MA20226: Statistics 2A

Simon Shaw
s.shaw@bath.ac.uk
2011/12 Semester I

Contents

List of Figures

Introduction

In this course, we shall be interested in modelling some stochastic system by a random quantity X with outcomes $x \in \Omega$, where Ω denotes the sample, or outcome, space.

If Ω is finite (or countable) then X is DISCRETE and we write $P(X = x)$ to denote the probability that the random quantity X is equal to the outcome x. Some examples of discrete random quantities are as follows.

Example 1 *Bernoulli: takes only two values:* 0 *and* 1*, so* $\Omega = \{0, 1\}$ *and*

$$P(X = x) \quad = \quad p^x(1-p)^{1-x}.$$

Example 2 *Geometric: series of independent trials, each trial is a "success" with probability* p *and* X *measures the total number of trials up to and including the first success. Thus,* $\Omega = \{1, 2, \ldots\}$ *and*

$$P(X = x) \quad = \quad p(1-p)^{x-1}.$$

Example 3 *Binomial:* n *independent trials, each trial is either a "success" with probability* p *or a "failure" with probability* $1 - p$*.* X *measures the total number of successes, so* $\Omega = \{0, 1, \ldots, n\}$ *and*

$$P(X = x) \quad = \quad \left(\begin{array}{c} n \\ x \end{array} \right) p^x(1-p)^{n-x}.$$

The random quantity X is CONTINUOUS if it can take any value in some (finite or infinite) interval of real numbers. We denote its probability density function (pdf) by $f(x)$. Some examples of continuous random quantities are as follows.

Example 4 *Uniform:* $\Omega = [a, b]$*, with* $a < b$*. This distribution may be thought of as a model for choosing a number at random between* a *and* b*. The pdf is given by*

$$f(x) \quad = \quad \left\{ \begin{array}{cc} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \textit{otherwise.} \end{array} \right.$$

Example 5 *Exponential: often used to model lifetimes or waiting times.* $\Omega = [0, \infty)$ *with*

$$f(x) \quad = \quad \begin{cases} \lambda \exp(-\lambda x) & x \geq 0, \\ 0 & \textit{otherwise.} \end{cases}$$

Example 6 *Normal: plays a central role in statistics, also known as the Gaussian distribution.* $\Omega = (-\infty, \infty)$ *and*

$$f(x) \quad = \quad \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}.$$

Each of these examples denotes a FAMILY OF DISTRIBUTIONS, varying in scale/location/ shape. The family is indexed by some parameter $\theta \in \Theta$; θ may be univariate or multivariate.

- The Bernoulli and Geometric, Examples 1 and 2, are both indexed by the parameter $p \in (0, 1)$.
- The Binomial, Example 3, has two parameters: $p \in (0, 1)$ and n, the sample size which is a positive integer. In most cases, n is known.
- In Example 4, the Uniform has two parameters a and b where $a < b$ are real numbers.
- The Exponential, Example 5, has a single parameter, $\lambda \in (0, \infty)$.
- Finally, in Example 6, the Normal has two: $\mu \in (-\infty, \infty)$ and $\sigma^2 \in (0, \infty)$.

We could choose to make this "family behaviour" explicit by writing $P_\theta(X = x)$ or $P(X = x \,|\, \theta)$ for discrete distributions and $f_\theta(x)$ and $f(x \,|\, \theta)$ in the continuous setting. In this course, we shall use the second of these two conventions. This choice also avoids confusion when we wish to make the random quantity over which the distribution is specified explicit.

Example 7 *If* $X \sim N(\mu, \sigma^2)$ *then* $X = \sigma Z + \mu$ *where* $Z \sim N(0, 1)$ *and we write*

$$f_X(x \,|\, \mu, \sigma^2) \quad = \quad \frac{1}{\sigma} f_Z(\tfrac{x-\mu}{\sigma})$$

where

$$f_Z(z) \quad = \quad \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{z^2}{2} \right\}.$$

Often θ may be unknown and learning about it, that is making INFERENCES about it, is the key issue. The course begins by discussing how we can estimate θ based upon a sample $x_1, \ldots, x_n$ of observations believed to come from the underlying distribution $f(x \,|\, \theta)$. That is, we are interested in PARAMETRIC INFERENCE assuming a known particular family.

Chapter 1

Point Estimation

1.1 Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed random quantities. Each X_i has probability density function $f_{X_i}(x \mid \theta) = f(x \mid \theta)$ if continuous and probability mass function $P(X_i = x \mid \theta)$ if discrete, where $\theta \in \Theta$ is a parameter indexing a family of distributions.

Example 8 *We are interested in whether a coin is fair. If we judge tossing a head to be a success then each toss of the coin constitutes a Bernoulli trial with parameter p unknown. The coin is fair if p equals one half.*

We take a sample of observations $x_1, \ldots, x_n$ where each $x_i \in \Omega$ and our aim is to determine from this data a number that can be taken to be an estimate of θ.

1.2 Estimators and estimates

It is important to distinguish between the method or rule of estimation, which is the ESTIMATOR, and the value to which it gives rise to in a particular case, the ESTIMATE.

Example 9 *Recall Example 8. The parameter p is the probability of tossing a head on a single toss. Suppose we observe n tosses, so* $X_1 = x_1, \ldots, X_n = x_n$ *where*

$$x_i = \begin{cases} 1 & \text{if } i\text{th toss is a head,} \\ 0 & \text{if } i\text{th toss is a tail.} \end{cases}$$

The observed sample mean

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

is an intuitive way to estimate p. Prior to observing the data,

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

is unknown and is thus a random quantity. It is the estimator of p.

Definition 1 *(Estimator, Estimate)*
The estimator for θ is a function of the random quantities, $X_1, \ldots, X_n$, denoted $T(X_1, \ldots, X_n)$, and is thus itself a random quantity. For a specific set of observations $x_1, \ldots, x_n$, the observed value of the estimator, $T(x_1, \ldots, x_n)$, is the point estimate of θ.

Our aim in this course is to find estimators: we do not choose/reject an estimator because it gives a good/bad result in a particular case. Rather, we should choose an estimator that gives good results in the long run. In particular, we look at properties of its SAMPLING DISTRIBUTION. The sampling distribution is the probability distribution of the estimator $T(X_1, \ldots, X_n)$.

Example 10 *Recall Examples 8 and 9. We observe $X_1 = x_1, \ldots, X_n = x_n$. Prior to observing the data, the probability, or likelihood, of the data taking this form is*

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n \,|\, p) &= \prod_{i=1}^{n} P(X_i = x_i \,|\, p) \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \\ &= p^{n\overline{x}}(1-p)^{n-n\overline{x}}. \end{aligned}$$

Note that to calculate this probability, n and $\overline{x}$ are sufficient for $x_1, \ldots, x_n$. If p is unknown, we could regard this probability as a function of p,

$$L(p) \;=\; P(X_1 = x_1, \ldots, X_n = x_n \,|\, p) \;=\; p^{n\overline{x}}(1-p)^{n-n\overline{x}}. \tag{1.1}$$

We could then choose p to make $L(p)$ as large as possible, that is to maximise the probability, or likelihood, of observing the data. This is the method of MAXIMUM LIKELIHOOD ESTIMATION.

1.3 Maximum likelihood estimation

The principle of maximum likelihood estimation is to choose the value of θ which makes the observed set of data most likely to occur. For ease of notation, we denote the probability density function by $f(x \,|\, \theta)$ indifferently for a discrete or continuous random quantity. The joint probability distribution, $f_{X_1 \ldots X_n}(x_1, \ldots, x_n \,|\, \theta)$, of the n random quantities $X_1, \ldots, X_n$ is , for fixed θ, a function of the observations, $x_1, \ldots, x_n$. If θ is unknown but the observations known then we may regard it as a function of θ.

Definition 2 *(Likelihood function)*
The joint probability distribution of $X_1, \ldots, X_n$,

$$L(\theta) \;=\; f_{X_1 \ldots X_n}(x_1, \ldots, x_n \,|\, \theta) \;=\; \prod_{i=1}^{n} f(x_i \,|\, \theta), \tag{1.2}$$

regarded as a function of θ *for given observations* $x_1, \ldots, x_n$ *is the likelihood function.*

Note that the simplification in equation (1.2) follows as the X_i are independent and identically distributed.

Definition 3 *(Maximum likelihood estimate/estimator)*
For a given set of observations, the maximum likelihood estimate is the value $\hat{\theta} \in \Theta$ *which maximises* $L(\theta)$. *If each* $x_1, \ldots, x_n$ *leads to a unique value of* $\hat{\theta}$, *then the procedure defines a function*

$$\hat{\theta} \;=\; T(x_1, \ldots, x_n)$$

and the corresponding random quantity $T(X_1, \ldots, X_n)$ *is the maximum likelihood estimator.*

Example 11 *Recall the coin tossing example, see Example 10, which utilises the Bernoulli distribution. From equation (1.1), the likelihood function is*

$$L(\theta) \;=\; L(p) \;=\; p^{n\overline{x}}(1-p)^{n-n\overline{x}}.$$

Notice that $L(0) = 0 = L(1)$ *and that the likelihood is always nonnegative as it is a probability and continuous. Thus,* $L(p)$ *has a maximum in* $(0, 1)$. *Differentiating* $L(p)$ *with respect to* p *gives*

$$\begin{aligned} L'(p) &= \{n\overline{x}(1-p) - (n - n\overline{x})p\}p^{n\overline{x}-1}(1-p)^{n-n\overline{x}-1} \\ &= n(\overline{x} - p)p^{n\overline{x}-1}(1-p)^{n-n\overline{x}-1}. \end{aligned}$$

Hence, solving $L'(p) = 0$ *for* $p \in (0, 1)$ *gives* $\hat{p} = \overline{x}$. *The maximum likelihood estimate is*

$$\hat{p} \;=\; T(x_1, \ldots, x_n) \;=\; \overline{x}.$$

The maximum likelihood estimator is

$$T(X_1, \ldots, X_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i \;=\; \overline{X}.$$

The sampling distribution of the maximum likelihood estimator is easy to obtain as $\sum_{i=1}^{n} X_i \sim Bin(n, p)$. *Thus,*

$$P(\overline{X} = x \,|\, p) \;=\; \binom{n}{nx} p^{nx}(1-p)^{n-nx},$$

for $x = 0, 1/n, 2/n, \ldots, 1$.

We now generalise the approach this example, firstly for the case when θ is univariate and then for the multivariate case.

1.3.1 Maximum likelihood estimation when θ is univariate

If $L(\theta)$ is a twice-differentiable function of θ then stationary values of θ will be given by the roots of

$$L'(\theta) \;=\; \frac{\partial L}{\partial \theta} \;=\; 0.$$

A sufficient condition that a stationary value, $\hat{\theta}$, be a local maximum is that

$$L''(\hat{\theta}) < 0.$$

The maximum likelihood estimate is the point at which the global maximum is attained which is either at a stationary value, or is at an extreme permissible value of θ. An example of this latter case can be seen in Question 2 of Question Sheet Two.

In practice, it is often simpler to work with the LOG-LIKELIHOOD,

$$l(\theta) \quad = \quad \log L(\theta).$$

Note that, as the logarithm is a monotonically increasing function, the likelihood, $L(\theta)$, and the log-likelihood, $l(\theta)$, have the same maxima.

Alternatively note that

$$l'(\theta) \quad = \quad \frac{L'(\theta)}{L(\theta)}$$

and $L(\theta) > 0$ so that $l(\theta)$ and $L(\theta)$ have the same stationary points. If $\hat{\theta}$ is such a stationary point then $l''(\hat{\theta}) = L''(\hat{\theta})/L(\hat{\theta})$ so $L''(\hat{\theta})$ and $l''(\hat{\theta})$ share the same sign.

The principle reason for working with the log-likelihood is that the differentiation will typically be easier as it avoids having to differentiate a product. From equation (1.2), the likelihood function is

$$L(\theta) \quad = \quad \prod_{i=1}^{n} f(x_i \,|\, \theta).$$

Noting that the 'log of a product is the sum of the logs' then the log-likelihood is given by

$$l(\theta) \quad = \quad \sum_{i=1}^{n} \log\{f(x_i \,|\, \theta)\},$$

so that

$$l'(\theta) \quad = \quad \sum_{i=1}^{n} \frac{f'(x_i \,|\, \theta)}{f(x_i \,|\, \theta)}.$$

Many distributions also involve exponentials so the taking of logs help remove these and simplify the differentiation further. The Poisson distribution provides an example of this in action.

Example 12 *Consider $X \sim Po(\lambda)$. Then, $P(X = x \,|\, \lambda) = \lambda^x \exp(-\lambda)/x!$ for $x = 0, 1, \ldots$. Suppose $x_1, \ldots, x_n$ are a random sample from $Po(\lambda)$. The likelihood function is*

$$L(\lambda) \;=\; \prod_{i=1}^{n} \frac{\lambda^{x_i} \exp(-\lambda)}{x_i!}.$$

To find the maximum likelihood estimator of λ we work with the corresponding log-likelihood,

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{n} \log\left\{\frac{\lambda^{x_i} \exp(-\lambda)}{x_i!}\right\} \\ &= (\sum_{i=1}^{n} x_i) \log\lambda - n\lambda - \sum_{i=1}^{n} \log x_i! \\ &= n\overline{x} \log\lambda - n\lambda - \sum_{i=1}^{n} \log x_i! \end{aligned}$$

Differentiating with respect to λ gives

$$l'(\lambda) \;=\; \frac{n\overline{x}}{\lambda} - n.$$

So, solving $l'(\lambda) = 0$ gives $\hat{\lambda} = \overline{x}$. Note that $L(0) = 0 = L(\infty)$ and $l''(\lambda) = -n\overline{x}/\lambda^2 < 0$ for all $\lambda > 0$, so that $\overline{x}$ is the maximum likelihood estimate and $\overline{X}$ the maximum likelihood estimator for λ.

1.3.2 Maximum likelihood estimation when θ is multivariate

We now consider the general case when more than one parameter is required to be estimated simultaneously. Let $\theta = \{\theta_1, \ldots, \theta_k\}$, where each θ_r is univariate. We wish to choose the $\theta_1, \ldots, \theta_k$ which make the likelihood function an absolute maximum. A necessary condition for a local turning point in the likelihood function is that

$$\frac{\partial}{\partial\theta_r} \log L(\theta_1, \ldots, \theta_k) \;=\; \frac{\partial l}{\partial\theta_r} \;=\; 0, \text{ for each } r = 1, \ldots, k.$$

In this case, a sufficient condition that a solution $\theta = \hat{\theta}$ is a maximum is that the matrix

$$A \;=\; \left(\frac{\partial^2 l}{\partial\theta_r \partial\theta_s}\right)\Big|_{\theta=\hat{\theta}},$$

so A is the $k \times k$ matrix whose (r, s)th entry is $\frac{\partial^2 l}{\partial\theta_r \partial\theta_s}$ evaluated at $\theta = \hat{\theta}$, is negative definite. That is, for all nonzero vectors $y = [y_1 \ldots y_k]^T$,

$$y^T A y \;<\; 0.$$

Example 13 *Consider the normal distribution with $\theta = \{\mu, \sigma^2\}$. If the independent observations $x_1, \ldots, x_n$ are assumed to come from a $N(\mu, \sigma^2)$ then the likelihood function is*

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

The log-likelihood is then

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

The first order partial derivatives are

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

Solving $\frac{\partial l}{\partial \mu} = 0$ *we find that, for* $\sigma^2 > 0$, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$. *Substituting this into* $\frac{\partial l}{\partial \sigma^2} = 0$ *gives*

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\overline{x})^2 = 0,$$

so that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$. *The second order partial derivatives are*

$$\begin{aligned}\frac{\partial^2 l}{\partial \mu^2} &= -\frac{n}{\sigma^2};\\ \frac{\partial^2 l}{\partial (\sigma^2)^2} &= \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}\sum_{i=1}^{n}(x_i-\mu)^2;\\ \frac{\partial^2 l}{\partial \mu \partial \sigma^2} &= -\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu) \;=\; \frac{\partial^2 l}{\partial \sigma^2 \partial \mu}.\end{aligned}$$

Evaluating these at $\hat{\mu} = \overline{x}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$ *gives*

$$\begin{aligned}\left.\frac{\partial^2 l}{\partial \mu^2}\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} &= -\frac{n}{\hat{\sigma}^2};\\ \left.\frac{\partial^2 l}{\partial (\sigma^2)^2}\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} &= \frac{n}{2(\hat{\sigma}^2)^2} - \frac{1}{(\hat{\sigma}^2)^3}\sum_{i=1}^{n}(x_i-\overline{x})^2 \;=\; -\frac{n}{2(\hat{\sigma}^2)^2};\\ \left.\frac{\partial^2 l}{\partial \mu \partial \sigma^2}\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} &= -\frac{1}{(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(x_i-\overline{x}) \;=\; 0.\end{aligned}$$

A sufficient condition for $L(\hat{\mu}, \hat{\sigma}^2)$ *to be a maximum is that the matrix*

$$A \;=\; \left.\begin{pmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l}{\partial (\sigma^2)^2} \end{pmatrix}\right|_{\mu=\hat{\mu},\sigma^2=\hat{\sigma}^2} \;=\; \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2(\hat{\sigma}^2)^2} \end{pmatrix}$$

is negative definite. For any vector $y = [y_1\, y_2]^T$ *we have*

$$y^T A y \;=\; -\frac{n y_1^2}{\hat{\sigma}^2} - \frac{n y_2^2}{2(\hat{\sigma}^2)^2} < 0$$

so that A *is negative definite. The maximum likelihood estimates are* $\hat{\mu} = \overline{x}$ *and* $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$. *The corresponding maximum likelihood estimators are* $\overline{X}$ *and* $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ *for* μ *and* σ^2 *respectively.*

Chapter 2

Evaluating point estimates

2.1 Bias

Definition 4 *(Biased/unbiased estimator)*
An estimator $T(X_1, \ldots, X_n)$ for parameter θ has bias defined by

$$b(T) \quad = \quad E(T \,|\, \theta) - \theta.$$

If $b(T) = 0$ then T is an unbiased estimator of θ; otherwise it is biased.

