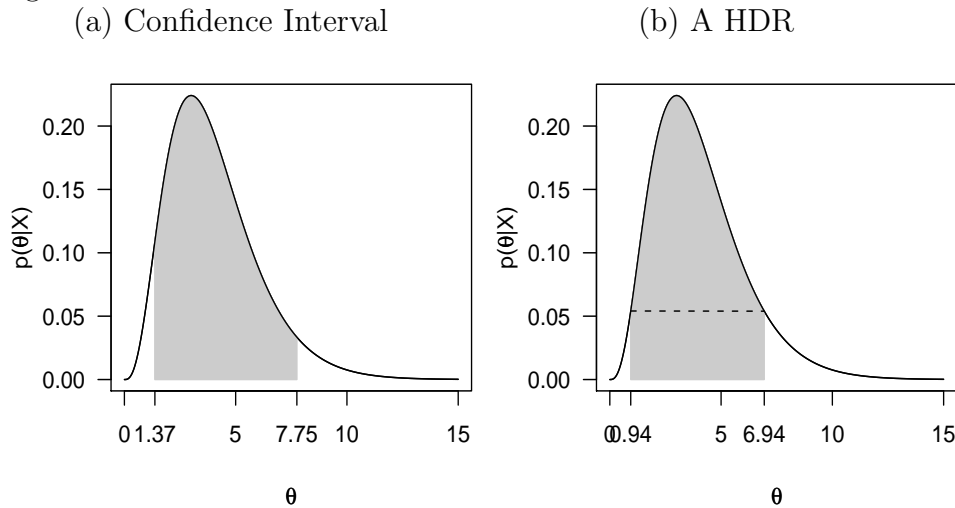
\mathbf{X} . This knowledge may come from previous sampling, (say). The *posterior* represents an update on $P(\theta)$ with the new information at hand, i.e. \mathbf{x} .

If the likelihood is weak due to insufficient sampling or wrong choice of likelihood function, the prior can dominate so that the posterior is just an adaptation of the prior. Alternatively, if the sample size is large so that the likelihood function is strong, the prior will not have much impact and the Bayesian analysis is the same as the maximum likelihood.

The output is a distribution, $p(\theta|\mathbf{X})$ and we may interpret it using summaries such as the median and an interval where the true value of θ would lie with a certain probability. The interval that we shall use is the *Highest Density Region*, or HDR. This is the interval for which the density of any point within it is higher than the density of any point outside.

Figure 1.3 depicts a density with shaded areas of 0.9 in 2 cases. In frame (a), observe that there are quantiles outside the interval (1.37, 7.75) for which the density is greater than quantiles within the interval. Frame (b) depicts the HDR as (0.94, 6.96).

Figure 1.3: Comparison of a 2 types of regions, a Confidence Interval and a Highest Density Region



The following example illustrates the principles.

Computer Exercise 1.7

Generate 10 samples of size 10 from a uniform distribution, $U(0, \theta)$ with $\theta = 10$. Estimate the 90% Highest Density Region (HDR) for θ from your samples using a prior $p(\theta) \propto \frac{1}{\theta}$.

We recap that the job is to use the simulated data ($X \sim U(0, \theta)$ ($\theta = 10$)), and estimate θ as if it were unknown. Then the estimate is compared with the true value.

Both the previous estimation methods have not been entirely satisfactory,

- moment estimation gave too many estimates way outside the true value,

- maximum likelihood estimates were biased because we could only use the maximum value and had no information regarding future samples which might exceed the maximum.

The Bayesian philosophy attempts to address these concerns.

For this example and further work, we require the *indicator function*,

$$I_A(x) = \begin{cases} 1 & (x \in A) \\ 0 & (x \notin A) \end{cases}$$

1.8.3 Likelihood

Denote the joint density of the observations by $p(\mathbf{x}|\theta)$. For $X \sim U(0, \theta)$,

$$p(\mathbf{x}|\theta) = \begin{cases} \theta^{-n} & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases} \quad (1.29)$$

The likelihood of the parameter θ is

$$\ell(\theta|\mathbf{x}) = \begin{cases} \theta^{-n} & \theta > x \\ 0 & \text{otherwise} \end{cases}$$

and if $M = \max(x_1, x_2, \dots, x_n)$,

$$\begin{aligned} \ell(\theta|\mathbf{x}) &= \begin{cases} \theta^{-n} & \theta > M \\ 0 & \text{otherwise} \end{cases} \\ &\propto \theta^{-n} I_{(M, \infty)}(\theta) \end{aligned}$$

1.8.4 Prior

When $X \sim U(0, \theta)$, a convenient prior distribution for θ is

$$\begin{aligned} p(\theta|\xi, \gamma) &= \begin{cases} \gamma \xi^\gamma \theta^{-\gamma-1} & \theta > \xi \\ 0 & \text{otherwise} \end{cases} \\ &\propto \theta^{-\gamma-1} I_{(\xi, \infty)}(\theta) \end{aligned}$$

This is a *Pareto* distribution. At this point, just accept that this prior works but later we shall consider how to choose priors.¹

¹ ξ is the Greek letter pronounced “xi”.

1.8.5 Posterior

By Bayes rule,

$$\begin{aligned}
 p(\theta|\mathbf{x}) &\propto p(\theta) \times \ell(\theta|\mathbf{x}) \\
 &\propto \underbrace{\theta^{-\gamma-1} I_{(\xi, \infty)}(\theta)}_{\text{prior}} \times \underbrace{\theta^{-n} I_{(M, \infty)}(\theta)}_{\text{likelihood}} \\
 &\propto \theta^{-(\gamma+n)-1} I_{(\xi', \infty)}(\theta)
 \end{aligned}$$

where $\xi' = \max(M, \xi)$.

Thus we combine the information gained from the data with of our prior beliefs to get a distribution of θ 's.

In this exercise, there is a fixed lower endpoint which is zero, $X \sim U(0, \theta)$.

The prior chosen is a Pareto distribution with $\xi = 0$, $\gamma = 0$ so that

$$p(\theta) = \theta^{-1} I_{(\xi, \infty)}(\theta)$$

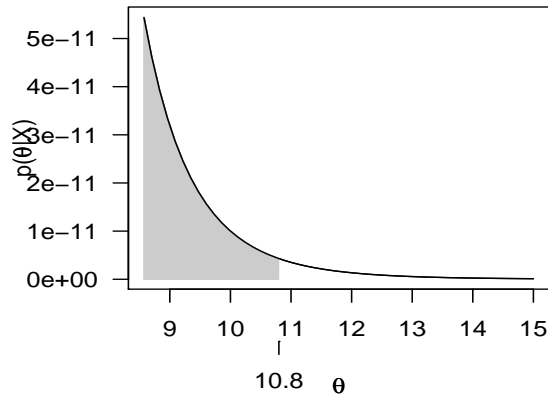
This is chosen so that the prior does not change very much over the region in which the likelihood is appreciable and does not take on large values outside that region. It is said to be *locally uniform*. We defer the theory about this, for now you may just accept that it is appropriate for this exercise.

The posterior density, $p(\theta|\mathbf{X})$ is

$$p(\theta|\mathbf{X}) \propto \theta^{-n-1} I_{(\xi', \infty)}(\theta)$$

The HDR will be as in Figure 1.4

Figure 1.4: The 90% HDR for $p(\theta|\mathbf{X})$



The lower end-point is $M = \max(x_1, x_2, \dots, x_n)$. This is the MLE and in that setting, there was no other information that we could use to address the point that $\theta \geq M$; M

had to do the job but we were aware that it was very possible that the true value of θ was greater than the maximum value of the sample.

The upper end-point is found from the distribution function. We require ν such that

$$\begin{aligned}\int_{\xi'}^{\nu} p(\theta|\mathbf{X})d\theta &= 0.9 \\ \left[1 - \left(\frac{M}{\nu}\right)^n\right] &= 0.9 \\ \nu &= \frac{M}{0.1^{\frac{1}{n}}}\end{aligned}$$

Likewise, we can compute the median of the posterior distribution,

$$Q_{0.5} = \frac{M}{0.5^{\frac{1}{n}}}$$

The following R program was used to estimate the median and HDR of $p(\theta|\mathbf{X})$.

```
#_____ UniformBayes.R _____
theta <- 10
sampsz <- 10
nsimulations <- 10

for (i in 1:nsimulations){
  xi <- max(runif(n=sampsz,min=0,max=theta) )
  Q0.9 <- xi/(0.1^(1/sampsz) )
  Q0.5 <- xi/(0.5^(1/sampsz) )
  cat("simulation ",i,"median = ",round(Q0.5,2),"90% HDR = (",round(xi,2),round(Q0.9,2),")\n")
}
```

```
simulation 1 median = 10.65 90% HDR = ( 9.94 12.51 )
simulation 2 median = 10.09 90% HDR = ( 9.42 11.85 )
simulation 3 median = 8.92 90% HDR = ( 8.32 10.48 )
simulation 4 median = 10.64 90% HDR = ( 9.93 12.5 )
simulation 5 median = 9.86 90% HDR = ( 9.2 11.59 )
simulation 6 median = 9.88 90% HDR = ( 9.22 11.61 )
simulation 7 median = 8.4 90% HDR = ( 7.84 9.87 )
simulation 8 median = 8.66 90% HDR = ( 8.08 10.18 )
simulation 9 median = 10.41 90% HDR = ( 9.71 12.22 )
simulation 10 median = 9.19 90% HDR = ( 8.57 10.79 )
```

1.9 Normal Prior and Likelihood

This section is included to demonstrate the process for modelling the posterior distribution of parameters and the notes shall refer to it in an example in Chapter 2.

- $x \sim N(\mu, \sigma^2)$
- $\mu \sim N(\mu_0, \sigma_0^2)$

$$\begin{aligned}
 p(x|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right\} \\
 p(\mu) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2} \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right\} \\
 p(\mu|x) &= p(x|\mu)p(\mu) \\
 &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right\} \times \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2} \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right\}
 \end{aligned} \tag{1.30}$$

$$\propto \exp\left\{-\frac{1}{2}\mu^2\left(\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}\right) + \mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2}\right)\right\} \tag{1.31}$$

Define the precision as the reciprocal of the variance,

$$\tau = \frac{1}{\sigma^2} \quad \tau_0 = \frac{1}{\sigma_0^2}$$

Addressing (1.31), put

$$\tau_1 = \tau + \tau_0 \tag{1.32}$$

$$\mu_1 = \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}\right) \times \left(\frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2}\right) \tag{1.33}$$

$$\tag{1.34}$$

Equation 1.32 states

$$\begin{aligned}
 \text{Posterior precision} &= \text{datum precision} + \text{prior precision} \quad \text{or} \\
 \frac{1}{\sigma_1^2} &= \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}
 \end{aligned}$$

Then (1.31) can be expressed as

$$p(\mu|x) \propto \exp\left\{-\frac{1}{2}\tau_1\mu + \tau_1\mu\mu_1\right\}$$

Add into the exponent the term $-\frac{1}{2}\tau_1\mu_1^2$ which is a constant as far as μ is concerned. Then

$$\begin{aligned} p(\mu|x) &\propto \exp \left\{ -\frac{1}{2}\tau_1(\mu - \mu_1)^2 \right\} \\ &= (2\pi\sigma_1^2)^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left(\frac{\mu - \mu_1}{\sigma_1} \right)^2 \right\} \end{aligned}$$

The last result containing the normalising constant $(2\pi\sigma_1^2)^{-\frac{1}{2}}$ comes from $\int p(\mu|x)dx = 1$.

Thus the posterior density is $\mu|x \sim N(\mu_1, \sigma_1^2)$, where

$$\begin{aligned} \frac{1}{\sigma_1^2} &= \frac{1}{\sigma^2} + \frac{1}{\sigma_0^2} \\ \mu_1 &= \mu_0 \frac{\tau_0}{\tau_0 + \tau} + x \frac{\tau}{\tau_0 + \tau} \\ &= \sigma_1^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{x}{\sigma^2} \right) \end{aligned}$$

Posterior mean is weighted mean of prior mean and datum value. The weights are proportional to their respective precisions.

Example 1.10

Suppose that $\mu_0 \sim N(370, 20^2)$ and that $x|\mu \sim N(421, 8^2)$. What is $p(\mu|x)$?

$$\begin{aligned} \tau_1 &= \frac{1}{20^2} + \frac{1}{8^2} \Rightarrow \\ \sigma_1^2 = \frac{1}{\tau_1} &= 55 \\ \mu_1 &= 55 \left(\frac{370}{20^2} + \frac{421}{8^2} \right) = 413 \\ \mu|x &\sim N(413, 55) \end{aligned}$$

1.10 Bootstrap Confidence Intervals

The *Bootstrap* is a Monte-Carlo method which uses (computer) simulation in lieu of mathematical theory. It is not necessarily simpler. Exercises with the bootstrap are mostly numerical although the underlying theory follows much of the analytical methods.

1.10.1 The empirical cumulative distribution function.

An important tool in non-parametric statistics is the *empirical cumulative distribution function* (acronym ecdf) which uses the ordered data as quantiles and probabilities are steps of $\frac{1}{(n+1)}$.

We have used the word *empirical* for this plot because it uses only the information in the sample. The values for cumulative area that are associated with each datum are determined by the following argument.

The sample as collected is denoted by x_1, x_2, \dots, x_n . The subscript represents the position in the list or row in the data file. Bracketed subscripts denote ordering of the data, where $x_{(1)}$ is the smallest, $x_{(2)}$ is the second smallest, $x_{(n)}$ is the largest. In general x_i is not the same datum $x_{(i)}$ but of course this correspondence *could* happen.

The n sample points are considered to divide the sampling interval into $(n + 1)$ sub-intervals,

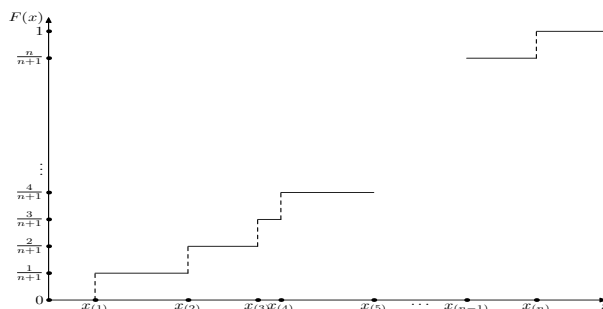
$$(0, x_{(1)}), (x_{(1)}, x_{(2)}), (x_{(2)}, x_{(3)}), \dots, (x_{(n-1)}, x_{(n)}), (x_{(n)}, \infty)$$

The total area under the density curve (area=1) has been subdivided into $(n + 1)$ sub-regions with individual areas approximated as $\frac{1}{(n+1)}$. The values of the cumulative area under the density curve is then approximated as:-

Interval	$(0, x_1)$	(x_1, x_2)	\dots	$(x_{(n-1)}, x_n)$	(x_n, ∞)
Cumulative area	$\frac{0}{(n+1)}$	$\frac{1}{(n+1)}$	\dots	$\frac{n}{(n+1)}$	1

A diagram of this is shown in Figure 1.5

Figure 1.5: The ecdf is a step function with step size $\frac{1}{(n+1)}$ between data points.



If there are tied data, say k of them, the step size is $\frac{k}{(n+1)}$.

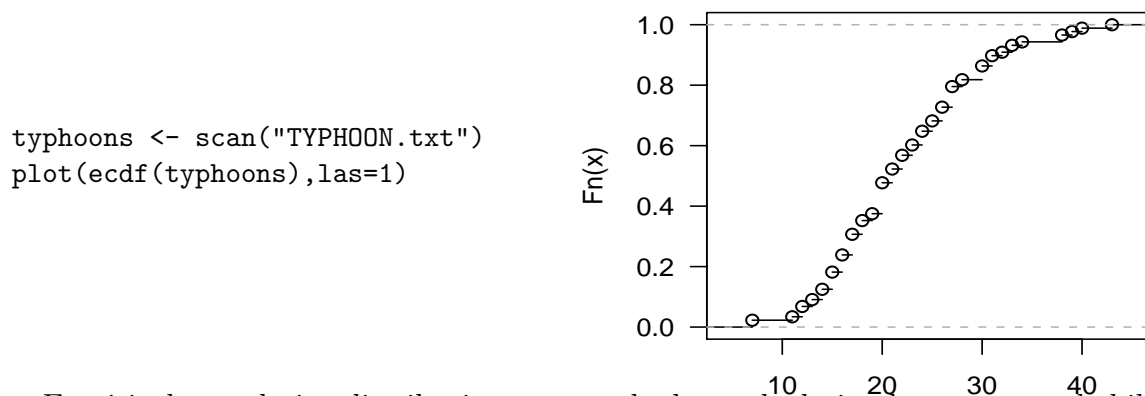
In R the required function is `ecdf()`.

The following data are the numbers of typhoons in the North Pacific Ocean over 88 years and assume that they are saved in a file called `TYPHOON.txt`

```
13 7 14 20 13 12 12 15 20 17 11 14 16 12 17 17
16 7 14 15 16 20 17 20 15 22 26 25 27 18 23 26 18 15 20 24
19 25 23 20 24 20 16 21 20 18 20 18 24 27 27 21 21 22 28 38
39 27 26 32 19 33 23 38 30 30 27 25 33 34 16 17 22 17 26 21
30 30 31 27 43 40 28 31 24 15 22 31
```

A plot of the ecdf shown in Figure 1.6 is generated with the following R code,

Figure 1.6: An empirical distribution function for typhoon data.



Empirical cumulative distributions are used when calculating bootstrap probabilities.

Example

Suppose that in Example 1.9, the data were

```
8 6 5 10 8 12 9 9 8 11 7 3 6 7 5 8 10 7 8 8 10 8 5 10 8 6 10 6 8 14
```

Denote this sample by x_1, x_2, \dots, x_n where $n = 30$. the summary statistics are

$$\sum_{i=1}^n x_i = 240 \quad \bar{X} = 8$$

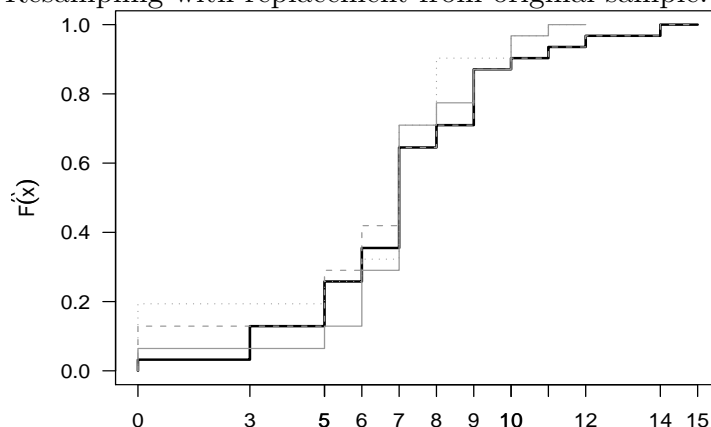
We shall use this example to illustrate (a) resampling, and (b) the bootstrap distribution.

The sample, x_1, x_2, \dots, x_n , are independently and identically distributed (i.i.d.) as $\text{Poisson}(\lambda)$ which means that each observation is as important as any other for providing information about the population from which this sample is drawn. That infers we can replace any number by one of the others and the “new” sample will still convey the same information about the population.

This is demonstrated in Figure 1.7. Three “new” samples have been generated by taking samples of size $n = 30$ with replacement from \mathbf{x} . The ecdf of \mathbf{x} is shown in bold and

the ecdf's of the “new” samples are shown with different line types. There is little change in the empirical distributions or estimates of quantiles. If a statistic (e.g. a quantile) were estimated from this process a large number of times, it would be a reliable estimate of the population parameter. The “new” samples are termed *bootstrap samples*.

Figure 1.7: Resampling with replacement from original sample.

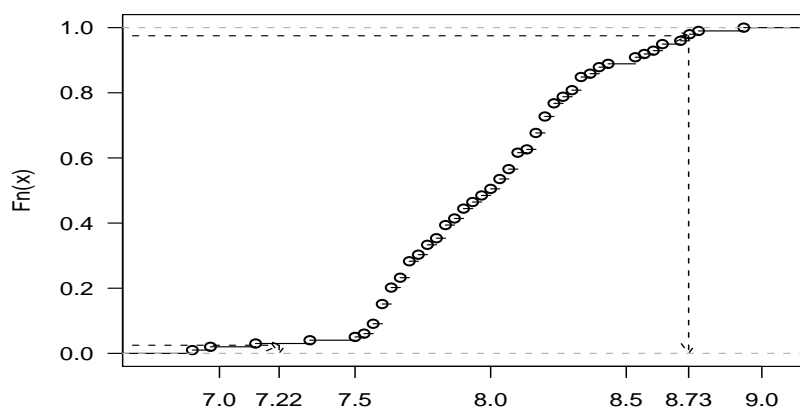


This is the bootstrap procedure for the CI for λ in the current example.

1. Nominate the number of bootstrap samples that will be drawn, e.g. $nBS=99$.
2. Sample with replacement from \mathbf{x} a bootstrap sample of size n , \mathbf{x}_1^* .
3. For each bootstrap sample, calculate the statistic of interest, $\hat{\lambda}_1^*$.
4. Repeat steps 2 and 3 nBS times.
5. Use the empirical cumulative distribution function of $\hat{\lambda}_1^*, \hat{\lambda}_2^*, \dots, \hat{\lambda}_{nBS}^*$ to get the Confidence Interval.

This is shown in Figure 1.8.

Figure 1.8: Deriving the 95% CI from the ecdf of bootstrap estimates of the mean



The bootstrap estimate of the 95% CI for λ is (7.22, 8.73). Note that although there is a great deal of statistical theory underpinning this (the ecdf, iid, a thing called order statistics etc.), there is no theoretical formula for the CI and it is determined numerically from the sample.

This is R code to generate the graph in Figure 1.8.

```
x <- c(8,6,5,10,8,12,9,9,8,11,7,3,6,7,5,8,10,7,8,8,10,8,5,10,8,6,10,6,8,14)
n <- length(x)
nBS <- 99                      # number of bootstrap simulations
BS.mean <- numeric(nBS)
i <- 1
while (i < (nBS+1) ){
  BS.mean[i] <- mean(sample(x,replace=T,size=n))
  i <- i + 1
}                                # end of the while() loop

Quantiles <- quantile(BS.mean,p = c(0.025,0.975))
cat(" 95% CI = ",Quantiles,"\n")
plot(ecdf(BS.mean),las=1)
```

The `boot` package in R has functions for bootstrapping. The following code uses that to get the same CI as above,

```
library(boot)
mnz <- function(z,id){mean(z[id])}    # user must supply this
bs.samples <- boot(data=x,statistic=mnz,R=99)
boot.ci(bs.samples,conf=0.95,type=c("perc","bca"))
```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 99 bootstrap replicates

CALL :

```
boot.ci(boot.out = bs.samples, conf = 0.95)
```

Intervals :

Level	Percentile	BCa
95%	(7.206, 8.882)	(7.106, 8.751)

It seems that the user must supply a function (e.g. `mnz` here) to generate the bootstrap samples. The variable `id` is recognised by R as a vector `1:length(z)` so that it can draw the samples.

Chapter 2

Hypothesis Testing

2.1 Introduction

Consider the following problems:

- (i) An engineer has to decide on the basis of sample data whether the true average lifetime of a certain kind of tyre is at least 22000 kilometres.
- (ii) An agronomist has to decide on the basis of experiments whether fertilizer A produces a higher yield of soybeans than fertilizer B.
- (iii) A manufacturer of pharmaceutical products has to decide on the basis of samples whether 90% of all patients given a new medication will recover from a certain disease.

These problems can be translated into the language of **statistical tests of hypotheses**.

- (i) The engineer has to test the assertion that if the lifetime of the tyre has pdf. $f(x) = \alpha e^{-\alpha x}$, $x > 0$, then the expected lifetime, $1/\alpha$, is at least 22000.
- (ii) The agronomist has to decide whether $\mu_A > \mu_B$ where μ_A , μ_B are the means of 2 normal distributions.
- (iii) The manufacturer has to decide whether p , the parameter of a binomial distribution is equal to .9.

In each case, it is assumed that the stated distribution correctly describes the experimental conditions, and that the hypothesis concerns the **parameter(s)** of that distribution. [A more general kind of hypothesis testing problem is where the **form** of the distribution is unknown.]

In many ways, the formal procedure for hypothesis testing is similar to the scientific method. The scientist formulates a theory, and then tests this theory against observation. In our context, the scientist poses a theory concerning the value of a parameter. He then samples the population and compares observation with theory. If the observations disagree strongly enough with the theory the scientist would probably reject his hypothesis. If not, the scientist concludes either that the theory is probably correct or that the sample he

considered did not detect the difference between the actual and hypothesized values of the parameter.

Before putting hypothesis testing on a more formal basis, let us consider the following questions. What is the role of statistics in testing hypotheses? How do we decide whether the sample value disagrees with the scientist's hypothesis? When should we reject the hypothesis and when should we withhold judgement? What is the probability that we will make the wrong decision? What function of the sample measurements should be used to reach a decision? Answers to these questions form the basis of a study of statistical hypothesis testing.

2.2 Terminology and Notation.

2.2.1 Hypotheses

A **statistical hypothesis** is an assertion or conjecture about the distribution of a random variable. We assume that the form of the distribution is known so the hypothesis is a statement about the value of a parameter of a distribution.

Let X be a random variable with distribution function $F(x; \theta)$ where $\theta \in \Omega$. That is, Ω is the set of all possible values θ can take, and is called the **parameter space**. For example, for the binomial distribution, $\Omega = \{p : p \in (0, 1)\}$. Let ω be a subset of Ω . Then a statement such as " $\theta \in \omega$ " is a statistical hypothesis and is denoted by H_0 . Also, the statement " $\theta \in \bar{\omega}$ " (where $\bar{\omega}$ is the complement of ω with respect to Ω) is called the **alternative** to H_0 and is denoted by H_1 . We write

$$H_0 : \theta \in \omega \text{ and } H_1 : \theta \in \bar{\omega} \text{ (or } \theta \notin \omega \text{)}.$$

Often hypotheses arise in the form of a claim that a new product, technique, etc. is better than the existing one. In this context, H is a statement that nullifies the claim (or represents the *status quo*) and is sometimes called a **null hypothesis**, but we will refer to it as **the hypothesis**.

If ω contains only one point, that is, if $\omega = \{\theta : \theta = \theta_0\}$ then H_0 is called a **simple hypothesis**. We may write $H_0 : \theta = \theta_0$. Otherwise it is called **composite**. The same applies to alternatives.

2.2.2 Tests of Hypotheses

A **test** of a statistical hypothesis is a procedure for deciding whether to "accept" or "reject" the hypothesis. If we use the term "accept" it is with reservation, because it implies stronger action than is really warranted. Alternative phrases such as "reserve judgement", "fail to reject" perhaps convey the meaning better. A **test** is a rule, or decision function, based on a sample from the given distribution which divides the sample space into 2 regions, commonly called

- (i) the **rejection region** (or **critical region**), denoted by R ;
- (ii) the **acceptance region** (or region of indecision), denoted by \bar{R} (complement of R).

If we compare two different ways of partitioning the sample space then we say we are comparing two tests (of the same hypothesis). For a sample of size n , the sample space is of course n -dimensional and rather than consider R as a subset of n -space, it's helpful to realize that we'll condense the information in the sample by using a statistic (for example \bar{x}), and consider the rejection region in terms of the range space of the random variable \bar{X} .

2.2.3 Size and Power of Tests

There are two types of errors that can occur. If we reject H when it is true, we commit a **Type I** error. If we fail to reject H when it is false, we commit a **Type II** error. You may like to think of this in tabular form.

		Our decision	
		do not reject H_0	reject H_0
Actual situation	H_0 is true	correct decision	Type I error
	H_0 is not true	Type II error	correct decision

Probabilities associated with the two incorrect decisions are denoted by

$$\alpha = P(H_0 \text{ is rejected when it is true}) = P(\text{Type I error}) \quad (2.1)$$

$$\beta = P(H_0 \text{ is not rejected when it is false}) = P(\text{Type II error}) \quad (2.2)$$

The probability α is sometimes referred to as the **size** of the critical region or the **significance level** of the test, and the probability $1 - \beta$ as the **power** of the test.

The roles played by H_0 and H_1 are not at all symmetric. From consideration of potential losses due to wrong decisions, the decision-maker is somewhat conservative for holding the hypothesis as true unless there is overwhelming evidence from the data that it is false. He believes that the consequence of wrongly rejecting H is much more severe to him than of wrongly accepting it.

For example, suppose a pharmaceutical company is considering the marketing of a newly developed drug for treatment of a disease for which the best available drug on the market has a cure rate of 80%. On the basis of limited experimentation, the research division claims that the new drug is more effective. If in fact it fails to be more effective, or if it has harmful side-effects, the loss sustained by the company due to the existing drug becoming obsolete, decline of the company's image, etc., may be quite severe. On the other hand, failure to market a better product may not be considered as severe a loss. In this problem it would be appropriate to consider $H_0 : p = .8$ and $H_1 : p > .8$. Note that H_0 is simple and H_1 is composite.

Ideally, when devising a test, we should look for a decision function which makes probabilities of Type I and Type II errors as small as possible, but, as will be seen in a later example, these depend on one another. For a given sample size, altering the decision rule to decrease one error, results in the other being increased. So, recalling that the Type I error is more serious, a possible procedure is to hold α fixed at a suitable level (say $\alpha = .05$ or $.01$) and then look for a decision function which minimizes β . The first solution for this was given by Neyman and Pearson for a simple hypothesis versus a simple alternative. It's often referred to as the Neyman-Pearson fundamental lemma. While the formulation of a general theory of hypothesis testing is beyond the scope of this unit, the following examples illustrate the concepts introduced above.

2.3 Examples

Example 2.1

Suppose that random variable X has a normal distribution with mean μ and variance 4. Test the hypothesis that $\mu = 1$ against the alternative that $\mu = 2$, based on a sample of size 25.

Solution: An unbiased estimate of μ is \bar{X} and we know that \bar{X} is distributed normally with mean μ and variance σ^2/n which in this example is $4/25$. We note that values of \bar{x} close to 1 support H whereas values of \bar{x} close to 2 support A. We could make up a decision rule as follows:

If $\bar{x} > 1.6$ claim that $\mu = 2$,

If $\bar{x} \leq 1.6$ claim that $\mu = 1$.

The diagram in Figure fig.CRUpperTail shows the sample space of \bar{x} partitioned into

- (i) the critical region, $R = \{\bar{x} : \bar{x} > 1.6\}$
- (ii) the acceptance region, $\bar{R} = \{\bar{x} : \bar{x} \leq 1.6\}$

Here, 1.6 is the critical value of \bar{x} .

We will find the probability of Type I and Type II error,

$$P(\bar{X} > 1.6 | \mu = 1, \sigma = \frac{2}{5}) = .0668. \quad (\text{pnorm}(q=1.6, \text{mean}=1, \text{sd}=0.4, \text{lower.tail}=F))$$

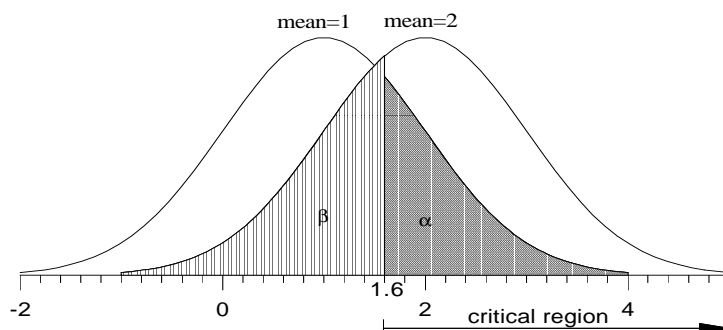
This is

$$P(H_0 \text{ is rejected} | H_0 \text{ is true}) = P(\text{Type I error}) = \alpha$$

Also

$$\begin{aligned} \beta = P(\text{Type II error}) &= P(H_0 \text{ is not rejected} | H_0 \text{ is false}) \\ &= P(\bar{X} \leq 1.6 | \mu = 2, \sigma = \frac{2}{5}) \\ &= .1587 \quad (\text{pnorm}(q=1.6, \text{mean}=2, \text{sd}=0.4, \text{lower.tail}=T)) \end{aligned}$$

Figure 2.1: Critical Region – Upper Tail



To see how the decision rule could be altered so that $\alpha = .05$, let the critical value be c . We require

$$P(\bar{X} > c | \mu = 1, \sigma = \frac{2}{5}) = 0.05$$

$$\Rightarrow c = 1.658 \quad (\text{qnorm}(p=0.05, \text{mean}=1, \text{sd}=0.4, \text{lower.tail}=T))$$

$$P(\bar{X} < c | \mu = 2, \sigma = \frac{2}{5}) = 0.196 \quad (\text{pnorm}(q=1.658, \text{mean}=2, \text{sd}=0.4, \text{lower.tail}=T))$$

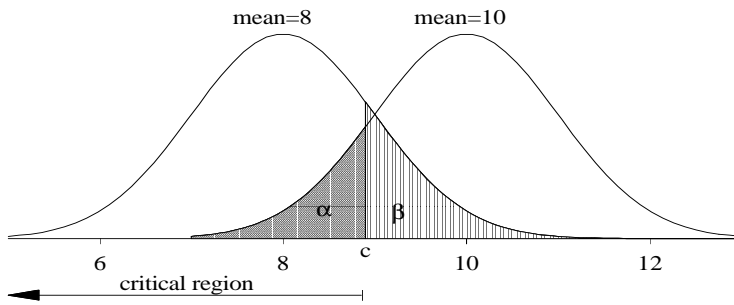
This value of c gives an α of 0.05 and a β of 0.196 illustrating that as one type of error (α) decreases the other (β) increases.

Example 2.2

Suppose we have a random sample of size n from a $N(\mu, 4)$ distribution and wish to test $H_0 : \mu = 10$ against $H_1 : \mu = 8$. The decision rule is to reject H_0 if $\bar{x} < c$. We wish to find n and c so that $\alpha = 0.05$ and $\beta \approx 1$.

Solution: In Figure 2.2 below, the left curve is $f(\bar{x}|H_1)$ and the right curve is $f(\bar{x}|H_0)$. The critical region is $\{\bar{x} : \bar{x} < c\}$, so α is the left shaded area and β is the right shaded area.

Figure 2.2: Critical Region – Lower Tail



Now

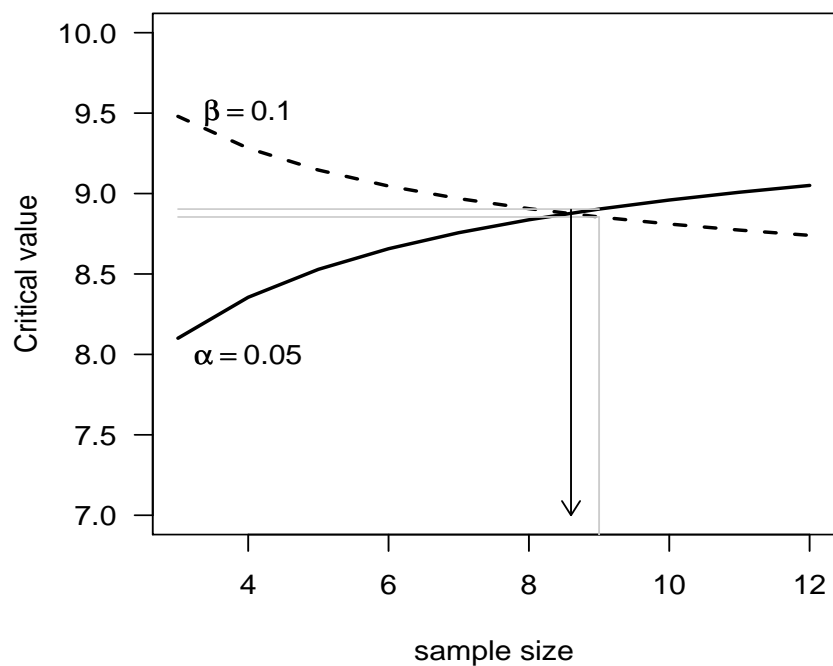
$$\alpha = 0.05 = P(\bar{X} < c | \mu = 10, \sigma = \frac{2}{\sqrt{n}}) \quad (2.3)$$

$$\beta = 0.1 = P(\bar{X} \geq c | \mu = 8, \sigma = \frac{2}{\sqrt{n}}) \quad (2.4)$$

$$(2.5)$$

We need to solve (2.3) and (2.4) simultaneously for n as shown in Figure 2.3

Figure 2.3: Solution for size and power of test



The R code for the above diagram is:-

```
n <- 3:12
alpha <- 0.05
beta <- 0.1
Acrit <- qnorm(mean=10,sd=2/sqrt(n),p=alpha)
Bcrit <- qnorm(mean=8,sd=2/sqrt(n),p=beta,lower.tail=F)

plot(Acrit ~ n,type='l',xlab="sample size",ylab="Critical value",las=1,ylim=c(7,10) ,lwd=2)
lines(n,Bcrit,lty=2,lwd=2)
```

A sample size $n = 9$ and critical value $c = 8.9$ gives $\alpha \approx 0.05$ and $\beta \approx 0.1$.

