Lent 2000 version of March 8, 2001 Richard Weber

C11: STATISTICS

Contents

Aims of the course

The aim of this course is to aquaint you with the basics of mathematical statistics: the ideas of estimation, hypothesis testing and statistical modelling.

After studying this material you should be familiar with

1. the notation and keywords listed on the following pages;
2. the definitions, theorems, lemmas and proofs in these notes;
3. examples in notes and examples sheets that illustrate important issues concerned with topics mentioned in the schedules.

Schedules

Estimation
Review of distribution and density functions, parametric families, sufficiency, Rao-Blackwell theorem, factorization criterion, and examples; binomial, Poisson, gamma. Maximum likelihood estimation. Confidence intervals. Use of prior distributions and Bayesian inference.

Hypothesis Testing
Simple examples of hypothesis testing, null and alternative hypothesis, critical region, size, power, type I and type II errors, Neyman-Pearson lemma. Significance level of outcome. Uniformly most powerful tests. Likelihood ratio, and the use of likelihood ratio to construct test statistics for composite hypotheses. Generalized likelihood-ratio test. Goodness-of-fit and contingency tables.

Linear normal models
The χ^2, t and F distribution, joint distribution of sample mean and variance, Student's t-test, F-test for equality of two variances. One-way analysis of variance.

Linear regression and least squares
Simple examples, *Use of software*.

Recommended books

M. H. De Groot, *Probability and Statistics*, 2nd edition, Addison-Wesley, 1986.
J. A. Rice, *Mathematical Statistics and Data Analysis*, 2nd edition, Duxbury Press, 1994.
G. Casella and J. O. Berger, *Statistical Inference*, Brooks Cole, 1990.
D. A. Berry and B. W. Lindgren, *Statistics, Theory and Methods*, Brooks Cole, 1990 (out of print).

Keywords

Notation

X	a scalar or vector random variable, $X = (X_1, \dots, X_n)$
$X \sim$	X has the distribution ...
$\mathbb{E}X$, $\mathrm{var}(X)$	mean and variance of X
μ, σ^2	mean and variance as typically used for $N(\mu, \sigma^2)$
RV, IID	'random variable', 'independent and identically distributed'
$\mathrm{beta}(m, n)$	beta distribution
$B(n, p)$	binomial distribution
χ^2_n	chi-squared distribution with n d.f.
$\mathcal{E}(\lambda)$	exponential distribution
$F_{m,n}$	F distribution with m and n d.f.
$\mathrm{gamma}(n, \lambda)$	gamma distribution
$N(\mu, \sigma^2)$	normal (Gaussian) distribution
$P(\lambda)$	Poisson distribution
$U[a, b]$	uniform distribution
t_n	Student's t distribution with n d.f.
Φ	distribution function of $N(0, 1)$
ϕ	density function of $N(0, 1)$
z_α, $t^{(n)}_\alpha$, $F^{(m,n)}_\alpha$	upper α points of $N(0, 1)$, t_n and $F_{m,n}$ distributions
θ	a parameter of a distribution
$\hat\theta(X)$, $\hat\theta(x)$	an estimator of θ, a estimate of θ.
MLE	'maximum likelihood estimator'
$F_X(x \mid \theta)$	distribution function of X depending on a parameter θ
$f_X(x \mid \theta)$	density function of X depending on a parameter θ
$f_\theta(x)$	density function depending on a parameter θ
$f_{X\mid Y}$	conditional density of X given Y
$p(\theta \mid x)$	posterior density of θ given data x
$x_1, \dots, x_n$	n observed data values
$x_{i\cdot}$, $x_{\cdot j}$, $x_{\cdot\cdot}$	$\sum_j x_{ij}$, $\sum_i x_{ij}$ and $\sum_{ij} x_{ij}$
$T(x)$	a statistic computed from $x_1, \dots, x_n$

H_0, H_1	null and alternative hypotheses
f_0, f_1	null and alternative density functions
$L_x(H_0), L_x(H_1)$	likelihoods of H_0 and H_1 given data x
$L_x(H_0, H_1)$	likelihood ratio $L_x(H_1)/L_x(H_0)$
$t_\alpha^{(n)}$, $F_\alpha^{(m,n)}$	points to the right of which lie $\alpha 100\%$ of T_n and $F_{m,n}$
C	critical region: reject H_0 if $T(x) \in C$.
$W(\theta)$	power function, $W(\theta) = \mathbb{P}(X \in C \mid \theta)$
α,β	probabilities of Type I and Type II error
	intercept and gradient of a regression line, $Y_i = \alpha + \beta w_i + \epsilon_i$
o_i, e_i, δ_i	observed and expected counts; $\delta_i = o_i - e_i$
$\bar{X}$	mean of $X_1, \dots, X_n$
S_{XX}, S_{YY}, S_{XY}	$\sum(X_i - \bar{X})^2$, $\sum(Y_i - \bar{Y})^2$, $\sum(X_i - \bar{X})(Y_i - \bar{Y})$
s.e.	'standard error', square root of an unbiased estimator of a variance.
R	residual sum of square in a regression model
s^2	unbiased estimate of the variance, $s^2 = S_{XX}/(n-1)$.
$d(X)$	decision function, $d(X) = a$.
$L(\theta, a)$	loss function when taking action a.
$R(\theta, d)$	risk function, $R(\theta, d) = \mathbb{E}[L(\theta, d(X))]$.
$B(d)$	Bayes risk, $\mathbb{E}[R(\theta, d)]$.

WWW site

There is a web page for this course, with copies of the lecture notes, examples sheets, corrections, past tripos questions, statistical tables and other additional material. It can be accessed as `http://www.statslab.cam.ac.uk/~rrw1/stats/`

1 Parameter estimation

Statisticians do it when it counts.

1.1 What is Statistics?

Statistics *is a collection of procedures and principles for gaining and processing information in order to make decisions when faced with uncertainty.*

This course is concerned with "Mathematical Statistics", i.e., mathematical ideas and tools that are used by statisticians to analyse data. We will study techniques for estimating parameters, fitting models, and testing hypotheses. However, as we study these techniques we should not lose sight of the fact that a practicing statistician needs more than simply a knowledge of mathematical techniques. The collection and interpretation of data is a subtle art. It requires common sense. It can sometimes raise philosophical questions. Although this course is primarily concerned with mathematical techniques, I will try — by means of examples and digressions — also to introduce you to some of the non-mathematical aspects of Statistics.

Statistics is concerned with *data analysis*: using data to make inferences. It is concerned with questions like 'what is this data telling me?' and 'what does this data suggest it reasonable to believe?' Two of its principal concerns are **parameter estimation** and **hypothesis testing**.

Example 1.1 *Suppose we wish to estimate the proportion p of students in Cambridge who have not showered or bathed for over a day.*

This is poses a number of questions. Who do we mean by students? Suppose time is limited and we can only interview 20 students in the street. Is it important that our survey be 'random'? How can we ensure this? Will students we question be embarrassed to admit if they have not bathed? And even if we can get truthful answers, will we be happy with our estimate if that random sample turns out to include no women, or if it includes only computer scientists?

Suppose we find that 5 have not bathed for over a day. We might estimate p by $\hat{p} = 5/20 = 0.25$. But how large an error might we expect $\hat{p}$ to have?

Many families of probability distributions depend on a small number of parameters; for example, the Poisson family depends on a single parameter λ and the Normal family on two parameters μ and σ. Unless the values of the parameters are known in advance, they must be estimated from data. One major themes of mathematical statistics is the theory of **parameter estimation** and its use in fitting probability distributions to data. A second major theme of Statistics is **hypothesis testing**.

Example 1.2 *A famous study investigated the effects upon heart attacks of taking an aspirin every other day.* The results after 5 years were

Condition	**Heart attack**	**No heart attack**	**Attacks per 1000**
Aspirin	104	10,933	9.42
Placebo	189	10,845	17.13

What can make of this data? Is it evidence for the hypothesis that aspirin prevents heart attacks?

The aspirin study is an example of a *controlled experiment*. The subjects were doctors aged 40 to 84 and none knew whether they were taking the aspirin or the placebo. Statistics is also concerned with analysing data from *observational studies*. For example, most of us make an intuitive statistical analysis when we use our previous experience to help us choose the shortest checkout line at a supermarket.

The data analysis of observational studies and experiments is a central component of **decision-making**, in science, medicine, business and government.

By the way: **data** is a plural noun referring to a collection of numbers or other pieces of information to which meaning has been attached.

The numbers 1.1, 3.0, 6.5 are not necessarily data. They become so when we are told that they are the muscle weight gains in kg of three athletes who have been trying a new diet.

1.2 RVs with values in $\mathbb{R}^n$ or $\mathbb{Z}^n$

In Statistics, our data are modelled by a vector of random variables

$$X = (X_1, X_2, \dots, X_n)$$

where X_i takes values in $\mathbb{Z}$ or $\mathbb{R}$.

To succeed in this course you should brush up on your knowledge of basic probability: of key distributions and how to make calculations with random variables. Let us review a few facts.

When our sample space Ω (a set of possible outcomes) is discrete (finite or countably infinite) we have a random variable (**RV**) X with values in $\mathbb{Z}$:

$$X \colon \Omega \to \mathbb{Z}.$$

RVs can also take values in $\mathbb{R}$ rather than in $\mathbb{Z}$ and the sample space Ω can be uncountable.

$$X \colon \Omega \to \mathbb{R}.$$

Since the outcome ω, $\omega \in \Omega$, is random, X is a function whose value, $X(\omega)$, is also random. E.g., to model the experiment of tossing a coin twice we might take $\Omega = \{hh, ht, th, th\}$. Then X might be the total number of heads.

In both cases the **distribution function** F_X of X is defined as:

$$F_X(x) := \mathbb{P}(X \le x) = \sum_{\{\omega\colon X(\omega) \le x\}} \mathbb{P}(\omega).$$

In the discrete case the **probability mass function** (pmf) f_X of X is

$$f_X(k) := \mathbb{P}(X = k), \qquad k \in \mathbb{Z}.$$

So

$$\mathbb{P}(X \in A) = \sum_{x \in A} f_X(x), \quad A \subseteq \mathbb{Z}.$$

In the continuous case we have the **probability density function** (pdf) f_X of X. In all cases we shall meet, X will have a piecewise smooth pdf such that

$$\mathbb{P}(X \in A) = \int_{x \in A} f_X(x)\, dx, \quad \text{for } \textit{nice} \text{ (measurable) subsets } A \subseteq \mathbb{R}.$$

Expectation of X: In the discrete case

$$\mathbb{E}(X) := \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) = \sum_{k \in \mathbb{Z}} k\, \mathbb{P}(X = k),$$

the first formula being the real definition. In the continuous case the calculation

$$E(X) = \int_\Omega X(\omega)\, \mathbb{P}(d\omega)$$

needs measure theory. However,

$$f_X(x) = \frac{d}{dx} F_X(x) \quad \text{except perhaps for finitely many } x.$$

Measure theory shows that for any nice function h on $\mathbb{R}$,

$$\mathbb{E}\, h(X) = \int_{\mathbb{R}} h(x) f_X(x)\, dx\ .$$

Variance of X: If $\mathbb{E}(X) = \mu$, then

$$\operatorname{var}(X) = \mathbb{E}(X - \mu)^2 = \mathbb{E}(X^2) - \mu^2.$$

1.3 Some important random variables

(a) We say that X has the **binomial distribution** $B(n,p)$, and write $X \sim B(n,p)$, if

$$\mathbb{P}(X=k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k \in \{0,\dots,n\}, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathbb{E}(X) = np$, $\text{var}(X) = np(1-p)$. This is the distribution of the number of successes in n independent trials, each of which has probability of success p.

(b) We say that X has the **Poisson distribution** with parameter λ, and write $X \sim P(\lambda)$, if

$$\mathbb{P}(X=k) = \begin{cases} e^{-\lambda}\lambda^k/k! & \text{if } k \in \{0,1,2,\dots\}, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\mathbb{E}(X) = \text{var}(X) = \lambda$. The Poisson is the limiting distribution of $B(n,p)$ as $n \to \infty$ and $p \to 0$ with $\lambda = np$.

(c) We say that X is **standard normal**, and write $X \sim N(0,1)$, if

$$f_X(x) = \varphi(x) := \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \qquad -\infty \le x \le \infty.$$

Then

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy = \Phi(x) := \int_{-\infty}^{x} \varphi(y)\,dy.$$

Then $\mathbb{E}(X) = 0$, $\text{var}(X) = 1$. Φ and φ are standard notations.

(d) We say that X is **Normal** with mean μ and variance σ^2 and write $X \sim N(\mu, \sigma^2)$ if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty \le x \le \infty.$$

Then $\mathbb{E}(X) = \mu$, $\text{var}(X) = \sigma^2$.

(e) We say that X is **uniform** on $[a,b]$, and write $X \sim U[a,b]$, if

$$f_X(x) = \frac{1}{b-a}, \qquad x \in [a,b].$$

Then $\mathbb{E}(X) = \frac{1}{2}(a+b)$, $\text{var}(X) = \frac{1}{12}(b-a)^2$.

1.4 Independent and IID RVs

Random variables $X_1, \dots, X_n$ are called **independent** if for all $x_1, \dots, x_n$

$$\mathbb{P}(X_1 \le x_1; \dots; X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n).$$

IID stands for independent identically distributed. Thus if $X_1, X_2, \dots, X_n$ are IID RVs, then they all have the same distribution function and hence the same mean and same variance.

We work with the probability mass function (pmf) of X in $\mathbb{Z}^n$ or probability density function (pdf) of X in $\mathbb{R}^n$: In most cases, $X_1, \dots, X_n$ are independent, so that if $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, then

$$f_X(x) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

1.5 Indicating dependence on parameters

If $X \sim N(\mu, \sigma^2)$, then we indicate the dependence of the pdf of X on μ and σ^2 by writing it as

$$f(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Or if $X = (X_1, \dots, X_n)$, where $X_1, \dots, X_n$ are IID $N(\mu, \sigma^2)$, then we would have

$$f(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\|x - \mu\mathbf{1}\|^2}{2\sigma^2}\right)$$

where $\mu\mathbf{1}$ denotes the vector $(\mu, \mu, \dots, \mu)^\top$.

In general, we write $f(x \mid \theta)$ to indicate that the pdf depends on a parameter θ. θ may be a vector of parameters. In the above $\theta = (\mu, \sigma^2)^\top$. An alternative notation we will sometimes employ is $f_\theta(x)$.

The set of distributions with densities $f_\theta(x)$, $\theta \in \Theta$, is called a **parametric family**. E.g.,, there is a parametric family of normal distributions, parameterised by values of μ, σ^2. Similarly, there is a parametric family of Poisson distributions, parameterised by values of λ.

1.6 The notion of a statistic

A **statistic**, $T(x)$, is any function of the data. E.g., given the data $x = (x_1, \dots, x_n)$, four possible statistics are

$$\frac{1}{n}(x_1 + \cdots + x_n), \quad \max_i x_i, \quad \frac{x_1 + x_3}{x_n} \log x_4, \quad 1997 + 10 \min_i x_i.$$

Clearly, some statistics are more natural and useful than others. The first of these would be useful for estimating μ if the data are samples from a $N(\mu, 1)$ distribution. The second would be useful for estimating θ if the data are samples from $U[0, \theta]$.