Thus, if T is unbiased it is "correct in expectation". If we know the sampling distribution, that is the pdf of T, $f(t \,|\, \theta)$ then

$$b(T) \quad = \quad \int_{-\infty}^{\infty} t f(t \,|\, \theta)\, dt - \theta.$$

Example 14 *If $X_1, \ldots, X_n$ are iid $U(0, \theta)$ where θ is unknown, then the maximum likelihood estimator of θ is $M = \max\{X_1, \ldots, X_n\}$. On Question 2 of Question Sheet Two, we find the sampling distribution of M to be*

$$f_M(m \,|\, \theta) \quad = \quad \begin{cases} n\frac{m^{n-1}}{\theta^n} & m \leq \theta, \\ 0 & \textit{otherwise}, \end{cases}$$

Hence,

$$E(M \,|\, \theta) \;=\; \int_0^{\theta} m \left(n \frac{m^{n-1}}{\theta^n} \right) dm \;=\; \left(1 - \frac{1}{n+1} \right) \theta$$

so that $b(M) = -\frac{1}{n+1}\theta$. M is thus a biased estimator of θ: it underestimates θ which is not surprising as we would not expect the observed sample maximum to be the global maximum. An unbiased estimator is $\frac{n+1}{n} M$.

In many situations, however, we do not need to know the full sampling distribution to determine whether or not an estimator is biased. Two important examples are the sample mean and the sample variance.

Example 15 ***The sample mean.*** *Suppose $X_1, \ldots, X_n$ are iid with pdf $f(x \,|\, \theta)$ where parameter $\theta_1 \subseteq \theta$ is such that $E(X_i \,|\, \theta) = \theta_1$. Then, the estimator $T = \overline{X}$ is unbiased as*

$$\begin{aligned} E(T \,|\, \theta) &= E\left(\frac{1}{n} \sum_{i=1}^{n} X_i \,\middle|\, \theta \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(X_i \,|\, \theta) \\ &= \frac{1}{n} \sum_{i=1}^{n} \theta_1 \;=\; \theta_1. \end{aligned}$$

An important example of this is when $X_i \sim N(\mu, \sigma^2)$, $\theta_1 = \mu$, $\theta_2 = \sigma^2$ and $\theta = \{\mu, \sigma^2\}$.

Example 16 ***The sample variance.*** *Let the parameter $\theta_2 \subseteq \theta$ be such that $Var(X_i \,|\, \theta) = \theta_2$. The estimator $T = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$ is not an unbiased estimator for θ_2.*

$$\begin{aligned} E(T \,|\, \theta) &= E\left(\frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \,\middle|\, \theta \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} E((X_i - \overline{X})^2 \,|\, \theta). \end{aligned}$$

Now, from Example 15, $E(X_i \,|\, \theta) = E(\overline{X} \,|\, \theta)$ so that $E(X_i - \overline{X} \,|\, \theta) = 0$ whence $Var(X_i - \overline{X} \,|\, \theta) = E((X_i - \overline{X})^2 \,|\, \theta)$. Thus,

$$\begin{aligned} E(T \,|\, \theta) &= \frac{1}{n} \sum_{i=1}^{n} Var(X_i - \overline{X} \,|\, \theta) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ Var(X_i \,|\, \theta) - 2Cov(X_i, \overline{X} \,|\, \theta) + Var(\overline{X} \,|\, \theta) \right\}. \end{aligned} \tag{2.1}$$

Now, $Var(X_i \,|\, \theta) = \theta_2$ and, as the X_is are iid,

$$Var(\overline{X} \,|\, \theta) \;=\; \frac{1}{n^2} \sum_{j=1}^{n} Var(X_j \,|\, \theta) \;=\; \frac{\theta_2}{n}.$$

Using the properties of covariance[1] *we have that*

$$\begin{aligned} Cov(X_i, \overline{X} \,|\, \theta) &= \frac{1}{n} \sum_{j=1}^{n} Cov(X_i, X_j \,|\, \theta) \\ &= \frac{1}{n} \left\{ Cov(X_i, X_i \,|\, \theta) + \sum_{j \neq i}^{n} Cov(X_i, X_j \,|\, \theta) \right\} \\ &= \frac{1}{n} \left\{ Var(X_i \,|\, \theta) + \sum_{j \neq i}^{n} 0 \right\} \;=\; \frac{\theta_2}{n}. \end{aligned}$$

[1] For random quantities X, Y and Z, constants a and b, note that $Var(X) = Cov(X, X)$, $Cov(X, Y) = Cov(Y, X)$ and $Cov(X, aY + bZ) = aCov(X, Y) + bCov(X, Z)$.

Substituting these three results into (2.1) gives

$$\begin{aligned} E(T \mid \theta) &= \frac{1}{n}\sum_{i=1}^{n}\left(\theta_2 - \frac{2\theta_2}{n} + \frac{\theta_2}{n}\right) \\ &= \left(1 - \frac{1}{n}\right)\theta_2 \;=\; \frac{n-1}{n}\theta_2. \end{aligned}$$

An unbiased estimator of θ_2 *is thus* $\frac{n}{n-1}T$ *which is*[2]

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

Example 17 *Let* $X_1, \ldots, X_n$ *be iid* $N(\mu, \sigma^2)$ *with* μ *and* σ^2 *both unknown. From Examples 15 and 16 (consider* $\theta = \{\mu, \sigma^2\}$*; so* $\theta_1 = \mu$ *and* $\theta_2 = \sigma^2$*) we observe that the maximum likelihood estimator of* μ*,* $\overline{X}$*, is unbiased whereas the maximum likelihood estimator of* σ^2 *(with* μ *unknown),* $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$*, is biased. An unbiased estimator of* σ^2 *is* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$*.*

2.2 Mean square error

Bias is just one facet of an estimator. Consider the following two scenarios.

1. The estimator T may be unbiased but the sampling distribution, $f(t \mid \theta)$, may be quite disperse: there is a large probability of being far away from θ, so for any $\epsilon > 0$ the probability that $P(\theta - \epsilon < T < \theta + \epsilon \mid \theta)$ is small.

2. The estimator T may be biased but the sampling distribution, $f(t \mid \theta)$, may be quite concentrated. So, for any $\epsilon > 0$ the probability that $P(\theta - \epsilon < T < \theta + \epsilon \mid \theta)$ is large.

In these cases, the biased estimator may be preferable to the unbiased one. We would like to know more than whether or not an estimator is biased. In particular, we wish to capture some idea of how concentrated the sampling distribution of the estimator T is around θ. Ideally we would like

$$P(|T - \theta| < \epsilon \mid \theta) \;=\; P(\theta - \epsilon < T < \theta + \epsilon \mid \theta)$$

to be large for all $\epsilon > 0$. This probability may be hard to evaluate, but we may make use of Chebyshev's inequality.

Theorem 1 *(Chebyshev's inequality)*
For any random quantity X *and any* $t > 0$*,*

$$P(|X - E(X)| \geq t) \leq \frac{Var(X)}{t^2}$$

so that $P(|X - E(X)| < t) \geq 1 - \frac{Var(X)}{t^2}$*.*

[2] In the same way as using $\overline{X}$ to denote $\frac{1}{n}\sum_{i=1}^{n} X_i$ it is common notation to use S^2 to denote $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Proof - (For completeness; not examinable) Let $R = \{x : |x - E(X)| \geq t\}$. Then, for $x \in R$,

$$\frac{(x - E(X))^2}{t^2} \quad \geq \quad 1 \tag{2.2}$$

and

$$P(|X - E(X)| \geq t) \;=\; \int_R f(x)\,dx \;\leq\; \int_R \frac{(x - E(X))^2}{t^2} f(x)\,dx \tag{2.3}$$

where the inequality follows from (2.2). Since $R \subseteq \Omega$ then

$$\int_R \frac{(x - E(X))^2}{t^2} f(x)\,dx \leq \int_{-\infty}^{\infty} \frac{(x - E(X))^2}{t^2} f(x)\,dx \;=\; \frac{Var(X)}{t^2}. \tag{2.4}$$

The result follows by joining (2.3) and (2.4). □

Suppose that T is an unbiased estimator of θ, so that $E(T \,|\, \theta) = \theta$, then from an application of Chebyshev's inequality we have

$$P(|T - \theta| < \epsilon \,|\, \theta) \;\geq\; 1 - \frac{Var(T \,|\, \theta)}{\epsilon^2} \;=\; 1 - \frac{E\{(T - \theta)^2 \,|\, \theta\}}{\epsilon^2}$$

so that if T is an unbiased estimator of θ then a small value of $E\{(T-\theta)^2 \,|\, \theta\}$ implies a large value of $P(|T - \theta| < \epsilon \,|\, \theta)$. A simple extension of Chebyshev's inequality (repeat the proof but with $R = \{x : |x - \theta| \geq \epsilon\}$) shows that, for all T,

$$P(|T - \theta| < \epsilon \,|\, \theta) \quad \geq \quad 1 - \frac{E\{(T - \theta)^2 \,|\, \theta\}}{\epsilon^2}. \tag{2.5}$$

Definition 5 *(Mean Square Error)*
The mean square error (MSE) of the estimator T is defined to be

$$MSE(T) \quad = \quad E\{(T - \theta)^2 \,|\, \theta\}.$$

By considering equation (2.5), we see that we may use the MSE as a measure of the concentration of the estimator $T = T(X_1, \ldots, X_n)$ around θ. Note that if T is unbiased then $MSE(T) = Var(T \,|\, \theta)$.

If we have a choice between estimators then we might prefer to use the estimator with the smallest MSE.

Definition 6 *(Relative Efficiency)*
Suppose that $T_1 = T_1(X_1, \ldots, X_n)$ and $T_2 = T_2(X_1, \ldots, X_n)$ are two estimators for θ. The efficiency of T_1 relative to T_2 is

$$\textit{RelEff}(T_1, T_2) \quad = \quad \frac{MSE(T_2)}{MSE(T_1)}.$$

Values of RelEff(T_1, T_2) close to 0 suggest a preference for the estimator T_2 over T_1 while large values (> 1) of RelEff(T_1, T_2) suggest a preference for T_1. Notice that if T_1 and T_2 are unbiased estimators then

$$\text{RelEff}(T_1, T_2) \;=\; \frac{Var(T_2 \,|\, \theta)}{Var(T_1 \,|\, \theta)}$$

and we choose the estimator with the smallest variance.

Example 18 *Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. Two unbiased estimators of μ are $T_1 = \overline{X}$ and $T_2 = median\{X_1, \ldots, X_n\}$. We have shown that $\overline{X} \sim N(\mu, \sigma^2/n)$. The sample median, T_2, is asymptotically normal with mean μ and variance $\pi\sigma^2/2n$. Consequently,*

$$\begin{aligned} \textit{RelEff}(T_1, T_2) &= \frac{Var(median\{X_1, \ldots, X_n\} \,|\, \theta)}{Var(\overline{X} \,|\, \theta)} \\ &= \frac{\pi\sigma^2/2n}{\sigma^2/n} \;=\; \frac{\pi}{2}. \end{aligned}$$

Thus, we prefer T_1 to T_2 as it is more concentrated around θ: we'd prefer to use the sample mean rather than the sample median as an estimator of θ under this criterion. Note that if we calculate the sample mean of n individuals, then we would have to calculate the median of a sample of size $n\pi/2$ for the two estimators to have the same variance.

How do we calculate the MSE, $E\{(T-\theta)^2 \,|\, \theta\}$, when T is biased? We use the result that

$$MSE(T) \;=\; E\{(T-\theta)^2 \,|\, \theta\} \;=\; Var(T \,|\, \theta) + b^2(T). \tag{2.6}$$

To derive this result note that, as we assuming θ is known, then

$$\begin{aligned} Var(T \,|\, \theta) &= Var(T - \theta \,|\, \theta) \\ &= E\{(T-\theta)^2 \,|\, \theta\} - E^2(T - \theta \,|\, \theta) \\ &= E\{(T-\theta)^2 \,|\, \theta\} - \{E(T \,|\, \theta) - \theta\}^2 \\ &= MSE(T) - b^2(T). \end{aligned}$$

Example 19 *Suppose that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. Letting $\theta = \{\mu, \sigma^2\}$, $T_1 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ is a biased estimator of σ^2 with*

$$E(T_1 \,|\, \theta) \;=\; \frac{(n-1)\sigma^2}{n};\; Var(T_1 \,|\, \theta) \;=\; \frac{2(n-1)\sigma^4}{n^2}.$$

Thus,

$$b(T_1) \;=\; \frac{(n-1)\sigma^2}{n} - \sigma^2 \;=\; -\frac{\sigma^2}{n}.$$

Hence,

$$\begin{aligned} MSE(T_1) &= Var(T_1 \,|\, \theta) + b^2(T_1) \\ &= \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} \;=\; \frac{(2n-1)\sigma^4}{n^2}. \end{aligned}$$

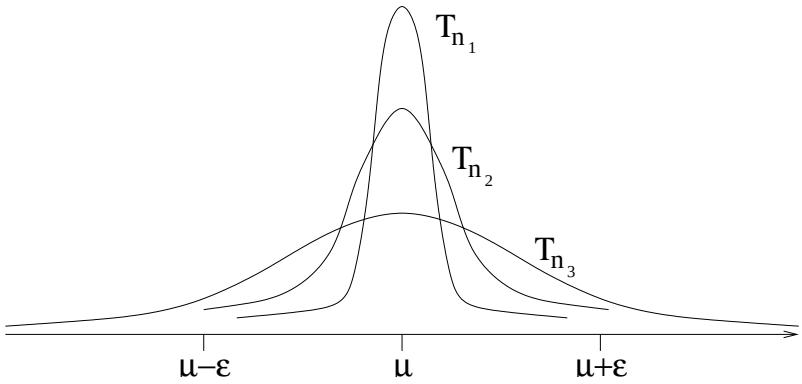

Figure 2.1: The pdf of $T_n \sim N(\mu, \sigma^2/n)$ for three values of n: $n_3 < n_2 < n_1$. As n increases, the distribution becomes more and more concentrated around μ.

An unbiased estimator of σ^2 *is* $T_2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ *and its second-order properties are*

$$E(T_2 \,|\, \theta) \;=\; \sigma^2;\; Var(T_2 \,|\, \theta) \;=\; \frac{2\sigma^4}{n-1}.$$

Thus, $MSE(T_2) = Var(T_2 \mid \theta)$ *and the relative efficiency of* T_1 *to* T_2 *is*

$$\begin{aligned} \textit{RelEff}(T_1, T_2) &= \frac{MSE(T_2)}{MSE(T_1)} \\ &= \frac{2\sigma^4/(n-1)}{(2n-1)\sigma^4/n^2} \\ &= \frac{2n^2}{(2n-1)(n-1)} > 1 \text{ if } n > 1/3. \end{aligned}$$

Although T_1 *is biased, it is more concentrated around* σ^2 *than* T_2.

2.3 Consistency

Bias and MSE are criteria for a fixed sample size n. We might also be interested in large sample properties. Let $T_n = T_n(X_1, \ldots, X_n)$ be an estimator for θ based on a sample of size n, $X_1, \ldots, X_n$. What can we say about T_n as $n \to \infty$? It might be desirable if, roughly speaking, the larger n is, the 'closer' T_n is to θ.

Example 20 *The maximum likelihood estimator for the parameter* μ *when the* X_i *are iid* $N(\mu, \sigma^2)$ *is* $T_n = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *and the sampling distribution of* $T_n \sim N(\mu, \sigma^2/n)$ *(see Q2 of Question Sheet One). As Figure 2.1 shows, as* n *increases, the probability that* $T_n \in (\mu - \epsilon, \mu + \epsilon)$ *gets large for any* $\epsilon > 0$*: the larger* n *is, the 'closer'* $T_n = \overline{X}$ *is to* μ.