2.4 One-sided and Two-sided Tests

Consider the problem where the random variable X has a binomial distribution with $P(\text{Success})=p$. How do we test the hypothesis $p = 0.5$. Firstly, note that we have an experiment where the outcome on an individual trial is *success* or *failure* with probabilities p and q respectively. Let us repeat the experiment n times and observe the number of successes.

Before continuing with this example it is useful to note that in most hypothesis testing problems we will deal with, H_0 is simple, but H_1 on the other hand, is composite, indicating that the parameter can assume a range of values. Examples 1 and 2 were more straightforward in the sense that H_1 was simple also.

If the range of possible parameter values lies entirely on the one side of the hypothesized value, the alternative is said to be **one-sided**. For example, $H_1 : p > .5$ is one-sided but $H_1 : p \neq .5$ is **two-sided**. In a real-life problem, the decision of whether to make the alternative one-sided or two-sided is not always clear cut. As a general rule-of-thumb, if parameter values in only one direction are physically meaningful, or are the only ones that are possible, the alternative should be one-sided. Otherwise, H_1 should be two-sided. Not all statisticians would agree with this rule.

The next question is what test statistic we use to base our decision on. In the above problem, since X/n is an unbiased estimator of p , that would be a possibility. We could even use X itself. In fact the latter is more suitable since its distribution is known. Recall that, the principle of hypothesis testing is that we will assume H_0 is correct, and our position will change only if the data show **beyond all reasonable doubt** that H_1 is true. The problem then is to define in quantitative terms what reasonable doubt means. Let us suppose that $n = 18$ in our problem above. Then the range space for X is $R_X = \{0, 1, \dots, 18\}$ and $E(X)=np= 9$ if H_0 is true. If the observed number of successes is close to 9 we would be obliged to think that H was true. On the other hand, if the observed value of X was 0 or 18 we would be fairly sure that H_0 was not true. Now **reasonable doubt** does not have to be as extreme as 18 cases out of 18. Somewhere between x-values of 9 and 18 (or 9 and 0), there is a point, c say, when for all practical purposes the credulity of H_0 ends and reasonable doubt begins. This point is called the **critical value** and it completely determines the decision-making process. We could make up a decision rule

$$\begin{aligned} \text{If } x &\geq c, \text{ reject } H_0 \\ \text{If } x &< c, \text{ conclude that } H_0 \text{ is probably correct.} \end{aligned} \tag{2.6}$$

In this case, $\{x : x \geq c\}$ is the rejection region, R referred to in §2.2.

We will consider appropriate tests for both one- and two-sided alternatives in the problem above.

2.4.1 Case(a) Alternative is one-sided

In the above problem, suppose that the alternative is $H_1 : p > .5$. Only values of x much **larger** than 9 would support this alternative and a decision rule such as (2.6) would be appropriate. The actual value of c is chosen to make α , the size of the critical region, suitably small. For example, if $c = 11$, then $P(X \geq 11) = .24$ and this of course is too large. Clearly we should look for a value closer to 18. If $c = 15$, $P(X \geq 15) = \sum_{x=15}^{18} \binom{18}{x} (.5)^{18} = 0.004$, on calculation. We may now have gone too far in the other extreme. Requiring 15 or more successes out of 18 before we reject $H_0 : p = 0.5$ means that only 4 times in a thousand would we reject H_0 wrongly. Over the years, a reasonable consensus has been reached as to how much evidence against H_0 is enough evidence. In many situations we define the beginning of **reasonable doubt** as the value of the test statistic that is equalled or exceeded by chance 5% of the time when H_0 is true. According to this criterion, c should be chosen so that $P(X \geq c | H_0 \text{ is true}) = 0.05$. That is c should satisfy

$$P(X \geq c | p = 0.5) = 0.05 = \sum_{x=c}^{18} \binom{18}{x} (0.5)^{18}.$$

A little trial and error shows that $c = 13$ is the appropriate value. Of course because of the discrete nature of X it will not be possible to obtain an α of exactly 0.05.

Defining the critical region in terms of the x -value that is exceeded only 5% of the time when H_0 is true is the most common way to quantify reasonable doubt, but there are others. The figure 1% is frequently used and if the critical value is exceeded only 1% of the time we say there is **strong evidence** against H_0 . If the critical value is only exceeded .1% of the time we may say that there is **very strong evidence** against H_0 .

So far we have considered a one-sided alternative. Now we'll consider the other case where the alternative is two-sided.

2.4.2 Case (b) Two-sided Alternative

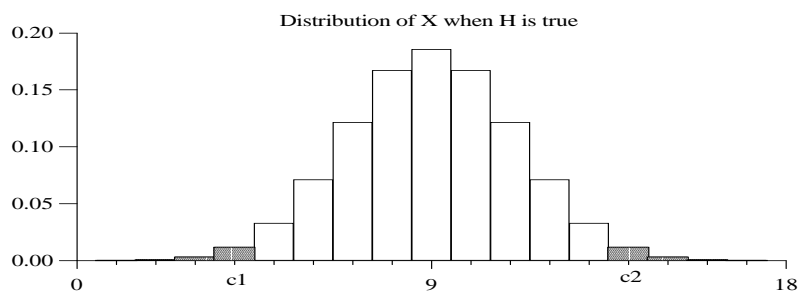
Consider now the alternative $H_1 : p \neq 0.5$. Values of x too large or too small would support this alternative. In this case there are two critical regions (or more correctly, *the* critical region consists of two disjoint sets), one in each 'tail' of the distribution of X . For a 5% critical region, there would be two critical values c_1 and c_2 such that

$$P(X \leq c_1 | H_0 \text{ is true}) \approx 0.025 \text{ and } P(X \geq c_2 | H_0 \text{ is true}) \approx 0.025.$$

This can be seen in Figure 2.4 below, where the graph is of the distribution of X when H_0 is true. (It can be shown that $c_1 = 4$ and $c_2 = 14$ are the critical values in this case.)

Tests with a one-sided critical region are called **one-tailed tests**, whereas those with a two-sided critical region are called **two-tailed tests**.

Figure 2.4: Critical Region – Twosided Alternative



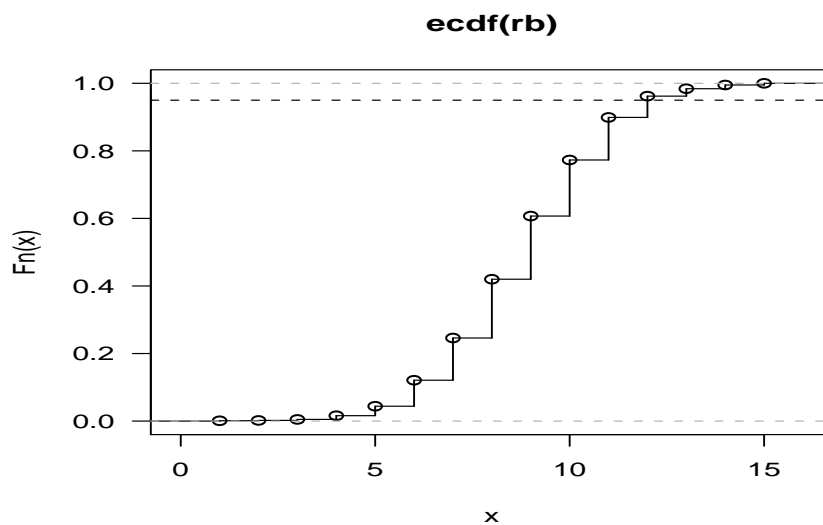
Computer Exercise 2.1 Use a simulation approach to estimate a value for c in (2.6) above.

Solution: Use the commands

```
#Generate 1000 random variables from a bin(18,0.5) distribution.
rb <- rbinom(n=1000,size=18,p=0.5)
table(rb) #Tabulate the results
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	3	11	28	77	125	174	187	166	126	63	22	11	5

Figure 2.5: Empirical cumulative distribution function for binomial rv's



This would indicate the onesided critical value should be $c = 13$ as the estimate of $P(X \geq 13)$ is 0.038. For a two sided test the estimated critical values are $c_1 = 4$ and $c_2 = 13$.

These results from simulation are in close agreement with theoretical results obtained in 2.4.1 and 2.4.2.

2.4.3 Two Approaches to Hypothesis Testing

It is worthwhile considering a definite procedure for hypothesis testing problems. There are two possible approaches.

- (i) See how the observed value of the statistic compares with that expected **if H_0 is true**. Find the probability, assuming H_0 to be true, of this event or others more extreme, that is, further still from the expected value. For a two-tailed test this will involve considering extreme values *in either direction*. If this probability is small (say, < 0.05), the event is an unlikely one **if H_0 is true**. So if such an event has occurred, doubt would be cast on the hypothesis.
- (ii) Make up a decision rule by partitioning the sample space of the statistic into a critical region, R , and its complement \bar{R} , choosing the critical value (or two critical values in the case of a two- tailed test) c , in such a way that $\alpha = 0.05$. We then note whether or not the observed value lies in this critical region, and draw the corresponding conclusion.

Example 2.3

Suppose we want to know whether a given die is biased towards 5 or 6 or whether it is “true”. To examine this problem the die is tossed 9000 times and it is observed that on 3164 occasions the outcome was 5 or 6.

Solution: Let X be the number of successes (5’s or 6’s) in 9000 trials. Then if $p = P(S)$, X is distributed $\text{bin}(9000, p)$. As is usual in hypothesis testing problems, we set up H_0 as the hypothesis we wish to “disprove”. In this case, it is that the die is “true”, that is, $p = 1/3$. If H_0 is not true, the alternative we wish to claim is that the die is biased towards 5 or 6, that is $p > 1/3$. In practice, one decides on this alternative before the experiment is carried out. We will consider the 2 approaches mentioned above.

Approach (i), probabilities

If $p = 1/3$ and $N = 9000$ then $E(X) = np = 3000$ and $\text{Var}(X) = npq = 2000$. The observed number of successes, 3164, was greater than expected if H_0 were true. So, assuming $p = 1/3$, the probability of the observed event together with others more extreme (that is, further still from expectation) is

$$P_B(X \geq 3164 | p = 1/3) = 0.0001 \quad (\text{pbinom}(q=3164, \text{size}=9000, \text{prob}=1/3, \text{lower.tail}=F)$$

This probability is small, so the event $X \geq 3164$ is an unlikely one if the assumption we’ve made ($p = 1/3$) is correct. Only about 1 times in 10000 would we expect such an

occurrence. Hence, if such an event did occur, we'd doubt the hypothesis and conclude that there is evidence that $p > 1/3$.

Approach (ii), quantiles

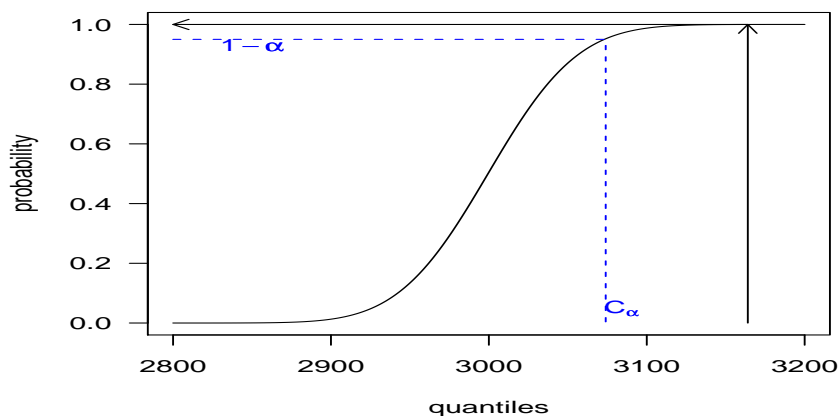
Clearly, large values of X support H_1 , so we'd want a critical region of the form $x \geq c$ where c is chosen to give the desired significance level, α . That is, for $\alpha = 0.05$, say, the upper tail 5% quantile of the binomial distribution with $p = \frac{1}{3}$ and $N = 9000$ is 3074. (`qbinom(size=N,prob=px,p=0.05,lower.tail=F)`)

The observed value 3164 exceeds this and thus lies in the critical region $[c, \infty]$. So we reject H_0 **at the 5% significance level**. That is, we will come to the conclusion that $p > 1/3$, but in so doing, we'll recognize the fact that the probability could be as large as 0.05 that we've rejected H_0 wrongly.

The 2 methods are really the same thing. Figure 2.6 shows the distribution function for $\text{Bin}(9000, \frac{1}{3})$ with the observed quantile 3164 and associated with it is $P(X > 3164)$. The dashed lines show the upper $\alpha = 0.05$ probability and the quantile $C_{1-\alpha}$. The event that $X > C_{1-\alpha}$ has a probability $p < \alpha$.

The rejection region can be defined either by the probabilities or the quantiles.

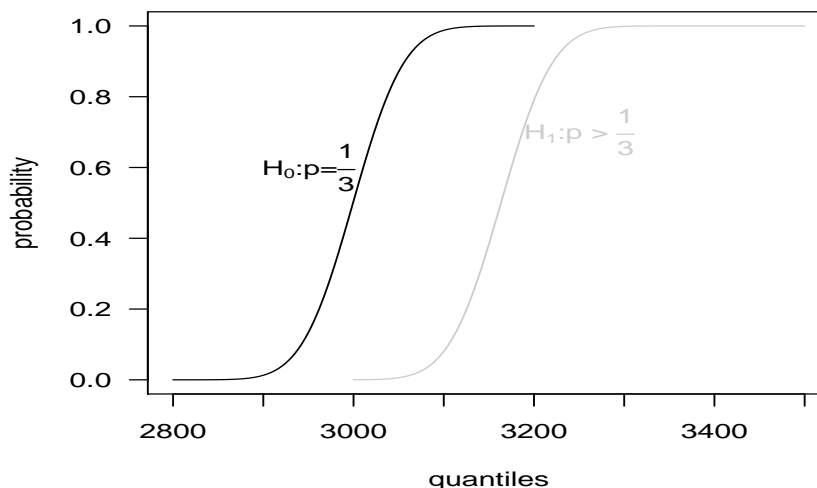
Figure 2.6: using either quantiles or probability to test the null hypothesis



In doing this sort of problem it helps to draw a diagram, or at least try to visualize the partitioning of the sample space as suggested in Figure 2.7.

If $x \in R$ it seems much more likely that the actual distribution of X is given by a curve similar to the one on the right hand side, with mean somewhat greater than 3000.

Figure 2.7: One Sided Alternative – Binomial.

**Computer Exercise 2.2**

The following random sample was drawn from a normal distribution with $\sigma = 5$. Test the hypothesis that $\mu = 23$.

18	14	23	23	18
21	22	16	21	28
12	19	22	15	18
28	24	22	18	13
18	16	24	26	35

Solution:

```
x <- c(18,14,23,23,18,21,22,16,21,28,12,19,22,15,18,28,24,22,18,13,18,16,24,26,35)
xbar <- mean(x)
n <- length(x)
> xbar
[1] 20.56
pnorm(q=xbar, mean=23,sd=5/sqrt(n))
[1] 0.007
qnorm(p=0.05,mean=23,sd=5/sqrt(n))
[1] 21
```

We can now use approach (i). For a two sided alternative calculated probability is $P = 0.015 (= 2 * 0.00734)$ so that the hypothesis is unlikely to be true.

For approach (ii) with $\alpha = 0.05$ the critical value is 21. The conclusion reached would therefore be the same by both approaches.

For testing $\mu = 23$ against the one sided alternative $\mu < 23$, $P = 0.0073$

2.5 Two-Sample Problems

In this section we will consider problems involving sampling from two populations where the hypothesis is a statement of equality of two parameters. The two problems are:

- (i) Test $H_0 : \mu_1 = \mu_2$ where μ_1 and μ_2 are the means of two normal populations.
- (ii) Test $H_0 : p_1 = p_2$ where p_1 and p_2 are the parameters of two binomial populations.

Example 2.4

Given independent random samples X_1, X_2, \dots, X_{n_1} from a normal population with unknown mean μ_1 and known variance σ_1^2 and Y_1, Y_2, \dots, Y_{n_2} from a normal population with unknown mean μ_2 and known variance σ_2^2 , derive a test for the hypothesis $H : \mu_1 = \mu_2$ against one-sided and two-sided alternatives.

Solution: Note that the hypothesis can be written as $H : \mu_1 - \mu_2 = 0$. An unbiased estimator of $\mu_1 - \mu_2$ is $\bar{X} - \bar{Y}$ so this will be used as the test statistic. Its distribution is given by

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

or, in standardized form, **if H_0 is true**

$$\frac{\bar{X} - \bar{Y}}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} \sim N(0, 1).$$

For a two-tailed test (corresponding to $H_1 : \mu_1 - \mu_2 \neq 0$) we have a rejection region of the form

$$\frac{|\bar{x} - \bar{y}|}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} > c \quad (2.7)$$

where $c = 1.96$ for $\alpha = .05$, $c = 2.58$ for $\alpha = .01$, etc.

For a one-tailed test we have a rejection region

$$\frac{\bar{x} - \bar{y}}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}} > c \quad \text{for } H_1 : \mu_1 - \mu_2 > 0 \quad (2.8)$$

$$< -c \quad \text{for } H_1 : \mu_1 - \mu_2 < 0 \quad (2.9)$$

where $c = 1.645$ for $\alpha = .05$, $c = 2.326$ for $\alpha = .01$, etc. Can you see what modification to make to the above rejection regions for testing $H_0 : \mu_1 - \mu_2 = \delta_0$, for some specified constant other than zero?

Example 2.5

Suppose that n_1 Bernoulli trials where $P(S) = p_1$ resulted in X successes and that n_2 Bernoulli trials where $P(S) = p_2$ resulted in Y successes. How do we test $H : p_1 = p_2$ ($= p$, say)?

Solution: Note that H_0 can be written $H_0 : p_1 - p_2 = 0$. Now X is distributed as $\text{bin}(n_1, p_1)$ and Y is distributed as $\text{bin}(n_2, p_2)$ and we have seen earlier than unbiased estimates of p_1, p_2 are respectively,

$$\tilde{p}_1 = x/n_1, \quad \tilde{p}_2 = y/n_2,$$

so an appropriate statistic to use to estimate $p_1 - p_2$ is $\frac{X}{n_1} - \frac{Y}{n_2}$.

For n_1, n_2 large, we can use the Central Limit Theorem to observe that

$$\frac{\frac{X}{n_1} - \frac{Y}{n_2} - E[\frac{X}{n_1} - \frac{Y}{n_2}]}{\sqrt{\text{Var}[\frac{X}{n_1} - \frac{Y}{n_2}]}} \sim \text{approximately } N(0, 1) \quad (2.10)$$

$$\begin{aligned} E\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) &= \frac{n_1 p_1}{n_1} - \frac{n_2 p_2}{n_2} = 0 \text{ under } H_0, \text{ and} \\ \text{Var}\left(\frac{X}{n_1} - \frac{Y}{n_2}\right) &= \frac{n_1 p_1 q_1}{n_1^2} + \frac{n_2 p_2 q_2}{n_2^2} = p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \text{ under } H_0 \end{aligned}$$

In (2.10) the variance is unknown, but we can replace it by an estimate and it remains to decide what is the best estimate to use. For the binomial distribution, the MLE of p is

$$\tilde{p} = \frac{X}{n} = \frac{\text{number of successes}}{\text{number of trials}}.$$

In our case, we have 2 binomial distributions with the same probability of success under H_0 , so intuitively it seems reasonable to “pool” the 2 samples so that we have $X + Y$ successes in $n_1 + n_2$ trials. So we will estimate p by

$$\tilde{p} = \frac{x + y}{n_1 + n_2}.$$

Using this in (2.10) we can say that to test $H_0 : p_1 = p_2$ against $H_1 : p_1 \neq p_2$ at the $100\alpha\%$ significance level, H_0 is rejected if

$$\frac{|(x/n_1) - (y/n_2)|}{\sqrt{\left(\frac{x+y}{n_1+n_2}\right) \left(1 - \frac{x+y}{n_1+n_2}\right) \left(\frac{n_1+n_2}{n_1 n_2}\right)}} > z_{\alpha/2}. \quad (2.11)$$

Of course the appropriate modification can be made for a one- sided alternative.

2.6 Connection between Hypothesis testing and CI's

Consider the problem where we have a sample of size n from a $N(\mu, \sigma^2)$ distribution where σ^2 is known and μ is unknown. An unbiased estimator of μ is $\bar{x} = \sum_{i=1}^n x_i/n$. We can use this information either

- (a) to test the hypothesis $H_0 : \mu = \mu_0$; or
- (b) to find a CI for μ and see if the value μ_0 is in it or not.

We will show that testing H_0 at the 5% significance level (that is, with $\alpha = .05$) against a 2-sided alternative is the same as finding out whether or not μ_0 lies in the 95% confidence interval.

- (a) For $H_1 : \mu \neq \mu_0$ we reject H_0 at the 5% significance level if

$$\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > 1.96 \quad \text{or} \quad \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -1.96. \quad (2.12)$$

That is, if

$$\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} > 1.96.$$

Or, using the “P-value”, if $\bar{x} > \mu_0$ we calculate the probability of a value as extreme or more extreme than this, in either direction. That is, calculate

$$P = 2 \times P(\bar{X} > \bar{x}) = 2 \times P_N \left(Z > \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right).$$

If $P < .05$ the result is significant at the 5% level. This will happen if $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -1.96$, as in (2.11).

- (b) A symmetric 95% confidence interval for μ is $\bar{x} \pm 1.96\sigma/\sqrt{n}$ which arose from considering the inequality

$$-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96$$

which is the event complementary to that in (2.11).

So, to reject H_0 at the 5% significance level is equivalent to saying that “the hypothesized value is not in the 95% CI”. Likewise, to reject H_0 at the 1% significance level is equivalent to saying that “the hypothesized value is not in the 99% CI”, which is equivalent to saying that “the P-value is less than 1%”.

If $1\% < P < 5\%$ the hypothesized value of μ will not be within the 95% CI but it will lie in the 99% CI.

This approach is illustrated for the hypothesis-testing situation and the confidence interval approach below.

Computer Exercise 2.3

Using the data in Computer Exercise 2.2, find a 99% CI for the true mean, μ .

Solution:

#Calculate the upper and lower limits for the 99% confidence interval.

```
CI <- qnorm(mean=xbar,sd=5/sqrt(25),p=c(0.005,0.995) )
```

```
> CI
```

```
[1] 18 23
```

So that the 99% CI is (18, 23).

Figure 2.8: Relationship between Non-significant Hypothesis Test and Confidence Interval

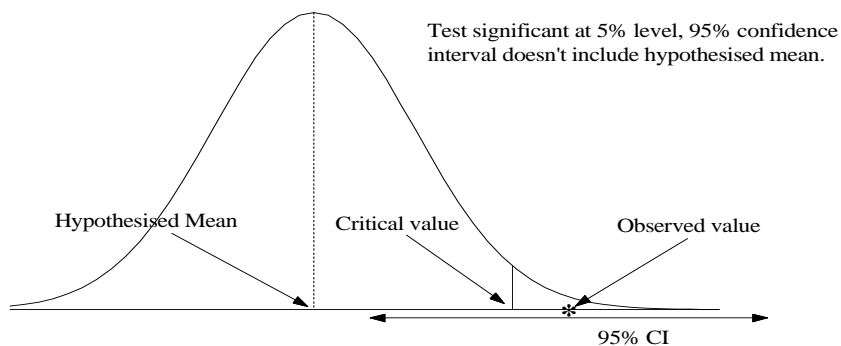
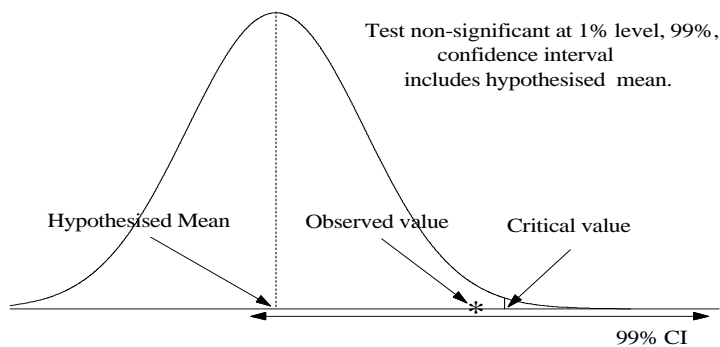


Figure 2.9: Relationship between Significant Hypothesis Test and Confidence Interval



2.7 Summary

We have only considered 4 hypothesis testing problems at this stage. Further problems will be dealt with in later chapters after more sampling distributions are introduced. The following might be helpful as a pattern to follow in doing examples in hypothesis testing.

1. State the hypothesis and the alternative. This must always be a statement about the unknown parameter in a distribution.
2. Select the appropriate *statistic* (function of the data). In the problems considered so far this is an unbiased estimate of the parameter or a function of it. State the distribution of the statistic and its particular form when H_0 is true.

Alternative Procedures

1. Find the critical region using the appropriate value of α (.05 or .01 usually).
 2. Find the observed value of the statistic (using the data).
 3. Draw conclusions. If the calculated value falls in the CR, this provides evidence against H_0 . You could say that the result is significant at the 5% (or 1% or .1% level).
1. Calculate P , the probability associated with values as extreme or more extreme than that observed. For a 2-sided H_1 , you'll need to double a probability such as $P(X \geq k)$.
 2. Calculate P , the probability associated with values as extreme or more extreme than that observed. For a 2-sided H_1 , you'll need to double a probability such as $P(X \geq k)$.
 3. Draw conclusions. For example, if $P < .1\%$ we say that there is *very strong* evidence against H_0 . If $.1\% < P < 1\%$ we say there is *strong* evidence. If $1\% < P < 5\%$ we say there is *some* evidence. For larger values of P we conclude that the event is not an unusual one **if H_0 is true**, and say that this set of data is consistent with H_0 .

2.8 Bayesian Hypothesis Testing

2.8.1 Notation

The notation that we have used in classical hypothesis testing is:-

$$\omega \cup \bar{\omega} = \Omega$$

$$H_0 : \theta \in \omega$$

$$H_1 : \theta \in \bar{\omega}$$

There is a set of observations x_1, x_2, \dots, x_n whose density is $p(x|\theta)$.

A test is decided by the rejection region

$$\mathbf{R} = \{\mathbf{x} \mid \text{observing } \mathbf{x} \text{ would lead to the rejection of } H_0\}$$

Decisions between tests are based on

$$\text{Size } \alpha = P(\mathbf{R}|\theta) \text{ for } \theta \in \omega \quad \text{Type I error}$$

$$\text{Power } \beta = 1 - P(\mathbf{R}|\theta) \text{ for } \theta \in \bar{\omega} \quad \text{Type II error}$$

The smaller the Type I error the larger the Type II error and vice versa. The rejection region is usually chosen as a balance between the 2 types of error.

2.8.2 Bayesian approach

We calculate the posterior probabilities,

$$p_0 = P(\theta \in \omega | \mathbf{x})$$

$$p_1 = P(\theta \in \bar{\omega} | \mathbf{x})$$

and decide between H_0 and H_1 using p_0 and p_1 .

Since $\omega \cup \bar{\omega} = \Omega$ and $\omega \cap \bar{\omega} = \Phi$, then $p_0 + p_1 = 1$.

We require prior probabilities

$$\pi_0 = P(\theta \in \omega)$$

$$\pi_1 = P(\theta \in \bar{\omega})$$

Thus $\pi_0 + \pi_1 = 1$.

The *prior odds* on H_0 against H_1 is $\frac{\pi_0}{\pi_1}$ and the *posterior odds* on H_0 against H_1 is $\frac{p_0}{p_1}$.

- If the prior odds is close to 1, then H_0 is approximately equally likely as H_1 *a priori*.
- If the prior odds ratio is large, H_0 is relatively likely.
- If the prior odds ratio is small, H_0 is relatively unlikely.
- The same remarks apply to the posterior odds.

Bayes factor

The Bayes factor, \mathbf{B} , is the odds in favour of H_0 against H_1 ,

$$\mathbf{B} = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{p_0\pi_1}{p_1\pi_0} \quad (2.13)$$

The posterior probability p_0 of H_0 can be calculated from its prior probability and the Bayes factor,

$$p_0 = \frac{1}{[1 + (\pi_1/\pi_0)\mathbf{B}^{-1}]} = \frac{1}{[1 + \{(1 - \pi_0)/\pi_0\} \mathbf{B}^{-1}]}$$

Simple Hypotheses

$$\begin{aligned} \omega &= \{\theta_0\} & \bar{\omega} &= \{\theta_1\} \\ p_0 &\propto \pi_0 p(\mathbf{x}|\theta_0) & p_1 &\propto \pi_1 p(\mathbf{x}|\theta_1) \end{aligned}$$

$$\begin{aligned} \frac{p_0}{p_1} &= \frac{\pi_0 p(\mathbf{x}|\theta_0)}{\pi_1 p(\mathbf{x}|\theta_1)} \\ \mathbf{B} &= \frac{p(\mathbf{x}|\theta_0)}{p(\mathbf{x}|\theta_1)} \end{aligned}$$

Example 2.6

Consider the following prior distribution, density and null hypothesis,

$$\begin{aligned} \mu &\sim N(82.4, 1.1^2) \\ x|\mu &\sim N(82.1, 1.7^2) \\ H_0 : x &< 83.0 \end{aligned}$$

From the results in section 1.9,

$$\begin{aligned} \tau_0 &= \frac{1}{\sigma_0^2} = \frac{1}{1.1^2} = 0.83 \\ \tau &= \frac{1}{\sigma^2} = \frac{1}{1.7^2} = 0.35 \\ \tau_1 &= \tau_0 + \tau = 1.18 \quad \Rightarrow \sigma_1^2 = (1.18)^{-1} = 0.85 \\ \mu_1 &= \left(\mu_0 \times \frac{\tau_0}{\tau_1} \right) + \left(\mu \times \frac{\tau}{\tau_1} \right) \\ &= \left(82.4 \times \frac{0.83}{1.18} \right) + \left(82.1 \times \frac{0.35}{1.18} \right) = 82.3 \end{aligned}$$

For $H_0 : x < 83$, and with π_0, p_0 being the prior and posterior probabilities under H_0 ,

$$\begin{aligned}\pi_0 &= P(x < 83 | \mu_0 = 82.4, \sigma_0 = 1.1) = 0.71 && \text{Use } \texttt{pnorm}(\texttt{mean}=82.4, \texttt{sd}=1.1, \texttt{q}=83) \\ \frac{\pi_0}{1 - \pi_0} &= \frac{0.71}{0.29} = 2.45 \\ p_0 &= P(x < 83 | \mu_1 = 82.3, \sigma_1 = \sqrt{0.85}) = 0.77 \\ \frac{p_0}{1 - p_0} &= \frac{0.77}{0.23} = 3.35\end{aligned}$$

The Bayes factor is

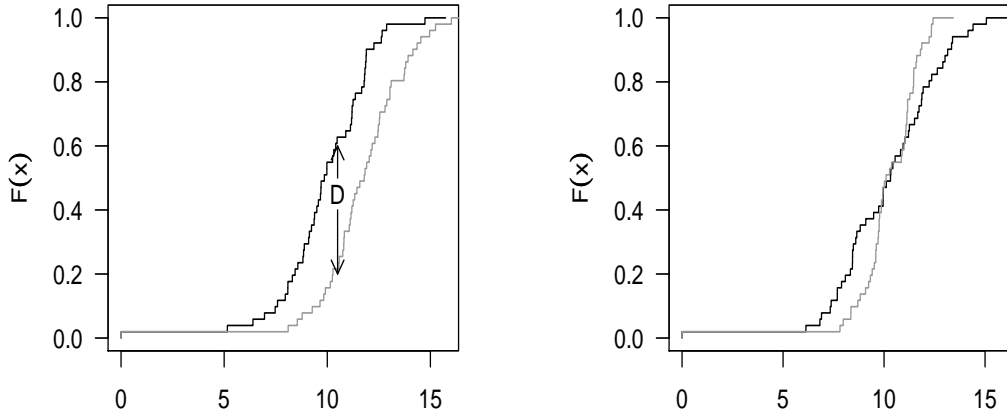
$$\mathbf{B} = \frac{p_0 \pi_1}{p_1 \pi_0} = \frac{3.35}{2.45} = 1.4$$

The data has not altered the prior beliefs about the mean, $\mathbf{B} \approx 1$.

2.9 Non-Parametric Hypothesis testing.

Figure 2.10 shows 2 ways in which distributions differ. The difference depicted in Figure 2.10 (a) is a shift in location (mean) and in Figure 2.10 (b) there is a shift in the scale (variance).

Figure 2.10: Distributions that differ due to shifts in (a) location and (b) scale.



2.9.1 Kolmogorov-Smirnov (KS)

The KS test is a test of whether 2 independent samples have been drawn from the same population or from populations with the same distribution. It is concerned with the agreement between 2 cumulative distribution functions. If the 2 samples have been drawn from the same population, then the cdf's can be expected to be close to each other and only differ by random deviations. If they are too far apart at any point, this suggests that the samples come from different populations.

The KS test statistics is

$$D = \max \left(|\hat{F}_1(x) - \hat{F}_2(y)| \right) \quad (2.14)$$

Exact sampling distribution

The exact sampling distribution of D under $H_0 : F_1 = F_2$ can be enumerated.

If H_0 is true, then $[(X_1, X_2, \dots, X_m), (Y_1, Y_2, \dots, Y_n)]$ can be regarded as a random sample from the same population with actual realised samples

$$[(x_1, x_2, \dots, x_m), (y_1, y_2, \dots, y_n)]$$

Thus (under H_0) an equally likely sample would be

$$[(y_1, x_2, \dots, x_m), (x_1, y_2, \dots, y_n)]$$

where x_1 and y_1 were swapped.

There are $\binom{m+n}{m}$ possible realisations of allocating the combined sample to 2 groups of sizes m and n and under H_0 the probability of each realisation is $\frac{1}{\binom{m+n}{m}}$. For each sample generated this way, a D^* is observed.

Now $F_1(x)$ is steps of $\frac{1}{m+1}$ and $F_2(y)$ is steps of $\frac{1}{n+1}$ so for given m and n , it would be possible to enumerate all $D_{m,n}^*$ if H_0 is true. From this enumeration the upper $100\alpha\%$ point of $\{D_{m,n}^*\}$, $\{D_{m,n}; \alpha\}$, gives the critical value for the α sized test. If the observed $\{D_{m,n}\}$ is greater than $\{D_{m,n}; \alpha\}$, reject H_0 .

2.9.2 Asymptotic distribution

If m and n become even moderately large, the enumeration is huge. In that case we can utilize the large sample approximation that

$$\chi^2 = \frac{4D^2(nm)}{n+m}$$

This was shown to be so by Goodman in 1954, *Psychological Bulletin* **51** 160-168.

Example 2.7

These data are the energies of sway signals from 2 groups of subjects, Normal group and Whiplash group. Whiplash injuries can lead to unsteadiness and the subject may not be able to maintain balance. each subject had their sway pattern measured by standing on a plate blindfolded. Does the distribution of energies differ between groups?

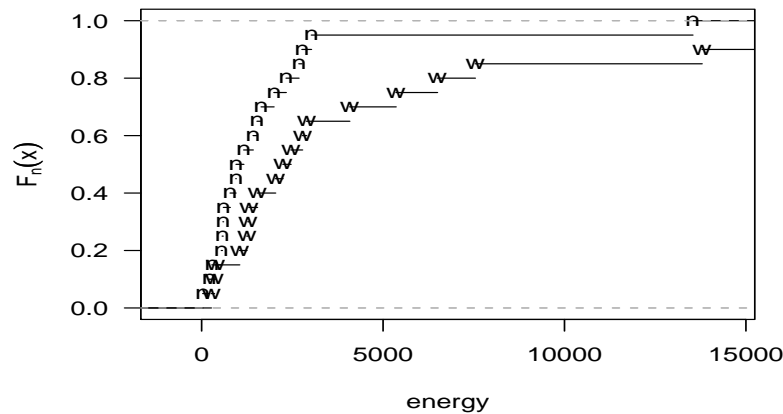
Table 2.1: Wavelet energies of the sway signals from normal subjects and subjects with whiplash injury.

Normal	33	211	284	545	570	591	602	786	945	951
	1161	1420	1529	1642	1994	2329	2682	2766	3025	13537
Whipl	269	352	386	1048	1247	1276	1305	1538	2037	2241
	2462	2780	2890	4081	5358	6498	7542	13791	23862	34734

The plots of the ecdf suggest a difference.

We apply the Kolmogorov-Smirnov test to these data.

Figure 2.11: The ecdf's of sway signal energies for N & W groups



```

N.energy <- c(33,211,284,545,570,591,602,786,945,951,1161,1420,
             1529,1642,1994,2329,2682,2766,3025,13537)
W.energy <- c(269,352,386,1048,1247,1276,1305,1538,2037,2241,2462,2780,
             2890,4081,5358,6498,754,1379,23862,34734)
KS <- ks.test(N.energy,W.energy,alternative="greater")
> KS

```

Two-sample Kolmogorov-Smirnov test

```

data: N.energy and W.energy
D+ = 0.35, p-value = 0.0863
alternative hypothesis: the CDF of x lies above that of y

```

```

# ----- the Asymptotic distribution -----
D <- KS$statistic
Chi <- 4*(KS$statistic^2)*m*n/(m+n)
P <- pchisq(q=Chi,df=2,lower.tail=F)

> cat("X2 = ",round(Chi,2),"P( > X2) = ",P,"\n")
X2 = 4.9 P( > X2) = 0.08629

```

1. The Kolmogorov-Smirnov test of whether the null hypothesis can be rejected is a *permutation test*.
2. The equality $F_1 = F_2$ means that F_1 and F_2 assign equal probabilities to all sets; $P_{F_1}(A) = P_{F_2}(A)$ for and A subset of the common sample space of x and y . If H_0 is true, there is no difference between the randomness of x or y .
3. The null hypothesis is set up to be rejected. If however, the data are such that the null hypothesis cannot be decisively rejected, then the experiment has not demonstrated a difference.

4. A hypothesis test requires a statistic, $\hat{\theta}$, for comparing the distributions. In the Kolmogorov-Smirnov test $\hat{\theta} = D$.
5. Having observed $\hat{\theta}$, the achieved significance level of the test is the probability of observing at least as large a value when H_0 is true, $P_{H_0}(\hat{\theta}^* \geq \hat{\theta})$. The observed statistic, $\hat{\theta}$ is fixed and the random variable $\hat{\theta}^*$ is distributed according to H_0 .
6. The KS test enumerated all permutations of elements in the samples. This is also termed *sampling without replacement*. Not all permutations are necessary but an accurate test does require a large number of permutations.
7. The permutation test applies to any test statistic. For the example in Figure 2.10(b), we might use $\hat{\theta} = \frac{\hat{\sigma}_x^2}{\hat{\sigma}_y^2}$.

2.9.3 Bootstrap Hypothesis Tests

The link between confidence intervals and hypothesis tests also holds in a bootstrap setting. The bootstrap is an approximation to a permutation test and a strategic difference is that bootstrap uses *sampling with replacement*.

A permutation test of whether $H_0 : F_1(x) = F_2(y)$ is true relies upon the ranking of the combined data set (\mathbf{x}, \mathbf{y}) . The data were ordered smallest to largest and each permutation was an allocation of the group labels to each ordered datum. In 1 permutation, the label x was ascribed to the first number and in another, the label y is given to that number and so on.

The test statistic can be a function of the data (it need not be an estimate of a parameter) and so denote this a $t(\mathbf{z})$.

The principle of bootstrap hypothesis testing is that if H_0 is true, a probability atom of $\frac{1}{m+n}$ can be attributed to each member of the combined data $\mathbf{z} = (\mathbf{x}, \mathbf{y})$.

The empirical distribution function of $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, call it $\hat{F}_0(z)$, is a non-parametric estimate of the common population that gave rise to \mathbf{x} and \mathbf{y} , assuming that H_0 is true.

Bootstrap hypothesis testing of H_0 takes these steps,

1. Get the observed value of t , e.g. $t_{\text{obs}} = \bar{x} - \bar{y}$.
2. Nominate how many bootstrap samples (replications) will be done, e.g. $B = 499$.
3. For b in $1:B$, draw samples of size $m + n$ *with replacement* from \mathbf{z} . Label the first m of these x_b^* and the remaining n be labelled y_b^* .
4. Calculate $t(z_b^*)$ for each sample. For example, $t(z_b^*) = \bar{x}_b^* - \bar{y}_b^*$
5. Approximate the probability of t_{obs} or greater by $\frac{\text{number of } t(\mathbf{z}_b^*) \geq t_{\text{obs}}}{B}$

Example

The data in Table 2.1 are used to demonstrate bootstrap hypothesis testing with the test statistic,

$$t(\mathbf{z}) = \frac{\bar{y} - \bar{x}}{\hat{\sigma} \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

The R code is written to show the required calculations more explicitly but a good program minimises the variables which are saved in the iterations loop.