1.7 Unbiased estimators

An **estimator** of a parameter θ is a function $T = T(X)$ which we use to estimate θ from an observation of X. T is said to be **unbiased** if

$$\mathbb{E}(T) = \theta.$$

The expectation above is taken over X. Once the actual data x is observed, $t = T(x)$ is the **estimate** of θ obtained via the estimator T.

Example 1.3 *Suppose $X_1, \dots, X_n$ are IID $B(1, p)$ and p is unknown. Consider the estimator for p given by $\hat{p}(X) = \bar{X} = \sum_i X_i/n$.* Then $\hat{p}$ is unbiased, since

$$\mathbb{E}\hat{p}(X) = \mathbb{E}\left[\frac{1}{n}(X_1 + \cdots + X_n)\right] = \frac{1}{n}(\mathbb{E}X_1 + \cdots + \mathbb{E}X_n) = \frac{1}{n}np = p\,.$$

Another possible unbiased estimator for p is $\tilde{p} = \frac{1}{3}(X_1 + 2X_2)$ (i.e., we ignore most of the data.) It is also unbiased since

$$\mathbb{E}\tilde{p}(X) = \mathbb{E}\left[\frac{1}{3}(X_1 + 2X_2)\right] = \frac{1}{3}(\mathbb{E}X_1 + 2\mathbb{E}X_2) = \frac{1}{3}(p + 2p) = p\,.$$

Intuitively, the first estimator seems preferable.

1.8 Sums of independent RVs

In the above calculations we have used the fact the expectation of a sum of random variables is the sum of their expectations. It is always true (even when $X_1, \dots, X_n$ are not independent) that

$$\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n),$$

and for linear combinations of RVs

$$\mathbb{E}(a_1 X_1 + \cdots + a_n X_n) = a_1\mathbb{E}(X_1) + \cdots + a_n\mathbb{E}(X_n).$$

If $X_1, X_2, \dots, X_n$ are independent, then

$$\mathbb{E}(X_1 X_2 \dots X_n) = \mathbb{E}(X_1)\mathbb{E}(X_2)\cdots\mathbb{E}(X_n),$$
$$\operatorname{var}(X_1 + \cdots + X_n) = \operatorname{var}(X_1) + \cdots + \operatorname{var}(X_n),$$

and for linear combinations of independent RVs

$$\operatorname{var}(a_1 X_1 + \cdots + a_n X_n) = a_1^2 \operatorname{var}(X_1) + \cdots + a_n^2 \operatorname{var}(X_n).$$

1.9 More important random variables

(a) We say that X is **geometric** with parameter p, if

$$\mathbb{P}(X=k)=\begin{cases} p(1-p)^{k-1} & \text{if } k\in\{1,2,\dots\},\\ 0 & \text{otherwise.}\end{cases}$$

Then $E(X)=1/p$ and $\operatorname{var}(X)=(1-p)/p^2$. X is the number of the toss on which we first observe a head if we toss a coin which shows heads with probability p.

(b) We say that X is **exponential** with **rate** λ, and write $X\sim\mathcal{E}(\lambda)$, if

$$f_X(x)=\begin{cases} \lambda e^{-\lambda x} & \text{if } x>0,\\ 0 & \text{otherwise.}\end{cases}$$

Then $\mathbb{E}(X)=\lambda^{-1}$, $\operatorname{var}(X)=\lambda^{-2}$.

The geometric and exponential distributions are discrete and continuous analogues. They are the unique 'memoryless' distributions, in the sense that $\mathbb{P}(X\geq t+s\mid X\geq t)=\mathbb{P}(X\geq s)$. The exponential is the distribution of the time between successive events of a Poisson process.

(c) We say that X is **gamma**(n,λ) if

$$f_X(x)=\begin{cases} \lambda^n x^{n-1}e^{-\lambda x}/(n-1)! & \text{if } x>0,\\ 0 & \text{otherwise.}\end{cases}$$

X has the distribution of the sum of n IID RVs that have distribution $\mathcal{E}(\lambda)$. So $\mathcal{E}(\lambda)=\text{gamma}(1,\lambda)$. $\mathbb{E}(X)=n\lambda^{-1}$ and $\operatorname{var}(X)=n\lambda^{-2}$.

This also makes sense for real $n>0$ (and $\lambda>0$), if we interpret $(n-1)!$ as $\Gamma(n)$, where $\Gamma(n)=\int_0^\infty x^{n-1}e^{-x}\,dx$.

(d) We say that X is **beta**(a,b) if

$$f_X(x)=\begin{cases} \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1} & \text{if } 0<x<1,\\ 0 & \text{otherwise.}\end{cases}$$

Here $B(a,b)=\Gamma(a)\Gamma(b)/\Gamma(a+b)$. Then

$$\mathbb{E}(X)=\frac{a}{a+b},\qquad \operatorname{var}(X)=\frac{ab}{(a+b+1)(a+b)^2}.$$

1.10 Laws of large numbers

Suppose $X_1,X_2,\dots$ is a sequence of IID RVs, each having finite mean μ and variance σ^2. Let

$$S_n:=X_1+X_2+\cdots+X_n,\ \text{ so that } \mathbb{E}(S_n)=n\mu, \operatorname{var}(S_n)=n\sigma^2.$$

The **weak law of large numbers** is that for $\epsilon > 0$,

$$\mathbb{P}(|S_n/n - \mu| > \epsilon) \to 0, \text{ as } n \to \infty\,.$$

The **strong law of large numbers** is that

$$\mathbb{P}(S_n/n \to \mu) = 1\,.$$

1.11 The Central Limit Theorem

Suppose $X_1, X_2, \ldots$ are as above. Define the **standardized** version S_n^* of S_n as

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}\,, \quad \text{so that } \mathbb{E}(S_n^*) = 0, \mathrm{var}(S_n^*) = 1.$$

Then for large n, S_n^* is approximately standard Normal: for $a < b$,

$$\lim_{n\to\infty} \mathbb{P}(a \le S_n^* \le b) = \Phi(b) - \Phi(a) = \lim_{n\to\infty} \mathbb{P}\left(n\mu + a\sigma\sqrt{n} \le S_n \le n\mu + b\sigma\sqrt{n}\right).$$

In particular, for large n,

$$\mathbb{P}(|S_n - n\mu| < 1.96\sigma\sqrt{n}) \doteqdot 95\%$$

since $\Phi(1.96) = 0.0975$ and $\Phi(-1.96) = 0.025$.

1.12 Poisson process of rate λ

The Poisson process is used to model a process of arrivals: of people to a supermarket checkout, calls at telephone exchange, etc.

Arrivals happen at times

$$T_1,\ T_1 + T_2,\ T_1 + T_2 + T_3,\ \ldots$$

where $T_1, T_2, \ldots$ are independent and each exponentially distributed with parameter λ. Numbers of arrivals in disjoint intervals are independent RVs, and the number of arrivals in any interval of length t has the $P(\lambda t)$ distribution. The time

$$S_n = T_1 + T_2 + \cdots + T_n$$

of the nth arrival has the gamma(n, λ) distribution, and $2\lambda S_n \sim \mathcal{X}^2_{2n}$.

2 Maximum likelihood estimation

When it is not in our power to follow what is true, we ought to follow what is most probable. (Descartes)

2.1 Maximum likelihood estimation

Suppose that the random variable X has probability density function $f(x \mid \theta)$. Given the observed value x of X, the **likelihood** of θ is defined by

$$\text{lik}(\theta) = f(x \mid \theta)\,.$$

Thus we are considering the density as a function of θ, for a fixed x. In the case of multiple observations, i.e., when $x = (x_1, \dots, x_n)$ is a vector of observed values of $X_1, \dots, X_n$, we assume, unless otherwise stated, that $X_1, \dots, X_n$ are IID; in this case $f(x_1, \dots, x_n \mid \theta)$ is the product of the marginals,

$$\text{lik}(\theta) = f(x_1, \dots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)\,.$$

It makes intuitive sense to estimate θ by whatever value gives greatest likelihood to the observed data. Thus the **maximum likelihood estimate** $\hat{\theta}(x)$ of θ is defined as the value of θ that maximizes the likelihood. Then $\hat{\theta}(X)$ is called the **maximum likelihood estimator** (**MLE**) of θ.

Of course, the maximum likelihood estimator need not exist, but in many examples it does. In practice, we usually find the MLE by maximizing $\log f(x \mid \theta)$, which is known as the **log-likelihood**.

Examples 2.1

(a) Smarties are sweets which come in k equally frequent colours. Suppose we do not know k. We sequentially examine 3 Smarties and they are red, green, red. The likelihood of this data, $x =$ *the second Smartie differs in colour from the first but the third Smartie matches the colour of the first*, is

$$\begin{aligned}\text{lik}(k) = p(x \mid k) &= \mathbb{P}(\text{2nd differs from 1st})\mathbb{P}(\text{3rd matches 1st}) = \left(\frac{k-1}{k}\right)\frac{1}{k}\\ &= (k-1)/k^2\,,\end{aligned}$$

which equals 1/4, 2/9, 3/16 for $k = 2, 3, 4$, and continues to decrease for greater k. Hence the maximum likelihood estimate is $\hat{k} = 2$.

Suppose a fourth Smartie is drawn and it is orange. Now

$$\text{lik}(k) = (k-1)(k-2)/k^3\,,$$

which equals 2/27, 3/32, 12/125, 5/54 for $k = 3, 4, 5, 6$, and decreases thereafter. Hence the likelihood estimate is $\hat{k} = 5$. Note that although we have seen only 3 colours the maximum likelihood estimate is that there are 2 colours we have not yet seen.

(b) $X \sim B(n, p)$, *n known, p to be estimated.*

Here

$$\log p(x \mid n, p) = \log \binom{n}{x} p^x (1-p)^{n-x} = \cdots + x \log p + (n - x) \log(1-p)\ .$$

This is maximized where

$$\frac{x}{\hat{p}} - \frac{n-x}{1-\hat{p}} = 0\,,$$

so the MLE of p is $\hat{p} = X/n$. Since $\mathbb{E}[X/n] = p$ the MLE is unbiased.

(c) $X \sim B(n, p)$, *p known, n to be estimated.*

Now we want to maximize

$$p(x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

with respect to n, $n \in \{x, x+1, \ldots\}$. To do this we look at the ratio

$$\frac{p(x \mid n+1, p)}{p(x \mid n, p)} = \frac{\binom{n+1}{x} p^x (1-p)^{n+1-x}}{\binom{n}{x} p^x (1-p)^{n-x}} = \frac{(1-p)(n+1)}{n+1-x}\,.$$

This is monotone decreasing in n. Thus $p(x \mid n, p)$ is maximized by the least n for which the above expression is ≤ 1, i.e., the least n such that

$$(1-p)(n+1) \leq n + 1 - x \iff n + 1 \geq x/p\,,$$

giving a MLE of $\hat{n} = [X/p]$. Note that if x/p happens to be an integer then both $n = x/p - 1$ and $n = x/p$ maximize $p(x \mid n, p)$. Thus the MLE need not be unique.

(d) $X_1, \ldots, X_n \sim$ geometric(p), *p to be estimated.*

Because the X_i are IID their joint density is the product of the marginals, so

$$\log f(x_1, \ldots, x_n \mid p) = \log \prod_{i=1}^{n} (1-p)^{x_i - 1} p = \left(\sum_{i=1}^{n} x_i - n \right) \log(1-p) + n \log p\,.$$

with a maximum where

$$-\frac{\sum_i x_i - n}{1 - \hat{p}} + \frac{n}{\hat{p}} = 0\ .$$

So the MLE is $\hat{p} = \bar{X}^{-1}$. This MLE is **biased**. For example, in the case $n = 1$,

$$\mathbb{E}[1/X_1] = \sum_{x=1}^{\infty} \frac{1}{x} (1-p)^{x-1} p = -\frac{p}{1-p} \log p > p\,.$$

Note that $\mathbb{E}[1/X_1]$ does not equal $1/\mathbb{E}X_1$.

2.2 Sufficient statistics

The MLE, if it exists, is always a function of a **sufficient statistic**. The informal notion of a sufficient statistic $T = T(X_1, \dots, X_n)$ is that it summarises all information in $\{X_1, \dots, X_n\}$ which is relevant to inference about θ.

Formally, the statistic $T = T(X)$ is said to be **sufficient** for $\theta \in \Theta$ if, for each t, $P_\theta\big(X \in \cdot \mid T(X) = t\big)$ does not depend on θ. I.e., the conditional distribution of $X_1, \dots, X_n$ given $T(X) = t$ does not involve θ. Thus to know more about x than that $T(x) = t$ is of no additional help in making any inference about θ.

Theorem 2.2 *The statistic T is sufficient for θ if and only if $f(x \mid \theta)$ can be expressed as*

$$f(x \mid \theta) = g\big(T(x), \theta\big)h(x).$$

This is called the **factorization criterion**.

Proof. We give a proof for the case that the sample space is discrete. A continuous sample space needs measure theory. Suppose $f(x \mid \theta) = \mathbb{P}_\theta(X = x)$ has the factorization above and $T(x) = t$. Then

$$\begin{aligned}\mathbb{P}_\theta\big(X = x \mid T(X) = t\big) &= \frac{\mathbb{P}_\theta(X = x)}{\mathbb{P}_\theta\big(T(X) = t\big)} = \frac{g\big(T(x), \theta\big)h(x)}{\sum_{x:T(x)=t} g\big(T(x), \theta\big)h(x)} \\ &= \frac{g(t, \theta)h(x)}{\sum_{x:T(x)=t} g(t, \theta)h(x)} = \frac{h(x)}{\sum_{x:T(x)=t} h(x)}\end{aligned}$$

which does not depend on θ. Conversely, if T is sufficient and $T(x) = t$,

$$\mathbb{P}_\theta(X = x) = \mathbb{P}_\theta\big(T(X) = t\big)\mathbb{P}_\theta\big(X = x \mid T(X) = t\big)$$

where by sufficiency the second factor does not depend on θ. So we identify the first and second terms on the r.h.s. as $g(t, \theta)$ and $h(x)$ respectively. ■

Examples 2.3

(a) $X_1, \dots, X_n \sim P(\lambda)$, *$\lambda$ to be estimated.*

$$f(x \mid \lambda) = \prod_{i=1}^{n} \{\lambda^{x_i} e^{-\lambda}/x_i!\} = \lambda^{\sum_i x_i} e^{-n\lambda} \Big/ \prod_{i=1}^{n} x_i!\,.$$

So $g\big(T(x), \lambda\big) = \lambda^{\sum_i x_i} e^{-n\lambda}$ and $h(x) = 1 \big/ \prod_i x_i!$. A sufficient statistic is $t = \sum_i x_i$.

Note that the sufficient statistic is not unique. If $T(X)$ is a sufficient statistic, then so are statistics like $T(X)/n$ and $\log T(X)$.

The MLE is found by maximizing $f(x \mid \lambda)$, and so

$$\frac{d}{d\lambda} \log f(x \mid \lambda)\bigg|_{\lambda=\hat{\lambda}} = \frac{\sum_i x_i}{\hat{\lambda}} - n = 0 \,.$$

Hence $\hat{\lambda} = \bar{X}$. It is easy to check that $\hat{\lambda}$ is unbiased.

Note that the MLE is always a function of the sufficient statistic. This is because the MLE is the value of θ which maximizes $f(x \mid \theta)$, and $f(x \mid \theta) = g\big(T(x), \theta\big)h(x)$. Thus the MLE is the θ which maximizes $g\big(T(x), \theta\big)$, and hence a function of $T(x)$.