Definition 7 *(Consistent)*
Let $\{T_n\}$ *denote the sequence of estimators* T_1, $T_2, \ldots$. *The sequence is consistent for* θ *if*

$$\lim_{n\to\infty} P(|T_n - \theta| < \epsilon \,|\, \theta) \;=\; 1 \quad \forall\, \epsilon > 0.$$

Thus, an estimator is consistent if it is possible to get arbitrarily close to θ by taking the sample size n sufficiently large. Now, from (2.5), we have a lower bound for $P(|T_n - \theta| < \epsilon \,|\, \theta)$, while 1 is an upper bound, so that

$$1 \;\geq\; P(|T_n - \theta| < \epsilon \,|\, \theta) \;\geq 1 - \frac{MSE(T_n)}{\epsilon^2}.$$

Hence, a sufficient condition for consistency of the estimator T_n is that

$$\lim_{n\to\infty} MSE(T_n) \;\;=\;\; 0.$$

If T_n is an unbiased estimator of θ then $MSE(T_n) = Var(T_n \,|\, \theta)$ so that the sufficient condition for consistency reduces to $\lim_{n\to\infty} Var(T_n \,|\, \theta) = 0$. For T_n biased, then, from (2.6), a sufficient condition that $\lim_{n\to\infty} MSE(T_n) = 0$ is that both

$$\lim_{n\to\infty} b(T_n) = 0 \quad \text{and} \quad \lim_{n\to\infty} Var(T_n \,|\, \theta) = 0.$$

Example 21 *For $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$, intuition suggests that $\overline{X}$ is a consistent estimator for μ. We can now confirm this intuition by noting that $\overline{X}$ is an unbiased estimator of μ with*

$$\lim_{n\to\infty} Var(\overline{X} \,|\, \mu, \sigma^2) \;=\; \lim_{n\to\infty} \frac{\sigma^2}{n} \;=\; 0,$$

so that the sufficient conditions for consistency are met.

2.4 Robustness

If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then $\overline{X}$ and $median\{X_1, \ldots, X_n\}$ are both unbiased estimators of μ. In Example 18, we showed that $\overline{X}$ was more efficient than $median\{X_1, \ldots, X_n\}$. Are there situations where we would ever use the median? The comparison in Example 18 depends upon the judgement that the X_i are iid $N(\mu, \sigma^2)$. What happens if this judgement of normality is flawed?

The sample mean and median are both examples of measures of location. If the observations $x_1, \ldots, x_n$ are believed to be different measurements of the same quantity, a measure of location - a measure of the centre of the observations - is often used as an estimate of the quantity.

An estimator which is fairly good (in some sense) in a wide range of situations is said to be **robust**. The median is more robust than the mean to 'outlying' values.

Example 22 *Suppose that we are interested in the 'true' heat of sublimation of platinum. Our strategy might be to perform an experiment n times and record the observed temperature of the sublimation and use this to estimate the 'true' heat. Our model may be*

$$X_i \;\;=\;\; \mu + \epsilon_i$$

where μ denotes the true heat of sublimation and the ϵ_i represent random errors (typically, the ϵ_i are assumed to be iid $N(0, \sigma^2)$ so that the X_i are iid $N(\mu, \sigma^2)$.) 26 observations are taken and the measurements displayed in the following stem-and-leaf plot (the decimal point is at the colon)

133 :	7						
134 :	1	3	4				
134 :	5	7	8	8	8	9	9
135 :	0	0	2	2	4	4	
135 :	8	8					
136 :	3						
136 :	6						

High/Outliers: 141.2 143.3 146.5 147.8 148.8

For this data, we have

$$\begin{aligned} \overline{x} &= \frac{3563.2}{26} = 137.05, \\ median(x_1, \ldots, x_{26}) &= \frac{135.0 + 135.2}{2} = 135.1. \end{aligned}$$

The median seems a more reasonable measure of the true heat: it is more robust than the mean to outlying values.

2.4.1 Trimmed mean

The sample mean is more efficient than the sample median but the sample median is more robust. The trimmed mean is an attempt to capture the efficiency of the sample mean while also improving its robustness to outlying values.

Definition 8 *(α-trimmed mean)*
If $x_1, \ldots, x_n$ are a series of measurements ordered as

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(\alpha)} \cdots \leq x_{(\alpha+1)} \leq \cdots \leq x_{(n-\alpha)} \leq x_{(n-\alpha+1)} \leq \cdots \leq \alpha_{(n)}$$

and we discard α observations at both extremes then the α-trimmed mean,

$$\overline{x}_\alpha = \frac{1}{n - 2\alpha} \sum_{i=\alpha+1}^{n-\alpha} x_{(i)},$$

is defined to be the mean of the remaining $n - 2\alpha$ values.

Note that if n is odd then the median is the $\frac{n-1}{2}$-trimmed mean while for n even, the median is the $(\frac{n}{2} - 1)$-trimmed mean.

Example 23 *We return to Example 22 and the sublimation of platinum. We find the* 5*-trimmed mean which discards the lower 19% and upper 19% of observations. It is*

$$\overline{x}_5 = \frac{2164.6}{16} = 135.29.$$

Chapter 3

Interval estimation

3.1 Principle of interval estimation

A simple point estimate $\hat{\theta} = T(x_1, \ldots, x_n)$ gives us no information about how accurate the corresponding estimator $T(X_1, \ldots, X_n)$ of θ is. However, sometimes, we can utilise information from the estimator's sampling distribution to help in this goal.

Example 24 *In Subsection 2.2 we looked at the mean square error of* $T(X_1, \ldots, X_n)$.

Example 25 *Suppose that* $X_1, \ldots, X_n$ *are iid* $N(\mu, \sigma^2)$ *and that* σ^2 *is known and we are interested in estimating* μ. $\overline{X}$ *is an unbiased estimator of* μ *with* $\overline{X} \sim N(\mu, \sigma^2/n)$. *Thus,*

$$P\left(-1.96 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 1.96 \,\middle|\, \mu, \sigma^2\right) \quad = \quad 0.95. \tag{3.1}$$

Equation (3.1) is a probability statement about $\overline{X}$. *Rearranging we have that*

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}} \,\middle|\, \mu, \sigma^2\right) \quad = \quad 0.95. \tag{3.2}$$

Equation (3.2) states that the RANDOM interval $(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}})$ *contains* μ *with probability 0.95. For observation* $\overline{X} = \overline{x}$, *we say we have 95% confidence that* μ *is in the interval* $(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}})$.

Construction of intervals like that in Example 25 is the goal of this chapter.

Definition 9 *(Pivot)*
Suppose $X_1, \ldots, X_n$ *are random quantities whose distribution has parameter* θ. *A function*

$$\phi(X_1, \ldots, X_n, \theta)$$

is called a pivot if its distribution, given θ, *does not depend upon* θ.

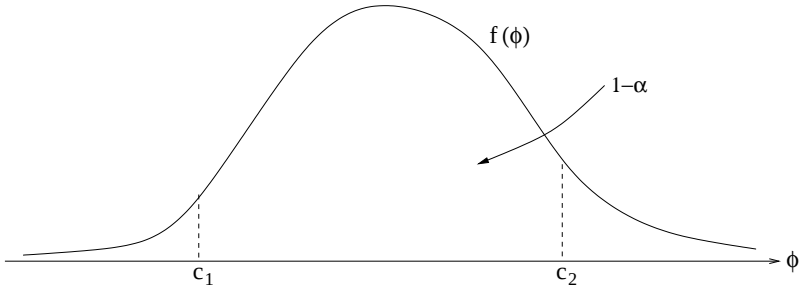

Figure 3.1: The probability density function $f(\phi)$ for a pivot ϕ. The probability of ϕ being between c_1 and c_2 is $1-\alpha$.

Example 26 *Suppose that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. There are (at most) two parameters: μ and σ^2. Note that, given μ and σ^2,*

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

and $N(0,1)$ does not depend upon either μ or σ^2. Thus, $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is a pivot.

If we have a pivot, ϕ, for θ, we can find its pdf $f(\phi)$. Moreover, see Figure 3.1, we can think of finding quantities c_1 and c_2 such that

$$P\{c_1 < \phi(X_1, \ldots, X_n, \theta) < c_2 \,|\, \theta\} \quad = \quad 1-\alpha.$$

If we can solve this for ϕ to get

$$P\{g_1(X_1, \ldots, X_n, c_1, c_2) < \theta < g_2(X_1, \ldots, X_n, c_1, c_2) \,|\, \theta\} \quad = \quad 1-\alpha.$$

then the random interval $(g_1(X_1, \ldots, X_n, c_1, c_2), g_2(X_1, \ldots, X_n, c_1, c_2))$ contains θ with probability $1-\alpha$. Moreover, having observed the data, we may compute a realisation of this interval.

Definition 10 *(Confidence interval)*
Suppose that the random interval $(g_1(X_1, \ldots, X_n, c_1, c_2), g_2(X_1, \ldots, X_n, c_1, c_2))$ contains θ with probability $1-\alpha$. A realisation of this random interval $(g_1(x_1, \ldots, x_n, c_1, c_2), g_2(x_1, \ldots, x_n, c_1, c_2))$ is called a $100(1-\alpha)\%$ confidence interval for θ.

It is important to note the following.

1. A confidence interval is **NOT** random; either it does or does not contain θ.
2. Thus, we **MUST NOT** talk about the probability that a confidence interval contains θ.
3. In the long run, $100(1-\alpha)\%$ of confidence intervals will contain θ.

3.2 Normal theory: confidence interval for μ when σ^2 is known

We now formalise Example 25. Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ where σ^2 is assumed to be known. $\overline{X}$ is an unbiased estimator of μ with sampling distribution $N(\mu, \sigma^2/n)$. Hence, given both μ and σ^2,

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

is a pivot. We may find constants c_1 and c_2 so that

$$P\left(c_1 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < c_2 \,\middle|\, \mu, \sigma^2\right) \quad = \quad 1 - \alpha.$$

Rearranging the inequality statement gives

$$P\left(\overline{X} - c_2\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - c_1\frac{\sigma}{\sqrt{n}} \,\middle|\, \mu, \sigma^2\right) \quad = \quad 1 - \alpha.$$

Thus, $(\overline{x} - c_2\frac{\sigma}{\sqrt{n}}, \overline{x} - c_1\frac{\sigma}{\sqrt{n}})$ is a $100(1-\alpha)\%$ confidence interval for μ. Note that this only works if σ^2 (and hence σ) is known so that both $\overline{x} - c_2\frac{\sigma}{\sqrt{n}}$ and $\overline{x} - c_1\frac{\sigma}{\sqrt{n}}$ are computable once we have observed the data. In the case when σ^2 is unknown, we construct an alternative confidence interval. We will tackle this in Section 3.4.

How do we choose c_1, c_2?

Let $Z \sim N(0,1)$. Typically, we choose c_1 and c_2 to form a symmetric interval around 0 so that $c_1 = -c_2$ with $c_2 = z_{(1-\frac{\alpha}{2})}$ where

$$P(Z \leq z_{(1-\frac{\alpha}{2})}) \quad = \quad 1 - \frac{\alpha}{2}.$$

Example 27 *$\alpha = 0.05$ to give a 95% confidence interval for μ. We have*

$$P(Z \leq 1.96) \;=\; 0.975 \;=\; 1 - \frac{0.05}{2}$$

(so $z_{0.975} = 1.96$) and the 95% confidence interval for μ is

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right).$$

Example 28 *$\alpha = 0.10$ to give a 90% confidence interval for μ. We have*

$$P(Z \leq 1.645) \;=\; 0.95 \;=\; 1 - \frac{0.10}{2}$$

(so $z_{0.95} = 1.645$) and the 90% confidence interval for μ is

$$\left(\overline{x} - 1.645\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.645\frac{\sigma}{\sqrt{n}}\right).$$

Example 29 *If $n = 20$ and $\sigma^2 = 10$ and we observe $\overline{x} = 107$ then a 95% confidence interval for μ is*

$$\left(107 - 1.96\sqrt{\frac{10}{20}}, 107 + 1.96\sqrt{\frac{10}{20}}\right) \quad = \quad (105.61, 108.39),$$

while a 90% confidence interval for μ is

$$\left(107 - 1.645\sqrt{\frac{10}{20}}, 107 + 1.645\sqrt{\frac{10}{20}}\right) \quad = \quad (105.84, 108.16).$$

3.3 Normal theory: confidence interval for σ^2

The confidence interval for σ^2 derives from the unbiased estimtor

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

for σ^2.

Theorem 2 *If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then $\overline{X}$ and S^2 are independent.*

Proof - It is beyond the scope of this course. The result essentially follows because $\overline{X}$ and, for each i, $X_i - \overline{X}$ are independent. You may verify that $Cov(\overline{X}, X_i - \overline{X} \,|\, \mu, \sigma^2) = 0$. □

Definition 11 *(Chi-squared distribution)*
If Z is a standard normal random quantity, then the distribution of $U = Z^2$ is called the chi-squared distribution with 1 degree of freedom. If $U_1, U_2, \ldots, U_n$ are independent chi-squared random quantities with 1 degree of freedom then $V = \sum_{i=1}^{n} U_i$ is called the chi-squared distribution with n degrees of freedom and is denoted χ^2_n.

Some properties of the chi-squared distribution

1. $E(\chi^2_n) = n$ and $Var(\chi^2_n) = 2n$. (We won't derive these: for future reference you may wish to note that χ^2_n may also be viewed as a Gamma distribution with parameters $n/2$ and $1/2$.)

2. If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then $\frac{X_1-\mu}{\sigma}, \ldots, \frac{X_n-\mu}{\sigma}$ are iid $N(0, 1)$ so that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \sim \chi^2_n.$$

Theorem 3 *If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then*

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Proof - Once again, the proof is omitted. If you want some insight into why this result is true note that

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\{(X_i-\overline{X})+(\overline{X}-\mu)\}^2\\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2+\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2.\end{aligned}$$

The left hand side is χ^2_n and $\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ is χ^2_1. The two components on the right hand side are, from Theorem 2, independent. □

Note that, from this result, we have

$$E(S^2\,|\,\mu,\sigma^2) = \frac{\sigma^2}{n-1}E(\chi^2_{n-1}) = \sigma^2$$

which verifies that S^2 is an unbiased estimator of σ^2 while

$$Var(S^2\,|\,\mu,\sigma^2) = \frac{\sigma^4}{(n-1)^2}Var(\chi^2_{n-1}) = \frac{2\sigma^4}{n-1}$$

which derives the variances used in Example 19.

From Theorem 3 we have that given σ^2, the distribution of $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ does not depend upon σ^2, so that it is a pivot for σ^2. We may find quantities c_1 and c_2 such that

$$P(c_1 < \chi^2_{n-1} < c_2) = 1-\alpha$$

and then using the pivot we have

$$P\left(c_1 < \frac{(n-1)S^2}{\sigma^2} < c_2 \,\middle|\, \mu,\sigma^2\right) = 1-\alpha.$$

Rearranging gives

$$P\left(\frac{(n-1)S^2}{c_2} < \sigma^2 < \frac{(n-1)S^2}{c_1} \,\middle|\, \mu,\sigma^2\right) = 1-\alpha.$$

Hence, $(\frac{(n-1)S^2}{c_2}, \frac{(n-1)S^2}{c_1})$ is a random interval which contains σ^2 with probability $1-\alpha$ so that a realisation of this interval,

$$\left(\frac{(n-1)s^2}{c_2}, \frac{(n-1)s^2}{c_1}\right),$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2$, is a $100(1-\alpha)\%$ confidence interval for σ^2.

How do we choose c_1 and c_2?

χ^2_{n-1} denotes the chi-squared distribution with $n-1$ degrees of freedom. The chi-squared

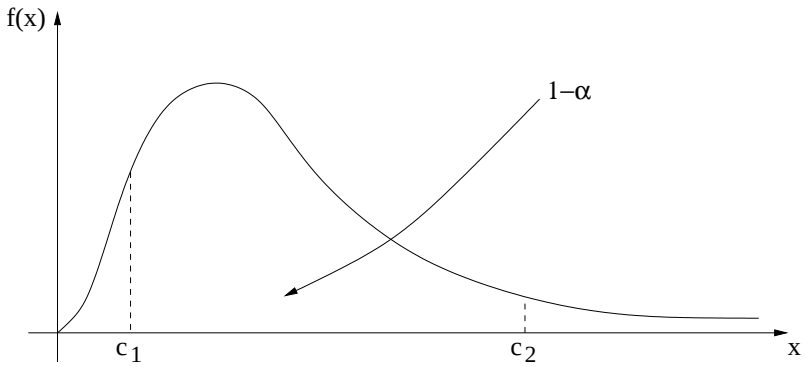

Figure 3.2: The probability density function for a chi-squared distribution. The probability of being between c_1 and c_2 is $1-\alpha$.

distribution is not symmetric, see Figure 3.2. The standard approach is to choose c_1 and c_2 such that

$$P(\chi^2_{n-1} < c_1) \;=\; \frac{\alpha}{2} \;=\; P(\chi^2_{n-1} > c_2).$$

The chi-squared tables gives values $\chi^2_{\nu,\alpha}$ where $P(\chi^2_\nu > \chi^2_{\nu,\alpha}) = \alpha$ for the chi-squared distribution with ν degrees of freedom. Thus, we choose

$$c_1 \;=\; \chi^2_{n-1,1-\frac{\alpha}{2}}, \qquad c_2 \;=\; \chi^2_{n-1,\frac{\alpha}{2}}$$

and our $100(1-\alpha)\%$ confidence interval for σ^2 is

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right).$$

Example 30 *If $n = 11$ then $\chi^2_{10,0.95} = 3.940$ and $\chi^2_{10,0.05} = 18.307$. A 90% confidence interval for σ^2 is*

$$\left(\frac{10s^2}{18.307}, \frac{10s^2}{3.940}\right).$$

Example 31 *In the production of synthetic fibres it is important that the fibres produced are consistent in quality. One aspect is that the tensile strength of the fibres should not vary too much. A sample of 8 pieces of fibre is taken and the tensile strength (in kg) of each fibre is tested. We find that $\overline{x} = 150.72kg$, $s^2 = 37.75kg^2$. Under the assumption that $X_1, \ldots, X_8$ are iid $N(\mu, \sigma^2)$ then $\frac{7S^2}{\sigma^2} \sim \chi^2_7$. We construct a 95% confidence interval for σ^2. Note that $\chi^2_{7,0.975} = 1.690$ and $\chi^2_{7,0.025} = 16.013$ so that a 95% confidence interval for σ^2 is*

$$\begin{aligned}\left(\frac{7s^2}{\chi^2_{7,0.025}}, \frac{7s^2}{\chi^2_{7,0.975}}\right) &= \left(\frac{7(37.75)}{16.013}, \frac{7(37.75)}{1.690}\right) \\ &= (16.502, 156.361)kg^2.\end{aligned}$$

Note that this interval is very wide.