```
#----- Bootstrap Hypothesis Test -----
N.energy <- c(33,211,284,545,570,591,602,786,945,951,1161,1420,
             1529,1642,1994,2329,2682,2766,3025,13537)
W.energy <- c(269,352,386,1048,1247,1276,1305,1538,2037,2241,2462,2780,
             2890,4081,5358,6498,754,1379,23862,34734)
Z <- c(N.energy,W.energy)
m <- length(N.energy)
n <- length(W.energy)
T.obs <- (mean(W.energy) - mean(N.energy))/(sd(Z)*sqrt(1/m + 1/n))

nBS <- 999

T.star <- numeric(nBS)

for (j in 1:nBS){
  z.star <- sample(Z,size=(m+n))
  w.star <- z.star[(m+1):(m+n)]
  n.star <- z.star[1:m]
  T.star[j] <- ( mean(w.star) - mean(n.star) )/( sd(z.star) * sqrt(1/m + 1/n) )
}
p1 <- sum(T.star >= T.obs)/nBS

cat( "P(T > ",round(T.obs,1),"|H0) = ",round(p1,2),"\\n")
```

The results are:-

$$T = 1.4$$

$$P(t > 1.4|H_0) = 0.09$$

Thus this statistic does not provide evidence that the 2 distributions are different.

Chapter 3 Chi-square Distribution

3.1 Distribution of S^2

Recall that if X_1, X_2, \dots, X_n is a random sample from a $N(\mu, \sigma^2)$ distribution then

$$S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$$

is an unbiased estimator of σ^2 . We will find the probability distribution of this random variable. Firstly note that the numerator of S^2 is a sum of n squares but they are not independent as each involves \bar{X} . This sum of squares can be rewritten as the sum of squares of $n - 1$ *independent* variables by the method which is illustrated below for the cases $n = 2, 3, 4$.

For $n = 2$,

$$\sum_{i=1}^2 (X_i - \bar{X})^2 = Y_1^2 \text{ where } Y_1 = (X_1 - X_2)/\sqrt{2};$$

for $n = 3$,

$$\sum_{i=1}^3 (X_i - \bar{X})^2 = \sum_{j=1}^2 Y_j^2 \text{ where } Y_1 = (X_1 - X_2)/\sqrt{2}, Y_2 = (X_1 + X_2 - 2X_3)/\sqrt{6};$$

for $n = 4$,

$$\sum_{i=1}^4 (X_i - \bar{X})^2 = \sum_{j=1}^3 Y_j^2 \text{ where } Y_1, Y_2 \text{ are as defined above and}$$

$$Y_3 = (X_1 + X_2 + X_3 - 3X_4)/\sqrt{12}.$$

Note that Y_1, Y_2, Y_3 are linear functions of X_1, X_2, X_3, X_4 which are mutually orthogonal with the sum of the squares of their coefficients equal to 1.

Consider now the properties of the X_i and the Y_j as random variables. Since Y_1, Y_2, Y_3 are mutually orthogonal linear functions of X_1, X_2, X_3, X_4 they are uncorrelated, and

since they are normally distributed (being sums of normal random variables), they are independent. Also,

$$E(Y_1) = 0 = E(Y_2) = E(Y_3)$$

and,

$$\text{Var}(Y_1) = \frac{1}{2} (\text{Var}(X_1) + \text{Var}(X_2)) = \sigma^2$$

$$\text{Var}(Y_2) = \frac{1}{6} \text{Var}(X_1) + \frac{1}{6} \text{Var}(X_2) + \frac{4}{6} \text{Var}(X_3) = \sigma^2.$$

Similarly, $\text{Var}(Y_3) = \sigma^2$.

In general the sum of n squares involving the X 's can be expressed as the sum of $n - 1$ squares involving the Y 's. Thus $\sum_{i=1}^n (X_i - \bar{X})^2$ can be expressed as

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{j=1}^{n-1} Y_j^2 = \sum_{j=1}^{\nu} Y_j^2$$

where $\nu = n - 1$ is called the number of **degrees of freedom** and

$$Y_j = \frac{X_1 + X_2 + \cdots + X_j - jX_{j+1}}{\sqrt{j(j+1)}}, \quad j = 1, 2, \dots, n-1.$$

The random variables Y_1, Y_2, \dots, Y_ν each have mean zero and variance σ^2 . So each $Y_j \sim N(0, \sigma^2)$ and the Y_j 's are independent.

Now write $S^2 = \frac{\sum_{j=1}^{\nu} Y_j^2}{\nu}$ and recall that

(i) If $X \sim N(\mu, \sigma^2)$ then $\frac{(X - \mu)^2}{2\sigma^2} \sim \text{Gamma}\left(\frac{1}{2}\right)$, [Statistics 260, (8.16)]

(ii) If X_1, X_2, \dots, X_ν are independent $N(\mu, \sigma^2)$ variates, then $\frac{\sum_{j=1}^{\nu} (X_j - \mu)^2}{2\sigma^2}$ is distributed as $\text{Gamma}\left(\frac{\nu}{2}\right)$ [Statistics 260, section 7.4].

Applying this to the Y_j where $\mu = 0$, $\frac{Y_j^2}{2\sigma^2} \sim \text{Gamma}\left(\frac{1}{2}\right)$ and

$$V = \frac{1}{2} \sum_{j=1}^{\nu} \frac{Y_j^2}{\sigma^2} \text{ is distributed as } \text{Gamma}\left(\frac{\nu}{2}\right). \quad (3.1)$$

Thus the pdf of V is given by

$$f(v) = \frac{1}{\Gamma(\frac{\nu}{2})} v^{(\frac{\nu}{2}-1)} e^{-v}, \quad v \in (0, \infty)$$

with V and S^2 being related by

$$S^2 = \frac{\sum_{j=1}^{\nu} Y_j^2}{\nu} = \frac{2\sigma^2 V}{\nu}$$

or

$$V = \frac{\nu}{2\sigma^2} S^2 \quad (3.2)$$

Now V is a strictly monotone function of S^2 so, by the change-of-variable technique, the pdf of S^2 is

$$\begin{aligned} g(s^2) &= f(v)|dv/ds^2| \\ &= \frac{e^{-\nu s^2/2\sigma^2}}{\Gamma(\nu/2)} \left(\frac{\nu s^2}{2\sigma^2}\right)^{(\nu/2)-1} \cdot \left(\frac{\nu}{2\sigma^2}\right), \quad s^2 \in (0, \infty) \\ &= \frac{1}{\Gamma(\frac{\nu}{2})} \times (s^2)^{(\frac{\nu}{2}-1)} \left(\frac{\nu}{2\sigma^2}\right)^{\nu/2} \exp\left\{-\frac{\nu}{2\sigma^2} s^2\right\} \end{aligned} \quad (3.3)$$

This is the pdf of S^2 derived from a $N(\mu, \sigma^2)$ distribution.

3.2 Chi-Square Distribution

Define the random variable W as

$$W = \nu S^2 / \sigma^2 = 2V,$$

where V is defined in (3.2). Note that W is a “sum of squares” divided by σ^2 , and can be thought of as a standardized sum of squares. Then the p.d.f. of W is

$$\begin{aligned} h(w) &= g(s^2) \left| \frac{ds^2}{dw} \right|, \quad \text{where } \frac{ds^2}{dw} = \frac{\sigma^2}{\nu} \\ &= \frac{e^{-w/2} w^{(\nu/2)-1}}{2^{\nu/2} \Gamma(\nu/2)}, \quad w \in [0, \infty]. \end{aligned} \quad (3.4)$$

A random variable W with this pdf is said to have a **chi-square distribution on ν degrees of freedom** (or with parameter ν) and we write $W \sim \chi_{\nu}^2$.

Notes: (a) $W/2 \sim \gamma(\nu/2)$.

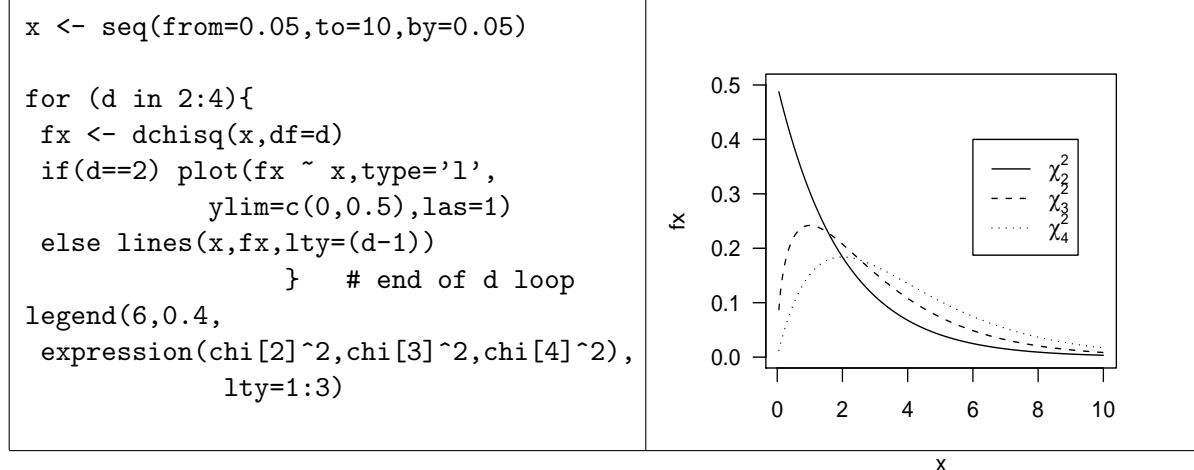
(b) This distribution can be thought of as a special case of the generalized gamma distribution.

(c) When $\nu = 2$, (3.4) becomes $h(w) = \frac{1}{2}e^{-w/2}$, $w \in [0, \infty]$, which is the exponential distribution.

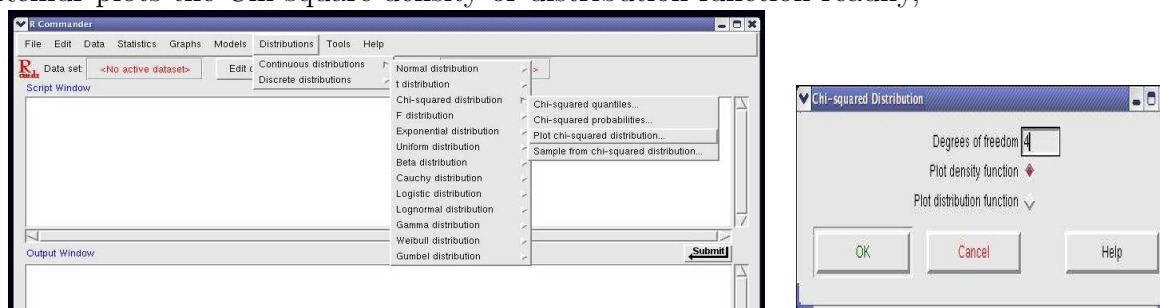
Computer Exercise 3.1

Graph the chi-square distributions with 2, 3 and 4 degrees of freedom for $x = 0.05, 0.1, 0.15, \dots, 10$, using one set of axes.

Solution:



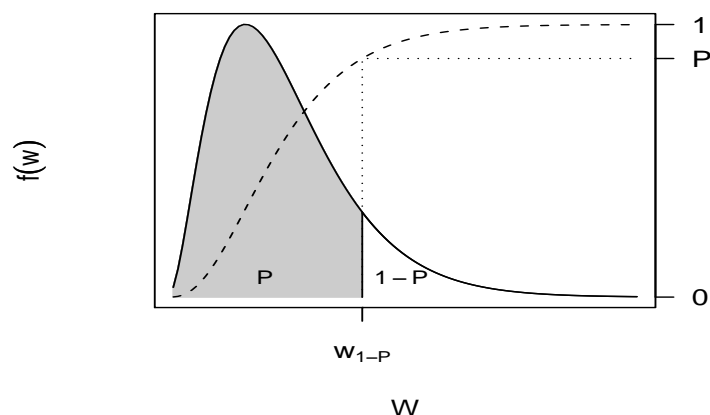
Rcmdr plots the Chi-square density or distribution function readily,

**Cumulative Distribution Function**

If $W \sim \chi^2_\nu$, percentiles (i.e 100P) of the chi-square distribution are determined by the inverse of the function

$$\frac{P}{100} = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^{w_{1-.01P}} w^{\frac{1}{2}\nu-1} e^{-w/2} dw = P(W \leq w_{1-.01P}).$$

Figure 3.1 depicts the tail areas corresponding to P (lower tail) and $1 - P$ (upper tail) for the density function and superimposed is the distribution function. The scales for the Y-axes of the density function (left side) and the distribution function (right side) are different.

Figure 3.1: Area corresponding to the $100P$ percentile of the χ^2 random variable w .

The R function for calculating tail area probabilities for given quantiles is
`pchisq(q= , df = , lower.tail= T (or F))`
 and for calculating quantiles corresponding to a probability, `qchisq(p = , df =)`

These functions are included in the Rcmdr menus.

The following example requires us to find a probability.

Example 3.1

A random sample of size 6 is drawn from a $N(\mu, 12)$ distribution.
 Find $P(2.76 < S^2 < 22.2)$.

Solution:

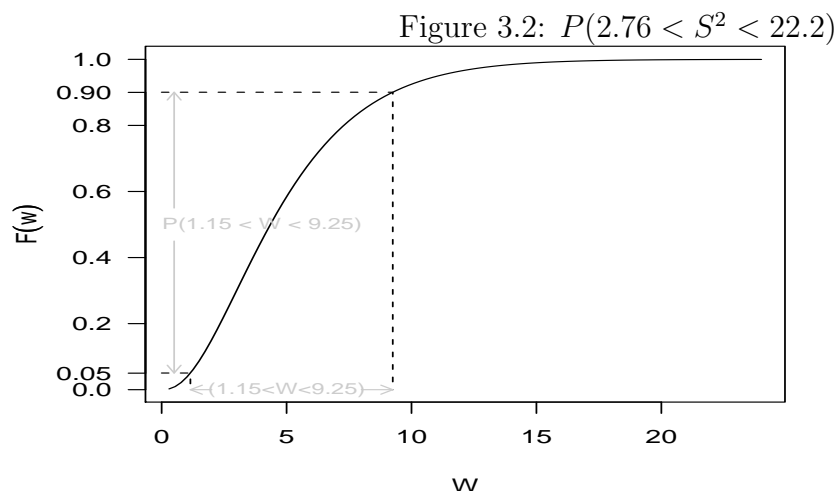
We wish to express this as a probability statement about the random variable W . That is,

$$\begin{aligned}
 P(2.76 < S^2 < 22.2) &= P\left(\frac{5}{12} \times 2.76 < \frac{\nu S^2}{\sigma^2} < \frac{5}{12} \times 22.2\right) \\
 &= P(1.15 < W < 9.25) \text{ where } W \sim \chi_5^2 \\
 &= P(W < 9.25) - P(W < 1.15)
 \end{aligned}$$

Solution:

```
#___ Pint.R _____
Q <- c(2.76,22.2)*5/12
Pint <- diff( pchisq(q=Q,df=5))
cat("P(2.76 < S2 < 22.2) = ",Pint,"\n")
```

```
> source("Pint.R")
P(2.76 < S2 < 22.2) = 0.85
```



Moments

As V (defined in (3.2)) has a gamma distribution it's mean and variance can be written down. That is, $V \sim \gamma(\nu/2)$, so that

$$E(V) = \nu/2 \text{ and } \text{Var}(V) = \nu/2.$$

Then since W is related to V by $W = 2V$

$$\begin{aligned} E(W) &= 2(\nu/2) = \nu \\ \text{Var}(W) &= 4(\nu/2) = 2\nu. \end{aligned} \tag{3.5}$$

Thus, a random variable $W \sim \chi_\nu^2$ has mean ν and variance 2ν .

Exercise: Find $E(W)$ and $\text{Var}(W)$ directly from $h(w)$.

Moment Generating Function

The MGF of a chi-square variate can be deduced from that of a gamma variate. Let $V \sim \gamma(\nu/2)$ and let $W = 2V$. We know $M_V(t) = (1 - t)^{-\nu/2}$ from Statistics 260, Theorem 4.4. Hence

$$M_W(t) = M_{2V}(t) = M_V(2t) = (1 - 2t)^{-\nu/2}.$$

So if $W \sim \chi_\nu^2$ then

$$M_W(t) = (1 - 2t)^{-\nu/2}. \tag{3.6}$$

Exercise: Find the MGF of W directly from the pdf of W . (Hint: Use the substitution $u = w(1 - 2t)/2$ when integrating.)

To find moments, we will use the power series expansion of $M_W(t)$.

$$\begin{aligned} M_W(t) &= 1 + \frac{\nu}{2} \cdot 2t + \frac{\nu}{2} \left(\frac{\nu}{2} + 1 \right) \frac{(2t)^2}{2!} + \frac{\nu}{2} \left(\frac{\nu}{2} + 1 \right) \left(\frac{\nu}{2} + 2 \right) \frac{(2t)^3}{3!} + \dots \\ &= 1 + \nu t + \nu(\nu + 2) \frac{t^2}{2!} + \nu(\nu + 2)(\nu + 4) \frac{t^3}{3!} + \dots \end{aligned}$$

Moments can be read off as appropriate coefficients here. Note that $\mu'_1 = \nu$ and $\mu'_2 = \nu(\nu + 2)$. The cumulant generating function is

$$\begin{aligned} K_W(t) &= \log M_W(t) = -\frac{\nu}{2} \log(1 - 2t) \\ &= -\frac{\nu}{2} \left[-2t - \frac{2^2 t^2}{2} - \frac{2^3 t^3}{3} - \frac{2^4 t^4}{4} - \dots \right] \\ &= \nu t + \frac{2\nu t^2}{2!} + \frac{8\nu t^3}{3!} + \frac{48\nu t^4}{4!} + \dots \end{aligned}$$

so the cumulants are

$$\kappa_1 = \nu, \kappa_2 = 2\nu, \kappa_3 = 8\nu, \kappa_4 = 48\nu.$$

We will now use these cumulants to find measures of skewness and kurtosis for the chi-square distribution.

Comparison with Normal

(i) Coefficient of skewness,

$$\begin{aligned} \gamma_1 = \kappa_3 / \kappa_2^{3/2} &= \frac{8\nu}{2\nu\sqrt{2\nu}} \text{ for the } \chi_\nu^2 \text{ distribution} \\ &\rightarrow 0 \text{ as } \nu \rightarrow \infty \end{aligned}$$

That is, the χ^2 distribution becomes symmetric for $\nu \rightarrow \infty$.

(ii) Coefficient of kurtosis,

$$\begin{aligned} \gamma_2 &= \kappa_4 / \kappa_2^2 \text{ for any distribution} \\ &= \frac{48\nu}{4\nu^2} \text{ for the } \chi^2 \text{ distribution} \\ &\rightarrow 0 \text{ as } \nu \rightarrow \infty. \end{aligned}$$

This is the value γ_2 has for the normal distribution.

Additive Property

Let $W_1 \sim \chi_{\nu_1}^2$ and W_2 (independent of W_1) $\sim \chi_{\nu_2}^2$. Then from (3.6) $W_1 + W_2$ has moment generating function

$$\begin{aligned} M_{W_1+W_2}(t) &= M_{W_1}(t)M_{W_2}(t) = (1-2t)^{-\nu_1/2}(1-2t)^{-\nu_2/2} \\ &= (1-2t)^{-(\nu_1+\nu_2)/2} \end{aligned}$$

This is also of the form (3.6); that is, we recognize it as the MGF of a χ^2 random variable on $(\nu_1 + \nu_2)$ degrees of freedom.

Thus if $W_1 \sim \chi_{\nu_1}^2$ and $W_2 \sim \chi_{\nu_2}^2$ and W_1 and W_2 are independent then

$$W_1 + W_2 \sim \chi_{\nu_1+\nu_2}^2$$

The result can be extended to the sum of k independent χ^2 random variables.

$$\text{If } W_1, \dots, W_k \text{ are independent } \chi_{\nu_1}^2, \dots, \chi_{\nu_k}^2 \text{ then } \sum_{i=1}^k W_i \sim \chi_{\nu}^2 \quad (3.7)$$

where $\nu = \sum \nu_i$. Note also that a χ_{ν}^2 variate can be decomposed into a sum of ν independent chi-squares each on 1 d.f.

Chi-square on 1 degree of freedom

For the special case $\nu = 1$, note that from (3.1) if $Y \sim N(0, \sigma^2)$ then $V = \frac{Y^2}{2\sigma^2} \sim \gamma(1/2)$ and $W = 2V = Y^2/\sigma^2 \sim \chi_1^2$.

Thus, if $Z = Y/\sigma$, it follows $Z \sim N(0, 1)$ and

$$Z^2 \sim \chi_1^2. \quad (3.8)$$

(The square of a $N(0, 1)$ random variable has a chi-square distribution on 1 df.)

Summary

You may find the following summary of relationships between χ^2 , gamma, S^2 and normal distributions useful.

Define $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)$, the X_i being independent $N(\mu, \sigma^2)$ variates, then

- (i) $W = \nu S^2 / \sigma^2 \sim \chi_{\nu}^2$ where $\nu = n-1$,
- (ii) $\frac{1}{2}W = \nu S^2 / 2\sigma^2 \sim \gamma(\nu/2)$,
- (iii) If $Z_i = \frac{X_i - \mu}{\sigma}$, (that is, $Z_i \sim N(0, 1)$) then

$$Z_i^2 \sim \chi_1^2 \text{ and } Z_1^2 + Z_2^2 + \dots + Z_k^2 \sim \chi_k^2.$$

3.3 Independence of \bar{X} and S^2

When \bar{X} and S^2 are defined for a sample from a normal distribution, \bar{X} and S^2 are statistically independent. This may seem surprising as the expression for S^2 involves \bar{X} .

Consider again the transformation from X 's to Y 's given in **3.1**. We've seen that $(n-1)S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$ can be expressed as $\sum_{j=1}^{\nu} Y_j^2$ where the Y_j defined by

$$Y_j = \frac{X_1 + X_2 + \cdots + X_j - jX_{j+1}}{\sqrt{j(j+1)}}, \quad j = 1, 2, \dots, n-1,$$

have zero means and variances σ^2 . Note also that the sample mean,

$$\bar{X} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

is a linear function of X_1, \dots, X_n which is orthogonal to each of the Y_j , and hence uncorrelated with each Y_j . Since the X_i are normally distributed, \bar{X} is thus independent of each of the Y_j and therefore independent of any function of them.

Thus when X_1, \dots, X_n are normally and independently distributed random variables \bar{X} and S^2 are statistically independent.

3.4 Confidence Intervals for σ^2

We will use the method indicated in **1.8** to find a confidence interval for σ^2 in a normal distribution, based on a sample of size n . The two cases (i) μ unknown; (ii) μ known must be considered separately.

Case (i)

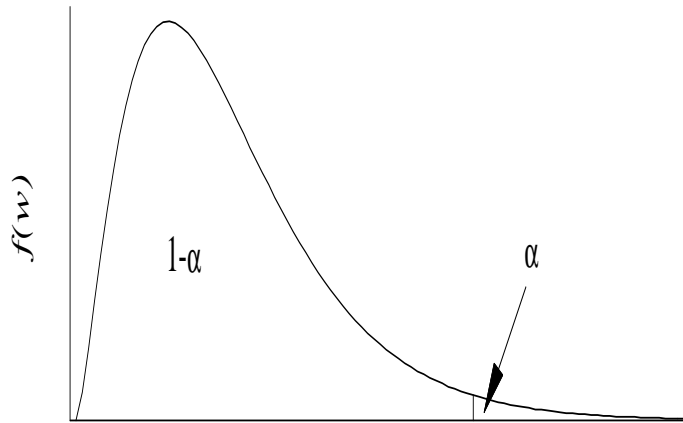
Let X_1, X_2, \dots, X_n be a random sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. It has been shown that S^2 is an unbiased estimate of σ^2 (Theorem 1.4) and we can find a confidence interval for σ^2 using the χ^2 distribution. Recall that $W = \nu S^2 / \sigma^2 \sim \chi_\nu^2$. By way of notation, let $w_{\nu, \alpha}$ be defined by $P(W > w_{\nu, \alpha}) = \alpha$, where $W \sim \chi_\nu^2$.

The quantile for the upper 5% region is obtained by:-

`qchisq(p=0.05,df=5,lower.tail=F)` or

`qchisq(p=0.95,df=5)`

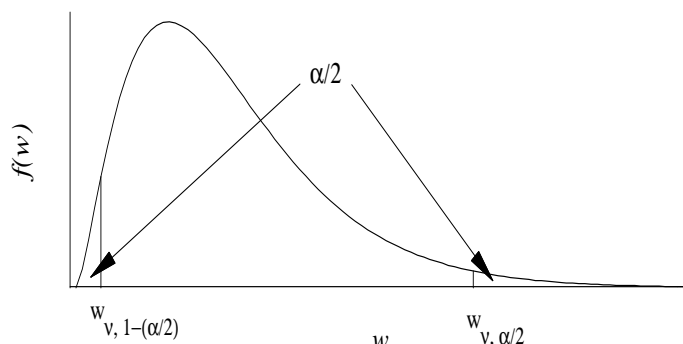
Figure 3.3: Area above $w_{\nu,\alpha}$



We find two values of W , $w_{\nu,\alpha/2}$ and $w_{\nu,1-(\alpha/2)}$, such that

$$P(w_{\nu,1-(\alpha/2)} < W < w_{\nu,\alpha/2}) = 1 - \alpha.$$

Figure 3.4: Upper and lower values for w



The event $w_{\nu,1-(\alpha/2)} < W < w_{\nu,\alpha/2}$ occurs if and only if the events

$$\sigma^2 < \nu S^2 / w_{\nu,1-(\alpha/2)}, \quad \sigma^2 > \nu S^2 / w_{\nu,\alpha/2}$$

occur. So

$$P(w_{\nu,1-(\alpha/2)} < W < w_{\nu,\alpha/2}) = P(\nu S^2 / w_{\nu,\alpha/2} < \sigma^2 < \nu S^2 / w_{\nu,1-(\alpha/2)})$$

and thus

$$\text{A } 100(1 - \alpha)\% \text{ CI for } \sigma^2 \text{ is } (\nu s^2 / w_{\nu,\alpha/2}, \nu s^2 / w_{\nu,1-(\alpha/2)}) \quad (3.9)$$

Example 3.2

For a sample of size $n = 10$ from a normal distribution s^2 was calculated and found to be 6.4. Find a 95% CI for σ^2 .

Solution: Now $\nu = 9$, and

```
qchisq(p=c(0.025,0.975),df=9,lower.tail=F)
[1] 19.0  2.7
```

$w_{9,.025} = 19$ and $w_{9,.975} = 2.7$.

Hence, $\nu s^2/w_{9,.025} = 3.02$, and $\nu s^2/w_{9,.975} = 21.33$.

That is, the 95% CI for σ^2 is (3.02, 21.33).

Case (ii)

Suppose now that X_1, X_2, \dots, X_n is a random sample from $N(\mu, \sigma^2)$ where μ is known and we wish to find a CI for the unknown σ^2 . Recall (Assignment 1, Question 4) that the maximum likelihood estimator of σ^2 (which we'll denote by S^{*2}) is

$$S^{*2} = \sum_{i=1}^n (X_i - \mu)^2 / n.$$

We can easily show that this is unbiased.

$$E(S^{*2}) = \sum_{i=1}^n \frac{E(X_i - \mu)^2}{n} = n \frac{1}{n} \sigma^2 = \sigma^2$$

The distribution of S^{*2} is found by noting that $nS^{*2}/\sigma^2 = \sum_{i=1}^n (X_i - \mu)^2/\sigma^2$ is the sum of squares of n independent $N(0,1)$ variates and is therefore distributed as χ_n^2 (using (3.8) and (3.7)). Proceeding in the same way as in Case (i) we find

$$\text{A } 100(1 - \alpha)\% \text{ CI for } \sigma^2 \text{ when } \mu \text{ is known is } \left(\frac{ns^{*2}}{w_{n,\alpha/2}}, \frac{ns^{*2}}{w_{n,1-(\alpha/2)}} \right) \quad (3.10)$$

3.5 Testing Hypotheses about σ^2

Again the cases (i) μ unknown; and (ii) μ known are considered separately.

Case (i)

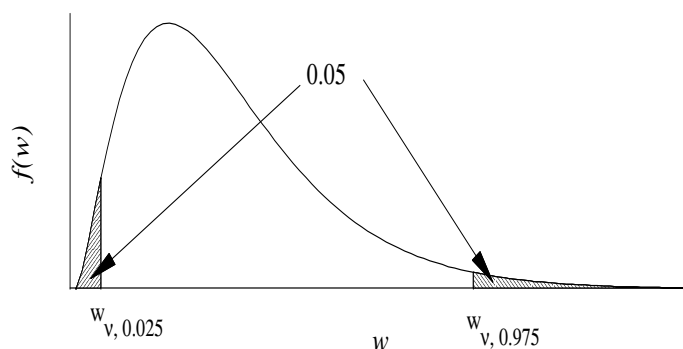
Let X_1, X_2, \dots, X_n be a random sample from a $N(\mu, \sigma^2)$ distribution where μ is unknown, and suppose we wish to test the hypothesis

$$H : \sigma^2 = \sigma_0^2 \text{ against } A : \sigma^2 \neq \sigma_0^2.$$

Under H , $\nu S^2/\sigma_0^2 \sim \chi_\nu^2$ and values of $\nu S^2/\sigma_0^2$ too large or too small would support A . For $\alpha = .05$, say, and equal-tail probabilities we have as critical region

$$R = \left\{ s^2 : \frac{\nu s^2}{\sigma_0^2} > w_{\nu,.025} \text{ or } \frac{\nu s^2}{\sigma_0^2} < w_{\nu,.975} \right\}.$$

Figure 3.5: Critical Region



Consider now a one-sided alternative. Suppose we wish to test

$$H : \sigma^2 = \sigma_0^2 \text{ against } A : \sigma^2 > \sigma_0^2.$$

Large values of s^2 would support this alternative. That is, for $\alpha = .05$, use as critical region

$$\{s^2 : \nu s^2 / \sigma_0^2 > w_{\nu, .05}\}.$$

Similarly, for the alternative $A : \sigma^2 < \sigma_0^2$, a critical region is

$$\{s^2 : \nu s^2 / \sigma_0^2 < w_{\nu, .95}\}.$$

Example 3.3

A normal random variable has been assumed to have standard deviation $\sigma = 7.5$. If a sample of size 25 has $s^2 = 95.0$, is there reason to believe that σ is greater than 7.5?

Solution: We wish to test $H: \sigma^2 = 7.5^2 (= \sigma_0^2)$ against $A: \sigma^2 > 7.5^2$.

Using $\alpha = .05$, the rejection region is $\{s^2 : \nu s^2 / \sigma_0^2 > 36.4\}$.

The calculated value of $\nu s^2 / \sigma^2$ is $\frac{24 \times 95}{56.25} = 40.53$.