(b) *$X_1, \dots, X_n \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$ to be estimated.*

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_i (x_i-\mu)^2/2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\left[\sum_i (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2)\right]/2\sigma^2} \end{aligned}$$

Thus, with $g\big(T(x), \theta\big)$ taken as the whole r.h.s. and $h(x) = 1$, the sufficient statistic for $\theta = (\mu, \sigma^2)$ is $T(x) = \big(\bar{x}, \sum_i (x_i - \bar{x})^2\big)$.

Note that sometimes the sufficient statistic is not just a single number, but as here, a vector $T(X) = \big(T_1(X), \dots, T_r(X)\big)$. This usually occurs when the parameter is a vector, as $\theta = (\mu, \sigma^2)$.

In this example, if σ^2 had been known, then $\bar{x}$ would have been sufficient for μ. If μ had been known, then $\sum_i (x_i - \mu)^2$ would have been sufficient for σ^2.

(c) *$X_1, \dots, X_k \sim U[0, \theta]$, θ to be estimated.*

$$f(x \mid \theta) = \prod_{i=1}^{n} 1\{0 \le x_i \le \theta\} \frac{1}{\theta} = 1\{0 \le \max_i x_i \le \theta\} \frac{1}{\theta^n} \,,$$

where $1\{\text{condition}\} = 1$ or 0 as 'condition' is true or false. Thus $g\big(T(x), \theta\big) = 1\{0 \le \max_i x_i \le \theta\}/\theta^n$, $h(x) = 1$ and $T(x) = \max_i x_i$ is sufficient for θ. The MLE is $\hat{\theta} = \max_i X_i$.

To find $\mathbb{E}\hat{\theta}$ we must find the distribution function of $\max_i x_i$. This is

$$F(t) = \mathbb{P}(\max_i x_i \le t) = \mathbb{P}(x_1 \le t) \cdots \mathbb{P}(x_n \le t) = (t/\theta)^n \,.$$

By differentiation, $f(t) = nt^{n-1}/\theta^n$, and hence

$$\mathbb{E} \max_i x_i = \int_0^\theta t \frac{nt^{n-1}}{\theta^n} \, dt = \frac{n}{n+1} \theta \,.$$

So $\hat{\theta}$ is biased.

However, $\mathbb{E}\hat{\theta} \to \theta$ as $n \to \infty$. We say that $\hat{\theta}$ is **asymptotically unbiased**. Under some mild assumptions, MLEs are always asymptotically unbiased. This is one reason why we like the maximum likelihood estimation procedure.

3 The Rao-Blackwell theorem

Variance is what any two statisticians are at.

3.1 Mean squared error

A good estimator should take values close to the true value of the parameter it is attempting to estimate. If $\hat{\theta}$ is an unbiased estimator of θ then $\mathbb{E}(\hat{\theta}-\theta)^2$ is the variance of $\hat{\theta}$. If $\hat{\theta}$ is a biased estimator of θ then $\mathbb{E}(\hat{\theta}-\theta)^2$ is no longer the variance of $\hat{\theta}$, but it is still useful as a measure of the **mean squared error** (**MSE**) of $\hat{\theta}$.

Example 3.1 Consider the estimators in Example 1.3. Each is unbiased, so its MSE is just its variance.

$$\operatorname{var}(\hat{p}) = \operatorname{var}\left[\frac{1}{n}(X_1+\cdots+X_n)\right] = \frac{\operatorname{var}(X_1)\cdots+\operatorname{var}(X_n)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$\operatorname{var}(\tilde{p}) = \operatorname{var}\left[\frac{1}{3}(X_1+2X_2)\right] = \frac{\operatorname{var}(X_1)+4\operatorname{var}(X_2)}{9} = \frac{5p(1-p)}{9}$$

Not surprisingly, $\operatorname{var}(\hat{p}) < \operatorname{var}(\tilde{p})$. In fact, $\operatorname{var}(\hat{p})/\operatorname{var}(\tilde{p}) \to 0$, as $n \to \infty$.

Note that $\hat{p}$ is the MLE of p. Another possible unbiased estimator would be

$$p^* = \frac{1}{\frac{1}{2}n(n+1)}(X_1+2X_2+\cdots+nX_n)$$

with variance

$$\operatorname{var}(p^*) = \frac{1}{\left[\frac{1}{2}n(n+1)\right]^2}\left(1+2^2+\cdots+n^2\right)p(1-p) = \frac{2(2n+1)}{3n(n+1)}p(1-p)\,.$$

Here $\operatorname{var}(\hat{p})/\operatorname{var}(p^*) \to 3/4$.

The next example shows that neither a MLE or an unbiased estimator necessarily minimizes the mean square error.

Example 3.2 Suppose $X_1,\dots,X_n \sim N(\mu,\sigma^2)$, μ and σ^2 unknown and to be estimated. To find the MLEs we consider

$$\log f(x \mid \mu,\sigma^2) = \log\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x_i-\mu)^2/2\sigma^2} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\,.$$

This is maximized where $\partial(\log f)/\partial\mu = 0$ and $\partial(\log f)/\partial\sigma^2 = 0$. So

$$(1/\hat{\sigma}^2)\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0, \qquad \text{and} \qquad -\frac{n}{2\hat{\sigma}^2}+\frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = 0,$$

and the MLEs are

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2 = \frac{1}{n}S_{XX} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

It is easy to check that $\hat{\mu}$ is unbiased. As regards $\hat{\sigma}^2$ note that

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - n\mathbb{E}(\mu - \bar{X})^2$$
$$= n\sigma^2 - n(\sigma^2/n) = (n-1)\sigma^2$$

so $\hat{\sigma}^2$ is biased. An unbiased estimator is $s^2 = S_{XX}/(n-1)$.

Let us consider an estimator of the form λS_{XX}. Above we see S_{XX} has mean $(n-1)\sigma^2$ and later we will see that its variance is $2(n-1)\sigma^4$. So

$$\mathbb{E}\left[\lambda S_{XX} - \sigma^2\right]^2 = \left[2(n-1)\sigma^4 + (n-1)^2\sigma^4\right]\lambda^2 - 2(n-1)\sigma^4\lambda + \sigma^4 .$$

This is minimized by $\lambda = 1/(n+1)$. Thus the estimator which minimizes the mean squared error is $S_{XX}/(n+1)$ and this is neither the MLE nor unbiased. Of course there is little difference between any of these estimators when n is large.

Note that $\mathbb{E}[\hat{\sigma}^2] \to \sigma^2$ as $n \to \infty$. So again the MLE is asymptotically unbiased.

3.2 The Rao-Blackwell theorem

The following theorem says that if we want an estimator with small MSE we can confine our search to estimators which are functions of the sufficient statistic.

Theorem 3.3 (Rao-Blackwell Theorem) *Let $\hat{\theta}$ be an estimator of θ with $\mathbb{E}(\hat{\theta}^2) < \infty$ for all θ. Suppose that T is sufficient for θ, and let $\theta^* = \mathbb{E}(\hat{\theta} \mid T)$. Then for all θ,*

$$\mathbb{E}(\theta^* - \theta)^2 \leq \mathbb{E}(\hat{\theta} - \theta)^2.$$

The inequality is strict unless $\hat{\theta}$ is a function of T.

Proof.

$$\mathbb{E}[\theta^* - \theta]^2$$
$$= \mathbb{E}\left[\mathbb{E}(\hat{\theta} \mid T) - \theta\right]^2 = \mathbb{E}\left[\mathbb{E}(\hat{\theta} - \theta \mid T)\right]^2 \leq \mathbb{E}\left[\mathbb{E}((\hat{\theta} - \theta)^2 \mid T)\right] = \mathbb{E}(\hat{\theta} - \theta)^2$$

The outer expectation is being taken with respect to T. The inequality follows from the fact that for any RV, W, $\mathrm{var}(W) = \mathbb{E}W^2 - (\mathbb{E}W)^2 \geq 0$. We put $W = (\hat{\theta} - \theta \mid T)$ and note that there is equality only if $\mathrm{var}(W) = 0$, i.e., $\hat{\theta} - \theta$ can take just one value for each value of T, or in other words, $\hat{\theta}$ is a function of T. ∎

Note that if $\hat{\theta}$ is unbiased then θ^* is also unbiased, since

$$\mathbb{E}\theta^* = \mathbb{E}\left[\mathbb{E}(\hat{\theta} \mid T)\right] = \mathbb{E}\hat{\theta} = \theta\,.$$

We now have a quantitative rationale for basing estimators on sufficient statistics: if an estimator is not a function of a sufficient statistic, then there is another estimator which is a function of the sufficient statistic and which is at least as good, in the sense of mean squared error of estimation.

Examples 3.4

(a) $X_1, \dots, X_n \sim P(\lambda)$, *$\lambda$ to be estimated.*

In Example 2.3 (a) we saw that a sufficient statistic is $\sum_i x_i$. Suppose we start with the unbiased estimator $\tilde{\lambda} = X_1$. Then 'Rao–Blackwellization' gives

$$\lambda^* = \mathbb{E}\left[X_1 \mid \textstyle\sum_i X_i = t\right]\ .$$

But

$$\sum_i \mathbb{E}\big[X_i \mid \textstyle\sum_i X_i = t\big] = \mathbb{E}\big[\textstyle\sum_i X_i \mid \sum_i X_i = t\big] = t\ .$$

By the fact that $X_1, \dots, X_n$ are IID, every term within the sum on the l.h.s. must be the same, and hence equal to t/n. Thus we recover the estimator $\lambda^* = \hat{\lambda} = \bar{X}$.

(b) $X_1, \dots, X_n \sim P(\lambda)$, *$\theta = e^{-\lambda}$ to be estimated.*

Now $\theta = \mathbb{P}(X_1 = 0)$. So a simple unbiased estimator is $\hat{\theta} = 1\{X_1 = 0\}$. Then

$$\begin{aligned}\theta^* &= \mathbb{E}\left[1\{X_1 = 0\} \,\middle|\, \sum_{i=1}^n X_i = t\right] = \mathbb{P}\left(X_1 = 0 \,\middle|\, \sum_{i=1}^n X_i = t\right)\\ &= \mathbb{P}\left(X_1 = 0; \sum_{i=2}^n X_i = t\right) \Big/ \mathbb{P}\left(\sum_{i=1}^n X_i = t\right)\\ &= e^{-\lambda}\,\frac{((n-1)\lambda)^t e^{-(n-1)\lambda}}{t!} \Big/ \frac{(n\lambda)^t e^{-n\lambda}}{t!} = \left(\frac{n-1}{n}\right)^t\end{aligned}$$

Since $\hat{\theta}$ is unbiased, so is θ^*. As it should be, θ^* is only a function of t. If you do Rao-Blackwellization and you do not get just a function of t then you have made a mistake.

(c) $X_1, \dots, X_n \sim U[0, \theta]$, *$\theta$ to be estimated.*

In Example 2.3 (c) we saw that a sufficient statistic is $\max_i x_i$. Suppose we start with the unbiased estimator $\tilde{\theta} = 2X_1$. Rao–Blackwellization gives

$$\theta^* = \mathbb{E}\left[2X_1 \mid \max_i X_i = t\right] = 2\left(\frac{1}{n}t + \frac{n-1}{n}(t/2)\right) = \frac{n+1}{n}t\ .$$

This is an unbiased estimator of θ. In the above calculation we use the idea that $X_1 = \max_i X_i$ with probability $1/n$, and if X_1 is not the maximum then its expected value is half the maximum. Note that the MLE $\hat{\theta} = \max_i X_i$ is biased.

3.3 Consistency and asymptotic efficiency*

Two further properties of maximum likelihood estimators are consistency and asymptotic efficiency. Suppose $\hat{\theta}$ is the MLE of θ.

To say that $\hat{\theta}$ is **consistent** means that

$$\mathbb{P}(|\hat{\theta} - \theta| > \epsilon) \to 0 \quad \text{as } n \to \infty\,.$$

In Example 3.1 this is just the weak law of large numbers:

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > \epsilon\right) \to 0\,.$$

It can be shown that $\text{var}(\tilde{\theta}) \geq 1/nI(\theta)$ for any unbiased estimate $\tilde{\theta}$, where $1/nI(\theta)$ is called the *Cramer-Rao lower bound.* To say that $\hat{\theta}$ is **asymptotically efficient** means that

$$\lim_{n\to\infty} \text{var}(\hat{\theta})/[1/nI(\theta)] = 1\,.$$

The MLE is asymptotically efficient and so asymptotically of minimum variance.

3.4 Maximum likelihood and decision-making

We have seen that the MLE is a function of the sufficient statistic, asymptotically unbiased, consistent and asymptotically efficient. These are nice properties. But consider the following example.

Example 3.5 You and a friend have agreed to meet sometime just after 12 noon. You have arrived at noon, have waited 5 minutes and your friend has not shown up. You believe that either your friend will arrive at X minutes past 12, where you believe X is exponentially distributed with an unknown parameter λ, $\lambda > 0$, or that she has completely forgotten and will not show up at all. We can associate the later event with the parameter value $\lambda = 0$. Then

$$\mathbb{P}(\text{data} \mid \lambda) = \mathbb{P}(\text{you wait at least 5 minutes} \mid \lambda) = \int_5^\infty \lambda e^{-\lambda t}\, dt = e^{-5\lambda}\,.$$

Thus the maximum likelihood estimator for λ is $\hat{\lambda} = 0$. If you base your decision as to whether or not you should wait a bit longer only upon the maximum likelihood estimator of λ, then you will estimate that your friend will never arrive and decide not to wait. This argument holds even if you have only waited 1 second.

The above analysis is unsatisfactory because we have not modelled the costs of either waiting in vain, or deciding not to wait but then having the friend turn up.

4 Confidence intervals

Statisticians do it with 95% confidence.

4.1 Interval estimation

Let $a(X)$ and $b(X)$ be two statistics satisfying $a(X) \leq b(X)$ for all X. Suppose that on seeing the data $X = x$ we make the inference $a(x) \leq \theta \leq b(x)$. Here $[a(x), b(x)]$ is called an **interval estimate** and $[a(X), b(X)]$ is called an **interval estimator**.

Previous lectures were concerned with making a **point estimate** for θ. Now we are being less precise. By giving up precision in our assertion about the value of θ we gain confidence that our assertion is correct. Suppose

$$P_\theta\Big(a(X) \leq \theta \leq b(X)\Big) = \gamma,$$

where γ does not depend on θ. Then the random interval $\big[a(X), b(X)\big]$ is called a $100\gamma\%$ **confidence interval** for θ. Typically γ is 0.95 or 0.99, so that the probability the interval contains θ is close to 1.

Given data x, we would call $[a(x), b(x)]$ a '$100\gamma\%$ confidence interval for θ'. Notice however, that θ is fixed, and therefore the interval either does or does not contain the true θ. However, if we repeat the procedure of sampling and constructing a a confidence interval many times, then our confidence interval will contain the true θ $100\gamma\%$ of the time. The point to understand here is that *it is the endpoints of the confidence interval that are random variables, not the parameter θ.*

Examples 4.1

(a) If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ independently, with μ unknown and σ^2 known, then

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad \text{and hence} \quad \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0,1)\,.$$

So if ξ and η are such that $\mathbb{P}(\xi \leq N(0,1) \leq \eta) = \gamma$, we have

$$P_{(\mu,\sigma^2)}\left(\xi \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \leq \eta\right) = \gamma\,,$$

which can be rewritten as

$$P_{(\mu,\sigma^2)}\left(\bar{X} - \frac{\eta\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - \frac{\xi\sigma}{\sqrt{n}}\right)\,.$$

Note that the choice of ξ and η is not unique. However, it is natural to try to make the length of the confidence interval as small as possible, so the symmetry of the normal distribution implies that we should take ξ and η symmetric about 0.