3.4 Normal theory: confidence interval for μ when σ^2 is unknown

In Section 3.2 we constructed a $100(1-\alpha)\%$ confidence interval for μ of the form

$$\left(\overline{x} - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}\right).$$

When σ^2 (and hence σ) is unknown an alternative approach is required as we cannot compute this interval.

Definition 12 *(t-distribution)*
If $Z \sim N(0,1)$ and $U \sim \chi^2_n$ and Z and U are independent, then the distribution of $Z/\sqrt{U/n}$ is called the t-distribution with n degrees of freedom.

Note that

$$\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right) \Big/ \sqrt{\frac{S^2}{\sigma^2}} \quad = \quad \frac{\overline{X}-\mu}{S/\sqrt{n}}$$

and $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ while $\sqrt{\frac{S^2}{\sigma^2}} = \sqrt{\frac{\chi^2_{n-1}}{n-1}}$. Also $\overline{X}$ and S^2 are independent (see Theorem 2). Consequently,

$$\frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

We may use this as a pivot for μ. In particular, we may find constants c_1 and c_2 such that

$$P(c_1 < t_{n-1} < c_2) \quad = \quad 1-\alpha$$

so that

$$P\left(c_1 < \frac{\overline{X}-\mu}{S/\sqrt{n}} < c_2 \,\middle|\, \mu, \sigma^2\right) \quad = \quad 1-\alpha.$$

Rearranging, we have

$$P\left(\overline{X} - c_2\frac{S}{\sqrt{n}} < \mu < \overline{X} - c_1\frac{S}{\sqrt{n}} \,\middle|\, \mu, \sigma^2\right) \quad = \quad 1-\alpha.$$

Hence, $(\overline{X} - c_2\frac{S}{\sqrt{n}}, \overline{X} - c_1\frac{S}{\sqrt{n}})$ is a random interval which contains μ with probability $1-\alpha$. A realisation of this,

$$\left(\overline{x} - c_2\frac{s}{\sqrt{n}}, \overline{x} - c_1\frac{s}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for μ.

How do we choose c_1 and c_2?

The t-distribution is symmetric around 0 and so we may choose $c_1 = -c_2$. Tables for the t-distribution give the value $t_{\nu,\alpha}$ where $P(t_\nu > t_{\nu,\alpha}) = \alpha$ for the t-distribution with ν degrees of freedom. We choose $c_2 = t_{n-1,\frac{\alpha}{2}}$ so that our $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\overline{x} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{x} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right).$$

Example 32 *Suppose that* $n = 15$. *Note that* $t_{14,0.05} = 1.761$ *so that* $(\overline{x} - 1.761\frac{s}{\sqrt{15}}, \overline{x} + 1.761\frac{s}{\sqrt{15}})$ *is a 90% confidence interval for* μ.

Example 33 *Return to the synthetic fibre example, Example 31. Note that* $t_{7,0.025} = 2.365$. *A 95% interval for the fibre strength is*

$$\begin{aligned}\left(\overline{x} - 2.365\frac{s}{\sqrt{8}}, \overline{x} + 2.365\frac{s}{\sqrt{8}}\right) &= \left(150.72 - 2.365\sqrt{\frac{37.75}{8}}, 150.72 + 2.365\sqrt{\frac{37.75}{8}}\right) \\ &= (145.583, 155.857)kg.\end{aligned}$$

Chapter 4

Hypothesis testing

4.1 Introduction

Statistical hypothesis testing is a formal means of distinguishing between probability distributions on the basis of observing random quantities generated from one of the distributions.

Example 34 *Suppose $X_1, \ldots, X_n$ are iid normal with known variance and mean either equal to μ_0 or μ_1. We must decide whether $\mu = \mu_0$ or $\mu = \mu_1$.*

The formal framework we discuss was developed by Neyman and Pearson. The ingredients are as follows.

NULL HYPOTHESIS, H_0

This represents the status quo. H_0 will be assumed to be true unless the data indicates otherwise.

versus

ALTERNATIVE HYPOTHESIS, H_1

This represents a change to the status quo. If the data suggests against H_0, we reject H_0 in favour of H_1.

Example 35 *In the case discussed in Example 34, H_0 might state that the distribution was $N(\mu_0, \sigma^2)$ with H_1 being that the distribution was $N(\mu_1, \sigma^2)$.*

TEST STATISTIC

We make a decision about whether or not to reject H_0 in favour of H_1 based on the value of a test statistic, $T(X_1, \ldots, X_n)$. A test statistic is just a function of the observations.

Example 36 *You have met two common examples of statistics: an estimator for a parameter θ (e.g. $\overline{X}$ for parameter μ when $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$) and a pivot for θ (e.g. $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ for μ when σ^2 is known and the X_i are iid $N(\mu, \sigma^2)$).*

We will need to know the sampling distribution of T in the case when H_0 is true and also when H_1 is true.

CRITICAL REGION

We may determine the values of the test statistic for which we reject H_0 in favour of H_1. These values form the critical region which we now define in a slightly more formal way.

Definition 13 *(Critical Region)*
Let Ω denote the sample space of the test statistic T. The region $C \subseteq \Omega$ for which H_0 is rejected in favour of H_1 is termed the critical (or rejection) region while the region $\Omega \setminus C$, where we accept H_0, is called the acceptance region.

4.2 Type I and Type II errors

Under this approach, two types of error may be incurred.

	ACCEPT H_0	REJECT H_0
H_0 TRUE	GOOD	BAD
H_1 TRUE	BAD	GOOD

Definition 14 *(Type I and Type II errors)*
A type I error occurs when H_0 is rejected when it is true. The probability of such an error is denoted by α so that

$$\alpha \;=\; P(\textit{Type I error}) \;=\; P(\textit{Reject } H_0 \,|\, H_0 \textit{ true}).$$

A type II error occurs when H_0 is accepted when it is false. The probability of such an error is denoted by β so that

$$\beta \;=\; P(\textit{Type II error}) \;=\; P(\textit{Accept } H_0 \,|\, H_1 \textit{ true}).$$

Example 37 *We return to Example 34 and assume $\mu_1 > \mu_0$. Under H_0, $\overline{X} \sim N(\mu_0, \sigma^2/n)$ while under H_1, $\overline{X} \sim N(\mu_1, \sigma^2/n)$. A large value of $\overline{X}$ may indicate that H_1 rather than H_0 is true. Intuitively, we may consider a critical region of the form*

$$C \;=\; \{(x_1, \ldots, x_n) : \overline{x} \geq c\}.$$

This critical region is shown in Figure 4.1 There are a number of immediate questions.

1. *How do we pick the constant c?*

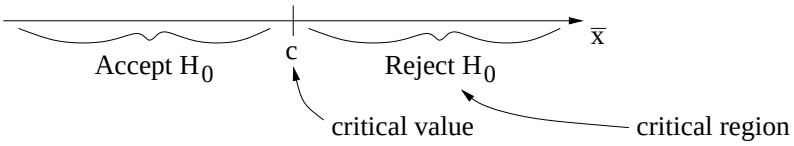

Figure 4.1: An illustration of the critical region $C = \{(x_1, \ldots, x_n) : \overline{x} \geq c\}$.

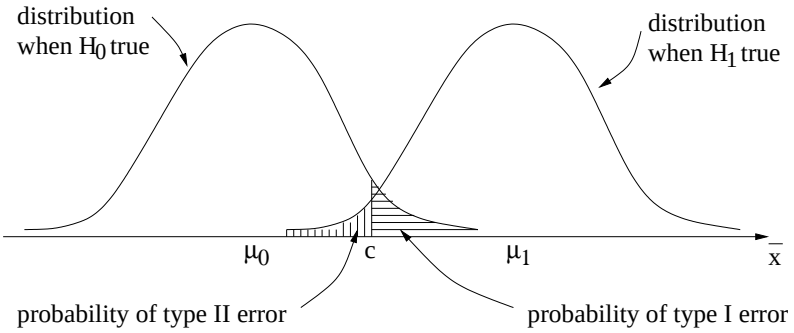

Figure 4.2: The errors resulting from the test with critical region $C = \{(x_1, \ldots, x_n) : \overline{x} \geq c\}$ of $H_0 : \mu = \mu_0$ versus $H_1 : \mu = \mu_1$ where $\mu_1 > \mu_0$.

2. *What if, by chance, H_0 is true but we happen to get a large value of $\overline{x}$?*

3. *What if, by chance, H_1 is true but we happen to get a small value of $\overline{x}$?*

4. *Was $\overline{X}$ the best test statistic to use anyway?*

We shall consider the middle two questions first. They help answer the first question. The answer to the fourth question will come later when we study a result known as the Neyman-Pearson Lemma.

We'd like to make the probability of either of these errors as small as possible. A possible scenario is shown in Figure 4.2.

- *If we INCREASE c then the probability of a type I error DECREASES but the probability of a type II error INCREASES.*
- *If we DECREASE c then the probability of a type I error INCREASES but the probability of a type II error DECREASES.*

It turns out, in practice, that this is always the case: in order to decrease the probability of a type I error, we must increase the probability of a type II error and vice versa. Recall that H_0 is the hypothesis which is taken to be true UNLESS the data suggests otherwise. We usually choose to fix α, the probability of a type I error, at some small value in advance.

e.g. $\alpha = 0.1, 0.05, 0.01, \ldots$

α is also known as the size or SIGNIFICANCE LEVEL of the test. Given a test statistic, fixing α determines the critical region.

Example 38 *In Example 37 we considered a critical region of the form*

$$C \;=\; \{(x_1,\ldots,x_n) : \overline{x} \geq c\}.$$

Now,

$$\begin{aligned} \alpha \;=\; P(\textit{Type I error}) &= P(\textit{Reject } H_0 \,|\, H_0 \textit{ true}) \\ &= P\{\overline{X} \geq c \,|\, \overline{X} \sim N(\mu_0, \sigma^2/n)\} \\ &= P\left(Z \geq \frac{c-\mu_0}{\sigma/\sqrt{n}}\right), \end{aligned}$$

where $Z \sim N(0,1)$.

Once α, the significance level, has been chosen, β is determined. It typically depends upon the sample size.

Example 39 *Using the critical region given in Example 38 we find*

$$\begin{aligned} \beta \;=\; P(\textit{Type II error}) &= P(\textit{Accept } H_0 \,|\, H_1 \textit{ true}) \\ &= P\{\overline{X} < c \,|\, \overline{X} \sim N(\mu_1, \sigma^2/n)\} \\ &= P\left(Z < \frac{c-\mu_1}{\sigma/\sqrt{n}}\right). \end{aligned}$$

Suppose we choose $\alpha = 0.05$. *Then*

$$\begin{aligned} P\left(Z \geq \frac{c-\mu_0}{\sigma/\sqrt{n}}\right) &= 0.05 \;\Rightarrow \\ \frac{c-\mu_0}{\sigma/\sqrt{n}} &= 1.645 \;\Rightarrow \\ c &= \mu_0 + 1.645\frac{\sigma}{\sqrt{n}}. \end{aligned}$$

Notice that as n *increases then* c *tends towards* μ_0. *The corresponding value of* β *is*

$$\begin{aligned} \beta \;=\; P\left(Z < \frac{c-\mu_1}{\sigma/\sqrt{n}}\right) &= p\left(Z < \frac{(\mu_0-\mu_1) + 1.645\sigma/\sqrt{n}}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(1.645 - \frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}\right) \end{aligned}$$

where $P(Z < z) = \Phi(z)$. *Note that as* n *increases then* β *decreases to zero.*

Definition 15 *(Power)*
The probability that H_0 *is rejected when it is false is called the power of the test. Thus,*

$$\textit{Power} \;=\; P(\textit{Reject } H_0 \,|\, H_1 \textit{ true}) \;=\; 1-\beta.$$

An important question In Examples 37 - 39, we worked with the test statistic $\overline{X}$ as the test statistic and $C = \{(x_1, \ldots, x_n) : \overline{x} \geq c\}$ as the critical region. Can we find a different test statistic T^* and critical region which, for the same sample size n, has the same value of $\alpha = P(\text{Type I error})$ but a SMALLER $\beta = P(\text{Type II error})$, equivalently, can we find the test with the largest power?

4.3 The Neyman-Pearson lemma

Consider the test of the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1$$

A 'best test' at significance level α would be the test with the greatest power. Our quest is to find such a test.

Suppose that $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots, x_n \,|\, \theta_0)$ under H_0 and $f(x_1, \ldots, x_n \,|\, \theta_1)$ under H_1. Define

$$\lambda(x_1, \ldots, x_n; \theta_0, \theta_1) \quad = \quad \frac{f(x_1, \ldots, x_n \,|\, \theta_0)}{f(x_1, \ldots, x_n \,|\, \theta_1)}. \tag{4.1}$$

Then $\lambda(x_1, \ldots, x_n; \theta_0, \theta_1)$ is the ratio of the likelihoods under H_0 and H_1. Let the critical region $C^* \subseteq \Omega$ be

$$C^* \quad = \quad \{(x_1, \ldots, x_n) : \lambda(x_1, \ldots, x_n; \theta_0, \theta_1) \leq k\} \tag{4.2}$$

where k is a constant chosen to make the test have significance level α, that is

$$P\{(X_1, \ldots, X_n) \in C^* \,|\, H_0 \text{ true}\} \quad = \quad \alpha.$$

Lemma 1 *(The Neyman-Pearson lemma)*
The test based on the critical region $C^* = \{(x_1, \ldots, x_n) : \lambda(x_1, \ldots, x_n; \theta_0, \theta_1) \leq k\}$ has the largest power (smallest type II error) of all tests with significance level α.

Proof - You will not be required to prove this lemma. If you want to see how to prove it then see Question Sheet Seven. □

Thus, among all tests with a given probability of a type I error, the likelihood ratio test minimises the probability of a type II error.

4.3.1 Worked example: Normal mean, variance known

Suppose that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ random quantities with σ^2 known. We shall apply the Neyman-Pearson lemma to construct the best test of the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu = \mu_1$$

where $\mu_1 > \mu_0$. From equation (4.1) we have that

$$\begin{aligned}
\lambda(x_1, \ldots, x_n; \mu_0, \mu_1) &= \frac{f(x_1, \ldots, x_n \,|\, \theta_0)}{f(x_1, \ldots, x_n \,|\, \theta_1)} \\
&= \frac{L(\mu_0)}{L(\mu_1)} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_1)^2\right\}} \\
&= \exp\left\{\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_0)^2\right)\right\}. \qquad (4.3)
\end{aligned}$$

Now,

$$\begin{aligned}
\sum_{i=1}^n (x_i - \mu_1)^2 - \sum_{i=1}^n (x_i - \mu_0)^2 &= \sum_{i=1}^n (x_i^2 - 2\mu_1 x_i + \mu_1^2) - \sum_{i=1}^n (x_i^2 - 2\mu_0 x_i + \mu_0^2) \\
&= -2\mu_1 n\overline{x} + n\mu_1^2 + 2\mu_0 n\overline{x} - n\mu_0^2 \\
&= n(\mu_1^2 - \mu_0^2) - 2n\overline{x}(\mu_1 - \mu_0). \qquad (4.4)
\end{aligned}$$

Substituting equation (4.4) into (4.3) gives

$$\lambda(x_1, \ldots, x_n; \mu_0, \mu_1) = \exp\left\{\frac{1}{2\sigma^2}\left(n(\mu_1^2 - \mu_0^2) - 2n\overline{x}(\mu_1 - \mu_0)\right)\right\}.$$

Using the Neyman-Pearson Lemma, see Lemma 1, the critical region of the most powerful test of significance level α for the test $H_0 : \mu = \mu_0 \text{versus} H_1 : \mu = \mu_1$ $(\mu_1 > \mu_0)$ is

$$\begin{aligned}
C^* &= \left\{(x_1, \ldots, x_n) : \exp\left\{\frac{1}{2\sigma^2}\left(n(\mu_1^2 - \mu_0^2) - 2n\overline{x}(\mu_1 - \mu_0)\right)\right\} \leq k\right\} \\
&= \left\{(x_1, \ldots, x_n) : n(\mu_1^2 - \mu_0^2) - 2n\overline{x}(\mu_1 - \mu_0) \leq 2\sigma^2 \log k\right\} \\
&= \left\{(x_1, \ldots, x_n) : -2n\overline{x}(\mu_1 - \mu_0) \leq 2\sigma^2 \log k + n(\mu_0^2 - \mu_1^2)\right\} \\
&= \left\{(x_1, \ldots, x_n) : \overline{x} \geq \frac{-\sigma^2}{n(\mu_1 - \mu_0)} \log k + \frac{(\mu_0 + \mu_1)}{2}\right\} \qquad (4.5) \\
&= \{(x_1, \ldots, x_n) : \overline{x} \geq k^*\}. \qquad (4.6)
\end{aligned}$$

Note that, in equation (4.5), we have utilised the fact that $\mu_1 - \mu_0 > 0$. The critical region given in equation (4.6) is identical to our intuitive interval derived in Example 37. For the test to be of significance α we choose (see Example 39)

$$k^* = \mu_0 + z_{(1-\alpha)}\frac{\sigma}{\sqrt{n}},$$

where $P(Z < z_{(1-\alpha)}) = 1-\alpha$. This is also written as $\Phi^{-1}(1-\alpha)$. $z_{(1-\alpha)}$ is the $(1-\alpha)$-quantile of Z, the standard normal distribution.