```
> pchisq(q=40.53,df=24,lower.tail=F)
```

```
[1] 0.019
```

When testing at the 5% level, there is evidence that the standard deviation is greater than 7.5.

Case (ii)

Let X_1, X_2, \dots, X_n be a random sample from $N(\mu, \sigma^2)$ where μ is **known**, and suppose we wish to test $H: \sigma^2 = \sigma_0^2$. Again we use the fact that **if H is true**, $nS^{*2} / \sigma_0^2 \sim \chi_n^2$ where $S^{*2} = \sum_{i=1}^n (X_i - \mu)^2 / n$, and the rejection region for a size- α 2-tailed test, for example, would be

$$\left\{ s^{*2} : \frac{n s^{*2}}{\sigma_0^2} > w_{n, \alpha/2} \text{ or } \frac{n s^{*2}}{\sigma_0^2} < w_{n, 1-(\alpha/2)} \right\}$$

3.6 χ^2 and Inv- χ^2 distributions in Bayesian inference

3.6.1 Non-informative priors

A prior which does not change very much over the region in which the likelihood is appreciable and does not take very large values outside that region is said to be locally uniform.

For such a prior,

$$p(\theta|y) \propto p(y|\theta) = \ell(\theta|y)$$

The term pivotal quantity was introduced in section 1.7.1 and now is defined for (i) location parameter and (ii) scale parameter.

- (i) If the density of y , $p(y|\theta)$, is such that $p(y - \theta|\theta)$ is a function that is free of y and θ , say $f(u)$ where $u = y - \theta$, then $y - \theta$ is a pivotal quantity and θ is a *location parameter*.

Example. If $(y|\mu, \sigma^2) \sim N(\mu, \sigma^2)$, then $(y - \mu|\mu, \sigma^2) \sim N(0, \sigma^2)$ and $y - \mu$ is a pivotal quantity.

- (ii) If $p(\frac{y}{\phi}|\phi)$ is a function free of ϕ and y , say $g(u)$ where $u = \frac{y}{\phi}$, then u is a pivotal quantity and ϕ is a *scale parameter*.

Example. If $(y|\mu, \sigma^2) \sim N(\mu, \sigma^2)$, then $\frac{y - \mu}{\sigma} \sim N(0, 1)$.

A non-informative prior for a **location** parameter, θ , would give $f(y - \theta)$ for the posterior distribution $p(y - \theta|y)$. That is under the posterior distribution, $(y - \theta)$ should still be a pivotal quantity.

Using Bayes' rule,

$$p(y - \theta|y) \propto p(\theta)p(y - \theta|\theta)$$

Thus $p(\theta) \propto \text{Constant}$.

For the case of a **scale** parameter, ϕ , Bayes' rule is

$$p(\frac{y}{\phi}|y) \propto p(\phi)p(\frac{y}{\phi}|\phi) \tag{3.11}$$

or

$$p(u|y) \propto p(\phi)p(u|\phi) \tag{3.12}$$

(The LHS of (3.11) is the posterior of a parameter say $\phi^* = \frac{y}{\phi}$ and the RHS is the density of a scaled variable $y^* = \frac{y}{\phi}$. Both sides are free of y and ϕ .)

$$p(y|\phi) = p(u|\phi) \left| \frac{du}{dy} \right| = \frac{1}{\phi} p(u|\phi)$$

$$p(\phi|y) = p(u|y) \left| \frac{du}{d\phi} \right| = \frac{y}{\phi^2} p(u|y)$$

Thus from (3.12), equate $p(u|y)$ to $p(u|\phi)$,

$$p(\phi|y) = \frac{y}{\phi} p(y|\phi)$$

so that the uninformative prior is

$$p(\phi) \propto \frac{1}{\phi} \quad (3.13)$$

3.7 The posterior distribution of the Normal variance

Consider normally distributed data,

$$y|\mu, \sigma^2 \sim N(\mu, \sigma^2)$$

The joint posterior density of parameters μ, σ^2 is given by

$$p(\mu, \sigma^2) \propto p(y|\mu, \sigma^2) \times p(\mu, \sigma^2) \quad (3.14)$$

To get the marginal posterior distribution of the variance, integrate with respect to μ ,

$$p(\sigma^2|y) = \int p(\mu, \sigma^2|y) d\mu \quad (3.15)$$

$$= \int p(\sigma^2|\mu, y) p(\mu|y) d\mu \quad (3.16)$$

Choose the prior

$$\begin{aligned} p(\mu, \sigma^2) &\propto p(\mu) p(\sigma^2) & (\mu \perp \sigma^2) \\ p(\mu, \sigma^2) &\propto (\sigma^2)^{-1} & (p(\mu) \propto \text{Const.}) \end{aligned} \quad (3.17)$$

Write the posterior density as

$$\begin{aligned} p(\mu, \sigma^2) &\propto \sigma^{-n-2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right\} \\ &= \sigma^{-n-2} \exp \left\{ -\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \right] \right\} \\ &= \sigma^{-n-2} \exp \left\{ -\frac{1}{2\sigma^2} [(n-1)S^2 + n(\bar{y} - \mu)^2] \right\} \end{aligned}$$

where $S^2 = \frac{\sum (y_i - \bar{y})^2}{(n-1)}$

Now integrate the joint density with respect to μ ,

$$\begin{aligned}
 p(\sigma^2|y) &\propto \int \sigma^{-n-2} \exp \left\{ -\frac{1}{2\sigma^2} [(n-1)S^2 + n(\bar{y} - \mu)^2] \right\} \\
 &= \sigma^{-n-2} \exp \left\{ -\frac{1}{2\sigma^2} (n-1)S^2 \right\} \int \exp \left\{ -\frac{1}{2\sigma^2/n} (\bar{y} - \mu)^2 \right\} d\mu \\
 &= \sigma^{-n-2} \exp \left\{ -\frac{1}{2\sigma^2} (n-1)S^2 \right\} \sqrt{2\pi\sigma^2/n} \\
 &= (\sigma^2)^{-\frac{n+1}{2}} \exp \left\{ -\frac{(n-1)S^2}{2\sigma^2} \right\}
 \end{aligned} \tag{3.18}$$

The pdf of S^2 was derived at (3.3),

$$\begin{aligned}
 g(s^2) &= \frac{1}{\Gamma(\frac{\nu}{2})} \times (s^2)^{(\frac{\nu}{2}-1)} \left(\frac{\nu}{2\sigma^2} \right)^{\nu/2} \exp \left\{ -\frac{\nu s^2}{2\sigma^2} \right\} \\
 &\propto (s^2)^{(\frac{\nu}{2}-1)} \exp \left\{ -\frac{\nu s^2}{2\sigma^2} \right\}
 \end{aligned}$$

with $\nu = (n-1)$ and this is a Gamma $\left(\frac{n-1}{2}, \frac{n-1}{2\sigma^2} \right)$ distribution.

3.7.1 Inverse Chi-squared distribution

Its Bayesian counterpart at (3.18) is a *Scaled Inverse Chi-squared distribution*. Since the prior was uninformative, similar outcomes are expected.

The inverse χ^2 distribution has density function

$$p(\sigma^2|\nu) = \frac{1}{\Gamma(\frac{\nu}{2})} \left(\frac{1}{2} \right)^{\frac{\nu}{2}} \left(\frac{1}{\sigma^2} \right)^{\frac{\nu}{2}+1} \exp \left\{ -\frac{1}{2\sigma^2} \right\} \times I_{(0,\infty)}(\sigma^2)$$

The scaled inverse chi-squared distribution has density

$$p(\sigma^2|\nu, s^2) = \frac{1}{\Gamma(\frac{\nu}{2})} \left(\frac{\nu}{2} \right)^{\frac{\nu}{2}} (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp \left\{ -\frac{\nu s^2}{2\sigma^2} \right\}$$

The prior $p(\sigma^2) \propto \frac{1}{\sigma^2}$ can be said to be an inverse chi-squared distribution on $\nu = 0$ degrees of freedom or sample size $n = 1$. Is there any value in it? Although uninformative, it ensures a mathematical “smoothness” and numerical problems are reduced.

The posterior density is Scaled Inverse Chi-squared with degrees of freedom $\nu = (n-1)$ and scale parameter s .

3.8 Relationship between χ_ν^2 and $\text{Inv-}\chi_\nu^2$

Recall that χ_ν^2 is $\text{Ga}\left(\frac{\nu}{2}, \frac{1}{2}\right)$.

The Inverse-Gamma distribution is also prominent in Bayesian statistics so we examine it first.

3.8.1 Gamma and Inverse Gamma

The densities of the Gamma and Inverse Gamma are:-

$$\text{Gamma } p(\theta|\alpha, \beta) = \frac{1}{\Gamma(\alpha)} \theta^{(\alpha-1)} \beta^\alpha \exp\{-\beta\theta\} \times I_{0,\infty}(\theta) \quad \alpha, \beta > 0 \quad (3.19)$$

$$\text{Inverse Gamma } p(\theta|\alpha, \beta) = \frac{1}{\Gamma(\alpha)} \theta^{-(\alpha+1)} \beta^\alpha \exp\left\{-\frac{\beta}{\theta}\right\} \times I_{0,\infty}(\theta) \quad \alpha, \beta > 0 \quad (3.20)$$

If $\theta^{-1} \sim \text{Ga}(\alpha, \beta)$, then $\theta \sim \text{InvGamma}(\alpha, \beta)$.

Put $\phi = \theta^{-1}$. then

$$\begin{aligned} f(\theta; \alpha, \beta) &= f(\phi^{-1}; \alpha, \beta) \left| \frac{d\phi}{d\theta} \right| \\ &= \frac{1}{\Gamma(\alpha)} \theta^{-(\alpha-1)} \beta^\alpha \exp\left\{-\frac{\beta}{\theta}\right\} \theta^{-2} \\ &= \frac{1}{\Gamma(\alpha)} \theta^{-(\alpha+1)} \beta^\alpha \exp\left\{-\frac{\beta}{\theta}\right\} \end{aligned}$$

3.8.2 Chi-squared and Inverse Chi-squared

If $Y = SX$ such that $Y^{-1} \sim S^{-1}\chi_\nu^2$, then Y is S times an inverse χ^2 distribution.

The Inverse- $\chi^2(\nu, s^2)$ distribution is a special case of the Inverse Gamma distribution with $\alpha = \frac{\nu}{2}$ and $\beta = \frac{\nu s^2}{2}$.

3.8.3 Simulating Inverse Gamma and Inverse- χ^2 random variables.

- **InvGa.** Draw X from $\text{Ga}(\alpha, \beta)$ and invert it.
- **ScaledInv- χ_{ν, s^2}^2 .** Draw X from χ_ν^2 and let $Y = \frac{\nu s^2}{X}$.

Example

Give a 90% HDR for the variance of the population from which the following sample is drawn.

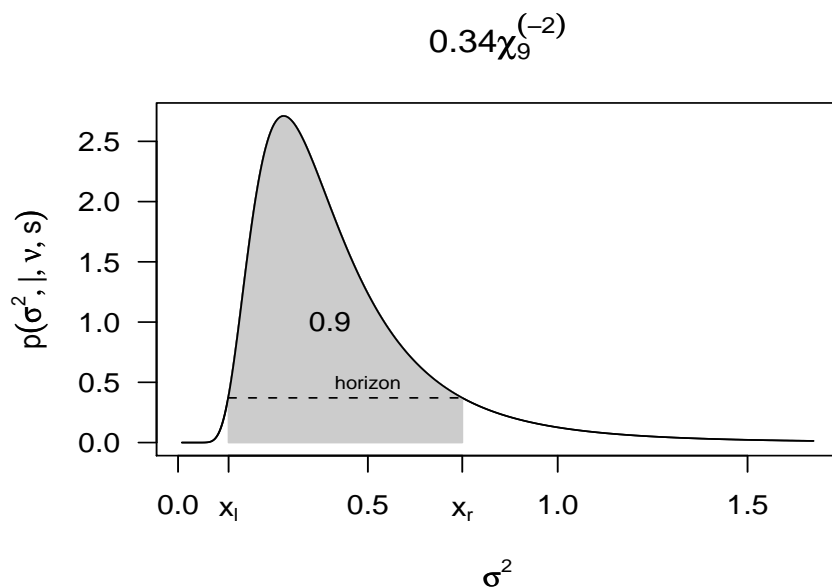
4.17 5.58 5.18 6.11 4.50 4.61 5.17 4.53 5.33 5.14

$$\begin{aligned} S^2 &= 0.34 \\ p(\sigma^2 | \nu, S^2) &= 0.34 \chi_9^{-2} \end{aligned}$$

The 90% CI for σ^2 is (0.18, 0.92). The mode of the posterior density of σ^2 is 0.28 and the 90% HDR for σ^2 is (0.13, 0.75).

The HDR was calculate numerically in this fashion,

1. Calculate the posterior density, (3.20)
2. Set an initial value for the “horizon”, estimate the abscissas (left and right of the mode) whose density is at the horizon. Call these x_l and x_r
3. Integrate the density function over (x_l, x_r) .
4. Adjust the horizon until this is 0.9. The HDR is then (x_l, x_r) at the current values.




```

#----- to calculate HDR of \sigma^2 -----
options(digits=2)
#----- functions to use later in the job -----
closest <- function(s,v){
delta <- abs(s-v)
p <- delta==min(delta)
return(p) }
#
IGamma <- function(v,a=df/2,b=0.5*df*S2){
p <- (1/gamma(a))* (v**(-(a+1)) ) * (b**a) * exp(-b/v)
return(p) }
#-----
wts <- c(4.17, 5.58, 5.18, 6.11, 4.50, 4.61, 5.17, 4.53, 5.33, 5.14) # the data
n <- length(wts); S2 <- var(wts); df <- n - 1 # statistics
cat("S-sq = ",S2,"\n")
# ----- 90% CI -----
Q <- qchisq(p=c(0.95,0.05),df=df)
CI <- df*S2/Q
cat("CI.sigma = ",CI,"\n")
# ----- Posterior -----
Ew <- df*S2/(df-2)
Vw <- (2*df^2*S2^2)/((df-2)^2*(df-4)^2)
w <- seq(0.01,(Ew+10*sqrt(Vw)),length=501)
ifw <- IGamma(v=w)
mode <- w[closest(max(ifw),ifw)]
# ----- deriving the HDR by numerical integration -----
PHDR <- 0.9 # this is the level of HDR we want
step <- 0.5; convergence.test <- 1e3; prop <- 0.9 # scalar variables for the numerical steps
while (convergence.test > 1e-3 ){ # iterate until the area is very close to 0.9
horizon <- max(ifw)*prop
left.ifw <- subset(ifw,subset=w < mode);lw <- w[w < mode]
right.ifw <- subset(ifw,subset=w > mode);rw <- w[w > mode]
xl <- lw[closest(horizon,left.ifw)]
xr <- rw[closest(horizon,right.ifw)]
Pint <- integrate(f=IGamma,lower=xl,upper=xr)
convergence.test <- abs(Pint$value - PHDR)
adjust.direction <- 2*(0.5 - as.numeric(Pint$value < PHDR)) # -1 if < +1 if >
prop <- prop+ adjust.direction*step*convergence.test
} # end of while loop

HDR <- c(xl,xr)
cat("HDR = ",HDR,"\n")

```

Chapter 4

F Distribution

4.1 Derivation

Definition 4.1

Suppose S_1^2 and S_2^2 are the sample variances for two samples of sizes n_1, n_2 drawn from normal populations with variances σ_1^2 and σ_2^2 , respectively. The random variable F is then defined as

$$F = S_1^2/S_2^2. \quad (4.1)$$

Suppose now that $\sigma_1^2 = \sigma_2^2 (= \sigma^2, \text{ say})$, then (4.1) can be written as

$$\frac{\nu_1 F}{\nu_2} = \frac{\nu_1 S_1^2/\sigma^2}{\nu_2 S_2^2/\sigma^2} = \frac{\nu_1 S_1^2/2\sigma^2}{\nu_2 S_2^2/2\sigma^2} \quad (4.2)$$

where the middle term is the ratio of 2 independent χ^2 variates on ν_1, ν_2 degrees of freedom, or equivalently, the ratio of 2 independent gamma variates with parameters $\frac{1}{2}\nu_1, \frac{1}{2}\nu_2$.

Thus, $Y = \frac{\nu_1 F}{\nu_2}$ has a derived beta distribution with parameters $\frac{1}{2}\nu_2, \frac{1}{2}\nu_1$. (Statistics 260 study guide, section 7.3.1.) Then (Example 7.5, from Statistics 260 study guide), Y has p.d.f.

$$f(y) = \frac{y^{(\nu_1/2)-1}}{(1+y)^{(\nu_1+\nu_2)/2} B(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2)}, \quad y \in [0, \infty)$$

and $g(F) = f(y) \left| \frac{dy}{dF} \right|$. So

$$g(F) = \frac{(\nu_1 F/\nu_2)^{(\nu_1/2)-1}}{\left(1 + \frac{\nu_1}{\nu_2} F\right)^{(\nu_1+\nu_2)/2} B(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2)} \frac{\nu_1}{\nu_2}, \quad F \in [0, \infty)$$

Thus

$$g(F) = \frac{\nu_1^{\nu_1/2} \nu_2^{\nu_2/2} F^{(\nu_1/2)-1}}{B(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2) (\nu_2 + \nu_1 F)^{(\nu_1+\nu_2)/2}}, \quad F \in [0, \infty) \quad (4.3)$$

This is the p.d.f. of a random variable with an F-distribution. A random variable F which can be expressed as

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \quad (4.4)$$

where $W_1 \sim \chi_{\nu_1}^2$, $W_2 \sim \chi_{\nu_2}^2$ and W_1 , and W_2 are independent random variables, is said to be distributed as $F(\nu_1, \nu_2)$, or sometimes as F_{ν_1, ν_2} . [Note that we have departed from the procedure of using a capital letter for the random variable and the corresponding small letter for its observed value, and will use F in both cases here.]

4.2 Properties of the F distribution

Mean

The mean could be found in the usual way, $E(F) = \int_0^\infty F g(F) dF$, but the rearrangement of the integrand to get an integral that can be recognized as unity, is somewhat messy, so we will use another approach.

For $W \sim \chi_\nu^2$, $E(W) = \nu$ and we will show that $E(W^{-1}) = \frac{1}{\nu - 2}$.

$$\begin{aligned} E(W^{-1}) &= \int_0^\infty \frac{w^{-1} e^{-w/2} w^{(\nu/2)-1} dw}{2^{(\nu/2)} \Gamma(\frac{1}{2}\nu)} \\ &= \frac{\Gamma(\frac{1}{2}\nu - 1)}{2\Gamma(\frac{1}{2}\nu)} \int_0^\infty \frac{e^{-w/2} w^{(\nu/2)-1-1} dw}{2^{(\nu/2)-1} \Gamma(\frac{1}{2}\nu - 1)} \\ &= \frac{\Gamma(\frac{1}{2}\nu - 1)}{2(\frac{1}{2}\nu - 1)\Gamma(\frac{1}{2}\nu - 1)} \\ &= \frac{1}{\nu - 2}. \end{aligned}$$

For independent random variables $W_1 \sim \chi_{\nu_1}^2$ and $W_2 \sim \chi_{\nu_2}^2$, define $F = \frac{W_1/\nu_1}{W_2/\nu_2} = \frac{\nu_2}{\nu_1} \frac{W_1}{W_2}$. Then,

$$\begin{aligned} E(F) &= \frac{\nu_2}{\nu_1} E(W_1) E(W_2^{-1}) \\ &= \frac{\nu_2}{\nu_1} \frac{\nu_1}{\nu_2 - 2} \\ &= \frac{\nu_2}{\nu_2 - 2}, \quad \text{for } \nu_2 > 2. \end{aligned} \quad (4.5)$$

Thus if a random variable $F \sim F(\nu_1, \nu_2)$ then

$$E(F) = \frac{\nu_2}{\nu_2 - 2}. \quad (4.6)$$

Notes:

1. The mean is independent of the value of ν_1 and is always greater than 1.
2. As $\nu_2 \rightarrow \infty$, $E(F) \rightarrow 1$.

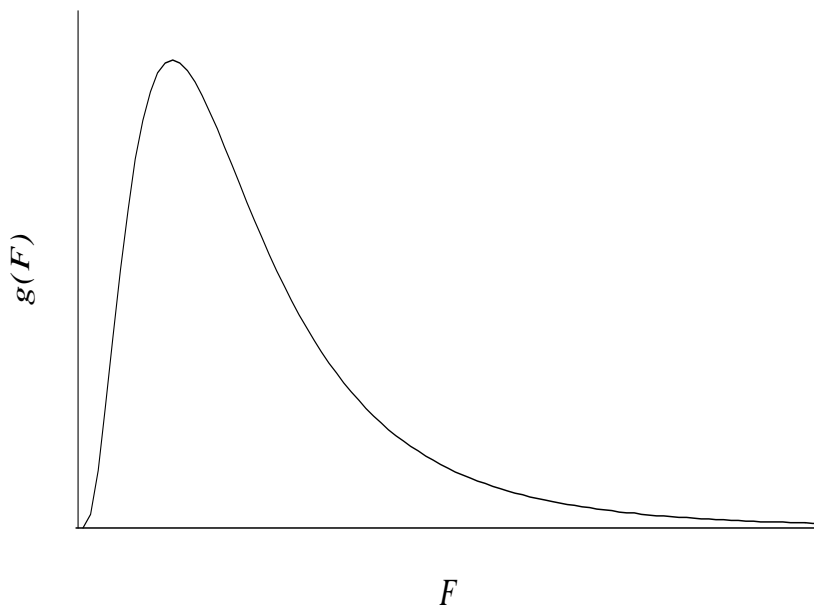
Mode

By differentiating $g(F)$ with respect to F it can be verified that the mode of the F distribution is at

$$F = \frac{\nu_2(\nu_1 - 2)}{\nu_1(\nu_2 + 2)} \quad (4.7)$$

which is always less than 1.

Figure 4.1: pdf of F -distribution

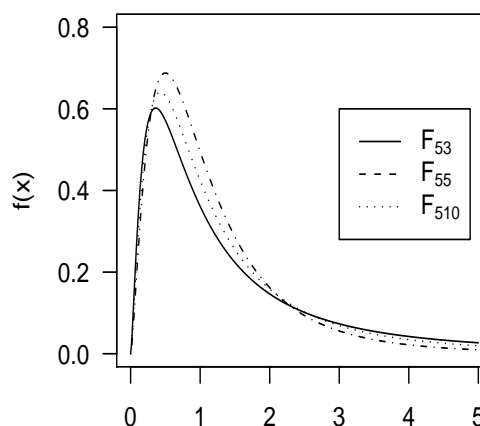


Computer Exercise 4.1

Examine the density function of the F -distribution. To do this plot the density function for the F -distribution for $\nu_1 = 5$ and $\nu_2 = 3, 5, 10$ for $x = 0, 0.01, 0.02, \dots, 5$. Overlay the plots on the same axes.

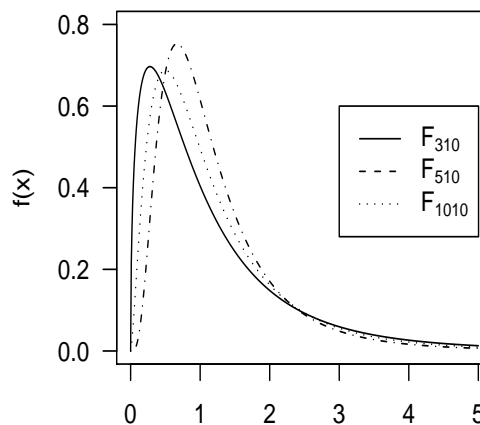
Solution:

```
#---- Fdensity.R -----
x <- seq(from=0,to=5,by=0.01)
l <- 1
for (d in c(3,5,10)){
  l <- l+1
  fx <- df(x=x,df1=5,df2=d)
  if(d==3) plot(fx ~ x,type='l',ylab="f(x)")
  else lines(x,fx,lty=1)
} # end of d loop
```



Now plot the density function for $\nu_2 = 10$ and $\nu_1 = 3, 5, 10$ again overlaying the plots on the same axes.

```
#---- Fdensity.R -----
x <- seq(from=0,to=5,by=0.01)
l <- 1
for (d in c(3,5,10)){
  l <- l+1
  fx <- df(x=x,df1=d,df2=10)
  if(d==3) plot(fx ~ x,type='l',ylab="f(x)")
  else lines(x,fx,lty=1)
} # end of d loop
```



Cumulative Distribution Function

The right-hand tail areas of the distribution are tabulated for various ν_1, ν_2 . For $P = 5, 2.5, 1, .1$, values of $F_{.01P}$ are given where

$$P/100 = \int_{F_{.01P}}^{\infty} g(F) dF.$$

Reciprocal of an F-variate

Let the random variable $F \sim F(\nu_1, \nu_2)$ and let $Y = 1/F$. Then Y has p.d.f.

$$\begin{aligned} f(y) &= g(F) \left| \frac{dF}{dy} \right| \\ &= \frac{\nu_1^{(\nu_1/2)} y^{1-(\nu_1/2)} \nu_2^{\nu_2/2} y^{(\nu_1+\nu_2)/2}}{B(\frac{1}{2}\nu_1, \frac{1}{2}\nu_2)(\nu_2 y + \nu_1)^{(\nu_1+\nu_2)/2}} \frac{1}{y^2} \\ &= \frac{\nu_2^{\nu_2/2} \nu_1^{\nu_1/2} y^{(\nu_2/2)-1}}{B(\frac{1}{2}\nu_2, \frac{1}{2}\nu_1)(\nu_1 + \nu_2 y)^{(\nu_1+\nu_2)/2}}, \quad y \in [0, \infty). \end{aligned}$$

$$\text{Thus if } F \sim F(\nu_1, \nu_2) \text{ and } Y = 1/F \text{ then } Y \sim F(\nu_2, \nu_1). \quad (4.8)$$

4.3 Use of F-Distribution in Hypothesis Testing

Let S_1^2 and S_2^2 be the sample variances of 2 samples of sizes n_1 and n_2 drawn from normal populations with variances σ_1^2 and σ_2^2 . Recall that (from (4.1), (4.2)) it is only if $\sigma_1^2 = \sigma_2^2$ ($= \sigma^2$, say) that S_1^2/S_2^2 has an F distribution. This fact can be used to test the hypothesis $H: \sigma_1^2 = \sigma_2^2$.

If the hypothesis H is *true* then,

$$S_1^2/S_2^2 \sim F(\nu_1, \nu_2) \text{ where } \nu_1 = n_1 - 1, \nu_2 = n_2 - 1.$$

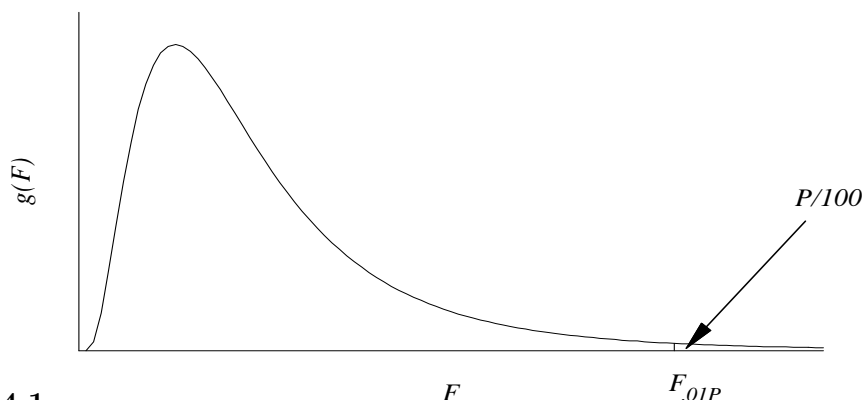
For the alternative

$$A: \sigma_1^2 > \sigma_2^2$$

only **large** values of the ratio s_1^2/s_2^2 would tend to support it, so a rejection region $\{F: F > F_{.01P}\}$ is used (Fig 4.2).

Since only the right hand tail areas of the distribution are tabulated it is convenient to always use $s_i^2/s_j^2 > 1$. That is, always put the larger sample variance in the numerator.

Figure 4.2: Critical region for F-distribution

**Example 4.1**

For two samples of sizes 8 and 12, the observed variances are .064 and .024 respectively. Let $s_1^2 = .064$ and $s_2^2 = .024$.

Solution: Test $H: \sigma_1^2 = \sigma_2^2$ against $A: \sigma_1^2 > \sigma_2^2$.

Then,

$$s_1^2/s_2^2 = .064/.024 = 2.67, \text{ and } \nu_1 = 7, \nu_2 = 11.$$

The 5% rejection region for $F(7,11)$ is $\{F : F > 3\}$. The observed value of 2.67 is less than 3 and so supports the hypothesis being true. (The observed value is *not significant* at the 5% level.)

Note: The exact $P(F \leq 2.67)$ can be found using *R*.

```
> qf(p=0.05,df1=7,df2=11,lower.tail=F)
[1] 3
> pf(q=2.67,df1=7,df2=11,lower.tail=F)
[1] 0.07
```

Thus $P(F > 2.67) = 0.07$ which agrees with the result above.

If the alternative is $\sigma_1^2 \neq \sigma_2^2$, then both tails of the distribution could be used for rejection regions, so it may be necessary to find the lower critical value. Let $F \sim F(\nu_1, \nu_2)$. That is we want find a value F_1 so that

$$\int_0^{F_1} g(F) dF = \alpha/2.$$

Put $Y = 1/F$ so that from (4.8), $Y \sim F(\nu_2, \nu_1)$. Then

$$\int_0^{F_1} g(F) dF = P(F \leq F_1) = P(Y > 1/F_1) = P(Y > F_2), \text{ say.}$$

Thus to find the lower $\frac{\alpha}{2}\%$ critical value, F_1 , first find the upper $\frac{\alpha}{2}\%$ critical value, F_2 from tables of $F(\nu_2, \nu_1)$, and then calculate F_1 as $F_1 = 1/F_2$.

Figure 4.3: Upper $\frac{\alpha}{2}\%$ point of F -distribution

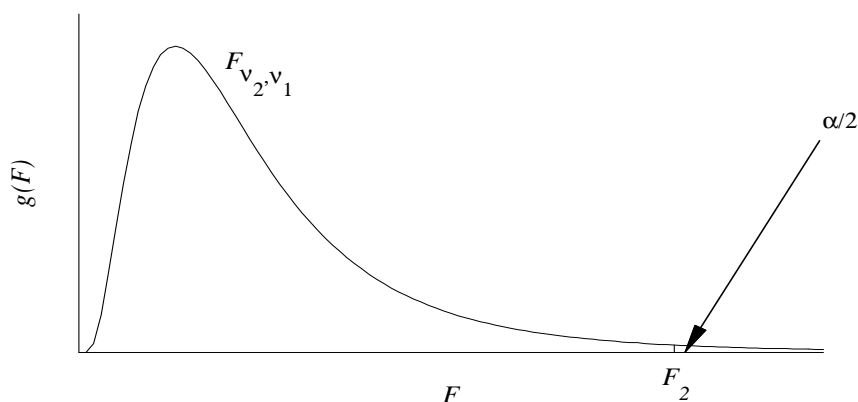
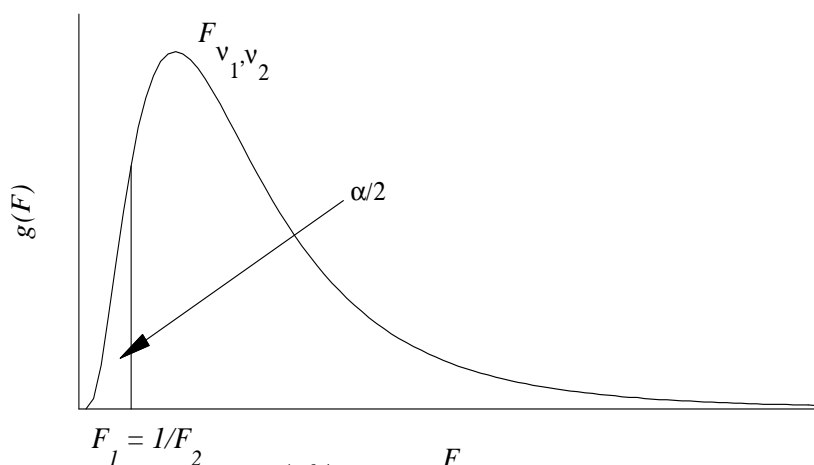


Figure 4.4: Lower $\frac{\alpha}{2}\%$ point of F -distribution



To find the **lower** $\frac{\alpha}{2}\%$ point of an F distribution with parameters ν_1, ν_2 , take the reciprocal of the **upper** $\frac{\alpha}{2}\%$ point of an F distribution with parameters ν_2, ν_1 . (4.9)

Example 4.2

Given $s_1^2 = 3.2^2$, $n_1 = 11$, $s_2^2 = 3.0^2$, $n_2 = 17$, test the hypothesis $H: \sigma_1^2 = \sigma_2^2$ against $A: \sigma_1^2 \neq \sigma_2^2$, at the 5% significance level.

Solution: Under H , $S_1^2/S_2^2 \sim F(10, 16)$.

From tables $F_{2.5\%}(10, 16) = 2.99 = F_2$.

The **lower** 2.5% critical point is then found by $F_1 = 1/F_{2.5\%}(16, 10) = 1/3.5 = .29$.

The calculated value of the statistic is $3.2^2/3.0^2 = 1.138$ which does not lie in the rejection region, and so is not significant at the 5% level. Thus the evidence supports the hypothesis that $\sigma_1^2 = \sigma_2^2$.

Of course, so long as we take s_1^2/s_2^2 to be greater than 1, we don't need to worry about the lower critical value. It will certainly be less than 1.

Computer Exercise 4.2

Use R to find the critical points in example 4.5.

Solution: We use the `qf` command.