Hence for a 95% confidence interval we would take $-\xi = \eta = 1.96$, as $\Phi(1.96) = 0.975$. The 95% confidence interval is

$$\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right]$$

For a 99% confidence interval, 1.96 would be replaced by 2.58, as $\Phi(2.58) = 0.995$.

(b) If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ independently, with μ and σ^2 both unknown, then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{S_{XX}/(n-1)}} \sim t_{n-1},$$

where t_{n-1} denotes the '**Student's *t*-distribution** on $n-1$ degrees of freedom' which will be studied later. So if ξ and η are such that $\mathbb{P}(\xi \leq t_{n-1} \leq \eta) = \gamma$, we have

$$P_{(\mu,\sigma^2)}\left(\xi \leq \frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\{S_{XX}/(n-1)\}}} \leq \eta\right) = \gamma,$$

which can be rewritten as

$$P_{(\mu,\sigma^2)}\left(\bar{X} - \eta\sqrt{S_{XX}/n(n-1)} \leq \mu \leq \bar{X} - \xi\sqrt{S_{XX}/n(n-1)}\right) = \gamma.$$

Again the choice of ξ and η is not unique, but it is natural to try to make the length of the confidence interval as small as possible. The symmetry of the t-distribution implies that we should choose ξ and η symmetrically about 0.

4.2 Opinion polls

Opinion polls are typically quoted as being accurate to ±3%. What does this mean and how many people must be polled to attain this accuracy?

Suppose we are trying to estimate p, the proportion of people who support the Labour party amongst a very large population. We interview n people and estimate p from $\hat{p} = \frac{1}{n}(X_1 + \cdots + X_n)$, where $X_i = 1$ if the ith person supports Labour and $X_i = 0$ otherwise. Then

$$\mathbb{E}\hat{p} = p \qquad \text{and} \qquad \operatorname{var}\hat{p} = \frac{p(1-p)}{n} \leq \frac{1}{4n},$$

where the inequality follows from the fact that $p(1-p)$ is maximized by $p = \frac{1}{2}$.

Let us approximate the distribution of $\hat{p}(X)$ by $N\big(p, p(1-p)/n\big)$. This is very good for n more than about 20. Then we have that approximately

$$(\hat{p} - p)/\sqrt{p(1-p)/n} \sim N(0,1).$$

So

$$\begin{aligned}
&\mathbb{P}(\hat{p} - 0.03 \leq p \leq \hat{p} + 0.03) \\
&\quad = \mathbb{P}\left(-\frac{0.03}{\sqrt{p(1-p)/n}} \leq \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \leq \frac{0.03}{\sqrt{p(1-p)/n}}\right) \\
&\quad \approx \Phi\left(0.03\sqrt{n/p(1-p)}\right) - \Phi\left(-0.03\sqrt{n/p(1-p)}\right) \\
&\quad \geq \Phi(0.03\sqrt{4n}) - \Phi(-0.03\sqrt{4n})
\end{aligned}$$

For this to be at least 0.95, we need $0.03\sqrt{4n} \geq 1.96$, or $n \geq 1068$.

Opinion polls typically use a sample size of about 1,100.

Example 4.2 *U.S. News and World Report* (Dec 19, 1994) reported on a telephone survey of 1,000 Americans, in which 59% said they believed the world would come to an end, and of these 33% believed it would happen within a few years or decades.

Let us find a confidence interval for the proportion of Americans who believe the end of the world in imminent. Firstly, $\hat{p} = 0.59(0.33) = 0.195$. The variance of $\hat{p}$ is $p(1-p)/590$ which we estimate by $(0.195)(0.805)/590 = 0.000266$. Thus an approximate 95% confidence interval is $0.195 \pm \sqrt{0.00266}(1.96)$, or $[0.163, 0.226]$.

Note that this is only approximately a 95% confidence interval. We have used the normal approximation, and we have approximated $p(1-p)$ by $\hat{p}(1-\hat{p})$. These are both good approximations and this is therefore a very commonly used analysis.

Sampling from a small population*

For small populations the formula for the variance of $\hat{p}$ depends on the total population size N. E.g., if we are trying to estimate the proportion p of $N = 200$ students in a lecture who support the Labour party and we take $n = 200$, so we sample them all, then clearly $\text{var}(\hat{p}) = 0$. If $n = 190$ the variance will be close to 0. In fact,

$$\text{var}(\hat{p}) = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}.$$

4.3 Constructing confidence intervals

The technique we have used in these examples is based upon finding some statistic whose distribution does not depend on the unknown parameter θ. This can be done when θ is a **location parameter** or **scale parameter**. In section 4.1 μ is an example of a location parameter and σ is an example of a scale parameter. We saw that the distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma$ does not depend on μ or σ.

In the following example we consider a scale parameter.

Example 4.3 *Suppose that $X_1, \dots, X_n$ are IID $\mathcal{E}(\theta)$.* Then

$$f(x \mid \theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta \sum_i x_i}$$

so $T(X) = \sum_i X_i$ is sufficient for θ. Also, $T \sim \text{gamma}(n, \theta)$ with pdf

$$f_T(t) = \theta^n t^{n-1} e^{-\theta t}/(n-1)!, \quad t > 0.$$

Consider $S = 2\theta T$. Now $\mathbb{P}(S \leq s) = \mathbb{P}(T \leq s/2\theta)$, so by differentiation with respect to s, we find the density of S to be

$$f_S(s) = f_T(s/2\theta)\frac{1}{2\theta} = \frac{\theta^n (s/2\theta)^{n-1} e^{-\theta(s/2\theta)}}{(n-1)!}\frac{1}{2\theta} = \frac{s^{n-1}(1/2)^n e^{-s/2}}{(n-1)!}, \quad s > 0\,.$$

So $S = 2\theta T \sim \text{gamma}\left(n, \frac{1}{2}\right) \equiv \chi^2_{2n}$.

Suppose we want a 95% confidence interval for the mean, $1/\theta$. We can write

$$\mathbb{P}(\xi \leq 2T\theta \leq \eta) = \mathbb{P}\left(2T/\eta \leq 1/\theta \leq 2T/\xi\right) = F_{2n}(\xi) - F_{2n}(\eta)\,,$$

where F_{2n} is the cdf of a χ^2_{2n} RV.

For example, if $n = 10$ we refer to tables for the χ^2_{20} distribution and pick $\xi = 34.17$ and $\eta = 9.59$, so that $F_{20}(\xi) = 0.975$, $F_{20}(\eta) = 0.025$ and $F_{20}(\xi) - F_{20}(\eta) = 0.95$. Then a 95% confidence interval for $1/\theta$ is

$$[2t/34.17\,,\; 2t/9.59\,]\,.$$

Along the same lines, a confidence interval for σ can be constructed in the circumstances of Example 4.1 (b) by using fact that $S_{XX}/\sigma^2 \sim \chi^2_{n-1}$. E.g., if $n = 21$ a 95% confidence interval would be

$$\left[\sqrt{S_{xx}/34.17}\,,\; \sqrt{S_{xx}/9.59}\,\right]\,.$$

4.4 A shortcoming of confidence intervals*

Confidence intervals are widely used, e..g, in reporting the results of surveys and medical experiments. However, the procedure has the problem that it sometimes fails to make the best interpretation of the data.

Example 4.4 *Suppose X_1, X_2 are two IID samples from $U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$.* Then

$$\mathbb{P}(\min_i x_i \leq \theta \leq \max_i x_i) = \mathbb{P}(X_1 \leq \theta \leq X_2) + \mathbb{P}(X_2 \leq \theta \leq X_1) = \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{1}{2}\,.$$

So $(\min_i x_i, \max_i x_i)$ is a 50% confidence interval for θ.

But suppose the data is $x = (7.4, 8.0)$. Then we know $\theta > 8.0 - 0.5 = 7.5$ and $\theta < 7.4 + 0.5 = 7.9$. Thus with certainty, $\theta \in (7.5, 7.9) \subset (7.4, 8.0)$, so we can be 100% certain, not 50% certain, that our confidence interval has captured θ. This happens whenever $\max_i x_i - \min_i x_i > \frac{1}{2}$.

5 Bayesian estimation

Bayesians probably do it.

5.1 Prior and posterior distributions

Bayesian statistics, (named after the Rev. Thomas Bayes, an amateur 18th century mathematician), represents a different approach to statistical inference. Data are still assumed to come from a distribution belonging to a known parametric family. However, whereas classical statistics considers the parameters to be *fixed but unknown*, the Bayesian approach treats them as random variables in their own right. Prior beliefs about θ are represented by the **prior distribution**, with a prior probability density (or mass) function, $p(\theta)$. The **posterior distribution** has posterior density (or mass) function, $p(\theta \mid x_1, \dots, x_n)$, and captures our beliefs about θ after they have been *modified* in the light of the observed data.

By Bayes' celebrated formula,

$$p(\theta \mid x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n \mid \theta) p(\theta)}{\int f(x_1, \dots, x_n \mid \phi) p(\phi)\, d\phi}\,.$$

The denominator of the above equation does not involve θ and so in practice is usually not calculated. Bayes' rule is often just written,

$$p(\theta \mid x_1, \dots, x_n) \propto p(\theta) f(x_1, \dots, x_n \mid \theta).$$

Example 5.1 Consider the Smarties example addressed in Example 2.1 (a) and suppose our prior belief is that the number of colours is either 5, 6, 7 or 8, with prior probabilities 1/10, 3/10, 3/10 and 3/10 respectively. On seeing the data $x =$‘red, green, red’ we have $f(x \mid k) = (k-1)/k^2$. Similarly, if the fourth Smartie is orange, $f(x \mid k) = (k-1)(k-2)/k^3$. Then

$x =$ **‘red, green, red’**

k	$p(k)$	$f(x \mid k)$	$p(k)f(x \mid k)$	$p(k \mid x)$
5	.1	.160	.016	.13
6	.3	.139	.042	.33
7	.3	.122	.037	.29
8	.3	.109	.033	.26

$x =$ **‘red, green, red, orange’**

k	$p(k)$	$f(x \mid k)$	$p(k)f(x \mid k)$	$p(k \mid x)$
5	.1	.096	.010	.11
6	.3	.093	.028	.31
7	.3	.087	.026	.30
8	.3	.082	.025	.28

There is very little modification of the prior. This analysis reveals, in a way that the maximum likelihood approach did not, that the data obtained from looking at just 4 Smarties is not very informative. However, as we sample more Smarties the posterior distribution will come to concentrate on the true value of k.

5.2 Conditional pdfs

The discrete case

Thus Bayesians statistics relies on calculation of conditional distributions. For two events A and B (measurable subsets of the sample space) with $\mathbb{P}(B) \neq 0$, we define

$$\mathbb{P}(A \mid B) := \mathbb{P}(A \cap B)/\mathbb{P}(B).$$

We can harmlessly agree to define $\mathbb{P}(A \mid B) := 0$ if $\mathbb{P}(B) = 0$.

If X and Y are RVs with values in $\mathbb{Z}$, and if $f_{X,Y}$ is their joint pmf:

$$\mathbb{P}(X = x; Y = y) = f_{X,Y}(x, y),$$

then we define

$$f_{X|Y}(x \mid y) := \mathbb{P}(X = x \mid Y = y) = \frac{\mathbb{P}(X = x; Y = y)}{\mathbb{P}(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

if $f_Y(y) \neq 0$. We can safely define $f_{X|Y}(x \mid y) := 0$ if $f_Y(y) = 0$. Of course,

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_x \mathbb{P}(X = x; Y = y) = \sum_x f_{X,Y}(x, y).$$

Example 5.2 *Suppose that X and R are independent RVs, where X is Poisson with parameter λ and R is Poisson with parameter μ. Let $Y = X + R$.*

Then

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^{y-x} e^{-\mu}}{(y-x)!} \bigg/ \sum_{x,r:x+r=y} \frac{\lambda^x e^{-\lambda}}{x!} \frac{\mu^r e^{-\mu}}{r!} \\ &= \frac{y!}{x!(y-x)!} \lambda^x \mu^{(y-x)} \bigg/ \sum_{x,r:x+r=y} \frac{y!}{x!r!} \lambda^x \mu^r \\ &= \binom{y}{x} \left(\frac{\lambda}{\lambda+\mu}\right)^x \left(\frac{\mu}{\lambda+\mu}\right)^{y-x}. \end{aligned}$$

Hence $(X \mid Y = y) \sim B(y, p)$, where $p = \lambda/(\lambda + \mu)$.

The continuous case

Let $Z = (X, Y)$ be a RV with values in $\mathbb{R}^{m+n}$, X having values in $\mathbb{R}^m$ and Y values in $\mathbb{R}^n$. Assume that Z has nice pdf $f_Z(z)$ and write

$$f_Z(z) = f_{X,Y}(x, y), \qquad (z = (x, y), x \in \mathbb{R}^m, y \in \mathbb{R}^n).$$

Then the pdf of Y is given by

$$f_Y(y) = \int_{\mathbb{R}^m} f_{X,Y}(x, y)\, dx.$$

We define $f_{X|Y}$, the conditional pdf of X given Y, by

$$f_{X|Y}(x \mid y) := \begin{cases} f_{X,Y}(x,y)/f_Y(y) & \text{if } f_Y(y) \neq 0, \\ 0 & \text{if } f_Y(y) = 0. \end{cases}$$

The intuitive idea is: $\mathbb{P}(X \in dx \mid Y \in dy) = \mathbb{P}(X \in dx; Y \in dy)/\mathbb{P}(Y \in dy)$.

Examples 5.3

(a) *A biased coin is tossed n times. Let x_i be 1 or 0 as the ith toss is or is not a head. Suppose we have no idea how biased the coin is, so we place a uniform prior distribution on θ, to give a so-called 'noninformative prior' of*

$$p(\theta) = 1, \quad 0 \leq \theta \leq 1\,.$$

Let t be the number of heads. Then the posterior distribution of θ is

$$p(\theta \mid x_1, \dots, x_n) = \theta^t (1-\theta)^{n-t} \times 1 \bigg/ \int_0^1 \phi^t (1-\phi)^{n-t} \times 1 \, d\phi \ .$$

We would usually not bother with the denominator and just write

$$p(\theta \mid x) \propto \theta^t (1-\theta)^{n-t}\,.$$

By inspection we recognise that if the appropriate constant of of proportionality is inserted on the r.h.s. then we have the density of beta$(t+1, n-t+1)$, so this is the posterior distribution of θ given x.

(b) *Suppose $X_1, \dots, X_n \sim N(\mu, 1)$, $p(\mu) \sim N(0, \tau^{-2})$ for known τ^{-2}.* Then

$$\begin{aligned} p(\mu \mid x_1, \dots, x_n) &\propto \exp\left\{ -\frac{1}{2}\left(\sum_{i=1}^n (x_i - \mu)^2 + \mu^2 \tau^2 \right) \right\} \\ &\propto \exp\left\{ -\frac{1}{2}(n+\tau^2)\left(\mu - \frac{\sum x_i}{n+\tau^2} \right)^2 \right\} \\ \mu \mid x_1, \dots, x_n &\sim N\left(\frac{\sum x_i}{n+\tau^2}, \frac{1}{n+\tau^2} \right) \end{aligned}$$

Note that as $\tau \to 0$ the prior distribution becomes less informative.