4.4 A Practical Example of the Neyman-Pearson lemma

Suppose that the distribution of lifetimes of TV tubes can be adequately modelled by an exponential distribution with mean θ so

$$f(x \mid \theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right)$$

for $x \geq 0$ and 0 otherwise. Under usual production conditions, the mean lifetime is 2000 hours but if a fault occurs in the process, the mean lifetime drops to 1000 hours. A random sample of 20 tube lifetimes is to taken in order to test the hypotheses

$$H_0 : \theta = 2000 \quad \text{versus} \quad H_1 : \theta = 1000.$$

Use the Neyman-Pearson lemma to find the most powerful test with significance level α.

Note that

$$\begin{aligned} L(\theta) = \prod_{i=1}^{20} f(x_i \mid \theta) &= \frac{1}{\theta^{20}} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{20} x_i\right) \\ &= \frac{1}{\theta^{20}} \exp\left(-\frac{20\overline{x}}{\theta}\right). \end{aligned}$$

Thus,

$$\begin{aligned} \lambda(x_1, \ldots, x_{20}; \theta_0, \theta_1) &= \frac{L(2000)}{L(1000)} \\ &= \frac{\frac{1}{2000^{20}} \exp\left(-\frac{20\overline{x}}{2000}\right)}{\frac{1}{1000^{20}} \exp\left(-\frac{20\overline{x}}{1000}\right)} \\ &= \left(\frac{1000}{2000}\right)^{20} \exp\left(-\frac{20\overline{x}}{2000} + \frac{20\overline{x}}{1000}\right) \\ &= \frac{1}{2^{20}} \exp\left(\frac{\overline{x}}{100}\right). \end{aligned}$$

Using the Neyman-Pearson lemma, the most powerful test of significance α has critical region

$$\begin{aligned} C^* &= \left\{(x_1, \ldots, x_2) : \frac{1}{2^{20}} \exp\left(\frac{\overline{x}}{100}\right) \leq k\right\} \\ &= \left\{(x_1, \ldots, x_2) : \frac{\overline{x}}{100} \leq \log 2^{20} k\right\} \\ &= \left\{(x_1, \ldots, x_2) : \overline{x} \leq k^*\right\}. \end{aligned}$$

That is, a test of the form reject H_0 if $\overline{x} \leq k_1$. To find k^*, we need to know the sampling distribution of $\overline{X}$ when $X_1, \ldots, X_{20}$ are iid exponentials with mean $\theta = 2000$ as

$$P(\overline{X} \leq k^* \mid \theta = 2000) = \alpha.$$

It turns out that the sum of n independent exponential random quantities with mean θ follows a distribution called the Gamma distribution with parameters n and $\frac{1}{\theta}$. We may use this to deduce k^*: $20k^*$ is the α-quantile of the $Gamma(20, \frac{1}{2000})$ distribution.

4.5 One-sided and two-sided tests

So far, we have been concerned with testing the hypotheses

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta = \theta_1.$$

Each of these hypotheses completely specifies the probability distribution and we can compute the corresponding values of α, the probability of a type I error, and β, the probability of a type II error.

Example 40 *In the Normal example, see Subsection 4.3.1, we wish to test the hypotheses*

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_1 : \mu = \mu_1$$

where $\mu_1 > \mu_0$ and σ^2 is known. Under H_0 the distribution of each X_i is completely specified as $N(\mu_0, \sigma^2)$ while under H_1 the distribution of each X_i is completely specified as $N(\mu_1, \sigma^2)$.

Example 41 *In the exponential example, see Section 4.4, we test the hypotheses*

$$H_0 : \theta = 2000 \quad \textit{versus} \quad H_1 : \theta = 1000.$$

Under H_0 the distribution of each X_i is completely specified as the exponential with mean 2000 while under H_1 the distribution of each X_i is completely specified as the exponential with mean 1000.

Definition 16 *(Simple/Composite hypothesis)*
If a hypothesis completely specifies the probability distribution of each X_i then it is said to be a simple hypothesis. If the hypothesis is not simple then it is said to be composite.

We shall consider examples when a hypothesis might only partially specify the value of a parameter of a known probability distribution. There are three particular tests of interest.

1. $H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$
2. $H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$
3. $H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$

In each case, the alternative hypothesis is not simple. In the first two cases, we have a *one-sided* alternative whilst in the latter case we have a *two-sided* alternative. The Neyman-Pearson lemma applies for a test of two simple hypotheses. How can we construct tests for the above three scenarios?

Example 42 *In the Normal example, see Subsection 4.3.1, the critical region of the most powerful test of the hypotheses*

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_1 : \mu = \mu_1$$

where $\mu_1 > \mu_0$ *and* σ^2 *is known is*

$$C^* \quad = \quad \{(x_1, \ldots, x_n) : \overline{x} \geq c\}$$

This region holds for ***all*** $\mu_1 > \mu_0$ *and so is the most powerful test for* ***every*** *simple hypothesis of the form* $H_1 : \mu = \mu_1$, $\mu_1 > \mu_0$. *The value of* μ_1 *only affects the power of the test. If* μ_1 *is close to* μ_0 *then we have a small power. The power increases as* μ_1 *increases: see Question 4 on Question Sheet Six for an example of this. We will explore this feature further in the next section.*

Example 43 *The critical region*

$$C^* \quad = \quad \{(x_1, \ldots, x_n) : \overline{x} \leq c\}$$

for the exponential hypotheses in Section 4.4 is the most powerful test for all $\theta_1 < \theta_0$ *and not just* $\theta_0 = 2000$ *and* $\theta_1 = 1000$. *Question 5 on Question Sheet Six demonstrates this.*

Definition 17 *(Uniformly Most Powerful Test)*
Suppose that H_1 *is composite. A test that is most powerful for every simple hypothesis in* H_1 *is said to be uniformly most powerful.*

Uniformly most powerful tests exist for some common one-sided alternatives.

Example 44 *If the* X_i *are iid* $N(\mu, \sigma^2)$ *with* σ^2 *known, then the test (see Example 42)*

$$C^* \quad = \quad \{(x_1, \ldots, x_n) : \overline{x} \geq c\}$$

is the most powerful for every $H_0 : \mu = \mu_0$ *versus* $H_1 : \mu = \mu_1$ *with* $\mu_1 > \mu_0$. *For a test with significance* α *we choose*

$$c \quad = \quad \mu_0 + z_{(1-\alpha)} \frac{\sigma}{\sqrt{n}}.$$

This test is uniformly most powerful for testing the hypotheses

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_1 : \mu > \mu_0,$$

with significance level α.

Example 45 *In a similar way to Subsection 4.3.1, we may show that if the* X_i *are iid* $N(\mu, \sigma^2)$ *with* σ^2 *known, then the test*

$$C^* \quad = \quad \{(x_1, \ldots, x_n) : \overline{x} \leq c\}$$

is the most powerful for every $H_0 : \mu = \mu_0$ *versus* $H_1 : \mu = \mu_1$ *with* $\mu_1 < \mu_0$. *For a test with significance* α *we choose*

$$c = \mu_0 - z_{(1-\alpha)}\frac{\sigma}{\sqrt{n}}.$$

This test is uniformly most powerful for testing the hypotheses

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_1 : \mu < \mu_0,$$

with significance level α.

Examples 44 and 45 provide uniformly most powerful tests for the two one-sided alternatives. Note that both of these tests are not the most powerful test for the two-sided alternative.

How could we construct a test for the two-sided alternative?

One approach is to combine the critical regions for testing the two one-sided alternatives. We shall develop this using the normal example but the approach is easily generalised for other parametric families. We consider that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ with σ^2 known. We wish to test the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

We combine the two one-sided tests to form a critical region of the form

$$C = \{(x_1, \ldots, x_n) : \overline{x} \leq k_2, \overline{x} \geq k_1\}.$$

Notice that

$$\begin{aligned} \alpha &= P(\{\overline{X} \leq k_2\} \cup \{\overline{X} \geq k_1\} \,|\, \overline{X} \sim N(\mu_0, \sigma^2/n)) \\ &= P\{\overline{X} \leq k_2 \,|\, \overline{X} \sim N(\mu_0, \sigma^2/n)\} + P\{\overline{X} \geq k_1 \,|\, \overline{X} \sim N(\mu_0, \sigma^2/n)\} \end{aligned}$$

One way to select k_1 and k_2 is to place $\alpha/2$ into each tail as shown in Figure 4.3. Then we have

$$\begin{aligned} k_1 &= \mu_0 + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}, \\ k_2 &= \mu_0 - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}, \end{aligned}$$

Thus, the test rejects for

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leq -z_{(1-\frac{\alpha}{2})} \quad \text{or} \quad \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{(1-\frac{\alpha}{2})}$$

which is equivalent to

$$\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} \geq z_{(1-\frac{\alpha}{2})}.$$

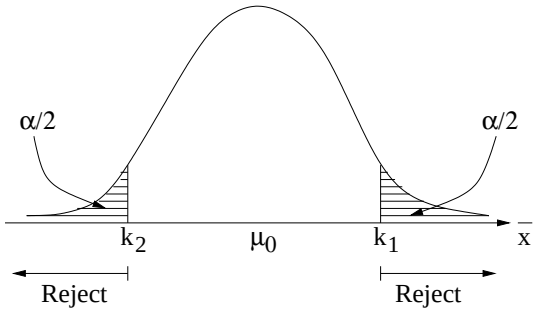

Figure 4.3: The critical region $C = \{(x_1, \ldots, x_n) : \overline{x} \leq k_2, \overline{x} \geq k_1\}$ for testing the hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$.

Hence, under H_0,

$$\begin{aligned} P\left(\frac{|\overline{X} - \mu_0|}{\sigma/\sqrt{n}} \geq z_{(1-\frac{\alpha}{2})} \,\middle|\, \mu_0, \sigma^2\right) &= \alpha \;\Rightarrow \\ P\left\{\left(\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}\right)^2 \geq z^2_{(1-\frac{\alpha}{2})} \,\middle|\, \mu_0, \sigma^2\right\} &= \alpha \;\Rightarrow \\ P(\chi^2_1 \geq z^2_{(1-\frac{\alpha}{2})}) &= \alpha. \end{aligned}$$

It is equivalent to reject for

$$\frac{n}{\sigma^2}(\overline{x} - \mu_0)^2 \geq \chi^2_{1,\alpha} \;=\; z^2_{(1-\frac{\alpha}{2})}.$$

We accept H_0 if

$$\begin{aligned} k_2 < \overline{x} < k_1 &\Rightarrow \\ \mu_0 - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < \overline{x} < \mu_0 + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} &\Rightarrow \\ \overline{x} - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < \mu_0 < \overline{x} + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} &\Rightarrow \\ \mu_0 \in (\overline{x} - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}). \end{aligned}$$

Recall, from Section 3.2, that $(\overline{x} - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}})$ is a $100(1-\alpha)\%$ confidence interval for μ. We see that μ_0 lies in the confidence interval if and only if the hypothesis test accepts H_0. Or, the confidence interval contains exactly those values of μ_0 for which we accept H_0. We have shown a **duality** between the hypothesis test and the confidence interval: the latter may be obtained by inverting the former and vice versa. The duality holds not just for this example but in all cases.

4.6 Power functions

Recall, from Definition 15, that the power of a test, $1 - \beta = P(\text{Reject } H_0 \,|\, H_1 \text{ true})$. When the alternative hypothesis is composite, the power will depend upon θ.

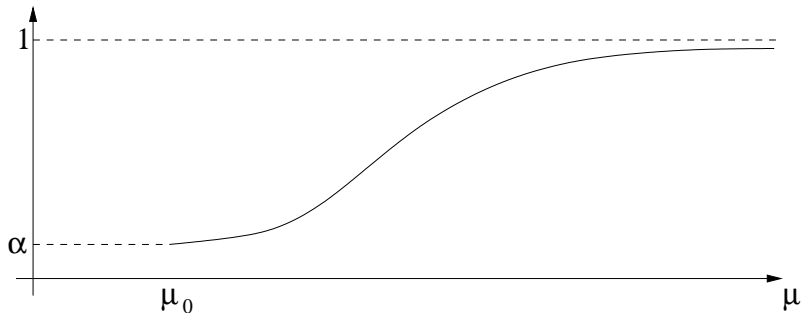

Figure 4.4: The power function, $\pi(\mu)$, for the uniformly most powerful test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$.

Example 46 *Recall Example 44. If the X_i are iid $N(\mu, \sigma^2)$ with σ^2 known, then the test*

$$C^* \quad = \quad \{(x_1, \ldots, x_n) : \overline{x} \geq c\}$$

with

$$c \quad = \quad \mu_0 + z_{(1-\alpha)} \frac{\sigma}{\sqrt{n}},$$

is uniformly most powerful for testing the hypotheses

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_1 : \mu > \mu_0,$$

with significance level α. In this case,

$$\begin{aligned} \beta &= P\left\{ \overline{X} < \mu_0 + z_{(1-\alpha)} \frac{\sigma}{\sqrt{n}} \,\middle|\, \overline{X} \sim N(\mu, \sigma^2/n) \right\} \\ &= \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{(1-\alpha)} \right). \end{aligned}$$

This is a function of μ. The corresponding power is also a function of μ, $\pi(\mu)$ say, where $\mu > \mu_0$. We have

$$\pi(\mu) \quad = \quad 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{(1-\alpha)} \right).$$

Figure 4.4 shows a sketch of $\pi(\mu)$. For μ arbitrarily close to μ_0 we have

$$\pi(\mu_0) \quad = \quad 1 - \Phi(z_{(1-\alpha)}) \; = \; \alpha.$$

As μ increases, $\Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{(1-\alpha)} \right)$ decreases so that $\pi(\mu)$ is an increasing function which tends to 1 as $\mu \to \infty$. Note that as $\mu \to \mu_0$ it is very hard to distinguish between the two hypotheses. Consequently, some authors talk of 'not rejecting H_0' rather than 'accepting H_0'.