```
> qf(p=c(0.975,0.025),df1=10,df2=16)
[1] 3.0 0.29
> pf(q=1.138,df1=10,df2=16,lower.tail=F)
[1] 0.27
```

4.4 Pooling Sample Variances

Given 2 unbiased estimates of σ^2 , s_1^2 and s_2^2 , it is often useful to be able to combine them to obtain a single unbiased estimate. Assume the new estimator, S^2 , is linear combination of s_1^2 and s_2^2 so that S^2 has the smallest variance of all such linear, unbiased estimates (that is it is said to have *minimum variance*). Let

$$S^2 = a_1 S_1^2 + a_2 S_2^2, \text{ where } a_1, a_2 \text{ are positive constants.}$$

Firstly, to be unbiased,

$$E(S^2) = a_1 E(S_1^2) + a_2 E(S_2^2) = \sigma^2(a_1 + a_2) = \sigma^2$$

which implies that

$$a_1 + a_2 = 1. \tag{4.10}$$

Secondly, if it is assumed that S_1^2 and S_2^2 are independent then

$$\begin{aligned} \text{Var}(S^2) &= a_1^2 \text{Var}(S_1^2) + a_2^2 \text{Var}(S_2^2) \\ &= a_1^2 \text{Var}(S_1^2) + (1 - a_1)^2 \text{Var}(S_2^2) \text{ using (4.10)} \end{aligned}$$

The variance of S^2 is minimised when, (writing $V(.)$ for $\text{Var}(.)$),

$$\frac{dV(S^2)}{da_1} = 2a_1 V(S_1^2) - 2(1 - a_1)V(S_2^2) = 0.$$

That is when,

$$a_1 = \frac{V(S_2^2)}{V(S_1^2) + V(S_2^2)}, \quad a_2 = \frac{V(S_1^2)}{V(S_1^2) + V(S_2^2)} \quad (4.11)$$

In the case where the X_i are normally distributed, $V(S_j^2) = 2\sigma^4/(n_j - 1)$ (see Assignment 3, Question 1). Then the pooled sample variance is

$$\begin{aligned} s^2 &= \frac{\frac{(n_1 - 1)s_1^2}{2\sigma^4} + \frac{(n_2 - 1)s_2^2}{2\sigma^4}}{\frac{n_1 - 1}{2\sigma^4} + \frac{n_2 - 1}{2\sigma^4}} \\ &= \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2} \end{aligned} \quad (4.12)$$

where $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$.

The above method can be extended to pooling k unbiased estimate s^2 of σ^2 . That is,

$$s^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2 + \cdots + \nu_k s_k^2}{\nu_1 + \nu_2 + \cdots + \nu_k}, \quad (4.13)$$

where S^2 is on $\sum_{i=1}^k \nu_i$ ($= \nu$, say) degrees of freedom, and $\nu S^2/\sigma^2$ is distributed as χ_ν^2 . Also the theory applies more generally to pooling unbiased estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ of a parameter θ .

$$\hat{\theta} = \frac{\frac{\hat{\theta}_1}{V(\hat{\theta}_1)} + \frac{\hat{\theta}_2}{V(\hat{\theta}_2)} + \cdots + \frac{\hat{\theta}_k}{V(\hat{\theta}_k)}}{\frac{1}{V(\hat{\theta}_1)} + \frac{1}{V(\hat{\theta}_2)} + \cdots + \frac{1}{V(\hat{\theta}_k)}}. \quad (4.14)$$

The estimator $\hat{\theta}$ thus obtained is unbiased and has minimum variance. Note the following

- (i) $s^2 = \frac{1}{2}(s_1^2 + s_2^2)$ if $\nu_1 = \nu_2$;
- (ii) $E(S^2) = \sigma^2$;
- (iii) The s^2 in (4.11) is on $\nu_1 + \nu_2$ degrees of freedom.

4.5 Confidence Interval for σ_1^2/σ_2^2

Given s_1^2, s_2^2 are unbiased estimates of σ_1^2, σ_2^2 derived from samples of size n_1, n_2 respectively, from two normal populations, find a $(1 - \alpha\%)$ confidence interval for σ_1^2/σ_2^2 .

Now $\nu_1 S_1^2/\sigma_1^2$ and $\nu_2 S_2^2/\sigma_2^2$ are distributed as independent $\chi_{\nu_1}^2, \chi_{\nu_2}^2$ variates, and

$$\frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \sim \frac{W_2/\nu_2}{W_1/\nu_1} \sim F(\nu_2, \nu_1).$$

So

$$P\left(F_{1-\frac{\alpha}{2}}(\nu_2, \nu_1) < \frac{S_2^2 \sigma_1^2}{S_1^2 \sigma_2^2} < F_{\frac{\alpha}{2}}(\nu_2, \nu_1)\right) = 1 - \alpha.$$

That is

$$P\left(\frac{S_1^2}{S_2^2} F_{1-\frac{\alpha}{2}}(\nu_2, \nu_1) < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} F_{\frac{\alpha}{2}}(\nu_2, \nu_1)\right) = 1 - \alpha.$$

Thus a $100(1 - \alpha)\%$ confidence interval for σ_1^2/σ_2^2 is

$$\left(\frac{s_1^2}{s_2^2} F_{1-\frac{\alpha}{2}}(\nu_2, \nu_1), \frac{s_1^2}{s_2^2} F_{\frac{\alpha}{2}}(\nu_2, \nu_1)\right). \quad (4.15)$$

4.6 Comparing parametric and bootstrap confidence intervals for σ_1^2/σ_2^2

Example 4.3

These data are sway signal energies ($\div 1000$) from subjects in 2 groups; **N**ormal and **W**hiplash injured. The data and a program to calculate the confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$, defined at equation (4.15), are listed in Table 4.1

Although the data are presented in 2 blocks, you imagine them in a single file called `test1E.txt` with the 'W' data under the 'N' data in columns.

We find that using (4.15)

$$P\left(0.52 < \frac{\sigma_1^2}{\sigma_2^2} < 3.6\right) = 0.95$$

The bootstrap CI is calculated in R by the script in Table 4.2.

By this method,

$$P\left(0.26 < \frac{\sigma_1^2}{\sigma_2^2} < 7.73\right) = 0.95$$

The CI's are much larger because the method has not relied upon the assumptions of (4.15) and uses only the information contained in the data.

Table 4.1: Confidence interval for variance ratio using the F quantiles

category	D1	category	D1	E1 <- read.table("test1E.txt",header=T)
N	0.028	W	0.048	Sigmas <- tapply(E1\$D1,list(E1\$category),var)
N	0.036	W	0.057	nu <- table(E1\$category) -1
N	0.041	W	0.113	
N	0.098	W	0.159	VR <- Sigmas[1]/Sigmas[2]
N	0.111	W	0.214	Falpha <- qf(p=c(0.975,0.025),df1=nu[1],df2=nu[2])
N	0.150	W	0.511	CI <- VR/Falpha
N	0.209	W	0.527	
N	0.249	W	0.635	> CI
N	0.360	W	0.702	[1] 0.52 3.60
N	0.669	W	0.823	
N	0.772	W	0.943	
N	0.799	W	1.474	
N	0.984	W	1.894	
N	1.008	W	2.412	
N	1.144	W	2.946	
N	1.154	W	3.742	
N	2.041	W	3.834	
N	3.606			
N	4.407			
N	5.116			

Table 4.2: R code to use the boot package to calculate CI of $\frac{\sigma_1^2}{\sigma_2^2}$

```
library(boot)
var.ratio <- function(E1,id){
  yvals <- E1[[2]][id]
  vr <- var(yvals[E1[[1]]=="N"]) /
    var(yvals[E1[[1]]=="W"])
  return(vr)
} # end of the user supplied function

doBS <- boot(E1,var.ratio,999)
bCI <- boot.ci(doBS,conf=0.95,type=c("perc","bca"))
print(bCI)

> bCI
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 999 bootstrap replicates

CALL :
boot.ci(boot.out = boot(E1, var.ratio, 999), conf = 0.95, type = c("perc","bca") )

Intervals :
Level      Percentile          BCa
95%    ( 0.26,  7.73 )    ( 0.58, 22.24 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
```

Chapter 5

t-Distribution

5.1 Derivation

Let X_1, X_2, \dots, X_n be a random sample from a $N(\mu, \sigma^2)$ distribution. Then, provided σ^2 is known, the random variable

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

is distributed $N(0,1)$.

In practice σ^2 is usually not known and is replaced by its unbiased estimate S^2 so that in place of Z we define

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

We need to find the probability distribution of this random variable. T can be written as

$$T = \frac{(\bar{X} - \mu)/\frac{\sigma}{\sqrt{n}}}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{W/\nu}}, \quad (5.1)$$

where $\nu = n - 1$, $Z \sim N(0, 1)$, $W \sim \chi_\nu^2$. Furthermore, Z and W are independent since \bar{X} and S^2 are independent (given the $X_i \sim N(\mu, \sigma^2)$). Now $T^2 = \frac{Z^2}{W/\nu}$ so that its distribution is the same as that of the ratio of two independent chi-squares (on 1 and ν degrees of freedom), that is, $F_{1,\nu}$. This fact enables us to find the pdf of the random variable T which is said to have **Student's t** distribution (after W.S. Gossett, who first derived the distribution (1908) and wrote under the pseudonym of Student). So we have the following definition:

Definition 5.1

A random variable has a t-distribution on ν degrees of freedom (or with parameter ν) if it can be expressed as the ratio of Z to $\sqrt{W/\nu}$ where $Z \sim N(0, 1)$ and W (independent of Z) $\sim \chi_\nu^2$.

Theorem 5.1

A random variable T which has a t-distribution on ν d.f. has pdf

$$f(t) = \frac{\Gamma[\frac{1}{2}(1 + \nu)]}{\sqrt{\nu\pi} \Gamma(\nu/2) [1 + (t^2/\nu)]^{(1+\nu)/2}}, \quad t \in (-\infty, \infty). \quad (5.2)$$

Proof Putting $\nu_1 = 1$, $\nu_2 = \nu$ in the pdf for the F-distribution (4.3), we have

$$g(F) = \frac{\nu^{\nu/2} F^{-1/2}}{B(\frac{1}{2}, \frac{\nu}{2})(\nu + F)^{(1+\nu)/2}}, \quad F \in [0, \infty).$$

Defining a r.v. T by $F = T^2$, with inverse $T = \pm\sqrt{F}$, it can be seen that to every value of F in $[0, \infty)$ there correspond 2 values of t in $(-\infty, \infty)$. So,

$$g(F) = 2 f(t) \left| \frac{dt}{dF} \right|$$

and

$$\begin{aligned} f(t) &= \frac{1}{2} g(F) \left| \frac{dF}{dt} \right| \\ &= \frac{1}{2} g(t^2) \cdot 2t \\ &= \frac{\nu^{\nu/2} t^{-1} t}{B(\frac{1}{2}, \frac{\nu}{2})(\nu + t^2)^{(1+\nu)/2}} \\ &= \frac{\Gamma[\frac{1}{2}(\nu + 1)]}{\sqrt{\nu} \Gamma(\frac{1}{2}) \Gamma(\frac{\nu}{2}) [1 + (t^2/\nu)]^{(1+\nu)/2}}. \end{aligned}$$

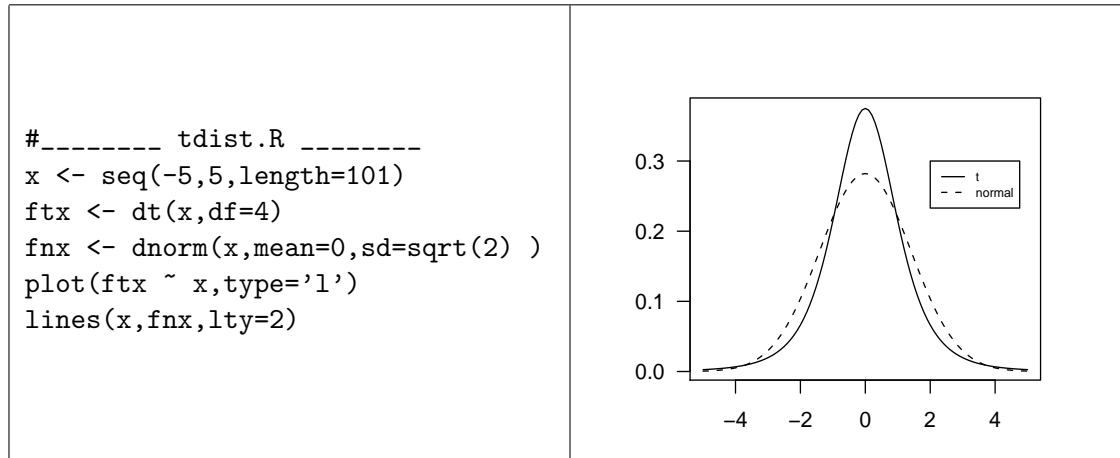
But $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ which completes the proof.

5.2 Properties of the t-Distribution

Graph

The graph of $f(t)$ is symmetrical about $t = 0$ since $f(t) = f(-t)$, unimodal, and $f(t) \rightarrow 0$ as $t \rightarrow \pm\infty$. It resembles the graph of the normal distribution but the tails are lower and the central peak higher than for a normal curve of the same mean and variance. This is illustrated in the figure 5.1 for $\nu = 4$.

Note: The density functions in Figure 5.1 were found and plotted using R .



Plots of the T distribution can also be done in Rcmdr,
Distributions → Continuous distributions → t distribution → Plot t distribution
You are required to enter the df and check whether to plot density or distribution function.

Special Cases

- (i) A special case occurs when $\nu = 1$. This is called the Cauchy distribution and it has pdf

$$f(t) = \frac{1}{\pi(1+t^2)}, \quad t \in (-\infty, \infty).$$

Check that the mean and variance of this distribution do not exist.

- (ii) It can be shown that as $\nu \rightarrow \infty$, $f(t) \rightarrow \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$, the pdf of a standardized normal distribution. To see this note that

$$\begin{aligned} \lim_{\nu \rightarrow \infty} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2} &= \lim_{\nu \rightarrow \infty} \left(1 + \frac{t^2}{\nu}\right)^{-1/2} \cdot \lim_{\nu \rightarrow \infty} \left(1 + \frac{t^2}{\nu}\right)^{-\nu/2} \\ &= 1 \cdot e^{-t^2/2} \end{aligned}$$

Then, using Stirling's approximation for $n!$, that is,

$$n! \simeq (2\pi)^{1/2} n^{n+\frac{1}{2}} e^{-n},$$

we have

$$\lim_{\nu \rightarrow \infty} \left[\Gamma\left(\frac{\nu+1}{2}\right) / \sqrt{\nu\pi} \Gamma(\nu/2) \right] = (2\pi)^{-1/2}.$$

Mean and Variance

Because of symmetry, the mean, median and mode coincide with $E(T) = 0$. Also, $\text{Var}(T) = E(T^2) = E(F) = \nu/(\nu-2)$ for $\nu > 2$. Note that, as $\nu \rightarrow \infty$ $\text{Var}(T) \rightarrow 1$.

Cumulative Distribution Function

The distribution function for the t-distribution is given by

$$\frac{P}{100} = \int_{-\infty}^{t_{1-.01P}} f(t) dt = P(T \leq t_{1-.01P})$$

Note that for $\nu = \infty$, t-distribution becomes the standard normal distribution.

Example 5.1

For $T \sim t_{10}$ find $P(T > 2.23)$.

```
pt(q=2.23,df=10,lower.tail=F)
[1] 0.025
```

In Rcmdr, the menu is

Distributions → Continuous distributions → t distribution → t probabilities
 Into the GUI you enter the quantile (i.e. 2.23 in this case) and the df.

Example 5.2

For $T \sim t_6$ find $P(|T| > 1.94)$.

```
> 2*pt(q=1.94,df=6,lower.tail=F)
[1] 0.1
```

Example 5.3

For $T \sim t_8$ find t_c such that $P(|T| > t_c) = .05$.

```
> qt(p=0.025,df=8,lower.tail=F)
[1] 2.3
```

Distributions → Continuous distributions → t distribution → t quantiles

Sampling Distributions

The χ^2 , F and t distributions are often referred to as **sampling distributions** because they are distributions of statistics arising when sampling from a normal distribution.

5.3 Use of t-Distribution in Interval Estimation

In Chapter 2 we studied the problems of getting a confidence interval for the mean μ , and testing hypotheses about μ when σ^2 was assumed known. In practice σ^2 is usually not known and must be estimated from the data and it is the t -distribution that must be used to find a confidence interval for μ and test hypotheses about μ .

In this section we will derive a $100(1 - \alpha)\%$ confidence interval for the unknown parameter μ .

One-sample Problem

Given X_1, X_2, \dots, X_n is a random sample from a $N(\mu, \sigma^2)$ distribution where σ^2 is unknown, then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}. \quad (5.3)$$

Then defining $t_{\nu, \alpha}$ as

$$P(T > t_{\nu, \alpha}) = \alpha \quad \text{where } T \sim t_{\nu}, \quad (5.4)$$

we have

$$P\left(t_{\nu, \frac{\alpha}{2}} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\nu, 1 - \frac{\alpha}{2}}\right) = 1 - \alpha.$$

That is,

$$P\left(\bar{X} - t_{\nu, \alpha/2} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\nu, \alpha/2} \frac{S}{\sqrt{n}}\right) = 1 - \alpha. \quad (5.5)$$

Now rearrange the terms on the LHS of (5.5) as follows,

$$\begin{aligned} & P\left(t_{\nu, \frac{\alpha}{2}} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\nu, 1 - \frac{\alpha}{2}}\right) \\ &= P\left(\frac{S}{\sqrt{n}} \times t_{\nu, \frac{\alpha}{2}} < \bar{X} - \mu < \frac{S}{\sqrt{n}} \times t_{\nu, 1 - \frac{\alpha}{2}}\right) \\ &= P\left(-\bar{X} + \frac{S}{\sqrt{n}} \times t_{\nu, \frac{\alpha}{2}} < -\mu < -\bar{X} + \frac{S}{\sqrt{n}} \times t_{\nu, 1 - \frac{\alpha}{2}}\right) \\ &= P\left(\bar{X} - \frac{S}{\sqrt{n}} \times t_{\nu, \frac{\alpha}{2}} > \mu > \bar{X} - \frac{S}{\sqrt{n}} \times t_{\nu, 1 - \frac{\alpha}{2}}\right) \quad \text{inequality directions not conventional} \\ &= P\left(\bar{X} - \frac{S}{\sqrt{n}} \times t_{\nu, 1 - \frac{\alpha}{2}} < \mu < \bar{X} - \frac{S}{\sqrt{n}} \times t_{\nu, \frac{\alpha}{2}}\right) \quad \text{inequality directions conventional} \end{aligned}$$

A $100(1 - \alpha)\%$ confidence interval for μ is

$$\left(\bar{x} - t_{\nu, 1 - \frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} - t_{\nu, \frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right). \quad (5.6)$$

Note how in (5.6), the upper tail quantile is subtracted from the sample mean to calculate the lower limit and the lower tail quantile is subtracted to calculate the upper limit. This arose by reversing the inequalities when making the transform $-\mu \rightarrow \mu$.

By the symmetry of the t-distribution, $t_{\nu, \frac{\alpha}{2}} = -t_{\nu, 1 - \frac{\alpha}{2}}$ and the lower tail quantile is a negative number, the upper tail quantile is the same magnitude but positive. So you would get the same result as (5.6) if you calculated

$$\left(\bar{x} - t_{\nu, 1 - \frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + t_{\nu, 1 - \frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$$

which is often how we think of it. However, it is very important that the true relationship be understood and known because it will be a critical point when we examine the bootstrap-t where the symmetry does not hold.

Example 5.4

The length (in cm) of skulls of 10 fossil skeletons of an extinct species of bird were measured with the following results.

5.22, 5.59, 5.61, 5.17, 5.27, 6.06, 5.72, 4.77, 5.57, 6.33.

Find a 95% CI for the true mean length of skulls of this species.

Solution: Computer Solution:(Ignore t -test output except for confidence interval.)

```
skulls <- data.frame(length=c(5.22,5.59,5.61,5.17,5.27,6.06,5.72,4.77,5.57,6.33) )
t.test(skulls$length,alternative="two.sided",mu=0,conf.level=0.95)
```

```
data: skulls$length
t = 38.69, df = 9, p-value = 2.559e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5.20 5.85
sample estimates:
mean of x
 5.53
```

A 95% CI is calculated by default. To obtain a CI with a different level of confidence (say 99%) we would use the command: `t.test(x, conf.level=0.99)`.

The Rcmdr menus mostly work only when there is an active data set and the data are organised into a data frame. To do the above in Rcmdr,

- (i) Make `skulls` the active data set either
 - Enter the data using `Data → New data set` or
 - Use a script such as above (`skulls <- data.frame(...)`), Submit and then `Data → Active data set → Select active data set`
- `Statistics → Means → Single-sample t-test`

Two-sample Problem

Let us now consider the two-sample problem where X_1, X_2, \dots, X_{n_1} and Y_1, Y_2, \dots, Y_{n_2} are independent random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ distributions respectively. Now the random variable $\bar{X} - \bar{Y}$ is distributed as $N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$. That is,

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1).$$

If σ^2 is unknown, its minimum variance unbiased estimate S^2 is given in (4.12). Consider the random variable T defined by

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad (5.7)$$

where

$$S^2 = \frac{\nu_1 S_1^2 + \nu_2 S_2^2}{\nu_1 + \nu_2}, \quad \text{and} \quad \frac{(\nu_1 + \nu_2) S^2}{\sigma^2} \sim \chi_{\nu_1 + \nu_2}^2. \quad (5.8)$$

Rewriting T with a numerator of $(\bar{X} - \bar{Y} - (\mu_1 - \mu_2))/\sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}$ and a denominator of $\sqrt{S^2/\sigma^2}$, we see that T can be expressed as the ratio of a $N(0,1)$ variate to the square root of an independent chi-square variate divided by its degree of freedom. Hence it has a t -distribution with $\nu_1 + \nu_2 = n_1 + n_2 - 2$ degrees of freedom.

We will now use (5.5) to find a confidence intervals for $\mu_1 - \mu_2$.

Given X_1, X_2, \dots, X_{n_1} is a random sample from $N(\mu_1, \sigma^2)$ and Y_1, Y_2, \dots, Y_{n_2} is an independent sample from $N(\mu_2, \sigma^2)$, and with $t_{\nu, \alpha}$ defined as in (5.3), we have

$$P \left(-t_{\nu, \alpha/2} < \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_2})}} < t_{\nu, \alpha/2} \right) = 1 - \alpha.$$

Rearranging, the $100(1 - \alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$\left(\bar{x} - \bar{y} - t_{\nu, \alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad \bar{x} - \bar{y} + t_{\nu, \alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \quad (5.9)$$

where S^2 is defined in (5.8) and $\nu = \nu_1 + \nu_2 = n_1 + n_2 - 2$.

Example 5.5

The cholesterol levels of seven male and 6 female turtles were found to be:

Male	226	228	232	215	223	216	223
Female	231	231	218	236	223	237	

Find a 99% CI for $\mu_m - \mu_f$.

Solution:

It will be assumed variances are equal. See chapter 4 for method of testing using R.

```
x <- c(226,228,232,215,223,216,223)
y <- c(231,231,218,236,223,237)
t.test(x,y,var.equal=T,conf.level=0.99)
```

Two Sample t-test

```
data: x and y
t = -1.6, df = 11, p-value = 0.1369
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -17.8  5.7
sample estimates:
mean of x mean of y
    223      229
```

A 99% CI for $\mu_1 - \mu_2$ is $(-17.8, 5.7)$

If you were to use rcmdr, you would need to organise the data into a data frame like this:-

```
x <- c(226,228,232,215,223,216,223)
y <- c(231,231,218,236,223,237)
turtles <- data.frame(chol=c(x,y),sex = c(rep("M",7),rep("F",6)))
> turtles
  chol sex
1  226  M
2  228  M
3  232  M
4  215  M
5  223  M
6  216  M
7  223  M
8  231  F
9  231  F
10 218  F
11 236  F
12 223  F
13 237  F
```

Make that data frame active and then use Statistics → Means → Independent samples t-test

5.4 Use of t-distribution in Hypothesis Testing

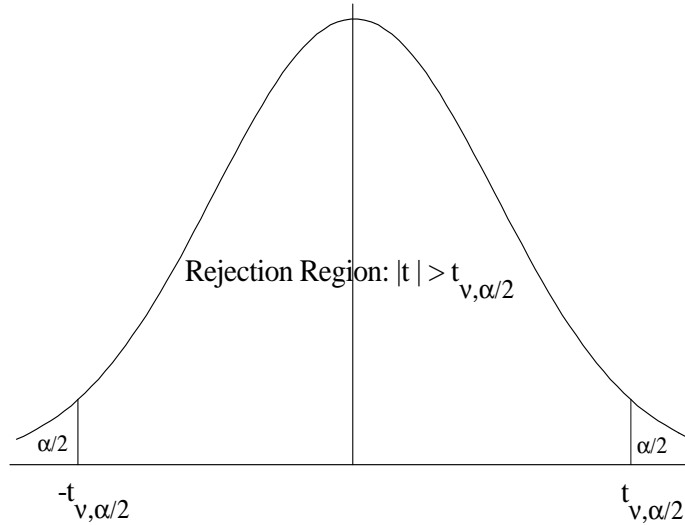
One-sample Problem

Given X_1, X_2, \dots, X_n is a random sample from $N(\mu, \sigma^2)$ where both parameters are unknown, we wish to test the hypothesis, $H : \mu = \mu_0$. Using (5.2) we can see that

- (a) for the alternative, $H_1 : \mu \neq \mu_0$, values of \bar{x} ‘close’ to μ_0 support the hypothesis being true while if $|\bar{x} - \mu_0|$ is too large there is evidence the hypothesis may be incorrect. That is, reject H_0 at the $100\alpha\%$ significance level if

$$\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} > t_{\nu, \alpha/2}.$$

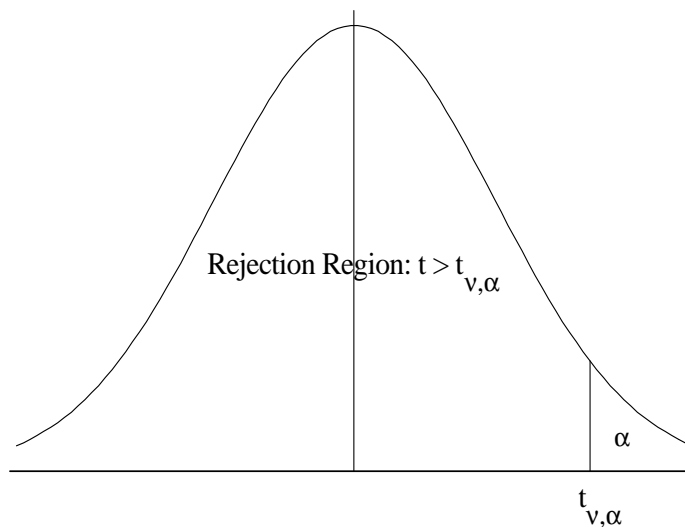
Figure 5.1: Critical Region for t -distribution: Two Sided



- (b) For $H_1 : \mu > \mu_0$, only *large* values of $(\bar{x} - \mu_0)$ tend to cast doubt on the hypothesis. That is, reject H_0 at the $100\alpha\%$ significance level if

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} > t_{\nu, \alpha}.$$

An alternative $H_1 : \mu < \mu_0$, would be treated similarly to (b) but with lower critical value $-t_{\nu, \alpha}$.

Figure 5.2: Critical Region for t -distribution: One Sided**Example 5.6**

A certain type of rat shows a mean weight gain of 65 gms during the first 3 months of life. A random sample of 12 rats were fed a particular diet from birth. After 3 months the following weight gains were recorded: 55, 62, 54, 57, 65, 64, 60, 63, 58, 67, 63, 61. Is there any reason to believe that the diet has resulted in a change of weight gain?

Solution: Let X be the weight gain in 3 months and assume that $X \sim N(\mu, \sigma^2)$. The hypothesis to be tested is $H : \mu = 65.0$ and the appropriate alternative, $H_1 : \mu \neq 65.0$. Then, $\bar{x} = 60.75$, $s^2 = 16.38$ and.

$$t = \frac{60.75 - 65.0}{\sqrt{16.38}/\sqrt{12}} = -3.64.$$

For a 2-tailed test with $\alpha = .05$, $t_{11,.025} \simeq 2.20$

```
wt <- c(55,62,54,57,65,64,60,63,58,67,63,61)
xbar <- mean(wt); s <- sd(wt); n <- length(wt)
tT <- (xbar-65)/(s/sqrt(n)); cat("t = ",tT,"\n")
t = -3.6
qt(p=0.025,df=(n-1))
[1] -2.2
pt(q=tT,df=(n-1))
[1] 0.0020
```

Our calculated value is less than $-t_{11,.025}$ and so is significant at the 5% level. Furthermore, $t_{11,.005} \simeq 3.11$ and our calculated value lies in the 1% critical region for the two-tailed test, so H is rejected at the 1% level. A better (and more modern) way to say this is that

if the hypothesis is true then the probability of an observed t -value as extreme (in either direction) as the one obtained is less than 1% . Thus there is strong evidence to suggest that the hypothesis is incorrect and that this diet has resulted in a change in the mean weight gained.

```
> t.test(wt, alternative="two.sided", mu=65)
      One Sample t-test
data:  wt
t = -3.6, df = 11, p-value = 0.003909
alternative hypothesis: true mean is not equal to 65
95 percent confidence interval:
 58 63
sample estimates:
mean of x
 61
```

Comment

The procedure adopted in the above example is a generally accepted one in hypothesis testing problems. That is, it is customary to start with $\alpha = .05$, and if the hypothesis is rejected at the 5% level (this is equivalent to saying that the observed value of the statistic is significant at the 5% level), then consider $\alpha = .01$. If the observed value is right out in the tail of the distribution, it may fall in the 1% critical region (one- or two-tailed, whichever is appropriate). To make a conclusion, claiming significance at the 1% level carries more weight than one claiming significance at the 5% level. This is because in the latter case we are in effect saying that, on the basis of the data we have, we will assert that H is not correct. In making such a statement we admit that 5 times in 100 we would reject H_0 wrongly. In the former case however, (significance at the 1% level) we realize that there is only 1 chance in 100 that we have rejected H wrongly. The commonly accepted values of α to consider are .05, .01, .001. For the t , F and χ^2 distributions, critical values can be read from the tables for both 1- and 2-tailed tests for these values of α .

Two-sample Problem

Given X_1, X_2, \dots, X_{n_1} and Y_1, Y_2, \dots, Y_{n_2} are independent random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively, we may wish to test $H : \mu_1 - \mu_2 = \delta_0$, say. Using (5.3) we can see that, under H_0 ,

$$\frac{\bar{X} - \bar{Y} - \delta_0}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

So H_0 can be tested against one- or two-sided alternatives.

Note however, that we have assumed that both populations have the same variance σ^2 , and this in general is not known. More generally, let X_1, X_2, \dots, X_{n_1} be a random

sample from $N(\mu_1, \sigma_1^2)$ and Y_1, Y_2, \dots, Y_{n_2} be an independent random sample from $N(\mu_2, \sigma_2^2)$ where $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ are unknown, and suppose we wish to test $H : \mu_1 - \mu_2 = \delta_0$. From the samples of sizes n_1, n_2 we can determine $\bar{x}, \bar{y}, s_1^2, s_2^2$. We first test the preliminary hypothesis that $\sigma_1^2 = \sigma_2^2$ and if evidence supports this, then we regard the populations as having a common variance σ^2 . So the procedure is:

- (i) Test $H_0 : \sigma_1^2 = \sigma_2^2 (= \sigma^2)$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$, using the fact that under $H_0, S_1^2/S_2^2 \sim F_{\nu_1, \nu_2}$. [This is often referred to as testing sample variances for compatibility.] A two-sided alternative and a two-tailed test is always appropriate here. We don't have any prior information about the variances. If this test is "survived" (that is, if H_0 is not rejected), proceed to (ii).
- (ii) Pool s_1^2 and s_2^2 using $s^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2}$ which is now an estimate of σ^2 based on $\nu_1 + \nu_2$ degrees of freedom.
- (iii) Test $H_0 : \mu_1 - \mu_2 = \delta_0$ against the appropriate alternative using the fact that, under H_0 ,

$$\frac{\bar{X} - \bar{Y} - \delta_0}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{\nu_1 + \nu_2}.$$

Example 5.7

A large corporation wishes to choose between two brands of light bulbs on the basis of average life. Brand 1 is slightly less expensive than brand 2. The company would like to buy brand 1 unless the average life for brand 2 is shown to be significantly greater. Samples of 25 lights bulbs from brand 1 and 17 from brand 2 were tested with the following results:

Brand 1 (X):

997, 973, 977, 1051, 1029, 934, 1007, 1020, 961, 948, 954, 939, 987, 956, 874, 1042, 1010, 942, 1011, 962, 993, 1042, 1058, 992, 979

Brand 2 (Y):

973, 970, 1018, 1019, 1004, 1009, 983, 1013, 968, 1025, 935, 1018, 1033, 992, 1037, 964, 1067

We want to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 < \mu_2$ where μ_1, μ_2 are the means from brands 1 and 2 respectively.

Solution: For the above data, $\bar{x} = 985.5$ hours, $\bar{y} = 1001.6$ hours, $s_1 = 43.2$, $s_2 = 32.9$.

- (i) Firstly test $H_0 : \sigma_1^2 = \sigma_2^2$ against a two-sided alternative noting that under $H_0, S_1^2/S_2^2 \sim F_{24, 16}$.
Then, $s_1^2/s_2^2 = 1.72$ and from the F-tables, the critical value for a two-tailed test with $\alpha = .05$ is $F_{2.5\%}(24, 16) = 2.63$. The calculated value is not significant (that is, does not lie in the critical region) so there is no reason to doubt H_0 .

(ii) Hence, pooling sample variances,

$$s^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2}{\nu_1 + \nu_2} = \frac{24 \times 1866.24 + 16 \times 1.82.41}{24 + 16} = 1552.71.$$

(iii) Assuming the hypothesis, $\mu_1 = \mu_2$ is true, $\frac{\bar{X} - \bar{Y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$. But,

$$t = \frac{-1001.6 + 985.5}{\sqrt{1552.71 \times .098824}} = \frac{-16.1}{12.387} = -1.30.$$

For a 1-tailed test with $\alpha = .05$ (with left- hand tail critical region), the critical value is $t_{40,.95} = -t_{40,.05} = -1.68$. The observed value is not in the critical region so is not significant at the 5% level and there is insufficient evidence to cast doubt on the truth of the hypothesis. That is, the average life for brand 2 is not shown to be significantly greater than that for brand 1.

Computer Solution:

```
x <- c(997, 973, 977, 1051, 1029, 934, 1007, 1020, 961, 948, 954, 939, 987,
      956, 874, 1042, 1010, 942, 1011, 962, 993, 1042, 1058, 992, 979)
y <- c(973, 970, 1018, 1019, 1004, 1009, 983, 1013, 968, 1025, 935, 1018, 1033,
      992, 1037, 964, 1067)
t.test(x,y,var.equal=T)
```

Two Sample t-test

```
data:  x and y
t = -1.3, df = 40, p-value = 0.2004
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -41.2    8.9
sample estimates:
mean of x mean of y
   986     1002
```

Notice here the actual P -value is given as 20%. The 95% confidence interval for $\mu_1 - \mu_2$ is also given. Both the t -test and the confidence interval are based on the pooled standard deviation which is not reported and would have to be calculated separately if needed.

Comment

When the population variances are **unequal and unknown**, the methods above for finding confidence intervals for $\mu_1 - \mu_2$ or for testing hypotheses concerning $\mu_1 - \mu_2$ are not appropriate. The problem of unequal variances is known as the **Behrens-Fisher problem**, and various approximate solutions have been given but are beyond the scope of this course. One such approximation (the Welch t -test) can be obtained in *R* by omitting the `var.equal=T` option from `t.test`.

5.5 Paired-sample t-test

Consider two random samples X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_n from normal distributions with means μ_1, μ_2 respectively, but where the two samples consist of n logically paired observations, (X_i, Y_i) . For example, X_i could be a person's pre diet weight while Y_i could be the weight of the same person after being on the diet for a specified period. (That is, the two samples are not independent.)

Define the random variables

$$D_i = X_i - Y_i, i = 1, \dots, n.$$

The hypothesis of interest is then

$$H_0 : \mu_{\text{diff}} = \mu_1 - \mu_2 = \delta_0.$$

Then, under the hypothesis D_1, D_2, \dots, D_n can be regarded as a random sample from a normal distribution with mean δ_0 and variance σ^2 (unknown). Let

$$\bar{D} = \sum_{i=1}^n D_i/n \quad \text{and} \quad s_d^2 = \frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}.$$

If the hypothesis H_0 is true, $E(\bar{D}) = \delta_0$ and $\text{Var}(\bar{D}) = \sigma^2/n$. So

$$\frac{\bar{D} - \delta_0}{\sqrt{s_d^2/n}} \sim t_{n-1}.$$

Thus the hypothesis is tested by comparing $\frac{\bar{d} - \delta_0}{s_d/\sqrt{n}}$ with the appropriate critical value from tables.

One important reason for using paired observations is to eliminate effects in which there is no interest. Suppose that two teaching methods are to be compared by using 50 students divided into classes with 25 in each. One way to conduct the experiment is to assign randomly 25 students to each class and then compare average scores. But if one group happened to have the better students, the results may not give a fair comparison of the two methods. A better procedure is to pair the students according to ability (as measured by IQ from some previous test) and assign at random one of each pair to each class. The conclusions then reached are based on differences of paired scores which measure the effect of the different teaching methods.

When extraneous effects (for example, student's ability) are eliminated, the scores on which the test is based are less variable. If the scores measure both ability and difference in teaching methods, the variance will be larger than if the scores reflect only teaching method, as each score then has two "sources" of variation instead of one.

Example 5.8

The following table gives the yield (in kg) of two varieties of apple, planted in pairs at eight (8) locations. Let X_i and Y_i represent the yield for varieties 1, 2 respectively at location $i = 1, 2, \dots, 8$.

i	1	2	3	4	5	6	7	8
x_i	114	94	64	75	102	89	95	80
y_i	107	86	70	70	90	91	86	77
d_i	7	8	-6	5	12	-2	9	3

Test the hypothesis that there is no difference in mean yields between the two varieties, that is test: $H_0 : \mu_X - \mu_Y = 0$ against $H_1 : \mu_X - \mu_Y > 0$.

Solution:

	Paired t-test
<pre>x <-c(114,94,64,75,102,89,95,80) y <-c(107,86,70,70,90,91,86,77) t.test(x,y,paired=T, alternative="greater")</pre>	<pre>data: x and y t = 2.1, df = 7, p-value = 0.03535 alternative hypothesis: true difference in means is greater than 0 95 percent confidence interval: 0.5 Inf sample estimates: mean of the differences 4.5</pre>

The probability of observing $T > 2.1$ under H_0 is 0.035 which is sufficiently small a probability to reject H_0 and conclude that the observed difference is due to variety and not just random sampling.

In Rcmdr, the data are organised in a data frame in pairs,

```
> apples <- data.frame(yld1=x,yld2=y)
> apples
  yld1 yld2
1  114  107
2   94   86
3   64   70
4   75   70
5  102   90
6   89   91
7   95   86
8   80   77
```

Make the data set active and then use **Statistics** \rightarrow **Means** \rightarrow **paired t-test**.

Comments

1. Note that for all the confidence intervals and tests in this chapter, it is assumed the samples are drawn from populations that have a **normal** distribution.
2. Note that the violation of the assumptions underlying a test can lead to incorrect conclusions being made.

5.6 Bootstrap T-intervals

We can make accurate intervals without depending upon the assumption of the assumption of normality made at (5.1) by using the bootstrap. The method is named *bootstrap-t*.

The procedure is as follows:-

1. Estimate the statistic $\hat{\theta}$ (e.g. the sample mean) and its standard error, \hat{se} , and determine the sample size, n .
2. Nominate the number of bootstrap samples B , e.g. $B = 199$.
3. Loop B times
 - Generate a bootstrap sample $\mathbf{x}_{(b)}^*$ by taking a sample of size n *with replacement*.
 - Calculate the bootstrap sample statistic, $\hat{\theta}_{(b)}^*$, and its standard error, $\hat{se}_{(b)}^*$
 - Calculate $T_{(b)}^* = \frac{\hat{\theta}_{(b)}^* - \hat{\theta}}{\hat{se}_{(b)}^*}$ and save this result
4. Estimate the bootstrap-t quantiles from $T_{(1)}^*, T_{(2)}^*, \dots, T_{(B)}^*$. Denote these as $\hat{t}_{\frac{\alpha}{2}}$ and $\hat{t}_{1-\frac{\alpha}{2}}$.
5. The $100\alpha\%$ CI for θ is

$$\left(\hat{\theta} - \hat{t}_{1-\frac{\alpha}{2}} \times \hat{se} \quad , \quad \hat{\theta} - \hat{t}_{\frac{\alpha}{2}} \times \hat{se} \right)$$

The point made at equation (5.6) about selecting the correct quantiles to make the confidence limits now becomes important because the symmetry no longer holds.

Example 5.9

In example 5.4 the 95% CI for the mean was calculated to be (5.2, 5.9). We now calculate the CI using bootstrap-t.