(c) *Suppose $X_1, \dots, X_n \sim$ IID $\mathcal{E}(\lambda)$, and the prior for λ is given by $\lambda \sim \mathcal{E}(\mu)$, for fixed and known μ.* Then

$$p(\lambda \mid x_1, \dots, x_n) \propto \mu e^{-\lambda\mu} \prod_i \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda(\mu + \sum_{i=1}^n x_i)}\,,$$

i.e., gamma$(n+1, \mu + \sum x_i)$.

5.3 Estimation within Bayesian statistics

The Bayesian approach to the parameter estimation problem is to use a **loss function** $L(\theta, a)$ to measure the loss incurred by estimating the value of a parameter to be a when its true value is θ. Then $\hat{\theta}$ is chosen to minimize $\mathbb{E}[L(\theta, \hat{\theta})]$, where this expectation is taken over θ with respect to the posterior distribution $p(\theta \mid x)$.

Loss functions for quadratic and absolute error loss

(a) $L(\theta, a) = (a - \theta)^2$ is the **quadratic error loss** function.

$$\mathbb{E}[L(\theta, a)] = \int L(\theta, a) p(\theta \mid x_1, \dots, x_n)\, d\theta = \int (a - \theta)^2 p(\theta \mid x_1, \dots, x_n)\, d\theta\,.$$

Differentiating with respect to a we get

$$2 \int (a - \theta) p(\theta \mid x_1, \dots, x_n)\, d\theta = 0 \implies a = \int \theta p(\theta \mid x_1, \dots, x_n)\, d\theta\ .$$

Therefore quadratic error loss is minimized by taking $\hat{\theta}$ to be the **posterior mean**.

(b) $L(\theta, a) = |a - \theta|$ is the **absolute error loss** function.

$$\begin{aligned}\mathbb{E}[L(\theta, a)] &= \int L(\theta, a) p(\theta \mid x_1, \dots, x_n)\, d\theta \\ &= \int_{\theta=-\infty}^{a} (a - \theta) p(\theta \mid x_1, \dots, x_n)\, d\theta + \int_{a}^{\infty} (\theta - a) p(\theta \mid x_1, \dots, x_n)\, d\theta\,.\end{aligned}$$

Differentiating with respect to a we find that the minimum is where

$$\int_{-\infty}^{a} p(\theta \mid x_1, \dots, x_n)\, d\theta - \int_{a}^{\infty} p(\theta \mid x_1, \dots, x_n)\, d\theta = 0\,.$$

The minimum is achieved when both integrals are equal to $\frac{1}{2}$, i.e., by taking $\hat{\theta}$ to be the **posterior median**.

Example 5.4 *Let $X_1, \dots, X_n \sim P(\lambda)$, $\lambda \sim \mathcal{E}(1)$ so that $p(\lambda) = e^{-\lambda}$, $\lambda \geq 0$.*

The posterior distribution is

$$p(\lambda \mid x_1, \dots, x_n) = e^{-\lambda} \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \propto e^{-\lambda(n+1)} \lambda^{\sum x_i}\,,$$

i.e., $\text{gamma}\big(\sum x_i + 1, (n+1)\big)$. So under quadratic error loss,

$$\hat{\theta} = \text{posterior mean} = \frac{\sum_{i=1}^{n} x_i + 1}{n+1}\,.$$

Under absolute error loss, $\hat{\theta}$ solves

$$\int_0^{\hat{\theta}} \frac{e^{-\lambda(n+1)} \lambda^{\sum x_i} (n+1)^{\sum x_i + 1}}{(\sum x_i)!}\, d\lambda = \frac{1}{2}\,.$$

6 Hypothesis testing

Statistics is the only profession which demands the right to make mistakes 5 per cent of the time – Thomas Huxley.

6.1 The Neyman–Pearson framework

The second major area of statistical inference is **hypothesis testing**. A statistical **hypothesis** is an assertion or conjecture about the distribution of one or more random variables, and a test of a statistical hypothesis is a rule or procedure for deciding whether to reject that assertion.

Example 6.1 It has been suggested that dying people may be able to postpone their death until after an important occasion. In a study of 1919 people with Jewish surnames it was found that 922 occurred in the week before Passover and 997 in the week after. Is there any evidence in this data to reject the hypothesis that a person is as likely to die in the week before as in the week after Passover?

Example 6.2 In one of his experiments, Mendel crossed 556 smooth, yellow male peas with wrinkled, green female peas. Here is what he obtained and its comparison with predictions based on genetic theory.

type	observed	predicted	expected
	count	frequency	count
smooth yellow	315	9/16	312.75
smooth green	108	3/16	104.25
wrinkled yellow	102	3/16	104.25
wrinkled green	31	1/16	34.75

Is there any evidence in this data to reject the hypothesis that theory is correct?

We follow here an approach developed by Neyman and Pearson. Suppose we have data $x = (x_1, x_2, \dots, x_n)$ from a density f. We have two hypotheses about f. On the basis of the data one is **accepted**, the other **rejected**. The two hypotheses have different philosophical status. The first, called the **null hypothesis**, and denoted by H_0, is a conservative hypothesis, not to be rejected unless evidence is clear. The second, the **alternative hypothesis**, denoted by H_1, specifies the kind of departure from the null hypothesis of interest to us.

It is often assumed that f belongs to a specified parametric family $f(\cdot \mid \theta)$ indexed by a parameter $\theta \in \Theta$ (e.g. $N(\theta, 1)$, $B(n, \theta)$). We might then want to test a parametric hypothesis

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1$$

with $\Theta_0 \cap \Theta_1 = \emptyset$. We may, or may not, have $\Theta_0 \cup \Theta_1 = \Theta$.

We will usually be concerned with testing a parametric hypothesis of this kind, but alternatively, we may wish to test

$$H_0 : f = f_0 \quad \text{against} \quad H_1 : f \neq f_0$$

where f_0 is a specified density. This is a '**goodness-of-fit**' test.

A third alternative is that we wish to test

$$H_0 : f = f_0 \quad \text{against} \quad H_1 : f = f_1$$

where f_0 and f_1 are specified, but do not necessarily belong to the same family.

6.2 Terminology

A hypothesis which specifies f completely is called **simple**, e.g., $\theta = \theta_0$. Otherwise, a hypothesis is **composite**, e.g., $\theta > \theta_0$.

Suppose we wish to test H_0 against H_1. A test is defined by a **critical region** C. We write $\bar{C}$ for the complement of C.

$$\text{If } x = (x_1, x_2, \dots, x_n) \in \begin{cases} C \text{ then } H_0 \text{ is rejected, and} \\ \bar{C} \text{ then } H_0 \text{ is accepted (not rejected).} \end{cases}$$

Note that when $x \in \bar{C}$ we might sometimes prefer to say 'not rejected', rather than 'accepted'. This is a minor point which need not worry us, except to note that sometimes 'not rejected' does more accurately express what we are doing: i.e., looking to see if the data provides any evidence to reject the null hypothesis. If it does not, then we might want to consider other things before finally 'accepting H_0'.

There are two possible types of error we might make:

H_0 might be rejected when it is true (**a type I error**), or

H_0 might be accepted when it is false (**a type II error**).

Since H_0 is conservative, a type I error is generally considered to be 'more serious' that a type II error. For example, the jury in a murder trial should take as its null hypothesis that the accused is innocent, since the type I error (that an innocent person is convicted and the true murderer is never caught) is more serious than the type II error (that a murderer is acquitted).

Hence, we fix (an upper bound on) the probability of type I error, e.g., 0.05 or 0.01, and define the critical region C by minimizing the type II error subject to this.

If H_0 is simple, $\Theta_0 = \{\theta_0\}$, the probability of a type I error is called the **size**, or **significance level**, of the test. If H_0 is composite, the size of the test is $\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}(X \in C \mid \theta)$. .

The **likelihood** of a simple hypothesis $H : \theta = \theta^*$ given data x is

$$L_x(H) = f_X(x \mid \theta = \theta^*).$$

If H is composite, $H : \theta \in \Theta$, we define

$$L_x(H) = \sup_{\theta \in \Theta} f_X(x \mid \theta).$$

The **likelihood ratio** for two hypotheses H_0, H_1 is

$$L_x(H_0, H_1) = \frac{L_x(H_1)}{L_x(H_0)}.$$

Notice that if $T(x)$ is a sufficient statistic for θ then by the factorization criterion $L_x(H_0, H_1)$ is simply a function of $T(x)$.

6.3 Likelihood ratio tests

A test given by a critical region C of the form $C = \{x : L_x(H_0, H_1) > k\}$, for some constant k, is called a **likelihood ratio test**. The value of k is determined by fixing the size α of the test, so that $\mathbb{P}(X \in C \mid H_0) = \alpha$.

Likelihood ratio tests are optimal for simple hypotheses. Most standard tests are likelihood ratio tests, though tests can be built from others statistics.

Lemma 6.3 (Neyman–Pearson Lemma) *$H_0 : f = f_0$ is to be tested against $H_1 : f = f_1$. Assume that f_0 and f_1 are > 0 on the same regions and continuous.*

Then, among all tests of size $\leq \alpha$, the test with smallest probability of type II error is given by $C = \{x : f_1(x)/f_0(x) > k\}$, where k is determined by

$$\alpha = \mathbb{P}(X \in C \mid H_0) = \int_C f_0(x)\, dx.$$

Proof. Consider any test with size $\leq \alpha$, i.e., with a critical region D such that $\mathbb{P}(X \in D \mid H_0) \leq \alpha$. Define

$$\phi_D(x) = \left\{ \begin{array}{l} 1 \\ 0 \end{array} \right. \text{as } x \begin{array}{c} \in \\ \notin \end{array} D$$

and let C and k be defined as above. Note that

$$0 \leq \big(\phi_C(x) - \phi_D(x)\big)\big(f_1(x) - k f_0(x)\big), \text{ for all } x.$$

since this is always the product of two terms with the same sign. Hence

$$\begin{aligned} 0 &\leq \int_x \big(\phi_C(x) - \phi_D(x)\big)\big(f_1(x) - k f_0(x)\big)\, dx \\ &= \mathbb{P}(X \in C \mid H_1) - \mathbb{P}(X \in D \mid H_1) - k\big[\mathbb{P}(X \in C \mid H_0) - \mathbb{P}(X \in D \mid H_0)\big] \\ &= \mathbb{P}(X \in C \mid H_1) - \mathbb{P}(X \in D \mid H_1) - k\big[\alpha - \mathbb{P}(X \in D \mid H_0)\big] \\ &\leq \mathbb{P}(X \in C \mid H_1) - \mathbb{P}(X \in D \mid H_1) \end{aligned}$$

This implies $\mathbb{P}(X \notin C \mid H_1) \leq \mathbb{P}(X \notin D \mid H_1)$ as required. ∎

6.4 Single sample: testing a given mean, simple alternative, known variance (z-test)

Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$, where σ^2 is known. We wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, where $\mu_1 > \mu_0$.

The Neyman–Pearson test is to reject H_0 if the likelihood ratio is large (i.e., greater than some k). The likelihood ratio is

$$\begin{aligned}\frac{f(x \mid \mu_1, \sigma^2)}{f(x \mid \mu_0, \sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \mu_1)^2/2\sigma^2\right]}{(2\pi\sigma^2)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \mu_0)^2/2\sigma^2\right]} \\ &= \exp\left[\sum_{i=1}^n \left\{(x_i - \mu_0)^2 - (x_i - \mu_1)^2\right\}/2\sigma^2\right] \\ &= \exp\left[n\left\{2\bar{x}(\mu_1 - \mu_0) + (\mu_0^2 - \mu_1^2)\right\}/2\sigma^2\right]\end{aligned}$$

It often turns out, as here, that the likelihood ratio is a monotone function of a sufficient statistic and we can immediately rephrase the critical region in more convenient terms. We notice that the likelihood ratio above is increasing in the sufficient statistic $\bar{x}$ (since $\mu_1 - \mu_0 > 0$). So the Neyman–Pearson test is equivalent to 'reject H_0 if $\bar{x} > c$', where we choose c so that $\mathbb{P}(\bar{X} > c \mid H_0) = \alpha$. There is no need to try to write c in terms of k.

However, under H_0 the distribution of $\bar{X}$ is $N(\mu_0, \sigma^2/n)$. This means that

$$Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0, 1) .$$

It is now convenient to rephrase the test in terms of Z, so that a test of size α is to reject H_0 if $z > z_\alpha$, where $z_\alpha = \Phi^{-1}(1 - \alpha)$ is the 'upper α point of $N(0, 1)$' i.e., the point such that $\mathbb{P}(N(0, 1) > z_\alpha) = \alpha$. E.g., for $\alpha = 0.05$ we would take $z_\alpha = 1.645$, since 5% of the standard normal distribution lies to the right of 1.645.

Because we reject H_0 only if z is in the upper tail of the normal distribution we call this a **one-tailed test**. We shall see other tests in which H_0 is rejected if the test statistic lies in either of two tails. Such a test is called a **two-tailed test**.

Example 6.4 *Suppose $X_1, \ldots, X_n$ are IID $N(\mu, \sigma^2)$ as above, and we want a test of size* 0.05 *of $H_0 : \mu = 5$ against $H_1 : \mu = 6$, with $\sigma^2 = 1$. Suppose the data is $x = (5.1, 5.5, 4.9, 5.3)$.* Then $\bar{x} = 5.2$ and $z = 2(5.2 - 5)/1 = 0.4$. Since this is less than 1.645 we do not reject $\mu = 5$.

Suppose the hypotheses are reversed, so that we test $H_0 : \mu = 6$ against $H_1 : \mu = 5$. The test statistic is now $z = 2(5.2 - 6)/1 = -1.6$ and we should reject H_0 for values of Z less than -1.645. Since z is more than -1.645, it is not significant and we do not reject $\mu = 6$.

This example demonstrates the preferential position given to H_0 and therefore that it is important to choose H_0 in a way that makes sense in the context of the decision problem with which the statistical analysis is concerned.

7 Further aspects of hypothesis testing

Statisticians do it with only a 5% chance of being rejected.

7.1 The p-value of an observation

The **significance level of a test** is another name for α, the size of the test. For a composite hypothesis $H_0 : \theta \in \Theta_0$ and rejection region C this is

$$\alpha = \sup_{\theta \in \Theta_0} \mathbb{P}(X \in C \mid \theta) .$$

For a likelihood ratio test the **p-value** of an observation x is defined to be

$$p^* = \sup_{\theta \in \Theta_0} P_\theta \bigl(L_X(H_0, H_1) \geq L_x(H_0, H_1) \bigr) .$$

The p-value of x is the probability under H_0 of seeing x or something at least as 'extreme', in the sense of containing at least as much evidence against H_0. E.g., in Example 6.4, the p-value is p^* where

$$p^* = \mathbb{P}(Z > z \mid \mu = \mu_0) = 1 - \Phi(z) .$$

A test of size α rejects H_0 if and only if $\alpha \geq p^*$. Thus *the p-value of x is the smallest value of α for which H_0 would be rejected on the basis of seeing x.* It is often more informative to report the p-value than merely to report whether or not the null hypothesis has been rejected.