Definition 18 *(Power function)*
The power function $\pi(\theta)$ of a test of $H_0 : \theta = \theta_0$ is

$$\begin{aligned} \pi(\theta) &= P(\textit{Reject } H_0 \mid \textit{True value of } \theta) \\ &= 1 - P(\textit{Type II error at } \theta). \end{aligned}$$

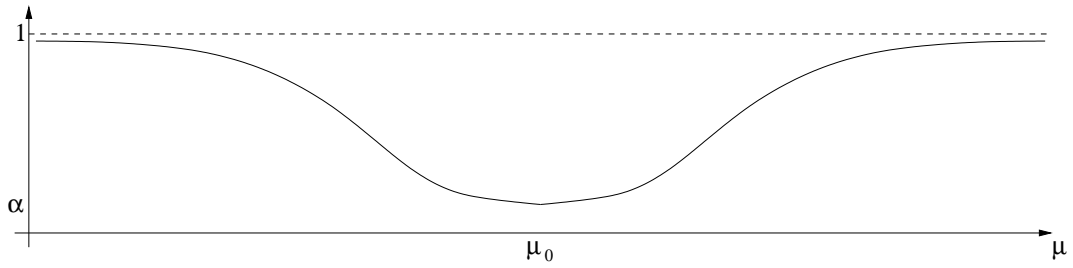

Figure 4.5: The power function, $\pi(\mu)$, for the test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. described in Example 47

Example 47 *A two-sided test of the hypotheses*

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_1 : \mu \neq \mu_0$$

where the X_i are iid $N(\mu, \sigma^2)$ with σ^2 known may be constructed, at significance level α, using the critical region

$$C = \{(x_1, \ldots, x_n) : \overline{x} \leq k_2, \overline{x} \geq k_1\}$$

where

$$\begin{aligned} k_1 &= \mu_0 + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}, \\ k_2 &= \mu_0 - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}. \end{aligned}$$

The corresponding power function is

$$\begin{aligned} \pi(\mu) &= P(\textit{Reject } H_0 \,|\, \textit{True value is } \mu) \\ &= P\{\overline{X} \leq k_2 \,|\, \overline{X} \sim N(\mu, \sigma^2/n)\} + P\{\overline{X} \geq k_1 \,|\, \overline{X} \sim N(\mu, \sigma^2/n)\} \\ &= \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{(1-\frac{\alpha}{2})}\right) + 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{(1-\frac{\alpha}{2})}\right). \end{aligned}$$

The power function is shown in Figure 4.5. Notice that the power function is symmetric about μ_0. To see this explicitly note that, for $\epsilon > 0$, we have

$$\pi(\mu_0 + \epsilon\sigma/\sqrt{n}) = \Phi(-\epsilon - z_{(1-\frac{\alpha}{2})}) + 1 - \Phi(-\epsilon + z_{(1-\frac{\alpha}{2})})$$

while

$$\begin{aligned} \pi(\mu_0 - \epsilon\sigma/\sqrt{n}) &= \Phi(\epsilon - z_{(1-\frac{\alpha}{2})}) + 1 - \Phi(\epsilon + z_{(1-\frac{\alpha}{2})}) \\ &= \{1 - \Phi(-\epsilon + z_{(1-\frac{\alpha}{2})})\} + 1 - \{1 - \Phi(-\epsilon - z_{(1-\frac{\alpha}{2})})\} \\ &= \pi(\mu_0 + \epsilon\sigma/\sqrt{n}). \end{aligned}$$

As $\mu \to \mu_0$ *then* $\pi(\mu) \to \pi(\mu_0)$ *where*

$$\begin{aligned} \pi(\mu_0) &= \Phi(-z_{(1-\frac{\alpha}{2})}) + 1 - \Phi(z_{(1-\frac{\alpha}{2})}) \\ &= 1 - \left(1 - \frac{\alpha}{2}\right) + 1 - \left(1 - \frac{\alpha}{2}\right) \\ &= \alpha \end{aligned}$$

while $\pi(\mu) \to 1$ *as both* $\mu \to \infty$ *and* $\mu \to -\infty$.

Chapter 5

Inference for normal data

In this chapter we shall assume that our observations come from normal distributions. In Sections 5.1 and 5.2 we consider data from a single sample before going on to compare paired and unpaired samples.

5.1 σ^2 in one sample problems

We assume that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. An unbiased point estimator of σ^2 is

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

with

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

In Section 3.3, we constructed a confidence interval for σ^2. Using the pivot $\frac{(n-1)S^2}{\sigma^2}$ for σ^2 we have that

$$P\left(\chi^2_{n-1,1-\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,\frac{\alpha}{2}} \,\middle|\, \mu, \sigma^2\right) = 1-\alpha$$

so that $\left(\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right)$ is a random interval which contains σ^2 with probability $1-\alpha$. Our $100(1-\alpha)\%$ confidence interval for σ^2 is a realisation of this random interval,

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}}, \frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right).$$

Let's consider hypothesis testing for σ^2. Suppose we want to test hypotheses of the form

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{versus} \quad \begin{cases} H_1 : \sigma^2 > \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2. \end{cases}$$

Notice that H_0 is not simple as we don't know μ. We will base tests on the statistic S^2.

5.1.1 $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$

Relative to σ_0^2, small values of s^2 support H_0 and large values H_1. We set a critical region of the form

$$C = \{(x_1, \ldots, x_n) : s^2 \geq k_1\}$$

where k_1 is chosen such that

$$\begin{aligned} \alpha &= P(S^2 \geq k_1 \mid H_0 \text{ true}) \\ &= P\left(\frac{(n-1)S^2}{\sigma_0^2} \geq \frac{(n-1)k_1}{\sigma_0^2} \,\middle|\, H_0 \text{ true}\right) \\ &= P\left(\chi^2_{n-1} \geq \frac{(n-1)k_1}{\sigma_0^2}\right) \end{aligned}$$

to give a test of significance α. Thus,

$$\frac{(n-1)k_1}{\sigma_0^2} = \chi^2_{n-1,\alpha} \quad \Rightarrow \quad k_1 = \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\alpha}.$$

5.1.2 $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 < \sigma_0^2$

Relative to σ_0^2, large values of s^2 support H_0 and small values H_1. We set a critical region of the form

$$C = \{(x_1, \ldots, x_n) : s^2 \leq k_2\}$$

where k_2 is chosen such that

$$\begin{aligned} \alpha &= P(S^2 \leq k_2 \mid H_0 \text{ true}) \\ &= P\left(\frac{(n-1)S^2}{\sigma_0^2} \leq \frac{(n-1)k_2}{\sigma_0^2} \,\middle|\, H_0 \text{ true}\right) \\ &= P\left(\chi^2_{n-1} \leq \frac{(n-1)k_2}{\sigma_0^2}\right) \end{aligned}$$

to give a test of significance α. Thus,

$$\frac{(n-1)k_2}{\sigma_0^2} = \chi^2_{n-1,1-\alpha} \quad \Rightarrow \quad k_2 = \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\alpha}.$$

5.1.3 $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 \neq \sigma_0^2$

If s^2 is 'close' to σ_0^2 then we have evidence for H_0. If s^2 is too small or too large, relative to σ_0^2, then we favour H_1. We combine the critical regions discussed in Subsections 5.1.1 and 5.1.2 to set a critical region of the form

$$C = \{(x_1, \ldots, x_n) : s^2 \leq k_2, s^2 \geq k_1\}$$

where k_2 and k_1 are chosen to give a test of significance α, that is so that

$$\alpha \;=\; P(S^2 \le k_2 \,|\, H_0 \text{ true}) + P(S^2 \ge k_1 \,|\, H_0 \text{ true}).$$

In the same way as the two-sided test in Section 4.5, we place $\alpha/2$ in each tail so that

$$\frac{\alpha}{2} \;=\; P(S^2 \le k_2 \,|\, H_0 \text{ true}) \;=\; P(S^2 \ge k_1 \,|\, H_0 \text{ true}).$$

Thus,

$$\begin{aligned} k_1 &= \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\frac{\alpha}{2}}; \\ k_2 &= \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\frac{\alpha}{2}}. \end{aligned}$$

Notice that we accept H_0 if

$$\begin{gathered} \tfrac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\frac{\alpha}{2}} < s^2 < \tfrac{\sigma_0^2}{n-1}\chi^2_{n-1,\frac{\alpha}{2}} \quad \Rightarrow \\ \tfrac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}} < \sigma_0^2 < \tfrac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} \end{gathered}$$

That is we accept H_0 if σ_0^2 is in the corresponding $100(1-\alpha)\%$ confidence interval for σ^2; a further illustration of the duality between hypothesis testing and confidence intervals.

5.1.4 Worked example

The weight of the contents of boxes of 'Honey Nut Loops' cereal is monitored by measuring 101 randomly selected boxes. The variance of the weight of the boxes was known to be $25g^2$ but Eagle-eyed Joe believes this is no longer the case and so wishes to test the hypotheses

$$H_0 : \sigma^2 = 25 \quad \text{versus} \quad H_1 : \sigma^2 \neq 25$$

at the 5% level. The critical region is, from Subsection 5.1.3,

$$\begin{aligned} C &= \left\{(x_1,\ldots,x_{101}) : s^2 \le \frac{25}{101-1}\chi^2_{101-1,1-\frac{0.05}{2}}, s^2 \ge \frac{25}{101-1}\chi^2_{101-1,\frac{0.05}{2}}\right\} \\ &= \left\{(x_1,\ldots,x_{101}) : s^2 \le \frac{1}{4}(74.222), s^2 \ge \frac{1}{4}(129.561)\right\} \\ &= \left\{(x_1,\ldots,x_{101}) : s^2 \le 18.5555, s^2 \ge 32.39025\right\}. \end{aligned}$$

Joe observes $s^2 = 31$ which does not lie in C. There is insufficient evidence to reject H_0 at the 5% level. Notice that the corresponding 95% confidence interval for σ^2 is

$$\begin{aligned} \left(\frac{100s^2}{\chi^2_{100,0.025}}, \frac{100s^2}{\chi^2_{100,0.975}}\right) &= \left(\frac{100(31)}{129.561}, \frac{100(31)}{74.222}\right) \\ &= (23.9270, 41.7666) \end{aligned}$$

which contains $\sigma_0^2 = 25$.

It is important to note that the decision to use a one-sided or two-sided test alternative **must** be made in light of the question of interest and **before** the test statistic is calculated.

e.g. You can't observe $s^2 > 25$ and then say right, I'll test $H_0 : \sigma^2 = 25$ versus $H_1 : \sigma^2 > 25$ as this affects the probability statements.

5.2 μ in one sample problems

We shall assume that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. An unbiased point estimator of μ is $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sampling distribution of $\overline{X}$ is $N(\mu, \sigma^2/n)$.

5.2.1 σ^2 known

In Section 4.5 we have constructed hypothesis tests under this scenario. We considered testing the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad \begin{cases} H_1 : \mu > \mu_0 \\ H_1 : \mu < \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

and used the respective critical regions

$$\begin{aligned} C^* &= \{(x_1, \ldots, x_n) : \overline{x} \geq \mu_0 + z_{(1-\alpha)}\tfrac{\sigma}{\sqrt{n}}\} \\ C^* &= \{(x_1, \ldots, x_n) : \overline{x} \leq \mu_0 - z_{(1-\alpha)}\tfrac{\sigma}{\sqrt{n}}\} \\ C &= \{(x_1, \ldots, x_n) : \overline{x} \leq \mu_0 - z_{(1-\frac{\alpha}{2})}\tfrac{\sigma}{\sqrt{n}}, \overline{x} \geq \mu_0 + z_{(1-\frac{\alpha}{2})}\tfrac{\sigma}{\sqrt{n}}\} \end{aligned}$$

for tests with significance level α.

Example 48 *A nutritionist thinks that the average person on low income gets less than the RDA of 800mg of calcium. To test this hypothesis, a random sample of 35 people with low income are monitored. With X_i denoting the calcium intake of the ith such person, we assume that $X_1, \ldots, X_{35}$ are iid $N(\mu, 250^2)$ and test the hypotheses*

$$H_0 : \mu = 800 \quad \textit{versus} \quad H_1 : \mu < 800.$$

We reject H_0 if

$$\overline{x} \leq 800 - z_{(1-\alpha)}\frac{250}{\sqrt{35}}.$$

For $\alpha = 0.1$, $z_{(1-\alpha)} = z_{0.9} = 1.282$ and we reject H_0 for $\overline{x} \leq 745.826$. For $\alpha = 0.05$, $z_{(1-\alpha)} = z_{0.95} = 1.645$ and reject for $\overline{x} \leq 730.486$. Suppose we observe $\overline{x} = 740$. There is sufficient evidence to reject the null hypothesis at the 10% level BUT insufficient evidence to reject H_0 at the 5% level. We might ask ***"At what level would our observed value have been on the reject/don't reject borderline?"*** *This value is termed the p-value.*

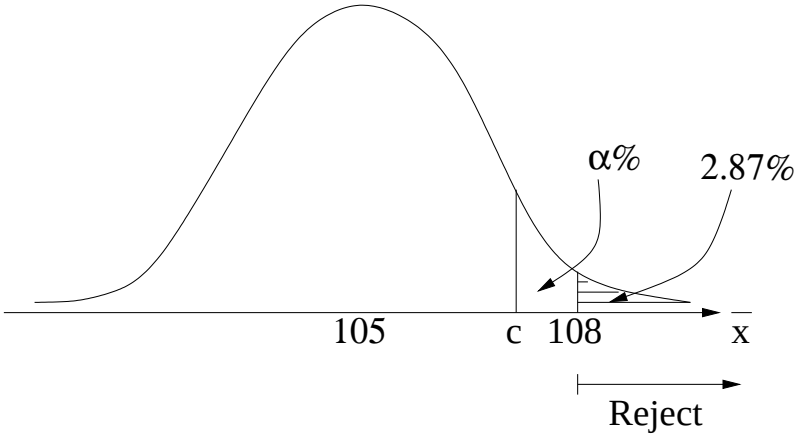

Figure 5.1: Illustration of the p-value corresponding to an observed value of $\overline{x} = 108$ in the test $H_0 : \mu = 105$ versus $H_0 : \mu > 105$. For all tests with significance level $\alpha\%$ larger than 2.87% we reject H_0 in favour of H_1.

5.2.2 The p-value

For any hypothesis test, rather than testing at significance level α, we could find the p-value of the test. The p-value enables us to determine at just what significance level our observed value of the test statistic would have been on the reject/don't reject borderline.

Definition 19 *(p-value)*
Suppose that $T(X_1, \ldots, X_n)$ is our test statistic for a hypothesis test and we observe $T(x_1, \ldots, x_n)$. The p-value is the probability that, if H_0 is true, we observe a value that is more extreme than $T(x_1, \ldots, x_n)$.

Example 49 *Suppose that $X_1, \ldots, X_{10}$ are iid $N(\mu, \sigma^2 = 25)$ and we wish to test the hypotheses*

$$H_0 : \mu = 105 \quad \textit{versus} \quad H_1 : \mu > 105.$$

Our test statistic is $\overline{X}$. Suppose we observe $\overline{x} = 108$. The p-value for this test is

$$\begin{aligned} P\{\overline{X} \geq 108 \,|\, \overline{X} \sim N(105, 2.5)\} &= P\left\{ \frac{\overline{X} - 105}{5/\sqrt{10}} \geq \frac{108 - 105}{5/\sqrt{10}} \;\middle|\; \frac{\overline{X} - 105}{5/\sqrt{10}} \sim N(0,1) \right\} \\ &= 1 - \Phi\left(\frac{108 - 105}{5/\sqrt{10}} \right\} \\ &= 1 - \Phi(1.90) \;=\; 0.0287. \end{aligned}$$

If we have a test at the 2.87% significance level, then $\overline{x} = 108$ is on the critical boundary. As Figure 5.1 illustrates, for all tests with significance level larger than 2.87% we reject H_0 in favour of H_1.

Example 50 *We compute the p-value corresponding to Example 48. It is*

$$\begin{aligned} P\{\overline{X} \leq 740 \,|\, \overline{X} \sim N(800, 250^2/35)\} &= P\left\{ \frac{\overline{X} - 800}{250/\sqrt{35}} \leq \frac{740 - 800}{250/\sqrt{35}} \,\middle|\, \frac{\overline{X} - 800}{250/\sqrt{35}} \sim N(0,1) \right\} \\ &= \Phi(-1.42) \\ &= 1 - 0.9222 \;=\; 0.0778. \end{aligned}$$

For all tests with significance level larger than 7.78% we reject H_0 in favour of H_1 which agrees with our finding in Example 48.

How do we compute the p-value in two-sided tests? Notice that we will observe a single value which will either be in the upper or lower tail of the distribution assumed true under H_0. Suppose we observe a value in the upper tail. We must also consider the corresponding extreme value in the lower tail. The p-value will be twice that to the corresponding observation in the one-sided test.

Example 51 *Suppose that $X_1, \ldots, X_{10}$ are iid $N(\mu, \sigma^2 = 25)$ and we wish to test the hypotheses*

$$H_0 : \mu = 105 \quad \textit{versus} \quad H_1 : \mu \neq 105.$$

Our test statistic is $\overline{X}$. Suppose we observe $\overline{x} = 108$. This value is in the upper tail of $\overline{X} \sim N(105, 25/10)$. We consider that it would just be as likely to observe the value 102 and

$$\begin{aligned} p &= P\{\overline{X} \leq 102 \,|\, \overline{X} \sim N(105, 2.5)\} + P\{\overline{X} \geq 108 \,|\, \overline{X} \sim N(105, 2.5)\} \\ &= 2(0.0287) \;=\; 0.0574. \end{aligned}$$

If we had a two-sided test of 5.74%, then our critical region would be

$$C \;=\; \{(x_1, \ldots, x_{10}) : \overline{x} \leq 102, \overline{x} \geq 108\}.$$

For all tests with significance level larger than 5.74% we reject H_0 in favour of H_1.

Notice that the symmetry of the Normal distribution makes it easy to find the equivalent extreme value in the opposite tail to the value we observed. For a general two-sided Normal hypothesis test we note that if we observe $\overline{x}$ and H_0 is true, then $\overline{x}$ is a realisation from a $N(0,1)$ distribution. The p-value is given by

$$p \;=\; P\left\{ \frac{|\overline{X} - \mu_0|}{\sigma/\sqrt{n}} \geq \frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} \,\middle|\, \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \right\}$$

However, if the underlying sampling distribution is not symmetric it is harder to find the equivalent extreme value. However, we do not need this if we only require the p-value: we just use the observation that the p-value for the two-sided test will be double the p-value for the corresponding observation in a one-sided test.