```
x <- c(5.22,5.59,5.61,5.17,5.27,6.06,5.72,4.77,5.57,6.33)
n <- length(x)
skulls <- data.frame(length=c(5.22,5.59,5.61,5.17,5.27,6.06,5.72,4.77,5.57,6.33) )
heta <- mean(skulls$length)
se.theta <- sd(skulls$length)/sqrt(n)

nBS <- 199
Tstar <- numeric(nBS)
i <- 1
while( i < (nBS+1)){          # looping 1 to nBS
x.star <- sample(skulls$length,size=n,replace=T)
Tstar[i] <- (mean(x.star) -heta ) / ( sd(x.star)/sqrt(n) )
i <- i+1                      }          # end of the while loop

bootQuantiles <- round(quantile(Tstar,p=c(0.025,0.975)),2)
cat("Bootstrap T quantiles = ",bootQuantiles,"\n")
CI <- theta - se.theta*rev(bootQuantiles)
cat("CI = ",CI,"\n")

Bootstrap T quantiles =  -2.37 2.41
CI =  5.2 5.83
```

Note in the code the use of the `rev()` function to reverse the quantiles for calculating the CI. Also observe that the quantiles are not of the same magnitude, symmetry is absent.

However, the asymmetry is not much.

Example 5.10

In example 5.7 the 95% CI for the mean difference was determined as $16 \pm 2 \times 39.4 \sqrt{\frac{1}{25} + \frac{1}{17}} = (-41, 9)$. What is the bootstrap-t CI for mean difference?

The `mndiff` function in the code below computes the mean and variance of the difference. The user must supply the variance calculations for `boot.ci` to calculate the bootstrap-t (or Studentized) CI's.

```

x <- c(997, 973, 977, 1051, 1029, 934, 1007, 1020, 961, 948, 954, 939, 987,
      956, 874, 1042, 1010, 942, 1011, 962, 993, 1042, 1058, 992, 979)
y <- c(973, 970, 1018, 1019, 1004, 1009, 983, 1013, 968, 1025, 935, 1018, 1033,
      992, 1037, 964, 1067)

lights <- data.frame(Brand=c(rep("X",length(x)),rep("Y",length(y))),hours=c(x,y) )

library(boot)
mndiff <- function(lights,id){
  yvals <- lights[[2]][id]
  delta <- mean(yvals[lights[[1]]=="X"]) - mean(yvals[lights[[1]]=="Y"])
  v <- var(yvals[lights[[1]]=="X"])/length(yvals[lights[[1]]=="X"]) +
      var(yvals[lights[[1]]=="Y"])/length(yvals[lights[[1]]=="Y"])
  return(c(delta,v))
}

doBS <- boot(lights,mndiff,999)
bCI <- boot.ci(doBS,conf=0.95,type=c("perc","stud"))
print(bCI)

```

Intervals :

Level	Studentized	Percentile
95%	(-55.81, -7.99)	(-24.34, 24.42)

Calculations and Intervals on Original Scale

Collating the results for the CI of the mean difference (and including the CI for variances not equal),

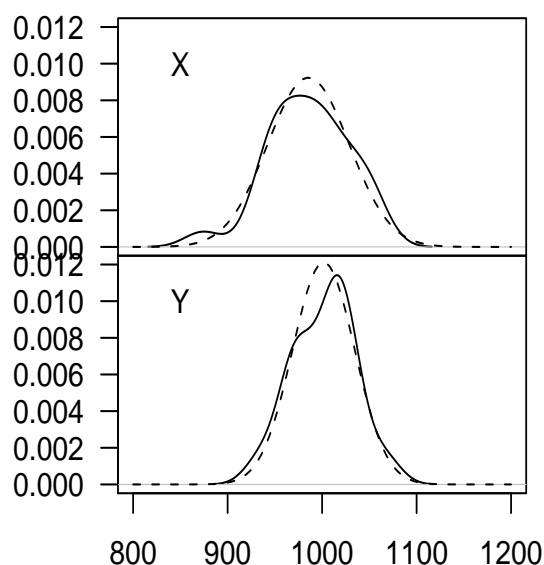
t with variances equal	(-41, 9)
t with variances unequal	(-40, 7.7)
bootstrap-t	(-56, 8)
percentile-t	(-24, 24)

Although the findings from each technique are the same, that there is insufficient evidence to conclude that the means are different, nevertheless there are disparities which we may attempt to understand.

Density plots of the 2 samples give some clues as to why the results might differ, Figure 5.3

Although an F-test supported the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$, the plots do indicate that the difference in variances might be an issue. Further, although the densities seem to be approximately normal, each could also be viewed as displaying skewness. The assumptions of normality and equal variances may not be justified.

Figure 5.3: Densities of lifetimes of brands X & Y light bulbs.



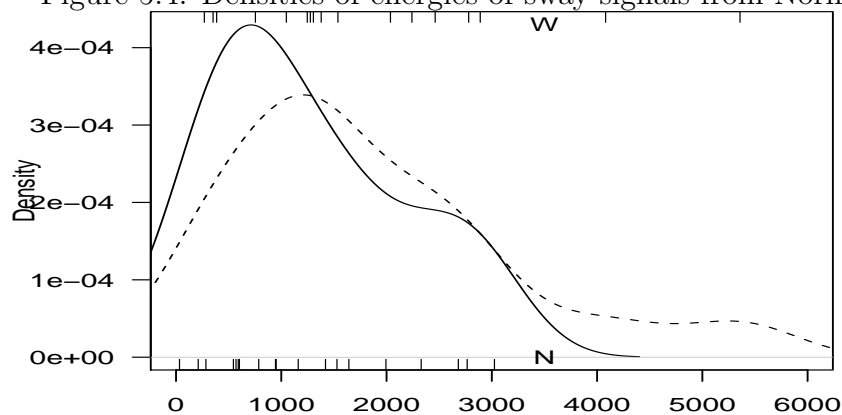
Normal densities are given by the dashed line.

Example 5.11

The densities of the data used in Example 2.7 are used to demonstrate how the assumption of normality is a strong condition for parametric t-tests.

The densities of the energies from 2 groups (N & W) are plotted in Figure 5.4.

Figure 5.4: Densities of energies of sway signals from Normal and whiplash subjects



The confidence intervals of mean difference are:-

parametric t with variances unequal $(-1385, 224)$

bootstrap-t $(-2075, -302)$

percentile-t $(-815, 791)$

Only the bootstrap-t suggests that the mean difference is unlikely to be zero.

Parametric-t loses out because the underlying assumptions of normality do not hold.

The percentile bootstrap is unreliable for small sample sizes (< 100).

Chapter 6 Analysis of Count Data

6.1 Introduction

This chapter deals with hypothesis testing problems where the data collected is in the form of **frequencies** or **counts**.

In section (6.2) we study a method for testing the very general hypothesis that a probability distribution takes on a certain form, for example, normal, Poisson, exponential, etc. The hypothesis may or may not completely prescribe the distribution. That is, it may specify the value of the parameter(s), or it may not. These are called *Goodness-of-Fit Tests*. Section (6.3) is then concerned with the analysis of data that is classified according to two attributes in a *Contingency Table*. Of interest here is whether the two attributes are associated.

6.2 Goodness-of-Fit Tests

Consider the problem of testing if a given die is unbiased. The first step is to conduct an experiment such as throwing the die n times and counting how many 1's, 2's, ..., 6's occur. If Y is the number that shows when the die is thrown once and *if the die is unbiased* then Y has a rectangular distribution. That is,

$$P(Y = i) = p_i = 1/6, \quad i = 1, 2, 3, 4, 5, 6.$$

Testing the hypothesis the die is unbiased is then equivalent to testing the hypothesis $H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$.

Let A_i be the event: i occurs (on a given throw). Then $P(A_i) = p_i = 1/6$, *under* H_0 (that is, if and only if, H_0 is true). The random variables $Y_i, i = 1, \dots, 6$ are now defined as follows. Let Y_i be the number of times in the n throws that A_i occurs. If the die is unbiased the distribution of (Y_1, \dots, Y_6) is then multinomial with parameters p_1, \dots, p_6 all equal to $1/6$.

Example 6.1

Suppose that the die is thrown 120 times and the **observed frequencies** in each category (denoted by o_1, \dots, o_6) are

i	1	2	3	4	5	6
o_i	15	27	18	12	25	23

Find the **expected frequencies**.

Solution: For a multinomial distribution, the **expected frequencies** (which will be denoted by e_i), are given by $E(Y_i) = np_i, i = 1, \dots, 6$, (Theorem 5.2, STAT260), and $E(Y_i) = 120 \times \frac{1}{6} = 20$ if H_0 is true. The question to be answered is whether the observed frequencies are close enough to those expected under the hypothesis, to be consistent with the given hypothesis. The following theorem provides a test to answer this question.

Theorem 6.1

Given

$$P(Y_1 = y_1, \dots, Y_k = y_k) = \frac{n!}{y_1! \dots y_k!} p_1^{y_1} p_2^{y_2} \dots p_k^{y_k}$$

where Y_i is the number of times in n trials that event A_i (which has probability p_i) occurs, $\sum_{i=1}^k p_i = 1$, $\sum_{i=1}^k y_i = n$, then the random variable X^2 defined by

$$X^2 = \sum_{i=1}^k \frac{(Y_i - np_i)^2}{np_i} \quad (6.1)$$

is distributed approximately as χ_{k-1}^2 for n large. This is often expressed as

$$X^2 = \sum_{i=1}^k \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}} = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}.$$

Outline of Proof (for $k = 2$)

We can write X^2 as

$$\begin{aligned} X^2 &= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2} \\ &= \frac{(Y_1 - np_1)^2}{np_1} + \frac{(n - Y_1 - n(1 - p_1))^2}{n(1 - p_1)} \\ &= \frac{(Y_1 - np_1)^2}{n} \left[\frac{1}{p_1} + \frac{1}{1 - p_1} \right] \\ &= \frac{(Y_1 - np_1)^2}{np_1 q_1} \end{aligned}$$

Now Y_1 is distributed as $\text{bin}(n, p_1)$. Thus (approximately)

$$X = \frac{Y_1 - np_1}{\sqrt{np_1 q_1}} \sim N(0, 1)$$

for n large. Hence (again approximately)

$$X^2 = \frac{(Y_1 - np_1)^2}{np_1 q_1} \sim \chi_1^2$$

for n large.

Comments

1. Note that the multinomial distribution only arises in the above problem in a secondary way when we count the number of occurrences of various events, A_1, A_2 , etc. where $\{A_1, \dots, A_k\}$ is a partition of the sample space.
2. If the underlying distribution is **discrete**, the event A_i usually corresponds to the random variable taking on a particular value in the range space (see MSW, example 14.2). When the underlying distribution is continuous the events $\{A_i\}$ have to be defined by subdividing the range space (see Examples 6.3 and 6.4 that follow). The method of subdivision is not unique but in order for the chisquare approximation to be reasonable the “cell boundaries” should be chosen so that $np_i \geq 5$ for all i . We want enough categories to be able to see what’s happening, but not so many that $np_i < 5$ for any i .

A case can be made for choosing equal-probability categories but this is only one possibility.

3. The fact that the X^2 defined in (6.1) has approximately a chi-square distribution with $k - 1$ degrees of freedom under H_0 , is only correct if the values of the parameters are **specified** in stating the hypothesis. (See MSW, end of 14.2.) If this is not so, a modification has to be made to the degrees of freedom. In fact, we can still say that

$$X^2 = \sum_{i=1}^k \frac{(Y_i - np_i)^2}{np_i} \sim \text{approximately as } \chi^2$$

when the cell probabilities depend on unknown parameters $\theta_1, \theta_2, \dots, \theta_r$ ($r < k$), provided that $\theta_1, \dots, \theta_r$ are replaced by their maximum likelihood estimates and provided that 1 degree of freedom is deducted for each parameter so estimated. That is, with k categories and r parameters estimated, the df would be $k - r - 1$.

4. Note that irrespective of the form of the alternative to the hypothesis only large values of X^2 provide evidence against the hypothesis. The closer the agreement between the expected values and observed values the smaller the values of X^2 . Small values of X^2 thus tend to indicate that the fit is good.

5. In R,

```
> obsv <- c(15,27,18,12,25,23)
> chisq.test(obsv,p=rep(1/6,6) )
```

```
Chi-squared test for given probabilities
data:  obsv
X-squared = 8.8, df = 5, p-value = 0.1173
```

Example 6.1(cont) Let us return to the first example and test the hypothesis, H_0 is $p_1 = p_2 = \dots = \frac{1}{6}$ against the alternative H_1 that $p_i \neq \frac{1}{6}$ for some i .

Solution: The observed frequencies (o_i) and those expected under H (e_i) are

i	1	2	3	4	5	6
o_i	15	27	18	12	25	23
e_i	20	20	20	20	20	20

So the observed value of X^2 is

$$x^2 = \sum_{i=1}^6 \frac{(o_i - e_i)^2}{e_i} = \frac{(-5)^2 + 7^2 + (-2)^2 + (-8)^2 + 5^2 + 3^2}{20} = 8.8.$$

Now the parameters p_1, \dots, p_6 are postulated by the hypothesis as $1/6$ so the df for χ^2 is $6 - 1 = 5$. Under H_0 , $X^2 \sim \chi_5^2$ and the hypothesis would be rejected for large values of x^2 .

The upper 5%ile is $\chi_{5,.05}^2 = 11.1$ (`qchisq(df=5,p=0.05,lower.tail=F)`) so the calculated value is not significant at the 5% level. There is insufficient evidence to cast doubt on the hypothesis so that we conclude the die is most likely unbiased.

Example 6.2

Merchant vessels of a certain type were exposed to risk of accident through heavy weather, ice, fire, grounding, breakdown of machinery, etc. for a period of 400 days. The number of accidents to each vessel, say Y , may be considered as a random variable. For the data reported below, is the assumption that Y has a Poisson distribution justified?

Number of accidents (y)	0	1	2	3	4	5	6
Number of vessels with y accidents	1448	805	206	34	4	2	1

Solution: Note that the parameter λ in the Poisson distribution is not specified and we have to estimate it by its mle, $\hat{\lambda} = \bar{y}$, which is the average number of accidents per vessel. Thus,

$$\begin{aligned}
 \bar{y} &= \frac{\text{total number of accidents}}{\text{total number of vessels}} \\
 &= \frac{(0 \times 1448) + (1 \times 805) + \dots + (5 \times 2) + (6 \times 1)}{1448 + 805 + \dots + 2 + 1} \\
 &= \frac{1351}{2500} \\
 &= .5404.
 \end{aligned}$$

We now evaluate $P(Y = y) = e^{-.5404}(.5404)^y/y!$ for $y = 0, 1, \dots$, to obtain

$$\begin{aligned}
 \hat{p}_0 &= P(Y = 0) = e^{-.5404} = .5825 \\
 \hat{p}_1 &= P(Y = 1) = .5404 \times e^{-.5404} = .3149
 \end{aligned}$$

Similarly, $\hat{p}_2 = .0851$, $\hat{p}_3 = .0153$, $\hat{p}_4 = 0.0021$, $\hat{p}_5 = .00022$, $\hat{p}_6 = .00002$

Recall that the χ^2 approximation is poor if the expected frequency of any cell is less than about 5. In our example, $E(Y_5) = 2500 \times 0.00022 = 0.55$ and $X(Y_6) = 0.05$. This means that the last 3 categories should be grouped into a category called $Y \geq 4$ for which $\hat{p}_4 = P(Y \geq 4) = .0022$.

The expected frequencies (under H_0) are then given by $E(Y_i) = 2500 \times \hat{p}_i$ and are tabulated below.

observed	1448	805	206	34	7
expected	1456.25	787.25	212.75	38.25	5.50

[**Note:** Do not round-off the expected frequencies to integers.]

$$x^2 = \sum \frac{(o - e)^2}{e} = \frac{(1448 - 1456.25)^2}{1456.25} + \dots + \frac{(1.5)^2}{5.5} = 1.54.$$

Since there are 5 categories and we estimated one parameter, the random variable X^2 is distributed approximately as a χ^2_3 . The upper 5% critical value is 7.81,

```
> qchisq(p=0.05,df=3,lower.tail=F)
[1] 7.81
```

so there is no reason to doubt the truth of H_0 and we would conclude that a Poisson distribution does provide a reasonable fit to the data.

Computer Solution: First enter the number of accidents into x and the observed frequencies into `counts`.

```
x <- 0:6
counts <- c(1448,805,206,34,4,2,1)
# Calculate rate
lambda <- sum(x*counts)/sum(counts)
# Merge cells with E(X) < 5
counts[5] <- sum(counts[5:7])
# Poisson probabilities
probs <- dpois(lam=lambda,x=0:4)
# ensure that the probabilities sum to 1, no rounding error
probs[5] <- 1- sum(probs[1:4])
# Chi square test of frequency for Poisson probabilities
chisq.test(counts[1:5],p= probs)
```

Chi-squared test for given probabilities

```
data: counts[1:5]
X-squared = 1.4044, df = 4, p-value = 0.8434
```

Notice the value for x^2 is slightly different. R uses more accurate values for the probabilities and also retains more decimal places for its calculations and so has less rounding error than we managed with a calculator. The conclusions reached are however the same.

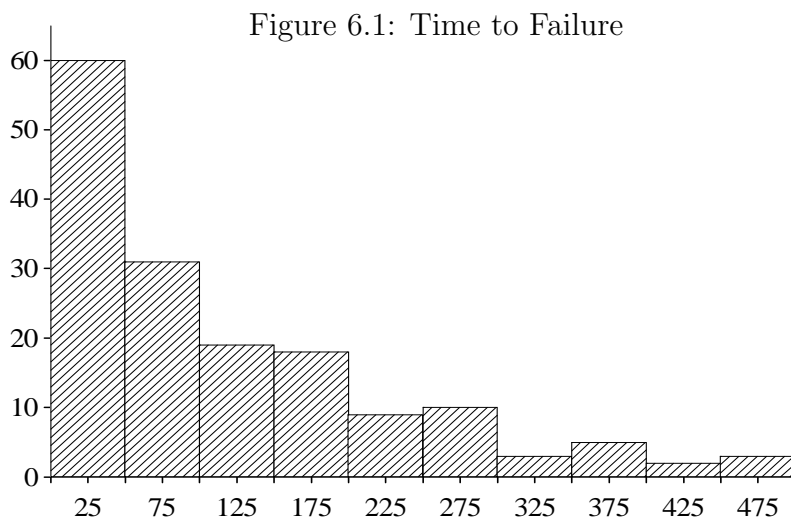
Example 6.3

Let random variable T be the length of life of a certain brand of light bulbs. It is hypothesised that T has a distribution with pdf

$$f(t) = \alpha e^{-\alpha t}, \quad t > 0.$$

Suppose that a 160 bulbs are selected at random and tested, the time to failure being recorded for each. That is, we have t_1, t_2, \dots, t_{150} . Show how to test that the data comes from an exponential distribution.

Solution: A histogram of the data might be used to give an indication of the distribution. Suppose this is as shown below.



The time axis is divided into 10 categories, with cell barriers at 50, 100, \dots , 500, and we might ask what are the expected frequencies associated with these categories, if T does have an exponential distribution. Let p_1, p_2, \dots, p_8 denote the probabilities of the categories, where

$$p_1 = \int_0^{50} \alpha e^{-\alpha t} dt = 1 - e^{-50\alpha}$$

$$p_2 = \int_{50}^{100} \alpha e^{-\alpha t} dt = e^{-50\alpha} - e^{-100\alpha}, \text{ etc.}$$

A value for α is needed to evaluate these probabilities.

- If α is assumed known, (for example it may be specified in the hypothesis where it may be stated that T has an exponential distribution with $\alpha = \alpha_0$) then the df for the χ^2 distribution is $k - 1$ where there are k categories.

- If α is not known it has to be estimated from the data. The mle of α is $1/\bar{t}$ and we use this value to calculate the p_i and hence the e_i . The degrees of freedom now become $k - 2$ since one parameter, α has been estimated.

Computer Exercise 6.1

Suppose 160 lightbulbs are tested with the following results.

Interval	0–50	50–100	100–150	150–200	200–250
Observed	60	31	19	18	9
Interval	250–300	300–350	350–400	400–450	>450
Observed	10	3	5	2	3

Test the hypothesis that the failure times follow an exponential distribution.

Solution: For the raw data, $\bar{t} = \frac{1}{\hat{\lambda}}$. Thus, $\lambda = 1/\bar{t}$, the parameter for the exponential distribution in R . We will approximate \bar{t} by assuming all observations fall at the midpoint of the interval they are in with the last three observations (greater than 450) being assumed to be at 475.

```
#_____ ChisqTestExp.R _____
obsv.freq <- c(60,31,19,18,9,10,3,5,2,3)
total <- sum(obsv.freq)
interval.ends <- seq(from=50,to=500,by=50)
med.times <- interval.ends - 25
tbar <- sum(med.times*obsv.freq)/sum(obsv.freq)
alpha <- 1/tbar
# _____ Cumulative probabilities at interval.ends _____
probs <- pexp(q=interval.ends,rate=alpha);
probs[10] <- 1 # ensure sum(probs)=1
probs.interval <- c(probs[1],diff(probs) ) # first prob is P(0<x<50)
expt.freq <- total*probs.interval
# _____ bulk the low expectation intervals _____
too.low <- expt.freq < 5
probs.interval[7] <- sum(probs.interval[too.low])
obsv.freq[7] <- sum(obsv.freq[too.low])
CHI.test <- chisq.test(obsv.freq[1:7],p=probs.interval[1:7])
P.gt.X2 <- pchisq(df=5,q=CHI.test$statistic,lower.tail=F)
cat("X2 = ",CHI.test$statistic,"          P(Chi > X2) = ",P.gt.X2,"\n")

X2 = 4.2          P(Chi > X2) = 0.52
```

Since α was estimated, X^2 has a chisquare distribution on 5 df and $P(X^2 > 4.2) = 0.52$. Hence it is likely the length of life of the light bulbs is distributed exponentially.

Example 6.4

Show how to test the hypothesis that a sample comes from a normal distribution.

Solution: If the parameters are not specified they must be estimated by

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / n$$

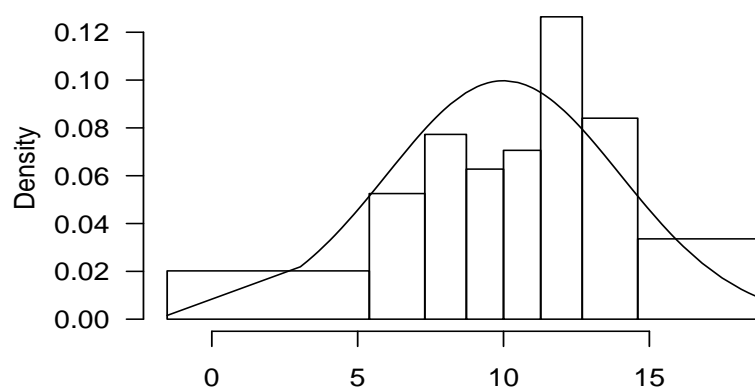
where n is the size of the sample.

The x-axis is then partitioned to form k categories and the number of observations counted in each. A suggested way is as follows

1. Define suitable probabilities for which quantiles can be calculated. These should allow for expected frequencies > 5 so if $n = 100$ (say), then bins corresponding to $p = 0.125, 0.25, \dots, 0.875$ will have sufficient resolution yet retain expected frequencies > 5 .
2. Calculate the quantiles for these probabilities using `qnorm()`.
3. Calculate the observed frequencies for bins defined by the quantiles with `hist(plot=F,)`.
4. Do a χ^2 goodness-of-fit test using the observed frequencies and the proposed probabilities for each bin.

For example, Figure 6.2 depicts a histogram, observed frequencies and the postulated normal distribution. The bins are chosen such that the expected probability under the normal distribution for each bin interval is $\frac{1}{8}$.

Figure 6.2: Partition with Equal Probabilities in Each Category



The following R script implements the above steps.

```
#           read the data
rx <- scan( )      # this bit is virtual
#   proposed normal distribution
xbar <- 10; s <- 4
#   select suitable probabilities
probs <- seq(0.125,0.875,0.125)
#   quantiles corresponding to these probabilities
qx <- qnorm(p=probs,mean=xbar,sd=s)
#           use hist(plot=F, ...) to get observed bin frequencies
#           note use of min(rx) and max(rx) to complete bin definition
histo <- hist(rx,breaks=c(min(rx),qx,max(rx)),plot=F)$counts
obsv.freq <- histo$counts
#   the chi square test
CHI.test <- chisq.test(obsv.freq,p=rep(1/8,8) )
print(CHI.test)
           Chi-squared test for given probabilities
data:  obsv.freq
X-squared = 4.8, df = 7, p-value = 0.6844
```

Since two parameters are estimated, the degrees of freedom for the χ^2 is $k - r - 1 = 8 - 2 - 1 = 5$.

The above output does not take into account that 2 parameters have been estimated so the correct χ^2 test needs to be done.

```
corrected.df <- length(obsv.freq)-3
pchisq(q=CHI.test$statistic,df=corrected.df,lower.tail=F)
X-squared
0.52
```

For this example, $P(\chi^2 > 4.8) = 0.52$.

6.3 Contingency Tables

Suppose a random sample of n individuals is classified according to 2 attributes, say A (rows) and B (columns). We wish to determine whether 2 such attributes are associated or not. Such a two-way table of frequencies is called a contingency table.

Example 6.5

Suppose a random sample of 300 oranges was classified according to **colour** (light, medium, dark) and **sweetness** (sweet or not sweet) then the resulting table

	light	medium	dark	
sweet	115	55	30	200
not sweet	35	45	20	100
	150	100	50	300

is an example of a contingency table.

For such a table it is of interest to consider the hypothesis H_0 : colour and sweetness are independent, against the alternative H_1 : colour and sweetness are associated.

6.3.1 Method

Assume that in the population there is a probability p_{ij} that an individual selected at random will fall in both categories A_i and B_j . The probabilities are shown in the following table.

	B_1	B_2	B_3	Sum
A_1	p_{11}	p_{12}	p_{13}	$p_{1.} = P(A_1)$
A_2	p_{21}	p_{22}	p_{23}	$p_{2.} = P(A_2)$
Sums	$p_{.1}$ $P(B_1)$	$p_{.2}$ $P(B_2)$	$p_{.3}$ $P(B_3)$	$p_{..} = 1$

The probabilities in the margins can be interpreted as follows:

$$p_{i.} = P(\text{item chosen at random will be in category } A_i) = \sum_j p_{ij}$$

$$p_{.j} = P(\text{item chosen at random will be in category } B_j) = \sum_i p_{ij}$$

for $i = 1, 2$ and $j = 1, 2, 3$. If the categories are independent then

$$P(\text{item is in both categories } A \text{ and } B) = P(\text{item in category } A)P(\text{item in category } B)$$

The hypothesis can now be written as

$$H_0 : p_{ij} = p_{i.}p_{.j}, \quad i = 1, 2; \quad j = 1, 2, 3. \quad (6.2)$$

Let the random variable N_{ij} be the number (out of n) in category $A_i \cap B_j$ and n_{ij} be its observed value. Let $n_{i.} = \sum_j n_{ij}$ and $n_{.j} = \sum_i n_{ij}$ so the table of observed frequencies is

n_{11}	n_{12}	n_{13}	$n_{1.}$
n_{21}	n_{22}	n_{23}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..} = n$

Now the set of random variables $\{N_{ij}\}$ have a multinomial distribution and if the p_{ij} are postulated, the expected frequencies will be $e_{ij} = E(N_{ij}) = np_{ij}$ and

$$X^2 = \sum_{l=1}^k \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(N_{ij} - np_{ij})^2}{np_{ij}}.$$

Under the hypothesis of independence, X^2 becomes

$$\sum_{i=1}^2 \sum_{j=2}^3 \frac{(N_{ij} - np_{i.}p_{.j})^2}{np_{i.}p_{.j}}$$

and will be distributed approximately as χ^2 on ν df where $\nu = 6 - 1$.

Usually the $p_{i.}$, $p_{.j}$ are not known and have to be estimated from the data. Now $p_{1.}$, $p_{2.}$, \dots , $p_{.3}$ are parameters in a multinomial distribution and the mle's are

$$\hat{p}_{1.} = n_{1.}/n, \quad \hat{p}_{2.} = n_{2.}/n$$

$$\hat{p}_{.1} = n_{.1}/n, \quad \hat{p}_{.2} = n_{.2}/n, \quad \hat{p}_{.3} = n_{.3}/n.$$

Under H and using the mle's, the expected frequencies take the form

$$e_{ij} = n\hat{p}_{ij} = n \frac{n_{i.}}{n} \frac{n_{.j}}{n}$$

and X^2 becomes

$$X^2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{\left(N_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}. \quad (6.3)$$

Now consider degrees of the freedom. Once three of the expected frequencies e_{ij} , have been determined the other expected frequencies can be determined from the marginal totals since they are assumed fixed. Thus the degrees of freedom is given by $6 - 1 - 3 = 2$.

In the more general case of r rows and c columns, the number of parameters to be estimated is $(r-1) + (c-1)$ so the degrees of freedom is $rc - 1 - (r-1 + c-1) = (r-1)(c-1)$.

Example 6.5(cont)

Test the hypothesis H_0 : colour and sweetness are independent.

Solution: The expected frequencies are:

	light	medium	dark
sweet	100	66.67	33.33
not sweet	50	33.33	16.67

The hypothesis can be stated as: P(sweet orange) is the same whether the orange is light, medium or dark.

Then,

$$x^2 = \sum \frac{(o_i - e_i)^2}{e_i} = \frac{(115 - 100)^2}{100} + \dots + \frac{(20 - 16.67)^2}{16.67} = 13.88.$$

The probability of getting χ^2 at least as large as 13.9 is

```
> pchisq(q=13.9,df=2,lower.tail=F)
[1] 0.00096
```

which indicates that the data suggest strongly ($p < 0.001$) that colour and sweetness are not independent.

Computer Solution: The observed values are entered into an data frame and the **xtabs** command used to make the 2-way table.

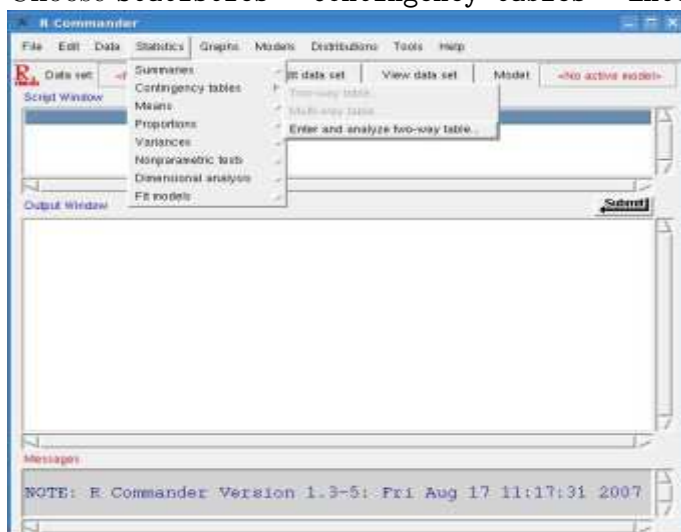
In making this 2-way table, the χ^2 test of independence of rows and columns is also calculated and saved in the summary.

```
#----- Oranges.R -----
Oranges <- expand.grid(sweet=c("Y","N"), colour=c("light","medium","dark"))
Oranges$frequncies <- c(115,35, 55,45, 30,20)
orange.tab <- xtabs(frequncies ~ sweet + colour ,data=Oranges)
print(summary(orange.tab))

      colour
sweet light medium dark
Y      115      55    30
N       35      45    20
Call: xtabs(formula = frequncies ~ sweet + colour, data = Oranges)
Number of cases in table: 300
Number of factors: 2
Test for independence of all factors:
      Chisq = 14, df = 2, p-value = 0.001
```

The Rcmdr menus are also very convenient for getting the χ^2 test of independence for factors in contingency tables.

Choose **Statistics** → **Contingency tables** → **Enter and analyze two-way table**



Change the numbers of rows and columns and provide row and column names.



A script similar to the above is generated and the output is the χ^2 test

```

      light medium dark
sweet    115    55   30
not sweet  35    45   20
> .Test <- chisq.test(.Table, correct=FALSE)
Pearson's Chi-squared test
data:  .Table
X-squared = 13.875, df = 2, p-value = 0.0009707

```

6.4 Special Case: 2×2 Contingency Table

While the 2×2 table can be dealt with as indicated in **6.3** for an $r \times c$ contingency table, it is sometimes treated as a separate case because the x^2 statistic can be expressed in a simple form without having to make up a table of expected frequencies. Suppose the observed frequencies are as follows:

	B_1	B_2	
A_1	a	b	a+b
A_2	c	d	c+d
	a+c	b+d	n

Under the hypothesis that the methods of classification are independent, the expected frequencies are

	B_1	B_2
A_1	$\frac{(a+b)(a+c)}{n}$	$\frac{(a+b)(b+d)}{n}$
A_2	$\frac{(a+c)(c+d)}{n}$	$\frac{(c+d)(b+d)}{n}$

The X^2 statistic is then given by,

$$\begin{aligned}
 x^2 &= \frac{\left(a - \frac{(a+b)(a+c)}{n}\right)^2}{\frac{(a+c)(a+b)}{n}} + \frac{\left(b - \frac{(a+b)(b+d)}{n}\right)^2}{\frac{(a+b)(b+d)}{n}} + \frac{\left(c - \frac{(a+c)(c+d)}{n}\right)^2}{\frac{(a+c)(c+d)}{n}} + \frac{\left(d - \frac{(c+d)(b+d)}{n}\right)^2}{\frac{(c+d)(b+d)}{n}} \\
 &= \frac{(ad - bc)^2}{n} \left\{ \frac{1}{(a+c)(a+b)} + \frac{1}{(a+b)(b+d)} + \frac{1}{(a+c)(c+d)} + \frac{1}{(c+d)(b+d)} \right\} \\
 &= \frac{(ad - bc)^2 \cdot n}{(a+b)(a+c)(b+d)(c+d)}, \text{ on simplification.}
 \end{aligned}$$

[Note that the number of degrees of freedom is $(r-1)(c-1)$ where $r = c = 2$.]

Yates Correction for Continuity

The distribution of the random variables $\{N_{ij}\}$ in a 2×2 contingency table is necessarily discrete whereas the chi-square distribution is continuous. It has been suggested that the approximation may be improved by using as the statistic,

$$X_c^2 = \frac{(|ad - bc| - \frac{1}{2}n)^2 \cdot n}{(a+b)(a+c)(c+d)(b+d)}, \quad (6.4)$$

which is distributed approximately as chi-square on 1 df. The $\frac{1}{2}n$ is known as Yates continuity correction and (6.4) arises by increasing or decreasing the observed values in the contingency table by $\frac{1}{2}$ as follows.

If $ad < bc$ replace a by $a + \frac{1}{2}$, d by $d + \frac{1}{2}$, b by $b - \frac{1}{2}$, c by $c - \frac{1}{2}$

Then we have, for the table of observed frequencies:

$$\begin{array}{cc} a + \frac{1}{2} & b - \frac{1}{2} \\ c - \frac{1}{2} & d + \frac{1}{2} \end{array}$$

If $ad > bc$ the $+$ and $-$ signs are reversed.

Note that the marginal totals are as before. Writing out $\sum (o_{ij} - e_{ij})^2 / e_{ij}$ leads to (6.4). However, there is no general agreement among statisticians that Yates continuity correction is useful as it does not necessarily improve the approximation and may be worse.

Example 6.6

A class of 200 students was classified as in the accompanying frequency table.

Test the hypothesis that the pass rate is the same for males and females.

	Passed	Failed	
Male	70	75	145
Female	35	20	55
	105	95	200

Solution: Now

$$x^2 = \frac{(ad - bc)^2 n}{(a + b)(c + d)(a + c)(b + d)} = \frac{(1400 - 2625)^2 200}{105 \times 95 \times 145 \times 55} = 3.77.$$

But if the continuity correction is used, we get $x_c^2 = 3.2$. Since $P(\chi_1^2 > 3.2) = 0.07$, our result is not significant and we conclude that there is no significant difference between the proportions of male and female students passing the examination.

Computer Solution

```
#----- Class.R -----
Class <- expand.grid(Gender=c("M","F"),Grade=c("P","F"))
Class$freq <- c(70,35, 75,20)
Two.way <- xtabs(freq ~ Gender + Grade,data=Class)

print(chisq.test(Two.way,correct=F))
print(chisq.test(Two.way,correct=T))
```

```
Pearson's Chi-squared test
data: Two.way
X-squared = 3.8, df = 1, p-value = 0.05209
```