The p-value is also sometimes called the **significance level** of x. Historically, the term arises from the practice of 'significance testing'. Here, we begin with an H_0 and a test statistic T, which need not be the likelihood ratio statistic, for which, say, large positive values suggest falsity of H_0. We observe value t_0 for the statistic: the significance level is $\mathbb{P}(T \geq t_0 \mid H_0)$. If this probability is small, H_0 is rejected.

7.2 The power of a test

For a parametric hypothesis about $\theta \in \Theta$, we define the **power function** of the test specified by the critical region C as

$$W(\theta) = \mathbb{P}(X \in C \mid \theta).$$

Notice that $\alpha = \sup_{\theta \in \Theta_0} W(\theta)$, and $1 - W(\theta) = \mathbb{P}(X \in \bar{C} \mid \theta) = \mathbb{P}(\text{type II error} \mid \theta)$ for $\theta \in \Theta_1$.

7.3 Uniformly most powerful tests

For the test of $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$, $\mu_1 > \mu_0$ described in section 6.4 the critical region turned out to be

$$C(\mu_0) = \left\{x : \sqrt{n}(\bar{x} - \mu_0)/\sigma > \alpha_z\right\}.$$

This depends on μ_0 but not on the specific value of μ_1. The test with this critical region would be optimal for any alternative $H_1 : \mu = \mu_1$, provided $\mu_1 > \mu_0$. This is the idea of a uniformly most powerful (UMP) test.

We can find the power function of this test.

$$\begin{aligned}
W(\mu) &= \mathbb{P}(Z > \alpha_z \mid \mu) \\
&= \mathbb{P}\big(\sqrt{n}(\bar{X} - \mu_0)/\sigma > \alpha_z \mid \mu\big) \\
&= \mathbb{P}\big(\sqrt{n}(\bar{X} - \mu)/\sigma + \sqrt{n}(\mu - \mu_0)\big)/\sigma > \alpha_z \mid \mu) \\
&= 1 - \Phi\big(\alpha_z - \sqrt{n}(\mu - \mu_0)/\sigma\big)
\end{aligned}$$

Note that $W(\mu)$ increases from 0 to 1 as μ goes from $-\infty$ to ∞ and $W(\mu_0) = \alpha$.

More generally, suppose $H_0 : \theta \in \Theta_0$ is to be tested against $H_1 : \theta \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset$. Suppose H_1 is composite. H_0 can be simple or composite. We want a test of size α. So we require $W(\theta) \le \alpha$ for all $\theta \in \Theta_0$, $W(\theta_0) = \alpha$ for some $\theta_0 \in \Theta_0$.

A **uniformly most powerful** (UMP) test of size α satisfies (i) it is of size α, (ii) $W(\theta)$ is as large as possible for every $\theta \in \Theta_1$.

UMP tests may not exist. However, likelihood ratio tests are often UMP.

Example 7.1 *Let $X_1, \dots, X_n$ be IID $N(\mu, \sigma^2)$, where μ is known. Suppose $H_0 : \sigma^2 \le 1$ is to be tested against $H_1 : \sigma^2 > 1$.*

We begin by finding the most powerful test for testing $H_0' : \sigma^2 = \sigma_0^2$ against $H_1' : \sigma^2 = \sigma_1^2$, where $\sigma_0^2 \le 1 < \sigma_1^2$. The Neyman–Pearson test rejects H_0' for large values of the likelihood ratio:

$$\begin{aligned}
\frac{f\left(x \mid \mu, \sigma_1^2\right)}{f\left(x \mid \mu, \sigma_0^2\right)} &= \frac{\left(2\pi\sigma_1^2\right)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \mu)^2/2\sigma_1^2\right]}{\left(2\pi\sigma_0^2\right)^{-n/2} \exp\left[-\sum_{i=1}^n (x_i - \mu)^2/2\sigma_0^2\right]} \\
&= (\sigma_0/\sigma_1)^n \exp\left[\left(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}\right) \sum_{i=1}^n (x_i - \mu)^2\right]
\end{aligned}$$

which is large when $\sum_i (x_i - \mu)^2$ is large. If $\sigma^2 = 1$ then

$$\sum_{i=1}^n (x_i - \mu)^2 \sim \chi_n^2 \, .$$

So a test of the form 'reject H_0 if $T := \sum_i (x_i - \mu)^2 > F_\alpha^{(n)}$', has size α where $F_\alpha^{(n)}$ is the upper α point of χ_n^2. That is, $\mathbb{P}\left(T > F_\alpha^{(n)} \mid \sigma^2 \le 1\right) \le \alpha$, for all $\sigma^2 \le 1$, with equality for $\sigma^2 = 1$. But this test doesn't depend on the value of σ_1^2, and hence is the UMP test of H_0 against H_1.

Example 7.2 Consider Example 6.1. Let p be the probability that if a death occurs in one of the two weeks either side of Passover it actually occurs in the week after the Passover. Let us test $H_0 : p = 0.5$ vs. $H_1 : p > 0.5$.

The distribution of the number of deaths in the week after Passover, say X, is $B(n,p)$, which we can approximate by $N\big(np, np(1-p)\big)$; under H_0 this is $N(0.5n, 0.25n)$. So a size 0.05 test is to reject H_0 if $z = \sqrt{n}\big(\bar{x} - \mu_0\big)/\sigma > 1.645$, where here $z = \sqrt{1919}\big(997/1919 - 0.5\big)/0.5 = 1.712$. So the data is just significant at the 5% level. We reject the hypothesis that death $p = 1/2$.

It is important to realise that this does not say anything about *why* this might be. It might be because people really are able to postpone their deaths to enjoy the holiday. But it might also be that deaths increase after Passover because of over-eating or stress during the holiday.

7.4 Confidence intervals and hypothesis tests

There is an interesting duality between confidence intervals and hypothesis tests. In the following, we speak of the **acceptance region**, i.e., the complement of the critical (or rejection) region C.

Theorem 7.3

(i) Suppose that for every θ_0 there is a size α test of $H_0 : \theta = \theta_0$ against some alternative. Denote the acceptance region by $A(\theta_0)$. Then $I(X) = \{\theta : X \in A(\theta)\}$ is a $100(1-\alpha)\%$ confidence interval for θ.

(ii) Conversely, if $I(X)$ is a $100(1-\alpha)\%$ confidence interval for θ then an acceptance region for a size α test of $H_0 : \theta = \theta_0$ is $A(\theta_0) = \{X : \theta_0 \in I(X)\}$.

Proof. The definitions in the theorem statement give

$$\mathbb{P}\big(X \in A(\theta_0) \mid \theta = \theta_0\big) = \mathbb{P}\big(\theta \in I(X) \mid \theta = \theta_0\big) .$$

By assumption the l.h.s. is $1 - \alpha$ in case (i) and the r.h.s. is $1 - \alpha$ in case (ii). ■

This duality can be useful. In some circumstances it can be easier to see what is the form of a hypothesis test and then work out a confidence interval. In other circumstances it may be easier to see the form of the confidence interval.

In Example 4.1 we saw that a 95% confidence interval for μ based upon $X_1, \dots, X_n$ being IID samples from $N(\mu, \sigma^2)$, σ^2 known, is

$$\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right] .$$

Thus $H_0 : \mu = \mu_0$ is rejected in a 5% level test against $H_1 : \mu \neq \mu_0$ if and only if μ_0 is not in this interval; i.e., if and only if

$$\sqrt{n}\big|\bar{X} - \mu_0\big|/\sigma > 1.96\,.$$

7.5 The Bayesian perspective on hypothesis testing

Suppose a statistician wishes to test the hypothesis that a coin is fair ($H_0 : p = 1/2$) against the alternative that it biased towards heads with probability p, ($H_1 : p = p_1 > 1/2$). He decides to toss it five times. It is easy to see that the best test is to reject H_0 if the total number of heads, say T, is large. Here $T \sim B(5, p)$. Suppose he observes `H,H,H,H,T`, so $T = 4$. The p-value (or significance level) of this result is the probability that under the null hypothesis he should see a result which is equally or more extreme, i.e.,

$$\mathbb{P}(T = 4 \text{ or } 5) = 5(0.5^4)(0.5) + 0.5^5 = 6/32 = 0.1875\,.$$

Another statistician wishes also wishes to test H_0 against H_1, but with a different experiment. He plans to toss the coin until he gets a tail. Now the best test is to reject H_0 if the number of tosses, say N, is large. Here $N \sim \text{geometric}(1-p)$. Suppose this statistician also observes `H,H,H,H,T`, so $N = 5$. He figures the p-value as the probability of seeing this result or one that is even more extreme and obtains

$$\mathbb{P}(N \geq 5) = 0.5^4 = 1/16 = 0.0625\,.$$

Thus the two statisticians come to different conclusions about the significance of what they have seen. This is disturbing! The coin knew nothing about the experimental procedure. Maybe there were four tosses simply because that was the point at which the experimenter spilt his coffee and so decided to stop tossing the coin. What then is the significance level?

The Bayesian perspective on hypothesis testing avoids this type of problem. Suppose the experimenter places prior probabilities of $\mathbb{P}(H_0)$ and $\mathbb{P}(H_1)$ on the truth of two mutually exclusive hypotheses. Having observed the data x, these are modified into posterior probabilities in the usual way, i.e.,

$$\frac{\mathbb{P}(H_1 \mid x)}{\mathbb{P}(H_0 \mid x)} = \frac{\mathbb{P}(x \mid H_1)}{\mathbb{P}(x \mid H_0)}\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} = L_x(H_0, H_1)\frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)}\,.$$

Thus the ratio of the posterior probabilities is just the ratio of the prior probabilities multiplied by the likelihood ratio.

Under both procedures above $x = \{$`H,H,H,H,T`$\}$, and the likelihood ratios is

$$L_x(H_0, H_1) = \frac{p_1}{0.5} \times \frac{p_1}{0.5} \times \frac{p_1}{0.5} \times \frac{p_1}{0.5} \times \frac{(1-p_1)}{0.5} = \frac{p_1^4(1-p_1)}{0.5^5}\,.$$

In general, the Bayesian analysis does not depend on how the data was obtained.

8 Generalized likelihood ratio tests

An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.

8.1 The χ^2 distribution

This distribution plays a huge role in Statistics. For n a positive integer, the χ_n^2 distribution is the distribution of

$$X_1^2 + \cdots + X_n^2, \text{ where } X_1, \dots, X_n \text{ are IID each } N(0,1).$$

It is not difficult to show that

$$\chi_n^2 = \text{gamma}\left(\tfrac{1}{2}n, \tfrac{1}{2}\right),$$

with pdf

$$f(t) = (\tfrac{1}{2})^{n/2} t^{n/2-1} e^{-t/2} \Big/ \Gamma(n/2), \quad t > 0\,.$$

We speak of the 'chi-squared distribution with (or on) n **degrees of freedom**'. If $X \sim \chi_n^2$, $\mathbb{E}(X) = n$, $\text{var}(X) = 2n$.

8.2 Generalised likelihood ratio tests

Tests with critical regions of the form $C = \{x : L_x(H_0, H_1) > k\}$ are intuitively sensible, and in certain cases optimal. But such a critical region may not reduce to a region depending only on the value of a simple function $T(x)$. Even if it does, the distribution of the statistic $T(X)$, required to determine k, may not be simple. However, the following theorem allows us to use a likelihood ratio test even in this case. We must take care in describing the circumstances for which the theorem is valid.

So far we have considered **disjoint** alternatives, but if our real interest is in testing H_0 and we are not interested in any specific alternative it is simpler to take (in the parametric framework) $\Theta_1 = \Theta$, rather than $\Theta_1 = \Theta \setminus \Theta_0$.

So we now suppose we are testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta$, i.e., a null hypothesis which restricts θ against a general alternative.

Suppose that with this formulation Θ_0 imposes p independent restrictions on θ, so that, for example, we have $\Theta = \{\theta : \theta = (\theta_1, \dots, \theta_k)\}$ and

$$\begin{aligned}
&H_0 : \theta_{i_1} = \alpha_1, \dots, \theta_{i_p} = \alpha_p, \quad \text{for given } \alpha_j\,; \text{ or}\\
&H_0 : A\theta = b, \quad \text{for given } A_{p\times k}, b_{p\times 1}\,; \text{ or}\\
&H_0 : \theta_i = \theta_i(\phi_1, \dots, \phi_{k-p}), \quad i = 1, \dots, k, \quad \text{for given } \theta_1(\cdot), \dots, \theta_k(\cdot)\\
&\qquad\qquad \text{and } \phi_1, \dots, \phi_{k-p} \text{ to be estimated.}
\end{aligned}$$

Θ_1 has k free parameters and Θ_0 has $k-p$ free parameters. We write $|\Theta_1| = k$ and $|\Theta_0| = k-p$. Then we have the following theorem (not to be proved.)

Theorem 8.1 *Suppose $\Theta_0 \subset \Theta_1$ and $|\Theta_1| - |\Theta_0| = p$. Then under certain conditions, as $n \to \infty$ with $X = (X_1, \dots, X_n)$ and X_i IID,*

$$2 \log L_X(H_0, H_1) \sim \chi_p^2,$$

if H_0 is true. If H_0 is not true, $2\log L_X$ tends to be larger. We reject H_0 if $2 \log L_x > c$, where $\alpha = \mathbb{P}\left(\chi_p^2 > c\right)$ to give a test of size approximately α.

We say that $2\log L_X(H_0, H_1)$ is asymptotically distributed as χ_p^2. The conditions required by the theorem hold in all the circumstances we shall meet in this course.

Lemma 8.2 *Suppose $X_1, \dots, X_n$ are IID $N(\mu, \sigma^2)$. Then*

(i) $\max_\mu f(x \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\sum_i (x_i - \bar{x})^2 / 2\sigma^2\right]$.

(ii) $\max_{\sigma^2} f(x \mid \mu, \sigma^2) = \left[2\pi \frac{\sum_i (x_i - \mu)^2}{n}\right]^{-n/2} \exp\left[-n/2\right]$.

(iii) $\max_{\mu, \sigma^2} f(x \mid \mu, \sigma^2) = \left[2\pi \frac{\sum_i (x_i - \bar{x})^2}{n}\right]^{-n/2} \exp\left[-n/2\right]$.

8.3 Single sample: testing a given mean, known variance (z-test)

Let $X_1, \dots, X_n$ be IID $N(\mu, \sigma^2)$, where σ^2 is known. We wish to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. The generalized likelihood ratio test suggests that we should reject H_0 if $L_x(H_0, H_1)$ is large, where

$$\begin{aligned}
L_x(H_0, H_1) &= \frac{\sup_\mu f\left(x \mid \mu, \sigma^2\right)}{f\left(x \mid \mu_0, \sigma^2\right)} \\
&= \frac{\left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\sum_i (x_i - \bar{x})^2 / 2\sigma^2\right]}{\left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\sum_i (x_i - \mu_0)^2 / 2\sigma^2\right]} \\
&= \exp\left[(1/2\sigma^2) \sum_{i=1}^n \left\{(x_i - \mu_0)^2 - (x_i - \bar{x})^2\right\}\right] \\
&= \exp\left[(1/2\sigma^2) n (\bar{x} - \mu_0)^2\right]
\end{aligned}$$

That is, we should reject H_0 if $(\bar{x} - \mu_0)^2$ is large.