Example 52 *On question 3. of Question Sheet 8 you consider the test of the hypotheses*

$$H_0 : \sigma^2 = 10 \quad \textit{versus} \quad H_1 : \sigma^2 \neq 10$$

The test statistic is s^2 *and* $E(S^2 \,|\, H_0 \textit{ true}) = 10$. *You observe* $s^2 = 9.506 < 10$ *so* s^2 *is in the lower tail. The p-value for this two-sided test is twice that of the p-value corresponding to the test of the hypotheses*

$$H_0 : \sigma^2 = 10 \quad \textit{versus} \quad H_1 : \sigma^2 < 10$$

That is

$$p \quad = \quad 2P(S^2 \leq 9.506 \,|\, H_0 \textit{ true}).$$

5.2.3 σ^2 unknown

If the variance σ^2 is unknown to us then we cannot use the approach of Subsection 5.2.1 as σ is explicitly required when computing the critical values. We mirror the approach of Section 3.4 and use the pivot

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

the t-distribution with $n-1$ degrees of freedom. We are interested in testing the hypotheses

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad \begin{cases} H_1 : \mu > \mu_0 \\ H_1 : \mu < \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

and we'll let

$$t(X_1, \ldots, X_n) \quad = \quad \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

be our test statistic. If H_0 is true then $t(x_1, \ldots, x_n)$ should be an observation from a t-distribution with $n-1$ degrees of freedom. Recall that t-tables give the value $t_{\nu,\alpha}$ where $P(t_\nu > t_{\nu,\alpha}) = \alpha$.

$H_0 : \mu = \mu_0$ **versus** $H_1 : \mu > \mu_0$

If H_0 is true then t should be close to zero, while large values support H_1. We set a critical region of the form

$$C \quad = \quad \left\{ (x_1, \ldots, x_n) : t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \geq k_1 \right\}$$

where k_1 is chosen such that

$$\begin{aligned} \alpha \quad &= \quad P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq k_1 \,\middle|\, H_0 \text{ true} \right) \\ &= \quad P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq k_1 \,\middle|\, \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \right) \end{aligned}$$

to give a test of significance α. Thus,

$$k_1 \quad = \quad t_{n-1,\alpha}.$$

$H_0 : \mu = \mu_0$ **versus** $H_1 : \mu < \mu_0$

If H_0 is true then t should be close to zero, while small values support H_1. We set a critical region of the form

$$C \quad = \quad \left\{(x_1, \ldots, x_n) : t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq k_2\right\}$$

where k_2 is chosen such that

$$\begin{aligned} \alpha &= P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq k_2 \,\middle|\, H_0 \text{ true}\right) \\ &= P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq k_2 \,\middle|\, \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}\right) \end{aligned}$$

to give a test of significance α. Thus,

$$k_2 \quad = \quad -t_{n-1,\alpha}.$$

$H_0 : \mu = \mu_0$ **versus** $H_1 : \mu \neq \mu_0$

If H_0 is true then t should be close to zero, while large or small values support H_1. We combine the critical regions of the two one-sided tests and set a critical region of the form

$$C \quad = \quad \left\{(x_1, \ldots, x_n) : t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq k_2, t \geq k_1\right\}$$

where k_1 and k_2 are chosen so that

$$\begin{aligned} \alpha &= P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq k_2 \,\middle|\, H_0 \text{ true}\right) + P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq k_1 \,\middle|\, H_0 \text{ true}\right) \\ &= P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \leq k_2 \,\middle|\, \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}\right) + P\left(\frac{\overline{X} - \mu_0}{S/\sqrt{n}} \geq k_1 \,\middle|\, \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}\right) \end{aligned}$$

to give a test of significance α. We place $\alpha/2$ in each tail so that

$$k_1 \;=\; t_{n-1,\frac{\alpha}{2}}, \qquad k_2 \;=\; -t_{n-1,\frac{\alpha}{2}}.$$

Example 53 *The manufacturer of a new car claims that a typical car gets 26mpg. Honest Joe believes that the manufacturer is over egging the pudding and that the true mileage is less than 26mpg. To test the claim, Joe takes a sample of 1000 cars and denotes by X_i the miles per gallon (mpg) of the ith car. He assumes that the X_i are iid $N(\mu, \sigma^2)$ and he wishes to test whether the true mean is less than the manufacturers claim. Thus, he tests the hypotheses*

$$H_0 : \mu = 26 \quad \textit{versus} \quad H_1 : \mu < 26.$$

Joe will reject H_0 at the significance level α if

$$t = \frac{\overline{x} - 26}{s/\sqrt{1000}} \leq -t_{999,\alpha}.$$

Joe observes $\overline{x} = 25.9$ and $s^2 = 2.25$ so that

$$t = \frac{25.9 - 26}{\sqrt{2.25/1000}} = -2.108.$$

Now,

$$\begin{aligned} -t_{999,0.025} &= -1.962341 \text{ (using the R command } \texttt{qt(0.975,999)} \text{ to find } t_{999,0.025}) \\ -t_{999,0.01} &= -2.330086 \text{ (using the R command } \texttt{qt(0.99,999)} \text{ to find } t_{999,0.01}) \end{aligned}$$

We reject H_0 at the 2.5% level but not at the 1% level. The p-value is thus between 0.01 and 0.025 and may be found using R: `pt(-2.108,999)` $= 0.0176$. There is some evidence to suggest that these cars achieve less than 26mpg, but the difference is 'small'. This example raises the question of

Statistical significance

versus

Practical significance

As the sample size is so large, the sample mean $\overline{x} = 25.9$ is probably pretty close to the population mean. At the 2.5% level we obtained a statistically significant result to reject the manufacturers claim that $\mu = 26$mpg but how practically significant is the difference between 25.9mpg and 26mpg?

5.3 Comparing paired samples

In many experiments, we may have paired observations.

Example 54 *Blood pressure of an individual before and after exercise.*

Example 55 *We might match subjects by age/weight/condition and ascribe one to a test/treatment group and the other to a control group.*

Suppose we have n pairs and let (X_i, Y_i) denote the measurements for the ith pair. Assume that the X_is are iid with mean μ_X and variance σ_X^2 and the Y_is are iid with mean μ_Y and variance σ_Y^2. The quantities X_i and Y_i are not independent. Suppose that $Cov(X_i, Y_i) = \sigma_{XY}$ and that the pairs (X_i, Y_i) are independent, so that $Cov(X_i, Y_j) = 0$ for $i \neq j$.

Let $D_i = X_i - Y_i$, $i = 1, \ldots, n$ denote the ith difference. The D_i are independent with

$$\begin{aligned} E(D_i) &= \mu_X - \mu_Y = \mu_D; \\ Var(D_i) &= Var(X_i) + Var(Y_i) - 2Cov(X_i, Y_i) \\ &= \sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY} = \sigma_D^2. \end{aligned}$$

Assume that the D_i are iid $N(\mu_D, \sigma_D^2)$.[1] Typically, we are concerned as to whether there is a difference between the two measurements, that is we are interested in testing the hypotheses

$$H_0 : \mu_D = 0 \quad \text{versus} \quad \begin{cases} H_1 : \mu_D > 0 \\ H_1 : \mu_D < 0 \\ H_1 : \mu_D \neq 0 \end{cases}$$

Now,

$$\begin{aligned} \overline{D} &= \frac{1}{n}\sum_{i=1}^{n} D_i \sim N(\mu_D, \sigma_D^2/n); \\ S_D^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2 \text{ so } \frac{(n-1)S_D^2}{\sigma_D^2} \sim \chi_{n-1}^2. \end{aligned}$$

As $\overline{D}$ and S_D^2 are independent then

$$\frac{\overline{D} - \mu_D}{S_D/\sqrt{n}} \sim t_{n-1}.$$

To test the hypotheses we may use the t-tests derived in Subsection 5.2.3.

5.3.1 Worked example

A paediatrician measured the blood cholesterol of her patients and was worried to note that some had levels over 200mg/100ml. To investigate whether dietary regulation would lower blood cholesterol, the paediatrician selected 10 patients at random. Blood tests were conducted on these patients both before and after they had undertaken a two month nutritional program. The results are shown below.

Patient	1	2	3	4	5	6	7	8	9	10
(X) Before	210	217	208	215	202	209	207	210	221	218
(Y) After	212	210	210	213	200	208	203	199	218	214
(D) Difference	-2	7	-2	2	2	1	4	11	3	4

For example, $(x_4, y_4) = (215, 213)$ and $d_4 = x_4 - y_4 = 215 - 213 = 2$. We observe $\overline{x} = 211.7$, $s_x^2 = 34.2$, $\overline{y} = 208.7$ and $s_y^2 = 38.9$. There is lots of variability so it is difficult to tell from these summaries whether there is a statistically significant difference. The question of interest is whether dietary regulation *lowers* blood cholesterol. The paediatrician tests the hypotheses

$$H_0 : \mu_D = 0 \quad \text{versus} \quad H_1 : \mu_D > 0$$

where $D_1, \ldots, D_{10}$ are assumed to be iid $N(\mu_D, \sigma_D^2)$. The observed estimates of μ_D and σ_D^2 are $\overline{d} = 3$ and $s_d^2 = 15.3$ respectively. Under H_0, $\frac{\overline{D}-0}{S_D/\sqrt{n}} \sim t_{n-1}$ and our critical region is

$$C = \left\{(d_1, \ldots, d_{10}) : t = \frac{\overline{d}}{s_d/\sqrt{10}} \geq t_{9,\alpha}\right\}$$

[1] A nonparametric version, where we do not assume normality, is the (Wilcoxon) signed-rank test: see http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

The paediatrician observes

$$t \;=\; \frac{3}{\sqrt{15.3/10}} \;=\; 2.425356.$$

From tables, $t_{9,0.025} = 2.26$ and $t_{9,0.01} = 2.82$. Hence, we reject H_0 at the 2.5% level but not at the 1% level. The p-value is thus between these values and is about 0.019. (using R: `1 - pt(2.425356,9)`)

There is fairly strong evidence to suggest that this dietary program is effective in lowering blood cholesterol level in these patients.

5.4 Investigating σ^2 for unpaired data

In many experiments, the two samples may be regarded as being independent of each other.

Example 56 *In a medical study, a random sample of subjects may be assigned to a treatment group and another independent sample to a control group. There is no pairing in the samples, indeed the samples may be, and frequently are, of different sizes.*

Assume that the sample $X_1, \ldots, X_n$ is drawn from a normal distribution with mean μ_X and variance σ_X^2, so the X_is are iid $N(\mu_X, \sigma_X^2)$. Additionally, assume that a second, independent, sample $Y_1, \ldots, Y_m$ is drawn from a normal distribution with mean μ_Y and variance σ_Y^2, so the Y_is are iid $N(\mu_Y, \sigma_Y^2)$. Note:

1. It is assumed that n and m need not be equal.
2. There is no notion of pairing: X_1 is no more related to Y_1 than to Y_{37} so that there is no notion of individual differences.
3. Interest centres upon whether μ_X differs from μ_Y and whether σ_X^2 differs from σ_Y^2. We shall tackle the latter question first.

Definition 20 *(F-distribution)*
Let U and V be independent chi-square random quantities with ν_1 and ν_2 degrees of freedom respectively. The distribution of

$$W \;=\; \frac{U/\nu_1}{V/\nu_2}$$

is called the F-distribution with ν_1 and ν_2 degrees of freedom, written F_{ν_1,ν_2}.

Note that:

$$\text{if } W \;=\; \frac{U/\nu_1}{V/\nu_2} \sim F_{\nu_1,\nu_2} \text{ then } \frac{1}{W} \;=\; \frac{V/\nu_2}{U/\nu_1} \sim F_{\nu_2,\nu_1}.$$

For our unpaired data case, the sample variance of the X_is is

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{5.1}$$

so that

$$U_X = \frac{n-1}{\sigma_X^2}S_X^2 \sim \chi_{n-1}^2. \tag{5.2}$$

Similarly, the sample variance of the Y_is is

$$S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2 \tag{5.3}$$

with

$$V_Y = \frac{m-1}{\sigma_Y^2}S_Y^2 \sim \chi_{m-1}^2. \tag{5.4}$$

Thus, dividing (5.1) by (5.2) and using (5.3) and (5.4) we have that

$$\frac{S_X^2}{S_Y^2} = \left(\frac{\sigma_X^2}{\sigma_Y^2}\right)\frac{U_X/(n-1)}{V_Y/(m-1)} = \frac{\sigma_X^2}{\sigma_Y^2}W \tag{5.5}$$

where, as S_X^2 and S_Y^2 are independent and using Definition 20, $W \sim F_{n-1,m-1}$.

Suppose we want to test hypotheses of the form

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad \begin{cases} H_1 : \sigma_X^2 > \sigma_Y^2 \\ H_1 : \sigma_X^2 < \sigma_Y^2 \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 \end{cases}$$

Let $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ denote the observed value of S_X^2 and $s_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i - \overline{y})^2$ denote the observed value of S_Y^2. If H_0 is true, then, from (5.5), s_x^2/s_y^2 should be a realisation from the F-distribution with $n-1$ and $m-1$ degrees of freedom. We use S_X^2/S_Y^2 as the test statistic and under H_0,

$$\frac{S_X^2}{S_Y^2} \sim F_{n-1,m-1}.$$

5.4.1 $H_0 : \sigma_X^2 = \sigma_Y^2$ versus $H_1 : \sigma_X^2 > \sigma_Y^2$

If H_1 is true then $\sigma_X^2/\sigma_Y^2 > 1$ and so, from (5.5), s_x^2/s_y^2 should be large. We set a critical region of the form

$$C = \{(x_1, \ldots, x_n, y_1, \ldots, y_m) : s_x^2/s_y^2 \geq k_1\}$$

where k_1 is chosen such that

$$P(S_X^2/S_Y^2 \geq k_1 \mid H_0 \text{ true}) = P(F_{n-1,m-1} \geq k_1) = \alpha$$

for a test of significance α.

5.4.2 $H_0 : \sigma_X^2 = \sigma_Y^2$ versus $H_1 : \sigma_X^2 < \sigma_Y^2$

If H_1 is true then $\sigma_X^2/\sigma_Y^2 < 1$ and so, from (5.5), s_x^2/s_y^2 should be small. We set a critical region of the form

$$C \quad = \quad \{(x_1, \ldots, x_n, y_1, \ldots, y_m) : s_x^2/s_y^2 \leq k_2\}$$

where k_2 is chosen such that

$$P(S_X^2/S_Y^2 \leq k_2 \mid H_0 \text{ true}) \;=\; P(F_{n-1,m-1} \leq k_2) \;=\; \alpha$$

for a test of significance α.

5.4.3 $H_0 : \sigma_X^2 = \sigma_Y^2$ versus $H_1 : \sigma_X^2 \neq \sigma_Y^2$

If H_1 is true then, from (5.5), s_x^2/s_y^2 should be either small or large. We set a critical region of the form

$$C \quad = \quad \{(x_1, \ldots, x_n, y_1, \ldots, y_m) : s_x^2/s_y^2 \leq k_2,\ s_x^2/s_y^2 \geq k_1\}$$

where k_1 and k_2 are chosen such that

$$\begin{aligned} P(S_X^2/S_Y^2 \geq k_1 \mid H_0 \text{ true}) \quad &= \quad P(F_{n-1,m-1} \geq k_1) \;=\; \alpha/2 \\ P(S_X^2/S_Y^2 \leq k_2 \mid H_0 \text{ true}) \quad &= \quad P(F_{n-1,m-1} \leq k_2) \;=\; \alpha/2 \end{aligned}$$

for a test of significance α.

F-tables give upper 5%, 2.5%, 1% and 0.5% of the F-distribution, denoted $F_{n-1,m-1,\alpha}$ for the requisite degrees of freedom and significance level. These enable us to find the k_1 values. For the k_2 values, we note that

$$P(F_{n-1,m-1} \leq k_2) \quad = \quad P(F_{m-1,n-1} \geq 1/k_2).$$

So, in Subsection 5.4.1, $k_1 = F_{n-1,m-1,\alpha}$. For Subsection 5.4.2, $k_2 = 1/F_{m-1,n-1,\alpha}$. In Subsection 5.4.3, $k_1 = F_{n-1,m-1,\alpha/2}$ and $k_2 = 1/F_{m-1,n-1,\alpha/2}$.

5.4.4 Worked example

The US National Centre for Health Statistics compiles data on the length of stay by patients in short-term hospitals. We are interested in whether the variability of stay length is the same for both men and women. To investigate this, we take independent samples of 41 male patient stay lengths, denoted $X_1, \ldots, X_{41}$ and 31 female patient stay lengths, denoted $Y_1, \ldots, Y_{31}$. We assume that the X_is are iid $N(\mu_X, \sigma_X^2)$ and the Y_is are iid $N\mu_Y, \sigma_Y^2)$ and test the hypotheses

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2$$

at the 10% level. From Subsection 5.4.3, the critical region is

$$C \;=\; \{(x_1,\ldots,x_{41},y_1,\ldots,y_{31}) : s_x^2/s_y^2 \le k_2,\ s_x^2/s_y^2 \ge k_1\}$$

where $k_1 = F_{40,30,0.05} = 1.79$ and $k_2 = 1/F_{30,40,0.05} = 1/1.74 = 0.57$. If we observe $s_x^2 = 56.25$ and $s_y^2 = 46.24$ then $s_x^2/s_y^2 = 1.22$.