```
Pearson's Chi-squared test with Yates' continuity correction
data: Two.way
X-squared = 3.2, df = 1, p-value = 0.07446
```

Note that in all this we assume that n individuals are chosen at random, or we have n independent trials, and then we observe in each trial which of the $r \times c$ events has occurred.

6.5 Fisher's Exact Test

The method for 2×2 contingency tables in 6.4 is really only appropriate for n large and the method described in this section, known as **Fisher's exact test** should be used for smaller values of n , particularly if a number of the expected frequencies are less than 5. (A useful rule of thumb is that no more than 10% of *expected frequencies* in a table should be less than 5 and *no* expected frequency should be less than 1.)

Consider now all possible 2×2 contingency tables with the same set of marginal totals, say $a + c$, $b + d$, $a + b$ and $c + d$, where $a + b + c + d = n$.

	B_1	B_2	
A_1	a	b	a+b
A_2	c	d	c+d
	a+c	b+d	n

We can think of this problem in terms of the hypergeometric distribution as follows. Given n observations which result in $(a + c)$ of type B_1 [and $(b + d)$ of type B_2]; $(a + b)$ of type A_1 [and $(c + d)$ of type A_2], what is the probability that the frequencies in the 4 cells will be

$$\begin{array}{cc} a & b \\ c & d \end{array}$$

This is equivalent to considering a population of size n consisting of 2 types: $(a + c)$ B_1 's and $(b + d)$ B_2 's. If we choose a sample of size $a + b$, we want to find the probability that the sample will consist of a B_1 's and b B_2 's. That is

$$P(a \text{ } B_1 \text{'s, } b \text{ } B_2 \text{'s}) = \frac{\binom{a+c}{a} \binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!}, \quad (6.5)$$

Now if the methods of classification are independent, the expected number of type $A_1 B_1$ is $\frac{(a+b)(a+c)}{n}$. Fisher's exact test involves calculating the probability of the observed set of frequencies and of others more extreme, that is, further from the expected value. The hypothesis H is rejected if the sum of these probabilities is significantly small. Due to the calculations involved it is really only feasible to use this method when the numbers in the cells are small.

Example 6.7

Two batches of experimental animals were exposed to infection under comparable conditions. One batch of 7 were inoculated and the other batch of 13 were not. Of the inoculated group, 2 died and of the other group 10 died. Does this provide evidence of the value of inoculation in increasing the chances of survival when exposed to infection?

Solution: The table of observed frequencies is

	Died	Survived	
Not inoculated	10	3	13
Inoculated	2	5	7
	12	8	20

The expected frequencies, under the hypothesis that inoculation has no effect are

	Died	Survived	
Not inoculated	7.8	5.2	13
Inoculated	4.2	2.8	7
	12	8	20

Note that $e_{21} = 4.2 (= \frac{12 \times 7}{20})$ and the remaining expected frequencies are calculated by subtraction from the marginal totals.

Now the number in row 1, column 1 (10) is greater than the expected value for that cell. Those more extreme in this direction are 11, 12 with the corresponding tables

11	2	12	1
1	6	0	7

Using (6.5) to find the probability of the observed frequencies or others more extreme in the one direction, we have

$$\begin{aligned}
 P &= \sum_{x=10}^{12} \binom{12}{x} \binom{8}{13-x} / \binom{20}{13} \\
 &= \left[\binom{12}{10} \binom{8}{3} + \binom{12}{11} \binom{8}{2} + \binom{8}{1} \right] / \binom{20}{13} \\
 &= \frac{404}{7752} \\
 &= .052.
 \end{aligned}$$

Thus, if H_0 is true, the probability of getting the observed frequencies or others more extreme **in the one direction**, is about 5 in 100. If we wished to consider the alternative as two-sided, we would need to double this probability.

Note that if the chi-square approximation (6.4) is used for a 2×2 contingency table, then that accounts for deviations from expectation in both directions since the deviations are squared. If we had used (6.4) in the above example we would expect to get a probability of about .10. Carrying out the calculations we get $x^2 = 2.65$ and from chi-square tables, $P(W > 2.65)$ is slightly more than 0.10 (where $W \sim \chi_1^2$).

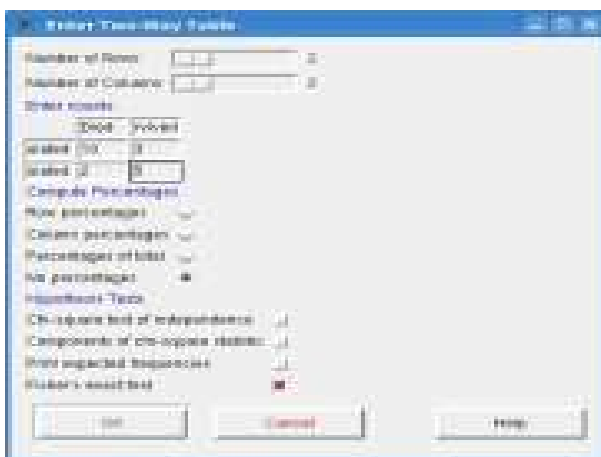
Computer Solution

```
#----- Fisher.R -----
Infection <- expand.grid(Inoculated=c("N","Y"),Survive=c("N","Y") )
Infection$freq <- c(10,2, 3,5)
Ftab <- xtabs(freq ~ Inoculated + Survive,data=Infection)

print(fisher.test(Ftab))
```

Fisher's Exact Test for Count Data

```
data: Ftab
p-value = 0.06233
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.74 117.26
sample estimates:
odds ratio
 7.3
```



6.6 Parametric Bootstrap- X^2

The theme of the previous sections was whether the distribution of observed counts could be considered as random samples from a multinomial distribution with known probabilities and total sample size. The test statistic was a measure of the difference between observed counts and the counts expected from the hypothesised multinomial distribution. This statistic was regarded as coming from a χ^2 distribution,

$$\begin{aligned} X^2 &= \sum \frac{(O - E)^2}{E} \\ X^2 &\sim \chi^2_\nu \end{aligned}$$

We can get a non-parametric estimation of the distribution of X^2 under H_0 and then compare the observed X^2 to decide whether it could have arisen with a reasonable probability ($p > 0.05$) if H_0 were true.

We term this procedure *parametric bootstrap* because the random sample is drawn from a parametric distribution, multinomial in this case, although the distribution of the test statistic X^2 is determined non-parametrically from the bootstrap samples.

The steps are:-

- (i) Calculate the test statistic from the observed data, X^2_{obs} .
- (ii) Determine the probabilities, π_{ij} , under H_0 .
- (iii) Make bootstrap samples and calculate X^2 for each sample.
(for j in $1:\text{nBS}$)
 1. Sample from the multinomial distribution with sample size N and probabilities.
 2. Calculate X_j^2 and record this statistic (i.e. save it in a vector).
- (iv) Plot the empirical distribution function of X^2 and estimate the 95% quantile.
- (v) Compare X^2_{obs} with the distribution of $X^2|H_0$.

Example 6.8

Revisit Example 6.1 where observed counts of faces from 120 throws were:-

i	1	2	3	4	5	6
o_i	15	27	18	12	25	23

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = \pi_5 = \pi_6 = \frac{1}{6}$$

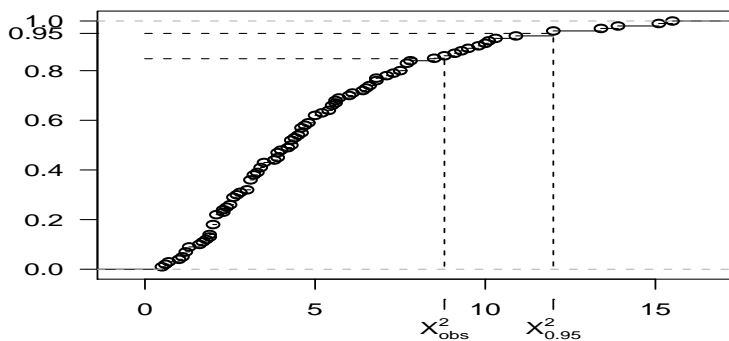
The bootstrap distribution of X^2 is found by

- (i) sampling from a `multinomial(size=120,p=rep(1,6)/6)` distribution and
- (ii) calculating X^2 for each sample.

```
Obs <- c(15,27,18,12,25,23)
Total <- sum(Obs)
p0 <- rep(1,6)/6
X2.obs <- as.numeric(chisq.test(Obs,p=p0)$statistic)
nBS <- 100
X2 <- numeric(nBS)
for (j in 1:nBS){
  simdata <- rmultinom(size=Total,p=p0,n=1)
  X2[j] <- chisq.test(simdata,p=p0)$statistic
} # end of the j loop
Q <- quantile(X2,prob=0.95 )
plot(ecdf(X2),las=1)
```

The results are shown in Figure 6.3. The plot indicates that $P(X^2 > 8.8|H_0) = 0.15$ compared with 0.12 for the χ^2 test.

Figure 6.3: Bootstrap distribution of X^2 for testing unbiasedness of die throws.



Example 6.9

The bootstrap test of independence of factors in a contingency table is illustrated using the data in Example 6.5.

	light	medium	dark
sweet	115	55	30
not sweet	35	45	20

The hypothesis of independence is $H_0 : p_{ij} = p_{i.}p_{.j}$ and the marginal probabilities are estimated from the data,

$$\hat{p}_{i.} = \frac{n_{i.}}{N} \text{ (row probabilities)} \quad \hat{p}_{.j} = \frac{n_{.j}}{N} \text{ (column probabilities)}$$

The cell probabilities can be calculated by matrix multiplication,

$$\begin{bmatrix} \hat{p}_{1.} \\ \hat{p}_{2.} \end{bmatrix} \begin{bmatrix} \hat{p}_{.1} & \hat{p}_{.2} & \hat{p}_{.3} \end{bmatrix} = \begin{bmatrix} \hat{p}_{11} & \hat{p}_{12} & \hat{p}_{13} \\ \hat{p}_{21} & \hat{p}_{22} & \hat{p}_{23} \end{bmatrix}$$

```

rnames <- c("sweet","not sweet")
cnames <- c("light","medium","dark")
Oranges <- expand.grid(Sweetness=rnames,Colour=cnames)
Oranges$counts <- c(115,35, 55,45, 30,20)
Ftab <- xtabs(counts ~ Sweetness + Colour ,data=Oranges)
X2.obs <- summary(Ftab)$statistic
#----- bootstrap -----
nBS <- 1000
X2 <- numeric(nBS)
Total <- sum(Ftab)
col.p <- margin.table(Ftab,margin=2)/Total
row.p <- margin.table(Ftab,margin=1)/Total
prob.indep <- row.p %*% t(col.p) # matrix multiplication of row & column probabilities
for (j in 1:nBS){
simdata <- matrix(rmultinom(prob=prob.indep,size=Total,n=1),nrow=length(rnames),ncol=length(cnames) )
X2[j] <- chisq.test(simdata)$statistic } # end of j loop
Q <- quantile(X2,prob=0.95 )
plot(ecdf(X2),las=1)

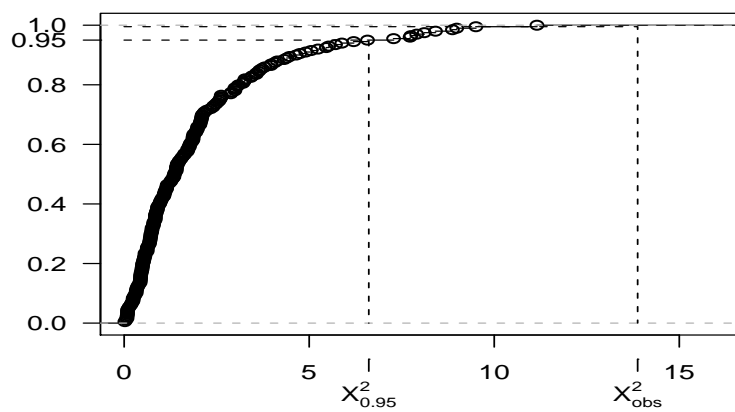
```

Figure 6.4 displays the bootstrap distribution function of $X^2|H_0$ with the observed value of X^2 and the 95%ile. This shows that $P(X^2 > 14|H_0) < 0.01$ as before.

Comment These examples do not show any advantage of the bootstrap over the parametric χ^2 tests. However, understanding of the technique is a platform for Bayesian Monte-Carlo Markov Chain methods (later on).

A Bayesian analysis is not presented here because the setting for that is called *log-linear models* which require some more statistical machinery. This will be encountered in the unit on Linear Models.

Figure 6.4: Bootstrap distribution of X^2 for testing independence of factors in “Oranges” contingency table.



Chapter 7 Analysis of Variance

7.1 Introduction

In chapter 5 the problem of comparing the population means of two normal distributions was considered when it was assumed they had a common (but unknown) variance σ^2 . The hypothesis that $\mu_1 = \mu_2$ was tested using the two sample t -test. Frequently, experimenters face the problem of comparing more than two means and need to decide whether the observed differences among the sample means can be attributed to chance, or whether they are indicative of real differences between the true means of the corresponding populations. The following example is typical of the type of problem we wish to address in this chapter.

Example 7.1

Suppose that random samples of size 4 are taken from three (3) large groups of students studying Computer Science, each group being taught by a different method, and that these students then obtain the following scores in an appropriate test.

Method A	71	75	65	69
Method B	90	80	86	84
Methods C	72	77	76	79

The means of these 3 samples are respectively 70, 85 and 76, but the sample sizes are very small. Does this data indicate a real difference in effectiveness of the three teaching methods or can the observed differences be regarded as due to chance alone?

Answering this and similar questions is the object of this chapter.

7.2 The Basic Procedure

Let μ_1, μ_2, μ_3 be the true average scores which students taught by the 3 methods should get on the test. We want to decide on the basis of the given data whether or not the hypothesis

$$H : \mu_1 = \mu_2 = \mu_3 \text{ against } A : \text{the } \mu_i \text{ are not all equal.}$$

is reasonable.

The three samples can be regarded as being drawn from three (possibly different) populations. It will be assumed in this chapter that the populations are normally distributed and have a common variance σ^2 . The hypothesis will be supported if the sample means are all ‘nearly the same’ and the alternative will be supported if the differences among the sample means are ‘large’. A precise measure of the discrepancies among the \bar{X} ’s is required and the most obvious measure is their variance.

Two Estimates of σ^2

Since each population is assumed to have a common variance the first estimate of σ^2 is obtained by “pooling” s_1^2, s_2^2, s_3^2 where s_i^2 is the i th sample variance. Recalling that we have $\bar{x}_1 = 70, \bar{x}_2 = 85, \bar{x}_3 = 76$, then

$$s_1^2 = \frac{(71 - 70)^2 + (75 - 70)^2 + (65 - 70)^2 + (69 - 70)^2}{3} = \frac{52}{3}.$$

Similarly, $s_2^2 = \frac{52}{3}, s_3^2 = \frac{26}{3}$. Pooling sample variances, we then have

$$s^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2 + \nu_3 s_3^2}{\nu_1 + \nu_2 + \nu_3} = 14.444.$$

Since this estimate is obtained from *within* each individual sample it will provide an unbiased estimate of σ^2 whether the hypothesis of equal means is true or false since it measures variation only within each population.

The second estimate of σ^2 is now found using the sample means. If the hypothesis that $\mu_1 = \mu_2 = \mu_3$ is true then the sample means can be regarded as a random sample from a normally distributed population with common mean μ and variance $\sigma^2/4$ (since $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ where n is the sample size). Then if H is true we obtain,

$$s_{\bar{x}}^2 = \widehat{\text{Var}}(\bar{X}) = \sum_{i=1}^3 (\bar{x}_i - \bar{x})^2 / (3 - 1) = \frac{(70 - 77)^2 + (85 - 77)^2 + (76 - 77)^2}{2}$$

which on evaluation is 57.

But $s_{\bar{x}}^2 = 57$ is an estimate of $\sigma^2/4$, so $4 \times 57 = 228$ is an estimate of σ^2 , the common variance of the 3 populations (provided the hypothesis is true). If the hypothesis is not true and the means are different then this estimate of σ^2 will be inflated as it will also be affected by the difference (spread) between the true means. (The further apart the true means are the larger we expect the estimate to be.)

We now have 2 estimates of σ^2 ,

$$\sum_i \nu_i s_i^2 / \sum_i \nu_i = 14.444 \quad \text{and} \quad n s_{\bar{x}}^2 = 228.$$

If the second estimate (based on variation **between** the sample means) is much larger than the first estimate (which is based on variation **within** the samples, and measures

variation that is due to chance alone) then it provides evidence that the means do differ and H should be rejected. In that case, the variation between the sample means would be **greater** than would be expected if it were due only to chance.

The comparison of these 2 estimates of σ^2 will now be put on a rigorous basis in **7.2** where it is shown that the two estimates of σ^2 can be compared by an F-test. The method developed here for testing $H : \mu_1 = \mu_2 = \mu_3$ is known as **Analysis of Variance** (often abbreviated to AOV).

7.3 Single Factor Analysis of Variance

Consider the set of random variables $\{X_{ij}\}$ where $j = 1, 2, \dots, n_i$ and $i = 1, 2, \dots, k$, and their observed values below, where x_{ij} is the j th observation in the i th group.

Group	Observations	Totals	Means	No. of observations
1	$x_{11} \ x_{12} \ \dots \ x_{1n_1}$	$T_1 = x_{1.}$	$\bar{x}_{1.}$	n_1
2	$x_{21} \ x_{22} \ \dots \ x_{2n_2}$	$T_2 = x_{2.}$	$\bar{x}_{2.}$	n_2
\vdots	\vdots	\vdots	\vdots	\vdots
k	$x_{k1} \ x_{k2} \ \dots \ x_{kn_k}$	$T_k = x_{k.}$	$\bar{x}_{k.}$	n_k

Notation

$$\begin{aligned}
 x_{i.} &= \sum_{j=1}^{n_i} x_{ij} = T_i = \text{Total for } i\text{th group} \\
 \bar{x}_{i.} &= T_i/n_i = \text{mean of the } i\text{th group} \\
 T &= \sum_{i=1}^k T_i = \sum_{i,j} x_{ij} \\
 n &= \sum_{i=1}^k n_i = \text{total number of observations} \\
 \bar{x}_{..} &= \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}/n = T/n = \text{grand mean.}
 \end{aligned}$$

Note:When a *dot* replaces a subscript it indicates summation over that subscript.

In example 7.1 the variation was measured in two ways. In order to do this the total sum of squares of deviations from the (grand) mean must be *partitioned* (that is, split up) appropriately. Theorem 7.1 shows how the total sum of squares can be partitioned.

Theorem 7.1 (Partitioning of the Sum of Squares)

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2 + \sum_{i=1}^k n_i (\bar{x}_{i.} - \bar{x}_{..})^2 \quad (7.1)$$

Proof

$$\begin{aligned} \sum_{i,j} (x_{ij} - \bar{x}_{..})^2 &= \sum_{i,j} (x_{ij} - \bar{x}_{i.} + \bar{x}_{i.} - \bar{x}_{..})^2 \\ &= \sum_{i,j} [(x_{ij} - \bar{x}_{i.})^2 + (\bar{x}_{i.} - \bar{x}_{..})^2 + 2(x_{ij} - \bar{x}_{i.})(\bar{x}_{i.} - \bar{x}_{..})] \\ &= \sum_{i,j} (x_{ij} - \bar{x}_{i.})^2 + \sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2 + 2 \sum_i (\bar{x}_{i.} - \bar{x}_{..}) \underbrace{\sum_j (x_{ij} - \bar{x}_{i.})}_{=0} \end{aligned}$$

Notes on the Partitioning:

1. $SS_T = \sum_{i,j} (x_{ij} - \bar{x}_{..})^2$ = Total sum of squares (of deviations from the grand mean)
2. Notice that $\sum_j (x_{ij} - \bar{x}_{i.})^2$ is just the total sum of squares of deviations from the mean in the i th sample and summing over these from all k groups then gives

$$SS_W = \sum_{i,j} (x_{ij} - \bar{x}_{i.})^2$$

This sum of squares is only affected by the variability of observations *within* each sample and so is called the *within subgroups* sum of squares.

3. The third term is the sum of squares obtained from the deviations of the sample means from the overall (grand) mean and depends on the variability *between* the sample means. That is,

$$SS_B = \sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2$$

and is called the *between subgroups* sum of squares.

4. We can think of Theorem 7.1 as a decomposition of a sum of squares into 2 parts, that is, $SS_T = SS_B + SS_W$, where SS_B results from variation between the groups and SS_W results from variation within the groups. It is these parts or “sources” of variation that will be compared with each other.

Our aim now is to relate equation (7.1) to random variables and the hypothesis-testing problem.

Assume $\{X_{ij}, j = 1, \dots, n_i\}$ are distributed $N(\mu_i, \sigma^2)$ and are mutually independent where $i = 1, 2, \dots, k$. Consider the hypothesis

$$H : \mu_1 = \mu_2 = \dots = \mu_k (= \mu \text{ say}), \quad (7.2)$$

(that is, there is no difference between group means) and the alternative,

$$A : \text{ the } \mu_i \text{ are not all equal.}$$

Note, this is not the same as saying $\mu_1 \neq \mu_2 \neq \dots \neq \mu_k$.

We will now find the probability distributions of SS_T , SS_B , SS_W under H.

Distributions of SS_T , SS_W and SS_B under H

Assume that H is true and that all $n (= \sum_{i=1}^k n_i)$ random variables are normally distributed with mean μ and variance σ^2 (that is, $X_{ij} \sim N(\mu, \sigma^2)$ for all i and j).

- (i) If all the data is treated as a single group coming from the same population then $\sum_{i,j} (x_{ij} - \bar{x}_{..})^2 / (n - 1)$ is an unbiased estimate of σ^2 . Then, using the result from chapter 3 that $\frac{\nu S^2}{\sigma^2} \sim \chi_\nu^2$,

$$\frac{1}{\sigma^2} \sum_{i,j} (X_{ij} - \bar{X}_{..})^2 \sim \chi_{n-1}^2. \quad (7.3)$$

- (ii) If only the i th group is considered then $s_i^2 = \sum_j (x_{ij} - \bar{x}_{i.})^2 / (n_i - 1)$ is also an unbiased estimate of σ^2 . The k unbiased estimates, $s_1^2, s_2^2, \dots, s_k^2$, can then be pooled to obtain another unbiased estimate of σ^2 , that is

$$s^2 = \sum_i \sum_j (x_{ij} - \bar{x}_{i.})^2 / (\sum_i n_i - k).$$

Thus, as in (i) above,

$$\frac{1}{\sigma^2} \sum_{i,j} (X_{ij} - \bar{X}_{i.})^2 \sim \chi_{n-k}^2. \quad (7.4)$$

- (iii) Now $\bar{X}_{i.} \sim N(\mu, \sigma^2/n_i)$, $i = 1, 2, \dots, k$, so that $\sqrt{n_i} \bar{X}_{i.} \sim N(\sqrt{n_i} \mu, \sigma^2)$. Regarding $\sqrt{n_1} \bar{X}_{1.}, \sqrt{n_2} \bar{X}_{2.}, \dots, \sqrt{n_k} \bar{X}_{k.}$ as a random sample from $N(\sqrt{n_i} \mu, \sigma^2)$, the sample variance is $\sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{X}_{..})^2 / (k - 1)$ and is another unbiased estimate of σ^2 . Then, again as in (i) above,

$$\frac{1}{\sigma^2} \sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{X}_{..})^2 \sim \chi_{k-1}^2. \quad (7.5)$$

It can be shown that the random variables in (7.4) and (7.5) are independent. (The proof is not given as it is beyond the scope of this unit.) Thus from (7.1) we have

$$\frac{1}{\sigma^2} \sum_{i,j} (X_{ij} - \bar{X}_{..})^2 = \frac{1}{\sigma^2} \sum_{i,j} (X_{ij} - \bar{X}_{i.})^2 + \frac{1}{\sigma^2} \sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{X}_{..})^2$$

which means we have expressed a chi-square rv on $n - 1$ df as the sum of two *independent* chi-square rv's on $n - k$ and $k - 1$ df's respectively. Note that the degrees of freedom of the rv on the RHS add to the degrees of freedom for the rv's on the LHS.

Thus, from (7.3), (7.4), (7.5) it follows that, **if H is true** we have 3 estimates of σ^2 :

- (i) $\sum_{i,j} (x_{ij} - \bar{x}_{..})^2 / (n - 1)$ which is based on the total variation in the whole sample;
- (ii) $\sum_{i,j} (x_{ij} - \bar{x}_{i.})^2 / (n - k)$ which is based on the variation occurring within each subsample;
- (iii) $\sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2 / (k - 1)$ which is based on the variation of the subsample means from the whole-sample mean.

If H is true, from (7.4) and (7.5) and the definition of F (equation 4.4)

$$\frac{\sum_i n_i (\bar{X}_{i.} - \bar{X}_{..})^2 / (k - 1)}{\sum_{i,j} (X_{ij} - \bar{X}_{i.})^2 / (n - k)} \sim F_{k-1, n-k}.$$

That is if

$$\frac{SS_B / (k - 1)}{SS_W / (n - k)} \sim F_{k-1, n-k}. \quad (7.6)$$

We can summarize the results in an “Analysis of Variance Table” as follows.

Source of Variation	Sum of Squares	df	Mean Square	F
Between gps	$\sum_i n_i (\bar{X}_{i.} - \bar{X}_{..})^2$	$k - 1$	$SS_B / (k - 1)$	$\frac{SS_B / (k-1)}{SS_W / (n-k)}$
Within gps	$\sum_{i,j} (X_{ij} - \bar{X}_{i.})^2$	$n - k$	$SS_W / (n - k)$	
Total	$\sum_{i,j} (X_{ij} - \bar{X}_{..})^2$	$n - 1$		

Note that the term **mean square (ms)** is used to denote (sums of squares)/df.

Method of Computation

In order to calculate the sums of squares in the AOV table it is convenient to express the sums of squares in a different form.

Total SS

$$SS_T = \sum_{i,j} (x_{ij} - \bar{x}_{..})^2 = \sum_{i,j} x_{ij}^2 - \frac{\sum x_{ij}^2}{n} == \sum_{i,j} x_{ij}^2 - \frac{T^2}{n}. \quad (7.7)$$

where

$\sum_{i,j} x_{ij}^2$ is called the *raw* sum of squares and

$\frac{T^2}{n}$ is called the *correction* term.

Between Groups SS

$$SS_B = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{n}. \quad (7.8)$$

since

$$\begin{aligned} SS_B &= \sum_i n_i (\bar{x}_{i.} - \bar{x}_{..})^2 \\ &= \sum_i n_i \bar{x}_{i.}^2 - 2 \bar{x}_{..} \sum_i n_i \bar{x}_{i.} + \underbrace{\bar{x}_{..}^2 \sum_i n_i}_n \\ &= \sum_i n_i \frac{T_i^2}{n_i^2} - 2 \frac{T}{n} \sum_i T_i + n \frac{T^2}{n^2} \\ &= \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{n} \end{aligned}$$

The same “correction term” is used here as appeared in the calculation of SS_T .

Within Groups SS

Since, $SS_T = SS_B + SS_W$, SS_W is found by subtracting SS_B from SS_T . Similarly the df for “within groups” is found by subtracting $k - 1$ from $n - 1$.

Testing the Hypothesis

We have so far considered the distributions of the various sums of squares assuming the hypothesis of equal means to be true. The expected values of these sums of squares are now found **when H is not true**. Recall that

$$\bar{X}_{i.} = \frac{1}{n_i} \sum_j X_{ij}$$

$$\bar{X}_{..} = \frac{1}{n} \sum_{i,j} X_{ij} =$$

The latter can also be written as

$$\bar{X}_{..} = \frac{1}{n} \sum_i n_i \bar{X}_i. \quad (7.9)$$

Note that $E(\bar{X}_i) = \mu_i, i = 1, 2, \dots, k$, and

$$E(\bar{X}_{..}) = \frac{1}{n} \sum_i n_i \mu_i = \bar{\mu}, \text{ say.} \quad (7.10)$$

Theorem 7.2

With SS_W and SS_B as defined earlier.

$$(a) \quad E(SS_W) = (n - k)\sigma^2.$$

$$(b) \quad E(SS_B) = (k - 1)\sigma^2 + \sum_i n_i (\mu_i - \bar{\mu})^2.$$

Proof of (a)

$$\begin{aligned} E(SS_W) &= E \sum_i \sum_j (X_{ij} - \bar{X}_i.)^2 \\ &= \sum_i E \sum_j (X_{ij} - \bar{X}_i.)^2 \\ &= \sum_{i=1}^k E(n_i - 1)S_i^2, \text{ where } S_i^2 \text{ is the sample variance of} \\ &\quad \text{the } i\text{th group, since } X_{ij} \sim N(\mu_i, \sigma^2) \\ &= \sum_{i=1}^k (n_i - 1)\sigma^2 \end{aligned}$$

Thus

$$E \left(\sum_{i,j} (X_{ij} - \bar{X}_i.)^2 \right) = (n - k)\sigma^2. \quad (7.11)$$

Proof of (b)

$$\begin{aligned} E(SS_B) &= E \left[\sum_i n_i (\bar{X}_i. - \bar{X}_{..})^2 \right] \\ &= E \left[\sum_i n_i \bar{X}_i.^2 - 2\bar{X}_{..} \underbrace{\sum_i n_i \bar{X}_i.}_{n\bar{X}_{..}} + n\bar{X}_{..}^2 \right] \\ &= E \left[\sum_i n_i \bar{X}_i.^2 - n\bar{X}_{..}^2 \right] \\ &= \sum_i n_i [\text{Var}(\bar{X}_i.) + (E(\bar{X}_i.))^2] - n[\text{Var}(\bar{X}_{..}) + (E(\bar{X}_{..}))^2] \end{aligned}$$

Now $E(\bar{X}_{i.}) = \mu_i$ and $E(\bar{X}_{..}) = \bar{\mu}$ from (7.10).

Also, $\text{Var}(\bar{X}_{i.}) = \sigma^2/n_i$ and $\text{Var}(\bar{X}_{..}) = \sigma^2/n$. So

$$E(SS_B) = \sum_i n_i \left[\frac{\sigma^2}{n_i} + \mu_i^2 \right] - n \left[\frac{\sigma^2}{n} + \bar{\mu}^2 \right] = k\sigma^2 + \sum_i n_i \mu_i^2 - \sigma^2 - n\bar{\mu}^2.$$

That is,

$$E \left(\sum_i n_i (\bar{X}_{i.} - \bar{X}_{..})^2 \right) = (k-1)\sigma^2 + \sum_i n_i (\mu_i - \bar{\mu})^2 \quad (7.12)$$

Note that sometimes Theorem 7.2 is stated in terms of the expected **mean squares** instead of expected sums of squares.

These results are summarized in the table below.

Source of Variation	Sum of Squares(SS)	df	Mean Square(MS)	E(Mean Square)
Between gps	SS_B	$k-1$	$SS_B/(k-1)$	$\sigma^2 + \frac{\sum_i n_i (\mu_i - \bar{\mu})^2}{k-1}$
Within gps	SS_W	$n-k$	$SS_W/(n-k)$	σ^2
Total	SS_T	$n-1$		

Now if H is true, that is, if $\mu_1 = \mu_2 = \dots = \mu_k$, then $\bar{\mu}$ (as defined in (7.10)) = μ then

$$\sum_i n_i (\mu_i - \bar{\mu})^2 = 0, \text{ and}$$

$$\frac{SS_B/(k-1)}{SS_W/(n-k)} \sim F_{k-1, n-k}.$$

However, if H is not true and the μ_i are not all equal then

$$\sum_i n_i (\mu_i - \bar{\mu})^2 / (k-1) > 0,$$

and the observed value of the F -ratio will tend to be large so that *large* values of F will tend to cast doubt of the hypothesis of equal means. That is if

$$\frac{SS_B/(k-1)}{SS_W/(n-k)} > F_\alpha(k-1, n-k),$$

where $F_\alpha(k-1, n-k)$ is obtained from tables. The significance level α , is usually taken as 5%, 1% or .1%.

Note: The modern approach is to find the probability that the observed value of F (or one larger) would have been obtained by chance under the assumption that the hypothesis is true and use this probability to make inferences. That is find

$$P(F \geq F_{\text{observed}})$$

and if it is small (usually less than 5%) claim that it provides evidence that the hypothesis is false. The smaller the probability the stronger the claim we are able to make. To use this approach ready access to a computer with suitable software is required. With tables we can only approximate this procedure since exact probabilities cannot in general be obtained.

Comments

1. $SS_W/(n - k)$, the “within groups” mean square, provides an unbiased estimate of σ^2 whether or not H is true.
2. When finding the F -ratio in an AOV, the “between groups” mean square *always* forms the **numerator**. This is because its expected value is always greater than or equal to the expected value of the “within groups” mean square (see 7.12). This is one case where one doesn’t automatically put the larger estimate of variance in the numerator. If H is true, both SS_B and SS_W are estimates of σ^2 and in practice either one may be the larger. However small values of F always support the hypothesis so that if $F < 1$ it is always non-significant.

Example 7.2

Suppose we have 4 kinds of feed (diets) and it is desired to test whether there is any significant difference in the average weight gains by certain animals fed on these diets. Twenty (20) animals were selected for the experiment and allocated *randomly* to the diets, 5 animals to each. The weight increases after a period of time were as follows.

Diet	Observations					$T_i = \sum_j x_{ij}$	\bar{x}_i	n_i
A	7	8	8	10	12	45	9.0	5
B	5	5	6	6	8	30	6.0	5
C	7	6	8	9	10	40	8.0	5
D	5	7	7	8	8	35	7.0	5
						$T = 150$		20

Solution: Let the random variable X_{ij} be the weight of the j th animal receiving the i th diet where $X_{ij}, j = 1, \dots, 5 \sim N(\mu_i, \sigma^2)$.

Test the hypothesis that all diets were equally effective, that is

$$H : \mu_1 = \mu_2 = \mu_3 = \mu_4 (= \mu, \text{ say}).$$

Calculations

$$\text{Total SS} = SS_T = \sum_{i,j} x_{ij}^2 - \frac{T^2}{n} = 7^2 + \dots + 8^2 - \frac{150^2}{20} = 63$$

$$\begin{aligned} \text{Between diets SS} &= SS_B = \sum_i (T_i^2/n_i) - \frac{T^2}{n} \quad \text{from (7.8)} \\ &= \frac{45^2}{5} + \frac{30^2}{5} + \frac{40^2}{5} + \frac{35^2}{5} - \frac{150^2}{20} = 25 \end{aligned}$$

$$\text{Within diets SS} = SS_W = 63 - 25 = 38.$$

The Analysis of Variance Table is as below.

Source of Variation	SS	df	MS	F
Between diets	25	3	8.333	3.51*
Within diets	38	16	2.375	
Total	63	19		

The 5% critical value of $F_{3,16}$ is 3.24, so the observed value of 3.51 is significant at the 5% level. Thus there is some reason to doubt the hypothesis that all the diets are equally effective level and conclude that there is a significant difference in weight gain produced by at least one of the diets when compared to the other diets.

Computer Solution:

```
#----- Diets.R -----
Feed <- expand.grid(unit=1:5,Diet=LETTERS[1:4])
Feed$wtgain <- c(7,8,8,10,12,5,5,6,6,8,7,6,8,9,10,5,7,7,8,8)
```

```
weight.aov <- aov(wtgain ~ Diet,data=Feed)
print(summary(weight.aov) )
```

```

      Df Sum Sq Mean Sq F value Pr(>F)
Diet      3   25.0     8.3   3.51  0.04
Residuals 16   38.0     2.4
---
```

R gives a P -value of 0.04 which indicates significance at the 4% level confirming the result ($P < 5\%$) obtained above.

7.4 Estimation of Means and Confidence Intervals

Having found a difference between the means our job is not finished. We want to try and find exactly where the differences are. First we want to estimate the means and their standard errors.

It is useful to find confidence intervals for these means. The best estimate for μ_i , the mean of the i th group, is given by

$$\bar{x}_i = \frac{\sum_j x_{ij}}{n_i}, \quad \text{for } i = 1, 2, \dots, k,$$

where n_i = number observations in the i th group. A $100(1 - \alpha)\%$ confidence interval for μ_i is then

$$\bar{x}_i \pm \frac{s}{\sqrt{n_i}} t_{\nu, \alpha/2}$$

where s^2 is the estimate of σ^2 given by the within groups (residual) mean square (in the AOV table) and is thus on $\nu = n - k$ degrees of freedom.

For straightforward data such as these, the means and their standard errors are calculated with `model.tables()`,

```
print(model.tables(weight.aov,type="means",se=T))
Tables of means
Grand mean
      7.5

Diet
A B C D
9 6 8 7
Standard errors for differences of means
      Diet
      0.9747
replic.      5
```

Note the output specifies that this is the *standard error of the differences of means* where $\text{s.e.m.} = \sqrt{2} \times \text{s.e.}$.

Therefore $\text{s.e.} = \frac{s}{\sqrt{n_i}} = \frac{\text{s.e.m.}}{\sqrt{2}}$. The standard error in the above case is $\frac{0.97}{\sqrt{2}}$ and it is this number which is multiplied by $t_{\nu, \alpha/2}$ to derive confidence limits.

7.5 Assumptions Underlying the Analysis of Variance

The assumptions required for the validity of the AOV procedure are that:

- (i) each of the k samples is from a normal population;
- (ii) each sample can be considered as being drawn randomly from one of the populations;
- (iii) samples are independent of each other;
- (iv) all k populations have a common variance (homogeneity of variance).

If these assumptions are violated then conclusions made from the AOV procedure may be incorrect. We need to be able to verify whether the assumptions are valid.

Assumption (i) may be tested using a chi-square “goodness-of-fit” test (Chapter 6), while careful planning of the experiment should ensure (ii) and (iii) holding.

There are several tests for (iv) three of which follow.

7.5.1 Tests for Equality of Variance

The F-max Test

For a quick assessment of heterogeneity in a set of sample variances we may compute the ratio of the largest sample variance to the smallest. This ratio is referred to as F_{max} . That is, the F_{max} statistic is defined by

$$F_{max} = S_{max}^2 / S_{min}^2.$$

The distribution depends on k and ν_i where k is the number of sample variances being compared and ν_i is the df of the i th s_i^2 . It is not the same as the ordinary F-distribution (except when $k = 2$) which was the ratio of 2 **independent** sample variances. Clearly we'd expect a greater difference between the largest and smallest of k ($= 6$, say) estimates of σ^2 than between 2 chosen at random.

While tables are available we will not use them. If F_{max} is small then it would seem there is no problem with the equality of variance assumption and no further action is required. If F_{max} is large enough to cause doubt then use either Levene's Test or Bartlett's test.

Bartlett's Test

Let $S_1^2, S_2^2, \dots, S_k^2$ be sample variances based on samples of sizes n_1, n_2, \dots, n_k . The samples are assumed to be drawn from normal distributions with variances $\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2$ respectively. Define

$$Q = \frac{(\sum_i \nu_i) \log_e S^2 - \sum_i \nu_i \log_e S_i^2}{1 + \frac{1}{3(k-1)} \left[\sum_i \left(\frac{1}{\nu_i} \right) - \frac{1}{\sum_i \nu_i} \right]} \quad (7.13)$$

where $S^2 = \sum_i \nu_i S_i^2 / \sum \nu_i$.

Then under the hypothesis

$$H : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2,$$

Q is distributed approximately as χ_{k-1}^2 . The approximation is not very good for small n_i .

The hypothesis is tested by calculating Q from (7.13) and comparing it with $w_{k-1, \alpha}$ found from tables of the chi-square distribution.

Example 7.3

Suppose we have 5 sample variances, 15.9, 6.1, 21.0, 3.8, 30.4, derived from samples of sizes 7, 8, 7, 6, 7 respectively. Test for the equality of the population variances.

Solution: $F_{max} = s_{max}^2 / s_{min}^2 = 30.4 / 3.8 = 8.0$.

This is probably large enough to require further checking.

For Bartlett's test, first pool the sample variances to get S^2 .

$$S^2 = \frac{\sum \nu_i S_i^2}{\sum \nu_i} = 15.5167$$

Then from (7.13) we obtain $Q = 7.0571$ which is distributed approximately as a chi-square on 4 df. This is non-significant ($P = 0.13$ using R). Hence we conclude that the sample variances are

compatible and we can regard them as 5 estimates of the one population variance, σ^2 .

It should be stressed that in both these tests the theory is based on the assumption that the k random samples are from normal populations. If this is not true, a significant value of F_{max} or Q may indicate departure from normality rather than heterogeneity of variance. Tests of this kind are more sensitive to departures from normality than the ordinary AOV. Levene's test (which follows) does appear to be robust to the assumption of normality, particularly when medians (instead of means as were used when the test was first proposed) are used in its definition.

Levene's Test

Let $V_{ij} = |X_{ij} - \nu_i|$ where ν_i is the median of the i th group, $i = 1, 2, \dots, k$, $j = 1, 2, \dots, n_i$. That is V_{ij} is the absolute deviation of the j th observation in the i th group from the median of the i th group. To test the hypothesis, $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$ against the alternative they are not all equal we carry out a oneway AOV using the V_{ij} as the data.

This procedure has proven to be quite robust even for small samples and performs well even if the original data is not normally distributed.

Computer Exercise 7.1

Use R to test for homogeneity of variance for the data in Example 7.2.

Solution:

Bartlett's Test

```
print(bartlett.test(wtgain ~ Diet,data=Feed) )
```

```
Bartlett test of homogeneity of variances
```

```
data:  wtgain by Diet
```

```
Bartlett's K-squared = 1.3, df = 3, p-value = 0.7398
```

Levene's Test The first solution derives Levene's test from first principles.