This is no surprise. For under H_0, $\bar{X} \sim N(\mu_0, \sigma^2/n)$, so that

$$Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0,1),$$

and a test of size α is to reject H_0 if $z > z_{\alpha/2}$ or if $z < -z_{\alpha/2}$, where $z_{\alpha/2}$ is the 'upper $\alpha/2$ point of $N(0,1)$' i.e., the point such that $\mathbb{P}(N(0,1) > z_{\alpha/2}) = \alpha/2$. This is an example of a **two-tailed test**.

Note that $2\log L_X(H_0,H_1) = Z^2 \sim \chi_1^2$. In this example H_0 imposes $p = 1$ constraint on the parameter space and the approximation in Theorem 8.1 is exact.

8.4 Single sample: testing a given variance, known mean (χ^2-test)

As above, let $X_1,\dots,X_n$ be IID $N(\mu,\sigma^2)$, where μ is known. We wish to test $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$. The generalized likelihood ratio test suggests that we should reject H_0 if $L_x(H_0,H_1)$ is large, where

$$L_x(H_0,H_1) = \frac{\sup_{\sigma^2} f\big(x \mid \mu,\sigma^2\big)}{f\big(x \mid \mu,\sigma_0^2\big)} = \frac{\left[\frac{2\pi\sum_i(x_i-\mu)^2}{n}\right]^{-n/2}\exp\left[-n/2\right]}{(2\pi\sigma_0^2)^{-n/2}\exp\big[-\sum_i(x_i-\mu)^2/2\sigma_0^2\big]}\,.$$

If we let $t = \sum_i(x_i-\mu)^2/n\sigma_0^2$ we find

$$2\log L_x(H_0,H_1) = n(t-1-\log t)\,,$$

which increases as t increases from 1 and t decreases from 1. Thus we should reject H_0 when the difference of t and 1 is large.

Again, this is not surprising, for under H_0,

$$T = \sum_{i=1}^{n}(X_i-\mu)^2/\sigma_0^2 \sim \chi_n^2\,.$$

So a test of size α is the two-tailed test which rejects H_0 if $t > F_{\alpha/2}^{(n)}$ or $t < F_{1-\alpha/2}^{(n)}$ where $F_{1-\alpha/2}^{(n)}$ and $F_{\alpha/2}^{(n)}$ are the lower and upper $\alpha/2$ points of χ_n^2, i.e., the points such that $\mathbb{P}\left(\chi_n^2 < F_{1-\alpha/2}^{(n)}\right) = \mathbb{P}\left(\chi_n^2 > F_{\alpha/2}^{(n)}\right) = \alpha/2$.

8.5 Two samples: testing equality of means, known common variance (z-test)

Let $X_1,\dots,X_m$ be IID $N(\mu_1,\sigma^2)$ and let $Y_1,\dots,Y_n$ be IID $N(\mu_2,\sigma^2)$, and suppose that the two samples are independent. It is required to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$. The likelihood ratio test is based on

$$\begin{aligned}
L_x(H_0,H_1) &= \frac{\sup_{\mu_1,\mu_2} f\big(x \mid \mu_1,\sigma^2\big) f\big(y \mid \mu_2,\sigma^2\big)}{\sup_{\mu} f\big(x \mid \mu,\sigma^2\big) f\big(y \mid \mu,\sigma^2\big)}\\
&= \frac{(2\pi\sigma^2)^{-(m+n)/2}\exp\left[-\sum_i(x_i-\bar{x})^2/2\sigma^2\right]\exp\left[-\sum_i(y_i-\bar{y})^2/2\sigma^2\right]}{(2\pi\sigma^2)^{-(m+n)/2}\exp\left[-\sum_i\left(x_i-\frac{m\bar{x}+n\bar{y}}{m+n}\right)^2/2\sigma^2\right]\exp\left[-\sum_i\left(y_i-\frac{m\bar{x}+n\bar{y}}{m+n}\right)^2/2\sigma^2\right]}\\
&= \exp\left[\frac{m}{2\sigma^2}\left(\bar{x}-\frac{m\bar{x}+n\bar{y}}{m+n}\right)^2 + \frac{n}{2\sigma^2}\left(\bar{y}-\frac{m\bar{x}+n\bar{y}}{m+n}\right)^2\right]\\
&= \exp\left[\frac{1}{2\sigma^2}\frac{mn}{(m+n)}(\bar{x}-\bar{y})^2\right]
\end{aligned}$$

So we should reject H_0 if $|\bar{x}-\bar{y}|$ is large. Now, $\bar{X} \sim N(\mu_1, \sigma^2/m)$ and $\bar{Y} \sim N(\mu_2, \sigma^2/n)$, and the samples are independent, so that, on H_0,

$$\bar{X}-\bar{Y} \sim N\left(0, \sigma^2\left(\frac{1}{m}+\frac{1}{n}\right)\right)$$

or

$$Z = (\bar{X}-\bar{Y})\left(\frac{1}{m}+\frac{1}{n}\right)^{-\frac{1}{2}}\frac{1}{\sigma} \sim N(0,1).$$

A size α test is the two-tailed test which rejects H_0 if $z > z_{\alpha/2}$ or if $z < -z_{\alpha/2}$, where $z_{\alpha/2}$ is, as in 8.3 the upper $\alpha/2$ point of $N(0,1)$. Note that $2\log L_X(H_0,H_1) = Z^2 \sim \chi_1^2$, so that for this case the approximation in Theorem 8.1 is again exact.

8.6 Goodness-of-fit tests

Suppose we observe n independent trials and note the numbers of times that each of k possible outcomes occurs, i.e., $(x_1, \ldots, x_k)$, with $\sum_j x_j = n$. Let p_i be the probability of outcome i. On the basis of this data we want to test the hypothesis that $p_1, \ldots, p_k$ take particular values. We allow that these values might depend on some unknown parameter θ (or parameters if θ is a vector). I.e., we want to test

$$H_0: \ p_i = p_i(\theta) \text{ for } \theta \in \Theta_0 \quad \text{against} \quad H_1: \ p_i \text{ are unrestricted.}$$

For example, H_0 might be the hypothesis that the trials are samples from a binomial distribution $B(k,\theta)$, so that under H_0 we would have $p_i(\theta) = \binom{k}{i}\theta^i(1-\theta)^{k-i}$.

This is called a **goodness-of-fit test**, because we are testing whether our data fit a particular distribution (in the above example the binomial distribution).

The distribution of $(x_1, \ldots, x_k)$ is the **multinomial distribution**

$$\mathbb{P}(x_1, \ldots, x_k \mid p) = \frac{n!}{x_1! \cdots x_k!}\, p_1^{x_1} \cdots p_k^{x_k},$$

for $(x_1, \ldots, x_k)$ s.t. $x_i \in \{0, \ldots, n\}$ and $\sum_{i=1}^k x_i = n$. Then we have

$$\sup_{H_1} \log f(x) = \text{const} + \sup\left\{\sum_{i=1}^k x_i \log p_i \;\middle|\; 0 \le p_i \le 1,\ \sum_{i=1}^k p_i = 1\right\}.$$

Now, $\sum_i x_i \log p_i$ may be maximised subject to $\sum_i p_i = 1$ by a Lagrangian technique and we get $\hat{p}_i = x_i/n$. Likewise,

$$\sup_{H_0} \log f(x) = \text{const} + \sup_{\theta}\left\{\sum_{i=1}^k x_i \log p_i(\theta)\right\}.$$

The generalized likelihood tells us to reject H_0 if $2L_x(H_0,H_1)$ is large compared to the chi-squared distribution with d.f. $|\Theta_1| - |\Theta_0|$. Here $|\Theta_1| = k-1$ and $|\Theta_0|$ is the number of independent parameters to be estimated under H_0.

9 Chi-squared tests of categorical data

A statistician is someone who refuses to play the national lottery,
but who does eat British beef. (anonymous)

9.1 Pearson's chi-squared statistic

Suppose, as in Section 8.6, that we observe $x_1, \dots, x_k$, the numbers of times that each of k possible outcomes occurs in n independent trials, and seek to make the **goodness-of-fit test** of

$$H_0 : \; p_i = p_i(\theta) \text{ for } \theta \in \Theta_0 \quad \text{against} \quad H_1 : \; p_i \text{ are unrestricted.}$$

Recall

$$2 \log L_x(H_0, H_1) = 2 \sum_{i=1}^k x_i \log \hat{p}_i - 2 \sum_{i=1}^k x_i \log p_i(\hat{\theta}) = 2 \sum_{i=1}^k x_i \log\bigl(\hat{p}_i / p_i(\hat{\theta})\bigr) \,,$$

where $\hat{p}_i = x_i/n$ and $\hat{\theta}$ is the MLE of θ under H_0. Let $o_i = x_i$ denote the number of time that outcome i occurred and let $e_i = np_i(\hat{\theta})$ denote the expected number of times it would occur under H_0. It is usual to display the data in k cells, writing o_i in cell i. Let $\delta_i = o_i - e_i$. Then

$$\begin{aligned}
2 \log L_x(H_0, H_1) &= 2 \sum_{i=1}^k x_i \log\bigl((x_i/n)/p_i(\hat{\theta})\bigr) \\
&= 2 \sum_{i=1}^k o_i \log(o_i/e_i) \\
&= 2 \sum_{i=1}^k (\delta_i + e_i) \log(1 + \delta_i/e_i) \\
&= 2 \sum_{i=1}^k (\delta_i + e_i)(\delta_i/e_i - \delta_i^2/2e_i^2 + \cdots) \\
&\doteqdot \sum_{i=1}^k \delta_i^2/e_i \\
&= \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \qquad (1)
\end{aligned}$$

This is called the **Pearson chi-squared statistic**.

For H_0 we have to choose θ. Suppose the optimization over θ has p degrees of freedom. For H_1 we have $k-1$ parameters to choose. So the difference of these

degrees of freedom is $k-p-1$. Thus, if H_0 is true the statistic (1) $\sim \chi^2_{k-p-1}$ approximately. A mnemonic for the d.f. is

$$\text{d.f.} = \#(\text{cells}) - \#(\text{parameters estimated}) - 1. \tag{2}$$

Note that

$$\sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \left[\frac{o_i^2}{e_i} - 2o_i + e_i \right] = \sum_{i=1}^{k} \frac{o_i^2}{e_i} - 2n + n = \sum_{i=1}^{k} \frac{o_i^2}{e_i} - n\,. \tag{3}$$

Sometimes (3) is easier to compute than (1).

Example 9.1 For the data from Mendel's experiment, the test statistic has the value 0.618. This is to be compared to χ^2_3, for which the 10% and 95% points are 0.584 and 7.81. Thus we certainly do not reject the theoretical model. Indeed, we would expect the observed counts to show even greater disparity from the theoretical model about 90% of the time.

Similar analysis has been made of many of Mendel's other experiments. The data and theory turn out to be too close for comfort. Current thinking is that Mendel's theory is right but that his data were massaged by somebody (Fisher thought it was Mendel's gardening assistant) to improve its agreement with the theory.

9.2 χ^2 test of homogeneity

Suppose we have a rectangular array of cells with m rows and n columns, with X_{ij} items in the (i,j) th cell of the array. Denote the row, column and overall sums by

$$X_{i\cdot} = \sum_{j=1}^{n} X_{ij}, \quad X_{\cdot j} = \sum_{i=1}^{m} X_{ij}, \quad X_{\cdot\cdot} = \sum_{i=1}^{m}\sum_{j=1}^{n} X_{ij}.$$

Suppose the row sums are fixed and the distribution of $(X_{i1}, \dots, X_{in})$ in row i is multinomial with probabilities $(p_{i1}, \dots, p_{in})$, independently of the other rows. We want to test the hypothesis that the distribution in each row is the same, i.e., $H_0 : p_{ij}$ is the same for all i, $(= p_j)$ say, for each $j = 1, \dots, n$. The alternative hypothesis is $H_1 : p_{ij}$ are unrestricted. We have

$$\log f(x) = \text{const} + \sum_i \sum_j x_{ij} \log p_{ij}\,, \quad \text{so that}$$

$$\sup_{H_1} \log f(x) = \text{const} + \sup \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} \log p_{ij} \;\middle|\; 0 \le p_{ij} \le 1, \; \sum_{j=1}^{n} p_{ij} = 1 \quad \forall i \right\}$$

Now, $\sum_j x_{ij} \log p_{ij}$ may be maximized subject to $\sum_j p_{ij} = 1$ by a Lagrangian technique. The maximum of $\sum_j x_{ij} \log p_{ij} + \lambda \left(1 - \sum_j p_{ij}\right)$ occurs when $x_{ij}/p_{ij} = \lambda$,

$\forall j$. Then the constraints give $\lambda = \sum_j x_{ij}$ and the corresponding maximizing p_{ij} is $\hat{p}_{ij} = x_{ij}/\sum_j x_{ij} = x_{ij}/x_{i\cdot}$. Hence,

$$\sup_{H_1} \log f(x) = \text{const} + \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \log(x_{ij}/x_{i\cdot}).$$

Likewise,

$$\begin{aligned}\sup_{H_0} \log f(x) &= \text{const} + \sup\left\{ \sum_i \sum_j x_{ij} \log p_j \;\middle|\; 0 \le p_j \le 1,\ \sum_j p_j = 1 \right\},\\ &= \text{const} + \sum_i \sum_j x_{ij} \log(x_{\cdot j}/x_{\cdot\cdot}).\end{aligned}$$

Here $\hat{p}_j = x_{\cdot j}/x_{\cdot\cdot}$. Let $o_{ij} = x_{ij}$ and write $e_{ij} = \hat{p}_j x_{i\cdot} = (x_{\cdot j}/x_{\cdot\cdot})x_{i\cdot}$ for the expected number of items in position (i,j) under H_0. As before, let $\delta_{ij} = o_{ij} - e_{ij}$. Then,

$$\begin{aligned}2\log L_x(H_0,H_1) &= 2\sum_i\sum_j x_{ij}\log(x_{ij}x_{\cdot\cdot}/x_{i\cdot}x_{\cdot j})\\ &= 2\sum_i\sum_j o_{ij}\log(o_{ij}/e_{ij})\\ &= 2\sum_i\sum_j(\delta_{ij}+e_{ij})\log(1+\delta_{ij}/e_{ij})\\ &\doteqdot \sum_i\sum_j \delta_{ij}^2/e_{ij}\\ &= \sum_i\sum_j(o_{ij}-e_{ij})^2/e_{ij}\ . \end{aligned} \tag{4}$$

For H_0, we have $(n-1)$ parameters to choose, for H_1 we have $m(n-1)$ parameters to choose, so the **degrees of freedom** is $(n-1)(m-1)$. Thus, if H_0 is true the statistic (4) $\sim \chi^2_{(n-1)(m-1)}$ approximately.

Example 9.2 The observed (and expected) counts for the study about aspirin and heart attacks described in Example 1.2 are

	Heart attack	**No heart attack**	**Total**
Aspirin	104 (146.52)	10,933 (10890.5)	11,037
Placebo	189 (146.48)	10,845 (10887.5)	11,034
Total	293	21,778	22,071

E.g., $e_{11} = \left(\frac{293}{22071}\right) 11037 = 146.52$. The χ^2 statistic is

$$\frac{(104-146.52)^2}{146.52} + \frac{(189-146.48)^2}{46.48} + \frac{(10933-10890.5)^2}{10890.5} + \frac{(10845-10887.5)^2}{10887.5} = 25.01\,.$$

The 95% point of χ_1^2 is 3.84. Since $25.01 > 3.84$, we reject the hypothesis that heart attack rate is independent of whether the subject did or did not take aspirin.