Now, $0.57 < 1.22 < 1.79$ so we do not reject H_0 at the 10% level (so the p-value is greater than 0.1). There is insufficient evidence to suggest a difference in variability of stay lengths for male and female patients.

5.5 Investigating the means for unpaired data when σ_X^2 and σ_Y^2 are known

As the X_is are iid $N(\mu_X,\sigma_X^2)$ then $\overline{X} \sim N(\mu_X,\sigma_X^2/n)$. Similarly, since the Y_is are iid $N(\mu_Y,\sigma_Y^2)$ then $\overline{Y} \sim N(\mu_Y,\sigma_Y^2/m)$. Our question of interest is typically 'are the means equal?'. We are interested in testing hypotheses of the form

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad \begin{cases} H_1 : \mu_X > \mu_Y \\ H_1 : \mu_X < \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

As $\overline{X}$ and $\overline{Y}$ are independent then $\overline{X}-\overline{Y} \sim N(\mu_X-\mu_Y, \sigma_X^2/n + \sigma_Y^2/m)$. If H_0 is true then

$$Z \;=\; \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

so that the observed value (which we can compute as σ_X^2 and σ_Y^2 are known) $(\overline{x}-\overline{y})/\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$ should be a realisation from a standard normal distribution if H_0 is true. We assess how extreme this observation is to perform the hypothesis tests outlined above.

5.6 Pooled estimator of variance (σ_X^2 and σ_Y^2 unknown)

If we accept $\sigma_X^2 = \sigma_Y^2$ then we should estimate the common variance σ^2. To do this we **pool** the two estimates s_x^2 and s_y^2 weighting them according to sample size.

> N.B. We can't pool the x_i, y_j together into one sample of size $n+m$ as the x_i have possibly different means to the y_j.

The pooled estimate of variance is the pooled sample variance,

$$s_p^2 \;=\; \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}. \tag{5.6}$$

The corresponding pooled estimator of variance is

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}.$$

Note that if $\sigma_X^2 = \sigma^2 = \sigma_Y^2$ then $\frac{n-1}{\sigma^2}S_X^2 \sim \chi_{n-1}^2$ and $\frac{m-1}{\sigma^2}S_Y^2 \sim \chi_{m-1}^2$. As S_X^2 and S_Y^2 are independent then

$$\frac{n+m-2}{\sigma^2}S_p^2 = \frac{n-1}{\sigma^2}S_X^2 + \frac{m-1}{\sigma^2}S_Y^2 \sim \chi_{n+m-2}^2.$$

N.B. As both S_X^2 and S_Y^2 are unbiased estimators of σ^2 then S_p^2 is also an unbiased estimator of σ^2.

5.7 Investigating the means for unpaired data when σ_X^2 and σ_Y^2 are unknown

We will restrict attention to the case where we can assume that $\sigma_X^2 = \sigma^2 = \sigma_Y^2$.

e.g. An F-test has not been able to conclude that $\sigma_X^2 \neq \sigma_Y^2$.

In this case, $\overline{X} - \overline{Y} \sim N(\mu_X - \mu_Y, \sigma^2(\frac{1}{n} + \frac{1}{m}))$.[2] We are interested in testing the hypotheses

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad \begin{cases} H_1 : \mu_X > \mu_Y \\ H_1 : \mu_X < \mu_Y \\ H_1 : \mu_X \neq \mu_Y \end{cases}$$

If H_0 is true then $(\overline{X} - \overline{Y})/\sqrt{\sigma^2(\frac{1}{n} + \frac{1}{m})} \sim N(0,1)$ and $\frac{n+m-2}{\sigma^2}S_p^2 \sim \chi_{n+m-2}^2$ so that

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

and $t = (\overline{x} - \overline{y})/\sqrt{s_p^2(\frac{1}{n} + \frac{1}{m})}$ should be a realisation from a t-distribution with $n+m-2$ degrees of freedom. We may thus use t to perform conventional t-tests, as described in Subsection 5.2.3.

5.7.1 Worked example

Recall the length of stay in hospital example of Subsection 5.4.4. We have 41 male patients, with length of stay $X_1, \ldots, X_{41}$ and 31 female patients, with length of stay $Y_1, \ldots, Y_{31}$. The observed sample variances were $s_x^2 = 56.25$ and $s_y^2 = 46.24$. An F-test concluded that there was no detectable difference in variability of stay lengths for male and female patients. We

[2] A nonparametric version, where we do not assume normality, is the Mann-Whitney U test: see http://en.wikipedia.org/wiki/Mann-Whitney_U

may assume that $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. We estimate σ^2 using the pooled sample variance s_p^2 as given by (5.6). We find that

$$s_p^2 \;=\; \frac{(41-1)56.25 + (31-1)46.24}{41+31-2} \;=\; 51.96$$

Suppose we test the hypothesis

$$H_0 : \mu_X = \mu_Y \quad \text{versus} \quad H_1 : \mu_X \neq \mu_Y$$

and observe $\overline{x} = 9.075$ and $\overline{y} = 7.114$. If H_0 is true then

$$t \;=\; \frac{9.075 - 7.114}{\sqrt{51.96(\frac{1}{41} + \frac{1}{31})}} \;=\; \frac{1.961}{1.7156} \;=\; 1.143$$

should be a realisation from a t-distribution with $41 + 31 - 2 = 70$ degrees of freedom. For this two-sided test, $t = 1.143$ corresponds to a p-value of $2\{1 - pt(1.143, 70)\} = 0.2569$. There is insufficient evidence to suggest a difference between the mean stay lengths of males and females.

Chapter 6

Goodness of fit tests

6.1 The multinomial distribution

Suppose that data can be classified into k mutually exclusive classes or categories and that the class i occurs with probability p_i where $\sum_{i=1}^{k} p_i = 1$. Suppose we take n observations and let X_i denote the number of observations in class i. Then

$$P(X_1 = x_1, \ldots, X_k = x_k \mid p_1, \ldots, p_k) \quad = \quad \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} \tag{6.1}$$

where $\sum_{i=1}^{k} x_i = n$ and $\frac{n!}{x_1! \ldots x_k!}$ is the number of ways that n objects can be grouped into k classes, with x_i in the ith class. The probability in (6.1) represents the probability density function of the **multinomial distribution**, a generalisation of the binomial distribution.

6.2 Pearson's chi-square statistic

In goodness of fit tests, we are interested in whether a given model fits the data. Consider the following two examples.

1. Is a die fair? We may roll the die n times and count the number of each score observed. If we let X_i be the number of is observed then we have a multinomial distribution with six classes, the ith class being that we roll an i and if the die is fair then $p_i = P(\text{Roll an } i) = \frac{1}{6}$ for each $i = 1, \ldots, 6$: the p_is are known if the model is true.

2. Emissions of alpha particles from radioactive sources are often modelled by a Poisson distribution. Suppose we construct intervals of 1-second length and count the number of emissions in each interval. We observe a total of $n = 12169$ such intervals and are observations are summarised in the following table.

Number of emissions	0	1	2	3	4	5
Observed	5267	4436	1800	534	111	21

If the model is correct then we may construct a multinomial distribution where the $i+1$st class represents that we observe i emissions in the interval of 1-second length and $p_1 = \exp(-\lambda)$, $p_2 = \lambda\exp(-\lambda)$, $p_3 = \frac{\lambda^2}{2!}\exp(-\lambda), \ldots$.

If λ is not given, then the p_i are unknown (although we have restricted them to be Poisson probabilities) and, in order to assess the fit of this model to the data, we must estimate a value of λ from the observed data. Typically, we use the maximum likelihood estimate. In this case, the maximum likelihood estimate of λ, $\hat{\lambda}$, is the sample mean,

$$\hat{\lambda} \;=\; \frac{(0\times 5267)+(1\times 4436)+\cdots+(5\times 21)}{12169} \;=\; \frac{10187}{12169} \;=\; 0.8371.$$

The goodness of fit of a model may be assessed by comparing the observed (O) counts with the expected counts (E) if the model was correct. Consider the two examples discussed above.

1. If the die was fair, we'd expect to observe $np_i = \frac{n}{6}$ in each of the six classes. If the actual observed counts differ widely from this, then we'd suspect the die wasn't fair.

2. In the Poisson model, we'd expect to observe $n\hat{p}_i$ in each class, where $\hat{p}_i$ is our estimated probability of being in that class assuming the Poisson model. Once again, discrepancies between the observed and expected counts are evidence against the assumed model.

For each class i, we have an observed value O_i and an expected value E_i. A widely used measure of the discrepancy between these two (and hence between the assumed model and the data) is **Pearson's chi-square statistic:**

$$X^2 \;=\; \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}. \tag{6.2}$$

The larger the value of X^2, the worse the fit of the model to the data. It can be shown that if the model is correct then the distribution of X^2 is approximately the chi-square distribution with ν degrees of freedom where

$$\nu \;=\; \text{number of classes} - \text{number of parameters fitted} - 1.$$

Intuitively, we can see the degrees of freedom by noting that we are free to obtain the expected counts in the first $k-1$ classes but the count in the final class is fixed as the total number of observations is n and then we lose a further degree of freedom for each parameter we fit. In the die example, we fit no parameters as the p_i are known under the assumed model. For the Poisson model, we fit a single parameter, λ.

If X^2 is observed to take the value x, then the p-value is

$$p^* \;=\; P(X^2 > x \mid \text{model is correct}) \;\approx\; P(\chi^2_\nu > x).$$

The approximation is best if we have large n and, as a rule of thumb, we group categories together to avoid any classes with $E_i < 5$.

6.2.1 Worked example

We return to the Poisson example. We have fitted $\hat{\lambda} = 10187/12169$ and

$$E_{i+1} = n\frac{\hat{\lambda}^i}{i!}\exp(-\hat{\lambda}) = \frac{\hat{\lambda}}{i}E_i.$$

$$\begin{aligned}
E_1 &= 12169\exp(-10187/12169) = 5268.600 \\
E_2 &= (10187/12169)E_1 = 4410.488 \\
E_3 &= \frac{10187/12169}{2}E_2 = 1846.069 \\
E_4 &= \frac{10187/12169}{3}E_3 = 515.132 \\
E_5 &= \frac{10187/12169}{4}E_4 = 107.808 \\
E_6 &= \frac{10187/12169}{5}E_5 = 18.050 \\
E_7 &= \frac{10187/12169}{6}E_6 = 2.518
\end{aligned}$$

and so on. Note that $E_i < 5$ for all $i \geq 7$. We pool these into the sixth class to ensure an expected count greater than 5 in all classes. This newly created sixth class corresponds to number of emissions greater than or equal to 5 so has probability $\sum_{i=5}^{\infty}\frac{\hat{\lambda}^i\exp(-\hat{\lambda})}{i!}$ and expected count $12169 - \sum_{i=1}^{5}E_i$. Hence, our observed and expected counts are as in the following table.

Number of emissions	0	1	2	3	4	5+
Observed	5267	4436	1800	534	111	21
Expected	5268.600	4410.488	1846.069	515.132	107.808	20.903

Using (6.2), we calculate the observed Pearson's chi-square statistic as

$$X^2 = \frac{(5267-5268.600)^2}{5268.600} + \frac{(4436-4410.488)^2}{4410.488} + \cdots + \frac{(21-20.903)^2}{20.903} = 2.0838.$$

If the Poisson model is true then 2.0838 should (approximately) be a realisation from a chi-square distribution with $6-1-1=4$ (we have 6 cells and have fitted one parameter) degrees of freedom. Now $P(\chi^2_4 > 2.0838) = 0.72$ so that the model fits the data very well.

Chapter 7

Appendix - Adding Independent Normals

We directly show that if $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ and X and Y are independent then $W = X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. It then follows by induction that the sum of n independent normal random quantities is also normal. This is for completeness and mirrors the work you did in §VI.b *Distributions of sums* of MA10212 when adding standard normal random quantities.

We shall utilise the following lemma.

Lemma 2 *The following identity is true.*

$$\frac{1}{\sigma_1^2}(x-\mu_1)^2 + \frac{1}{\sigma_2^2}(w-x-\mu_2)^2 = \frac{1}{\sigma_3^2}(x-\mu_3)^2 + \frac{1}{\sigma_1^2+\sigma_2^2}\{w-(\mu_1+\mu_2)\}^2$$

where

$$\mu_3 = \frac{\sigma_2^2\mu_1 + \sigma_1^2(w-\mu_2)}{\sigma_1^2+\sigma_2^2}; \tag{7.1}$$

$$\sigma_3^2 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}. \tag{7.2}$$

Proof -

$$\begin{aligned}
\frac{1}{\sigma_1^2}(x-\mu_1)^2 + \frac{1}{\sigma_2^2}(w-x-\mu_2)^2 &= \frac{1}{\sigma_1^2}(x-\mu_1)^2 + \frac{1}{\sigma_2^2}\{(w-\mu_2)-x\}^2 \\
&= \frac{(\sigma_1^2+\sigma_2^2)x^2 - 2\{\sigma_2^2\mu_1 + \sigma_1^2(w-\mu_2)\}x + \sigma_2^2\mu_1^2 + \sigma_1^2(w-\mu_2)^2}{\sigma_1^2\sigma_2^2} \qquad (7.3) \\
&= \frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2}\left[\left\{x - \frac{\sigma_2^2\mu_1 + \sigma_1^2(w-\mu_2)}{\sigma_1^2+\sigma_2^2}\right\}^2 + \frac{\sigma_2^2\mu_1^2 + \sigma_1^2(w-\mu_2)^2}{\sigma_1^2+\sigma_2^2} - \right. \\
&\qquad \left. \frac{\{\sigma_2^2\mu_1 + \sigma_1^2(w-\mu_2)\}^2}{(\sigma_1^2+\sigma_2^2)^2}\right] \qquad (7.4)
\end{aligned}$$

where (7.4) follows from (7.3) by completing the square for x. Now,

$$\begin{aligned}(\sigma_1^2+\sigma_2^2)\{\sigma_2^2\mu_1^2+\sigma_1^2(w-\mu_2)^2\}-\{\sigma_2^2\mu_1+\sigma_1^2(w-\mu_2)\}^2 &= \\ \sigma_1^2\sigma_2^2\{\mu_1^2-2\mu_1(w-\mu_2)+(w-\mu_2)^2\} &= \\ \sigma_1^2\sigma_2^2\{w-(\mu_1+\mu_2)\}^2. & \qquad (7.5)\end{aligned}$$

Substituting (7.5) into (7.4) and using (7.1) and (7.2) gives the result. □

Theorem 4 *If $X \sim N(\mu_1,\sigma_1^2)$ and $Y \sim N(\mu_2,\sigma_2^2)$ and X and Y are independent then $W = X+Y \sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$.*

Proof - We consider the region R where $W \leq w$. For a given $X = x$, $W \leq w$ provided $Y \leq w-x$. Thus, $R = \{-\infty \leq x \leq \infty, -\infty \leq y \leq w-x\}$ and

$$P(W \leq w) = \int_{-\infty}^{\infty}\int_{-\infty}^{w-x} f_{X,Y}(x,y)\,dy\,dx, \qquad (7.6)$$

where $f_{X,Y}(x,y)$ is the joint pdf of X and Y. Since X and Y are independent then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Using this and making the change of variables $y = v-x$ we may write (7.6) as

$$P(W \leq w) = \int_{-\infty}^{\infty}\int_{-\infty}^{w} f_X(x)f_Y(v-x)\,dv\,dx = \int_{-\infty}^{w}\int_{-\infty}^{\infty} f_X(x)f_Y(v-x)\,dx\,dv.$$

Differentiating both sides with respect to w gives

$$\begin{aligned}f_W(w) &= \int_{-\infty}^{\infty} f_X(x)f_Y(w-x)\,dx \\ &= \int_{-\infty}^{\infty}\frac{1}{2\pi\sigma_1\sigma_2}\exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma_1^2}(x-\mu_1)^2+\frac{1}{\sigma_2^2}(w-x-\mu_2)^2\right\}\right]dx \qquad (7.7)\end{aligned}$$

where (7.7) follows since $X \sim N(\mu_1,\sigma_1^2)$ and $Y \sim N(\mu_2,\sigma_2^2)$. From Lemma 2, (7.7) becomes

$$\begin{aligned}f_W(w) &= \int_{-\infty}^{\infty}\frac{1}{2\pi\sigma_1\sigma_2}\exp\left[-\frac{(x-\mu_3)^2}{2\sigma_3^2}-\frac{\{w-(\mu_1+\mu_2)\}^2}{2(\sigma_1^2+\sigma_2^2)}\right]dx \\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left[-\frac{\{w-(\mu_1+\mu_2)\}^2}{2(\sigma_1^2+\sigma_2^2)}\right]\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_3}\exp\left\{-\frac{(x-\mu_3)^2}{2\sigma_3^2}\right\}dx \qquad (7.8) \\ &= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}\exp\left[-\frac{\{w-(\mu_1+\mu_2)\}^2}{2(\sigma_1^2+\sigma_2^2)}\right] \qquad (7.9)\end{aligned}$$

where (7.9) follows from (7.8) as the integral is over the pdf of a $N(\mu_3,\sigma_3^2)$ random quantity and thus is equal to 1. We immediately identify (7.9) as the pdf of a $N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$ random quantity. □