```
#Calculate the medians for each group.
attach(Feed,pos=3)
med <- tapply(wtgain,index=Diet,FUN=median)
med <- rep(med, rep(5,4) )
> med
A A A A B B B B C C C C D D D D
8 8 8 8 6 6 6 6 8 8 8 8 7 7 7 7
# Find v, the absolute deviations of each observation from the group median.
v <- abs(wtgain-med)
# Analysis of variance using v (Levene's Test).
levene <- aov(v~diet)
summary(levene)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diet	3	1.350	0.450	0.3673	0.7776
Residuals	16	19.600	1.225		

There is a function `levene.test()` which is part of the `car` library of R. It would be necessary to download this library from CRAN to use the function.

```
library(car)
print(levene.test(Feed$wtgain,Feed$Diet))
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group 3    0.37  0.78
      16
```

Bartlett's Test, ($P = 0.744$), and Levene's test, ($P = 0.7776$), both give non-significant results, so there appears no reason to doubt the hypothesis, $\sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$.

7.6 Estimating the Common Mean

If an AOV results in a non-significant F-ratio, we can regard the k samples as coming from populations with the same mean μ (or coming from the same population). Then it is desirable to find both point and interval estimates of μ . Clearly the best estimate of μ is $\bar{x} = \sum_{i,j} x_{ij}/n$ where $n = \sum_i n_i$. A $100(1 - \alpha)\%$ confidence interval for μ is

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{\nu, \alpha/2},$$

where s^2 is the estimate of σ^2 given by the “within gps” mean square (in the AOV table) and is thus on $\nu = n - k$ degrees of freedom.

Chapter 8 Simple Linear Regression

8.1 Introduction

Frequently an investigator observes two variables X and Y and is interested in the relationship between them. For example, Y may be the concentration of an antibiotic in a persons blood and X may be the time since the antibiotic was administered. Since the effectiveness of an antibiotic depends on its concentration in the blood, the objective may be to predict how quickly the concentration decreases and/or to predict the concentration at a certain time after administration.

In problems of this type the value of Y will depend on the value of X and so we will observe the random variable Y for each of n different values of X , say x_1, x_2, \dots, x_n , which have been determined in advance and are assumed known. Thus the data will consist of n pairs, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. The random variable Y is called the *dependent* variable while X is called the *independent* variable or the *predictor* variable. (Note this usage of the word independent has no relationship to the probability concept of independent random variables.) It is important to note that in simple linear regression problems the values of X are assumed to be known and so X is not a random variable.

The aim is to find the relationship between Y and X . Since Y is a random variable its value at any one X value cannot be predicted with certainty. Different determinations of the value of Y for a particular X will almost surely lead to different values of Y being obtained. Our initial aim is thus to predict the $E(Y)$ for a given X . In general there are many types of relationship that might be considered but in this course we will restrict our attention to the case of a straight line. That is we will assume that the mean value of Y is related to the value of X by

$$\mu_Y = E(Y) = \alpha + \beta X \quad (8.1)$$

where the parameters α and β are constants which will in general be unknown and will need to be determined from the data. Equivalently,

$$Y = \alpha + \beta X + \epsilon \quad (8.2)$$

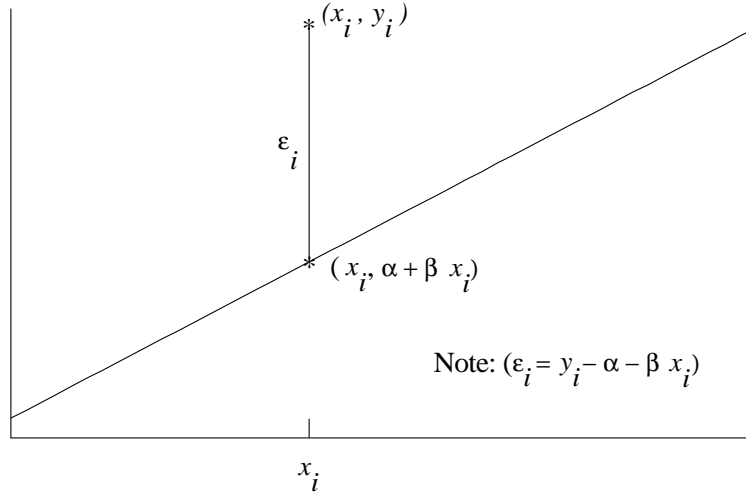
where ϵ is a random variable and is the difference between the observed value of Y and its expected value. ϵ is called the *error* or the *residual* and is assumed to have mean 0 and variance σ^2 for all values of X .

Corresponding to x_i the observed value of Y will be denoted by y_i and they are then related by

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, 2, \dots, n, \quad (8.3)$$

A graphical representation is given in figure 8.1.

Figure 8.1: Simple Linear Regression



Now α and β are unknown parameters and the problem is to estimate them from the sample of observed values $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Two methods of estimation, (least squares and maximum likelihood), will be considered.

8.2 Estimation of α and β .

It is easy to recognize that α is the intercept of the line with the y-axis and β is the slope. A diagram showing the n points $\{(x_i, y_i), i = 1, 2, \dots, n\}$ is called a **scatter plot**, or **scatter diagram**. One simple method to obtain approximate estimates of the parameters is to plot the observed values and draw in roughly the line that best seems to fit the data from which the intercept and slope can be obtained. This method while it may sometimes be useful has obvious deficiencies. A better method is required.

One such method is the method of least squares.

Method of Least Squares

One approach to the problem of estimating α and β is to minimise the sum of squares of the errors from the fitted line. That is the value of

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2 \quad (8.4)$$

is minimised by differentiating with respect to α and β and putting the resulting expressions equal to zero. The results are stated in Theorem 8.1.

Theorem 8.1

The Least Squares Estimates of α and β are

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2} \quad (8.5)$$

$$\hat{\alpha} = \bar{y} - b\bar{x}. \quad (8.6)$$

Proof

For convenience we will rewrite (8.3) as

$$y_i = \alpha' + \beta(x_i - \bar{x}) + \epsilon_i, \quad (8.7)$$

where

$$\alpha = \alpha' - \beta\bar{x}, \quad (8.8)$$

and the minimization will be with respect to α' and β . From (8.7)

$$\sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n [y_i - \alpha' - \beta(x_i - \bar{x})]^2.$$

Thus, taking partial derivative,

$$\begin{aligned} \frac{\partial \sum_{i=1}^n \epsilon_i^2}{\partial \alpha'} &= -2 \sum_{i=1}^n [y_i - \alpha' - \beta(x_i - \bar{x})] \\ \frac{\partial \sum_{i=1}^n \epsilon_i^2}{\partial \beta} &= -2 \sum_{i=1}^n [y_i - \alpha' - \beta(x_i - \bar{x})](x_i - \bar{x}). \end{aligned}$$

Equating the last two expressions to zero we obtain

$$\sum_{i=1}^n [y_i - \hat{\alpha}' - \hat{\beta}(x_i - \bar{x})] = 0 \quad (8.9)$$

$$\sum_{i=1}^n [y_i - \hat{\alpha}' - \hat{\beta}(x_i - \bar{x})](x_i - \bar{x}) = 0 \quad (8.10)$$

where $\hat{\alpha}'$ and $\hat{\beta}$ are the solutions of the equations. Equations (8.9) and (8.10) are referred to as the **normal equations**.

From (8.9) we have

$$\sum_{i=1}^n y_i = n\hat{\alpha}' + \underbrace{\hat{\beta} \sum_{i=1}^n (x_i - \bar{x})}_{=0}.$$

So

$$\hat{\alpha}' = \frac{\sum_i y_i}{n} = \bar{y}.$$

Then from (8.10) we have

$$\sum_{i=1}^n y_i(x_i - \bar{x}) = \hat{\alpha}' \underbrace{\sum_i (x_i - \bar{x})}_{=0} + \hat{\beta} \sum_i (x_i - \bar{x})^2,$$

giving

$$\hat{\beta} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}.$$

Then using (8.8) the estimate of α is

$$\hat{\alpha} = \bar{y} - b\bar{x}.$$

Comments

1. No assumptions about the distribution of the errors, ϵ_i , were made (or needed) in the proof of Theorem 8.1. The assumptions will be required to derive the properties of the estimators and for statistical inference.
2. For convenience, the *least squares estimators*, $\hat{\alpha}$, $\hat{\alpha}'$ and $\hat{\beta}$ will sometimes be denoted by a , a' and b . This should not cause confusion.
3. The estimators of α and β derived here are the Best Linear Unbiased Estimators (known as the BLUE's) in that they are
 - (i) **linear** combinations of y_1, y_2, \dots, y_n ,
 - (ii) **unbiased**;
 - (iii) of all possible linear estimators they are **best** in the sense of having minimum variance.

Method of Maximum Likelihood

Assume that $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, 2, \dots, n$ and that they are mutually independent. Then the Y_i are independent and are normally distributed with means $\alpha + \beta X_i$ and common variance σ^2 .

The likelihood for the errors is given by

$$L = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2} \left(\frac{\epsilon_i}{\sigma}\right)^2\right]}.$$

Since $\epsilon_i = y_i - \alpha' - \beta(x_i - \bar{x})$ the likelihood can be written as

$$L = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2} \left(\frac{y_i - \alpha' - \beta(x_i - \bar{x})}{\sigma}\right)^2\right]}$$

Logging the likelihood gives

$$\log(L) = -n \cdot \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^n \frac{1}{2} \left(\frac{y_i - \alpha' - \beta(x_i - \bar{x})}{\sigma} \right)^2$$

Differentiating $\log(L)$ with respect to α', β and σ^2 and setting the resultant equations equal to zero gives

$$\begin{aligned} \sum_{i=1}^n [y_i - \hat{\alpha}' - \hat{\beta}(x_i - \bar{x})] &= 0 \\ \sum_{i=1}^n (x_i - \bar{x}) [y_i - \hat{\alpha}' - \hat{\beta}(x_i - \bar{x})] &= 0 \\ -n + \frac{\sum_{i=1}^n (y_i - \hat{\alpha}' - \hat{\beta}x_i)^2}{\hat{\sigma}^2} &= 0 \end{aligned}$$

The first two of these equations are just the normal equations, (8.9) and (8.10) obtained previously by the method of least squares. Thus the maximum likelihood estimates of α and β are identical to the estimates (8.5) and (8.6) obtained by the method of least squares. The maximum likelihood estimate of σ^2 ,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\alpha}' - \hat{\beta}x_i)^2}{n}.$$

This estimate is biased.

Comments

1. The **fitted line**, $E(Y) = \hat{\alpha}' + \hat{\beta}(x - \bar{x})$ is called the **regression line** of Y on X .
2. The regression line passes through the point (\bar{x}, \bar{y}) .
3. In our notation we will not distinguish between the random variables $\hat{\alpha}$ and $\hat{\beta}$ and their observed values.
4. The estimate of σ^2 can be written as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n}$$

where

$$e_i = y_i - \hat{\alpha} - \hat{\beta}x_i \tag{8.11}$$

is an estimate of the true error (residual), ϵ .

5. For the purpose of calculation, it is often convenient to use an alternate form of (8.5),

$$\hat{\beta} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}, \tag{8.12}$$

where all sums are over $i = 1, 2, \dots, n$. To verify that these are equivalent, note that

$$\begin{aligned}
 \sum_i (x_i - \bar{x})(y_i - \bar{y}) &= \sum_i (x_i - \bar{x})y_i - \underbrace{\bar{y} \sum_i (x_i - \bar{x})}_{=0} \\
 &= \sum_i x_i y_i - \bar{x} \sum_i y_i \\
 &= \sum_i x_i y_i - \frac{\sum_i x_i \sum_i y_i}{n}.
 \end{aligned}$$

Mean, Variance and Covariance of Regression Coefficients

Theorem 8.2

If $\hat{\alpha}'$ and $\hat{\beta}$ are the least squares estimates of the regression coefficients α' and β then if $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are independently distributed with mean zero and variance σ^2 then,

$$E(\hat{\alpha}') = \alpha' \quad \text{and} \quad E(\hat{\beta}) = \beta, \quad (8.13)$$

$$\text{Var}(\hat{\alpha}') = \sigma^2/n, \quad \text{and} \quad \text{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}, \quad (8.14)$$

$$\text{cov}(\hat{\alpha}', \hat{\beta}) = 0. \quad (8.15)$$

Proof First write $\hat{\alpha}'$ and $\hat{\beta}$ as linear functions of the random variables Y_i . First,

$$\hat{\alpha}' = \frac{1}{n}Y_1 + \frac{1}{n}Y_2 + \dots + \frac{1}{n}Y_n$$

so that

$$E(\hat{\alpha}') = \frac{1}{n} \sum_i E(Y_i) = \frac{1}{n} \sum_i (\alpha' + \beta(x_i - \bar{x})) = \alpha' + \underbrace{\frac{\beta}{n} \sum_i (x_i - \bar{x})}_{=0}.$$

Secondly,

$$\hat{\beta} = \frac{x_1 - \bar{x}}{\sum (x_i - \bar{x})^2} Y_1 + \frac{x_2 - \bar{x}}{\sum (x_i - \bar{x})^2} Y_2 + \dots + \frac{x_n - \bar{x}}{\sum (x_i - \bar{x})^2} Y_n$$

giving

$$\begin{aligned}
 E(\hat{\beta}) &= \frac{1}{\sum (x_i - \bar{x})^2} [(x_1 - \bar{x})(\alpha' + \beta(x_1 - \bar{x})) + \dots + (x_n - \bar{x})(\alpha' + \beta(x_n - \bar{x}))] \\
 &= \frac{1}{\sum (x_i - \bar{x})^2} \left[\alpha' \underbrace{\sum (x_i - \bar{x})}_{=0} + \beta \sum_i (x_i - \bar{x})^2 \right] = \beta.
 \end{aligned}$$

Next,

$$\begin{aligned}\text{Var}(\hat{\alpha}') &= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n} \quad \text{and} \\ \text{Var}(\hat{\beta}) &= \frac{1}{[\sum_i (x_i - \bar{x})^2]^2} [(x_1 - \bar{x})^2 \sigma^2 + \dots + (x_n - \bar{x})^2 \sigma^2] \\ &= \frac{\sigma^2}{[\sum_i (x_i - \bar{x})^2]^2} \sum_i (x_i - \bar{x})^2 = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}.\end{aligned}$$

To show that a' and b are uncorrelated, first write $\epsilon_i = Y_i - E(Y_i)$, then

$$\begin{aligned}\text{cov}(Y_i, Y_j) &= E[(Y_i - E(Y_i))(Y_j - E(Y_j))] \\ &= E(\epsilon_i \epsilon_j) \\ &= \begin{cases} 0 & \text{if } i \neq j, \text{ since the } \epsilon_i \text{ are independent} \\ \sigma^2 & \text{if } i = j. \end{cases}\end{aligned}$$

Since $\hat{\alpha}'$ and $\hat{\beta}$ are linear forms in the Y_i we then have

$$\text{cov}(\hat{\alpha}', \hat{\beta}) = \sigma^2 \left[\frac{1}{n \sum_i (x_i - \bar{x})^2} \underbrace{[(x_1 - \bar{x}) + \dots + (x_n - \bar{x})]}_{=0} \right] = 0.$$

We then have the following corollary to Theorem 8.2.

Corollary 8.2.1

$$E(\hat{\alpha}) = \alpha \text{ and } \text{Var}(\hat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right)$$

Proof: Since $\hat{\alpha} = \alpha' - \hat{\beta}\bar{x}$ and using the results of Theorem 8.2 we have

$$E(\hat{\alpha}) = \alpha' - \beta\bar{x} = \alpha,$$

and

$$\text{Var}(\hat{\alpha}) = \text{Var}(\alpha') + \text{Var}(\hat{\beta}\bar{x}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right)$$

8.3 Estimation of σ^2

Theorem 8.3

Assuming that $E(Y) = \alpha + \beta X$ and $\text{Var}(Y) = \sigma^2$, then

$$\tilde{\sigma}^2 = \frac{1}{(n-2)} \sum_i (y_i - \bar{y} - b(x_i - \bar{x}))^2 \tag{8.16}$$

is an unbiased estimate of σ^2 .

Proof We will need the following:

(i)

$$\begin{aligned}
 \text{Var}(Y_i - \bar{Y}) &= \text{Var} \left[-\frac{Y_1}{n} - \frac{Y_2}{n} - \dots + \left(1 - \frac{1}{n}\right)Y_i - \frac{Y_{i+1}}{n} - \dots - \frac{Y_n}{n} \right] \\
 &= (n-1) \frac{1}{n^2} \sigma^2 + \left(1 - \frac{1}{n}\right)^2 \sigma^2 \\
 &= \left(1 - \frac{1}{n}\right) \sigma^2
 \end{aligned} \tag{8.17}$$

(ii) $Y_i = \alpha + \beta x_i + \epsilon_i$, and

$$\bar{Y} = \frac{\sum_i Y_i}{n} = \alpha + \beta \bar{x} + \frac{\sum_j \epsilon_j}{n} \text{ so that}$$

$$Y_i - \bar{Y} = \beta(x_i - \bar{x}) + \epsilon_i - \frac{\sum_j \epsilon_j}{n}. \text{ Then, } E(Y_i - \bar{Y}) = \beta(x_i - \bar{x}) + 0.$$

To prove the theorem, write

$$\begin{aligned}
 \sum_i (y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}))^2 &= \sum_i (y_i - \bar{y})^2 - 2\hat{\beta} \sum_i (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}^2 \sum_i (x_i - \bar{x})^2 \\
 &= \sum_i (y_i - \bar{y})^2 - 2\hat{\beta}^2 \sum_i (x_i - \bar{x})^2 + \hat{\beta}^2 \sum_i (x_i - \bar{x})^2 \\
 &= \sum_i (y_i - \bar{y})^2 - \hat{\beta}^2 \sum_i (x_i - \bar{x})^2.
 \end{aligned} \tag{8.18}$$

Consider (8.18) in terms of random variables. For the RHS,

$$\begin{aligned}
 E \sum_i (Y_i - \bar{Y})^2 &= \sum_i E(Y_i - \bar{Y})^2 = \sum_i [\text{Var}(Y_i - \bar{Y}) + (E(Y_i - \bar{Y}))^2] \\
 &= \sum_i \left[\left(1 - \frac{1}{n}\right) \sigma^2 + \beta^2 (x_i - \bar{x})^2 \right] \\
 &= (n-1) \sigma^2 + \beta^2 \sum_i (x_i - \bar{x})^2.
 \end{aligned}$$

Also

$$\begin{aligned}
 E(\hat{\beta}^2 \sum_i (x_i - \bar{x})^2) &= \sum_i (x_i - \bar{x})^2 [\text{Var}(\hat{\beta}) + (E(\hat{\beta}))^2] \\
 &= \sigma^2 + \sum_i (x_i - \bar{x})^2 \beta^2.
 \end{aligned}$$

So, from (8.18),

$$E \sum_i (Y_i - \bar{Y} - \hat{\beta}(x_i - \bar{x}))^2 = (n-1) \sigma^2 - \sigma^2 = (n-2) \sigma^2.$$

Thus $\tilde{\sigma}$ given by (8.16) is an unbiased estimate of σ^2 .

8.4 Inference about $\hat{\alpha}$, β and μ_Y

So far we have not assumed any particular probability distribution of the ϵ_i or equivalently, the Y_i . To find confidence intervals for α' , β and μ_Y let us now assume that the ϵ_i are **normally and independently distributed**. (with means 0 and variances σ^2 .) Since

$$Y_i = \alpha' + \beta(x_i - \bar{x}) + \epsilon_i = \alpha + \beta x_i + \epsilon_i = \mu_{Y_i} + \epsilon_i,$$

it follows that the Y_i are independently distributed $N(\alpha + \beta x_i, \sigma^2)$. Since $\hat{\alpha}'$ and $\hat{\beta}$ are linear combinations of the Y_i , and both $\hat{\alpha}$ and $\hat{\mu}_{Y_i}$ are linear combinations of $\hat{\alpha}'$ and $\hat{\beta}$, then each of $\hat{\alpha}'$, $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\mu}_{Y_i}$ are normally distributed. The means and variances of $\hat{\alpha}'$ and $\hat{\beta}$ are given in Theorem 8.2. Means and variances for $\hat{\alpha}$ are given in Corollary 8.2.1.

Now it can be shown that $\hat{\alpha}'$ and $\hat{\beta}$ are independent of $\tilde{\sigma}^2$ given in (8.16), so hypotheses about these parameters may be tested and confidence intervals for them found in the usual way, using the t-distribution. Thus to test the hypothesis, H: $\beta = \beta_0$, we use the fact that:

$$\frac{\hat{\beta} - \beta_0}{\sqrt{\text{Var}(\hat{\beta})}} \sim t_{n-2}.$$

A $100(1 - \alpha)\%$ CI for β is given by:

$$\hat{\beta} \pm t_{n-2, \alpha/2} \sqrt{\text{Var}(\hat{\beta})}$$

Similarly, to test H: $\alpha = \alpha_0$ we use

$$\frac{\hat{\alpha} - \alpha_0}{\sqrt{\text{Var}(\hat{\alpha})}} \sim t_{n-2}.$$

A $100(1 - \alpha)\%$ CI for α can be found using:

$$\hat{\alpha} \pm t_{n-2, \alpha/2} \sqrt{\text{Var}(\hat{\alpha})}$$

For $\hat{\mu}_{Y_i}$

$$E(\hat{\mu}_{Y_i}) = \alpha' + \beta(x_i - \bar{x}) \quad (8.19)$$

and,

$$\begin{aligned} \text{Var}(\hat{\mu}_{Y_i}) &= \text{Var}(\hat{\alpha}') + (x_i - \bar{x})^2 \text{Var}(\hat{\beta}) \quad \text{since } \text{cov}(\hat{\alpha}', \hat{\beta}) = 0 \\ &= \frac{\sigma^2}{n} + \frac{(x_i - \bar{x})^2 \sigma^2}{\sum_i (x_i - \bar{x})^2} \\ &= \sigma^2 \left\{ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right\} \end{aligned} \quad (8.20)$$

so that a $100(1 - \alpha)\%$ confidence interval for it is given by $\hat{\mu}_{Y_i} \pm t_{n-2, \alpha/2} \sqrt{\text{Var}(\hat{\mu}_{Y_i})}$. That is

$$\hat{\alpha}' + \hat{\beta}(x - \bar{x}) \pm t_{n-2, \alpha/2} \tilde{\sigma} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}}. \quad (8.21)$$

Comment

Notice that in (8.18), $y_i - \bar{y} - \hat{\beta}(x_i - \bar{x}) = e_i$ is an estimate of the (true) error, so the first term is called the Error Sum of Squares (Error SS). The first term on the right hand side, $\sum (y_i - \bar{y})^2$, is the total variation in y (Total SS), and $\hat{\beta}^2 \sum (x_i - \bar{x})^2$ is the Sum of Squares due to deviations from the regression line, which we will call the Regression SS. Thus in words, (8.18) can be expressed in the form

$$\text{Error SS} = \text{Total SS} - \text{Regression SS}.$$

This information can be summarised in an Analysis of Variance Table, similar in form to that used in the single factor analysis of variance.

Source	df	SS	MS
Regression	1	$\hat{\beta}^2 \sum (x_i - \bar{x})^2$	$\hat{\beta}^2 \sum (x_i - \bar{x})^2$
Error SS	$n - 2$	$\sum (y_i - \bar{y})^2 - \hat{\beta}^2 \sum (x_i - \bar{x})^2$	$\frac{1}{n - 2} \left[\sum (y_i - \bar{y})^2 - \hat{\beta}^2 \sum (x_i - \bar{x})^2 \right]$
Total	$n - 1$	$\sum (y_i - \bar{y})^2$	

It can be shown that the ratio of the Regression MS to the Error MS has an F distribution on 1 and $n - 2$ df and provides a test of the hypothesis, $H: \beta = 0$ which is equivalent to the t test above.

Question: Why should you expect these two tests to be equivalent?

Example 8.1

The following data refer to age (x) and bloodpressure (y) of 12 women.

x	56	42	72	36	63	47	55	49	38	42	68	60
y	147	125	160	118	149	128	150	145	115	140	152	155

Assuming that Y has a normal distribution with mean $\alpha + \beta x$ and variance σ^2 ,

- find the least squares estimates in the regression equation;
- test the hypothesis that the slope of the regression line is zero;
- find 95% confidence limits for the mean blood pressure of women aged 45.

Solution: We have $\sum x = 628$, $\sum y = 1684$, $\sum xy = 89894$, $\sum x^2 = 34416$, $\sum y^2 = 238822$, and $n = 12$.

- Regression coefficients in the equation $\hat{\mu}_Y = \hat{\alpha}' + \hat{\beta}(x - \bar{x})$ are given by

$$\begin{aligned} \hat{\alpha}' &= \bar{y} = 1684/12 = 140.33 \\ \hat{\beta} &= \frac{89894 - \frac{1057552}{12}}{34416 - \frac{394384}{12}} = \frac{1764.667}{1550.667} = 1.138. \end{aligned}$$

The regression equation is

$$\hat{\mu}_Y = 140.33 + 1.138(x - 52.333) = 80.778 + 1.138x.$$

(ii) To test the hypothesis $\beta = 0$ we need to calculate

$$\begin{aligned}\sum (y_i - \bar{y})^2 &= \sum y^2 - \frac{(\sum y)^2}{n} = 238822 - 236321.33 = 2500.67 \\ \hat{\beta}^2 \sum_i (x_i - \bar{x})^2 &= 1.295 \times 1550.667 = 2008.18 \\ \tilde{\sigma}^2 &= \frac{2500.67 - 2008.19}{10} = 49.26.\end{aligned}$$

Hence

$$\frac{\hat{\beta} - 0}{\text{estimated sd of } b} = \frac{1.138}{\sqrt{49.25}/\sqrt{1550.667}} = 6.4.$$

Comparing this with the critical value from the t-distribution on 10 degrees of freedom, we see that our result is significant at the .1% level.

(iii) For $x = 45$, $\hat{\mu}_Y = 80.778 + 1.138 \times 45 = 132.00$. Now the 95% confidence limits for the **mean** blood pressure of women aged 45 years is

$$132.00 \pm 2.228 \sqrt{49.25 \left\{ \frac{1}{12} + \frac{(45 - 52.33)^2}{1550.67} \right\}} = 132.00 \pm 5.37.$$

Computer Solution: Assuming the data for age (x) and blood pressure (y) is in the text file, bp.txt,

	#----- bp.R -----
	options(digits=2)
	bp <- read.table("bp.txt",header=T)
	bp.lm <- lm(bloodpressure ~ age,data=bp)
age bloodpressure	
56 147	
42 125	print(anova(bp.lm))
72 160	
36 118	bp.summ <- summary(bp.lm)
63 149	print(bp.summ)
47 128	print(confint(bp.lm))
55 150	
49 145	VB <- bp.summ\$sigma^2 * bp.summ\$cov.unscaled
38 115	print(VB)
42 140	print(sqrt(diag(VB)))
68 152	
60 155	newdata <- data.frame(age=c(45,60))
	preds <- predict(bp.lm,new=newdata,interval="confidence")
	newdata <- cbind(newdata,preds)
	print(newdata)

In order the outputs to be interpreted are

1. AOV

```
print(anova(bp.lm))
Analysis of Variance Table
```



```

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
age      1 2008.20  2008.20  40.778 7.976e-05 ***
Residuals 10  492.47    49.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

2. We extract the coefficients for the regression line by using the `summary()` command.

```

> summary(bp.lm)

Call:
lm(formula = bloodpressure ~ age, data = bp)

Residuals:
    Min       1Q   Median       3Q      Max
-9.02  -4.35  -3.09   6.11  11.43

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   80.778     9.544    8.46 7.2e-06 ***
age           1.138     0.178    6.39 8.0e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 7 on 10 degrees of freedom
Multiple R-Squared:  0.803,    Adjusted R-squared:  0.783
F-statistic: 40.8 on 1 and 10 DF,  p-value: 7.98e-05

```

You will notice that the `summary` command provides us with t -tests of the hypotheses $\alpha = 0$ and $\beta = 0$ as well as the residual standard error (s) and R^2 . The F -test of the hypothesis $\beta = 0$ is also reported and of course is identical to the F -test in the AOV table.

3. Confidence intervals for the regression coefficients

```

print(confint(bp.lm))

              2.5 % 97.5 %
(Intercept) 59.51 102.0
age         0.74  1.5

```

4. the variance-covariance of the regression coefficients may be needed for further work.

```

VB <- bp.summ$sigma^2 * bp.summ$cov.unscaled
print(VB)

```

```
print(sqrt(diag(VB)) )

              (Intercept)      age
(Intercept)      91.1 -1.662
age              -1.7  0.032
```

A quick check shows the connection between the variance-covariance matrix of the regression coefficients and the standard errors of the regression coefficients.

The diagonal of the matrix is (91.1, 0.032) and the square root of these numbers gives (9.54, 0.18) which are the s.e.'s of the regression coefficients.

5. We can now use our model to predict the blood pressure of 45 and 60 year old subjects. When the model is fitted, there are estimates of the regression coefficients, i.e. $\hat{\alpha} = 80.8$ and $\hat{\beta} = 1.14$. Given a new value of x (say $x = 45$), the predicted value is $\hat{y} = 80.8 + 1.14 \times 45$. The standard error and CI for this predicted value is also able to be calculated.

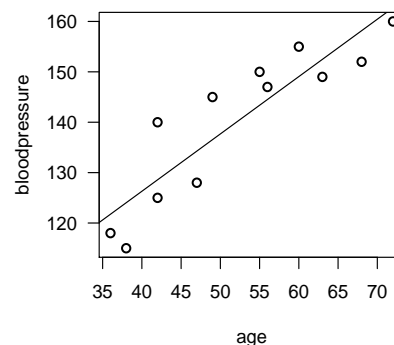
This is achieved by supplying a 'new' data frame of explanatory variables and calculating predictions with `predict()` and appropriate arguments.

```
newdata <- data.frame(age=c(45,60) )
preds <- predict(bp.lm,new=newdata,interval="confidence")
newdata <- cbind(newdata,preds)
print(newdata)
  age fit lwr upr
1  45 132 127 137
2  60 149 144 155
```

(You should make sure you can match this output up with the calculations made in example 8.1. For example $\tilde{\sigma}^2 = 49.2 = \text{Error MS}$. Also, from the AOV table, the F value is $40.78 = 6.39^2$, where 6.39 is the value of the t statistic for testing the hypothesis $\beta = 0$.)

The fitted model and the code that does it looks like this:-

```
plot(bloodpressure ~ age,data=bp,las=1)
abline(lsfilt(bp$age,bp$bloodpressure))
```



8.5 Correlation

Recall that in a bivariate normal distribution the correlation coefficient, ρ is defined by

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \text{ and } -1 \leq \rho \leq 1.$$

In practice, ρ is an unknown parameter and has to be estimated from a sample. Consider $(x_1, y_1), \dots, (x_n, y_n)$ as a random sample of n pairs of observations and define

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}. \quad (8.22)$$

This is called Pearson's correlation coefficient, and it is the maximum likelihood estimate of ρ in the bivariate normal distribution.

We will consider testing the hypothesis $H : \rho = 0$ against 1-sided or 2-sided alternatives, using r . It can be shown that, if H is true, and a sample of n pairs is taken from a bivariate normal distribution, then

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}. \quad (8.23)$$

Alternatively, a table of percentage points of the distribution of Pearson's correlation coefficient r when $\rho = 0$, may be used.

Example 8.2

Suppose a sample of 18 pairs from a bivariate normal distribution yields $r = .32$, test $H_0 : \rho = 0$ against $H_1 : \rho > 0$.

Solution: Now

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.32 \times 4}{\sqrt{1-.1024}} = 1.35.$$

The probability of getting a value at least as large as 1.35 is determined from the t-distribution on 18 degrees of freedom,

```
> pt(df=16,q=1.35,lower.tail=F)
[1] 0.098
```

so there is no reason to reject H .

Example 8.3

Suppose a sample of 10 pairs from a bivariate normal distribution yields $r = .51$. Test $H : \rho = 0$ against $A : \rho > 0$.

Solution: The critical value of t is

```
> qt(df=8,p=0.05,lower.tail=F)
[1] 1.9
```

A value of r which leads to $t < 1.9$ will be interpreted as insufficient evidence that $\rho > 0$.

The critical value of r is found by inverting equation (8.23), $r = \frac{t}{\sqrt{d+t^2}}$. The 5% (for a 1-tailed test) critical value of r is .55. Our observed value is not as large as this, so the hypothesis of zero correlation cannot be rejected at the 5% significance level.

Computer Exercise 8.1

Use R to find the correlation between the age and bloodpressure of the 12 women in Example 8.1 and test the hypothesis, $H_0 : \rho = 0$ against the alternative, $H_1 : \rho \neq 0$.

Solution: To find the correlation:

```
> cor(age,bp)
[1] 0.8961394
```

Notice that in the regression output, $R^2 = .803$ which is just 0.896^2 . Now calculate

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{.896\sqrt{10}}{1-0.896^2} = 6.39$$

Notice this is exactly the same t -value as obtained in the R output (and in Example 8.1) for testing the hypothesis, $H:\beta = 0$. **These tests are equivalent.** That is if the test of $\beta = 0$ is significant (non-significant), the test of $\rho = 0$ will also be significant (non-significant) with exactly the same P -value.