Note that if there had been only a tenth as many subjects, but the same percentages in each in cell, the statistic would have been 2.501 and not significant.

9.3 χ^2 test of row and column independence

This χ^2 test is similar to that of Section 9.2, but the hypotheses are different. Again, observations are classified into a $m \times n$ rectangular array of cells, commonly called a **contingency table**. The null hypothesis is that the row into which an observation falls is independent of the column into which it falls.

Example 9.3 *A researcher pretended to drop pencils in a lift and observed whether the other occupant helped to pick them up.*

	Helped	**Did not help**	**Total**
Men	370 (337.171)	950 (982.829)	1,320
Women	300 (332.829)	1,003 (970.171)	1,303
Total	670	1,953	2,623

To test the independence of rows and columns we take

$$H_0 : p_{ij} = p_i q_j \text{ with } 0 \le p_i, q_j \le 1, \ \sum_i p_i = 1, \ \sum_j q_j = 1\,;$$

$$H_1 : p_{ij} \text{ arbitrary s.t. } 0 \le p_{ij} \le 1, \ \sum_{i,j} p_{ij} = 1\,.$$

The same approach as previously gives MLEs under H_0 and H_1 of

$$\hat{p}_i = x_{i\cdot}/x_{\cdot\cdot}, \quad \hat{q}_j = x_{\cdot j}/x_{\cdot\cdot}, \quad e_{ij} = \hat{p}_i \hat{q}_j x_{\cdot\cdot} = (x_{i\cdot} x_{\cdot j}/x_{\cdot\cdot}), \quad \text{and} \quad \hat{p}_{ij} = x_{ij}/x_{\cdot\cdot}\,.$$

The test statistic can again be show to be about $\sum_{ij}(o_{ij} - e_{ij})^2/e_{ij}$. The e_{ij} are shown in parentheses in the table. E.g., $e_{11} = \hat{p}_1 \hat{q}_1 n = \left(\frac{1320}{2623}\right)\left(\frac{670}{2623}\right) 2623 = 337.171$. The number of free parameters under H_1 and H_0 are $mn - 1$ and $(m-1) + (n-1)$ respectively. The difference of these is $(m-1)(n-1)$, so the statistic is to be compared to $\chi^2_{(m-1)(n-1)}$. For the data above this is 8.642, which is significant compared to χ^2_1.

We have now seen Pearson χ^2 tests in three different settings. Such a test is appropriate whenever the data can be viewed as numbers of times that certain outcomes have occurred and we wish to test a hypothesis H_0 about the probabilities with which they occur. Any unknown parameter is estimated by maximizing the likelihood function that pertains under H_0 and e_i is computed as the expected number of times outcome i occurs if that parameter is replaced by this MLE value. The statistic is (1), where the sum is computed over all cells. The d.f. is given by (2).

10 Distributions of the sample mean and variance

Statisticians do it. After all, it's only normal.

10.1 Simpson's paradox

Example 10.1 *These are some Cambridge admissions statistics for 1996.*

	Women			Men		
	applied	accepted	%	applied	accepted	%
Computer Science	26	7	27	228	58	25
Economics	240	63	26	512	112	22
Engineering	164	52	32	972	252	26
Medicine	416	99	24	578	140	24
Veterinary medicine	338	53	16	180	22	12
Total	1184	274	23	2470	584	24

In all five subjects women have an equal or better success rate in applications than do men. However, taken overall, 24% of men are successful but only 23% of women are successful! This is called **Simpson's paradox** (though it was actually discovered by Yule 50 years earlier). It can often be found in real data. Of course it is not a paradox. The explanation here is that women are more successful in each subject, but tend to apply more for subjects that are hardest to get into (e.g., Veterinary medicine). This example should be taken as a warning that pooling contingency tables can produce spurious associations. The correct interpretation of this data is that, for these five subjects, women are significantly more successful in gaining entry than are men.

In order to produce an example of Simpson's paradox I carefully selected five subjects from tables of 1996 admissions statistics. Such 'data snooping' is cheating; a similar table that reversed the roles of men and women could probably be constructed by picking different subjects.

10.2 Transformation of variables

The rest of this lecture is aimed at proving some important facts about distribution of the statistics $\bar{X}$ and $S_{XX} = \sum_i (X_i - \bar{X})^2$, when $X_1, \dots, X_n$ are IID $N(\mu, \sigma^2)$. We begin by reviewing some ideas about transforming random variables.

Suppose the joint density of $X_1, \dots, X_n$ is f_X, and there is a 1–1 mapping between $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ such that $X_i = x_i(Y_1, \dots, Y_n)$. Then the joint density of

$Y_1, \ldots, Y_n$ is

$$f_Y(y_1, \ldots, y_n) = f_X(x_1(y), \ldots, x_n(y)) \begin{vmatrix} \frac{\partial x_1(y)}{\partial y_1} & \cdots & \frac{\partial x_1(y)}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial x_n(y)}{\partial y_1} & \cdots & \frac{\partial x_n(y)}{\partial y_n} \end{vmatrix}$$

where the **Jacobian** $:= J(y_1, \ldots, y_n)$ is the absolute value of the determinant of the matrix $(\partial x_i(y)/\partial y_j)$.

The following example is an important one, which also tells us more about the beta distribution.

Example 10.2 Let $X_1 \sim \text{gamma}(n_1, \lambda)$ and $X_2 \sim \text{gamma}(n_2, \lambda)$, independently. Let $Y_1 = X_1/(X_1 + X_2)$, $Y_2 = X_1 + X_2$. Since X_1 and X_2 are independent we multiply their pdfs to get

$$f_X(x) = \frac{\lambda^{n_1} x_1^{n_1-1}}{(n_1-1)!} e^{-\lambda x_1} \times \frac{\lambda^{n_2} x_2^{n_2-1}}{(n_2-1)!} e^{-\lambda x_2}.$$

Then $x_1 = y_1 y_2$, $x_2 = y_2 - y_1 y_2$, so

$$J(y_1, y_2) = \begin{vmatrix} \frac{\partial x_1(y)}{\partial y_1} & \frac{\partial x_1(y)}{\partial y_2} \\ \frac{\partial x_2(y)}{\partial y_1} & \frac{\partial x_2(y)}{\partial y_2} \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ -y_2 & 1 - y_1 \end{vmatrix} = y_2$$

Hence making the appropriate substitutions and arranging terms we get

$$f_Y(y) = \frac{(n_1+n_2-1)!}{(n_1-1)!(n_2-1)!} y_1^{n_1-1}(1-y_1)^{n_2-1} \times \frac{\lambda^{n_1+n_2} y_2^{n_1+n_2-1}}{(n_1+n_2-1)!} e^{-\lambda y_2}$$

from which it follows that Y_1 and Y_2 are independent RVs (since their joint density function factors into marginal density functions) and $Y_1 \sim \text{beta}(n_1, n_2)$, $Y_2 \sim \text{gamma}(n_1 + n_2, \lambda)$.

10.3 Orthogonal transformations of normal variates

Lemma 10.3 *Let $X_1, \ldots, X_n$, be independently distributed with distributions $N(\mu_i, \sigma^2)$ respectively. Let $A = (a_{ij})$ be an orthogonal matrix, so that $A^\top A = AA^\top = I$. Then the elements of $\boldsymbol{Y} = A\boldsymbol{X}$ are independently distributed, and $Y_i \sim N\big((A\mu)_i, \sigma^2\big)$, where $\mu = (\mu_1, \ldots, \mu_n)^\top$.*

Proof. The joint density of $X_1, \ldots, X_n$ is

$$\begin{aligned} f_X(x_1, \ldots, x_n \mid \mu, \sigma^2) &= \prod_i f_{X_i}(x_i \mid \mu_i, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_i (x_i - \mu_i)^2/2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(\mathbf{x}-\mu)^\top(\mathbf{x}-\mu)/2\sigma^2} \end{aligned}$$

Since $\mathbf{x} = A^\top \mathbf{y}$, we have $\partial x_i/\partial y_j = a_{ji}$ and hence $J(y_1, \ldots, y_n) = |\det(A^\top)| = 1$. Thus

$$\begin{aligned}
f_Y(y_1, \ldots, y_n \mid \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-(A^\top\mathbf{y} - \mu)^\top (A^\top\mathbf{y} - \mu)/2\sigma^2\right] \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-(A^\top\mathbf{y} - A^\top A\mu)^\top (A^\top\mathbf{y} - A^\top A\mu)/2\sigma^2\right] \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-(\mathbf{y} - A\mu)^\top A A^\top (\mathbf{y} - A\mu)/2\sigma^2\right] \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-(\mathbf{y} - A\mu)^\top (\mathbf{y} - A\mu)/2\sigma^2\right]
\end{aligned}$$

■

Remark. An alternative proof can be given using moment generating functions. For $\theta \in \mathbb{R}^n$, the mgf of the joint distribution is

$$\begin{aligned}
\mathbb{E}\exp\left[\theta^\top \mathbf{Y}\right] &= \mathbb{E}\exp\left[\theta^\top A\mathbf{X}\right] \\
&= \mathbb{E}\exp\left[(A^\top\theta)^\top \mathbf{X}\right] \\
&= \exp\left[(A^\top\theta)^\top \mu + \tfrac{1}{2}\sigma^2 (A^\top\theta)^\top (A^\top\theta)\right] \\
&= \exp\left[\theta^\top A\mu + \tfrac{1}{2}\sigma^2 \theta^\top\theta\right]
\end{aligned}$$

which we recognise as the mgf of independent RVs with distributions $N\big((A\mu)_i, \sigma^2\big)$.

10.4 The distributions of $\bar{X}$ and S_{XX}

Lemma 10.4 *Let $X_1, \ldots, X_n$ be IID $N(\mu, \sigma^2)$ and let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$, $S_{XX} = \sum_{i=1}^n (X_i - \bar{X})^2$. Then:*

(i) $\bar{X} \sim N(\mu, \sigma^2/n)$ and $n(\bar{X} - \mu)^2 \sim \sigma^2\chi_1^2$.

(ii) $X_i - \mu \sim N(0, \sigma^2)$, so $\sum_{i=1}^n (X_i - \mu)^2 \sim \sigma^2\chi_n^2$.

(iii) $\sum_{i=1}^n (X_i - \mu)^2 = S_{XX} + n(\bar{X} - \mu)^2$.

(iv) $S_{XX}/(n-1)$ is an unbiased estimator of σ^2.

(v) $\bar{X}$ and S_{XX} are independent.

(vi) $S_{XX} \sim \sigma^2\chi_{n-1}^2$.

Proof.

(i) and (ii) are immediate from the fact that linear combinations of normal RVs are normally distributed and the definition of χ_n^2. To prove (iii) and (iv) we note that

$$\begin{aligned}\sum_{i=1}^n (X_i-\mu)^2 &= \sum_{i=1}^n \big([X_i-\bar X]+[\bar X-\mu]\big)^2\\ &= \sum_{i=1}^n \big([X_i-\bar X]^2+2[X_i-\bar X][\bar X-\mu]+[\bar X-\mu]^2\big)\\ &= S_{XX}+n[\bar X-\mu]^2\end{aligned}$$

Let A be an orthogonal matrix such that

$$\mathbf{Y}=A\big(\mathbf{X}-\mu 1)=\big(\sqrt{n}(\bar X-\mu),Y_2,\dots,Y_n\big)\,.$$

I.e., we take

$$A=\begin{pmatrix}1/\sqrt{n} & 1/\sqrt{n} & \cdots & 1/\sqrt{n}\\ \vdots & \vdots & & \vdots\\ \cdot & \cdot & \cdots & \cdot\end{pmatrix}$$

where the rows below the first are chosen to make the matrix orthogonal. Then $Y_1=\sqrt{n}(\bar X-\mu)\sim N(0,\sigma^2)$ and Y_1 is independent of $Y_2,\dots,Y_n$. Since $\sum_{i=1}^n Y_i^2=\sum_i (X_i-\mu)^2$, we must have

$$\sum_{i=2}^n Y_i^2=\sum_{i=1}^n (X_i-\mu)^2-n(\bar X-\mu)^2=S_{XX}.$$

Hence S_{XX} and Y_1 (and equivalently S_{XX} and $\bar X$) are independent. This gives (v).

Finally, (vi) follows from $S_{XX}=\sum_{i=2}^n Y_i^2$ and the fact that $Y_2,\dots,Y_n$ are IID $N(0,\sigma^2)$. ■

10.5 Student's t-distribution

If $X\sim N(0,1)$, $Y\sim\chi^2_n$, independently of X, then

$$Z=X/(Y/n)^{\frac{1}{2}}\sim t_n,$$

where t_n is the Student's t-**distribution** with (or on) n degrees of freedom. Like the normal distribution, this distribution is symmetric about 0, and bell-shaped, but has more probability in its tails, i.e., for all $t>0$, $\mathbb{P}(Z>t)>\mathbb{P}(X>t)$.

From Lemma 10.4 we have $\sqrt{n}(\bar X-\mu)\sim\sigma N(0,1)$ and $S_{XX}\sim\sigma^2\chi^2_{n-1}$, independently. So from these and the definition of the t-distribution follows the important fact that if $X_1,\dots,X_n$ are IID $N(\mu,\sigma^2)$, then

$$\frac{\sqrt{n}(\bar X-\mu)}{\sqrt{S_{XX}/(n-1)}}\sim t_{n-1}\,.$$

11 The t-test

Statisticians do it with two-tail T tests.

11.1 Confidence interval for the mean, unknown variance

Suppose $X_1, \dots, X_n$ IID $N(\mu, \sigma^2)$, but now σ^2 is unknown. Recall

$$T = \frac{\sqrt{n}(\bar{X} - \mu)}{\hat{\sigma}} \sim t_{n-1} \, .$$

where $\hat{\sigma}^2 = S_{XX}/(n-1)$. A $100(1-\alpha)\%$ confidence interval for μ follows from

$$1 - \alpha = \mathbb{P}\left(-t_{\alpha/2}^{(n-1)} \leq \frac{\sqrt{n}(\bar{X}-\mu)}{\hat{\sigma}} \leq t_{\alpha/2}^{(n-1)}\right) = \mathbb{P}\left(\bar{X} - \frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{t_{\alpha/2}\hat{\sigma}}{\sqrt{n}}\right)$$

where $t_{\alpha/2}^{(n-1)}$ is the 'upper $\alpha/2$ point of a t-distribution on $n-1$ degrees of freedom', i.e., such that $\mathbb{P}\left(T > t_{\alpha/2}^{(n-1)}\right) = \alpha/2$.

Example 11.1 *In 'Sexual activity and the lifespan of male fruitflies',* Nature*, 1981, Partridge and Farquhar report experiments which examined the cost of increased reproduction in terms of reduced longevity for male fruitflies. They kept numbers of male flies under different conditions. 25 males in one group were each kept with 1 receptive virgin female. 25 males in another group were each kept with 1 female who had recently mated. Such females will refuse to remate for several days. These served as a control for any effect of competition with the male for food or space. The groups were treated identically in number of anaesthetizations (using CO2) and provision of fresh food.*

To verify 'compliance' two days per week throughout the life of each experimental male, the females that had been supplied as virgins to that male were kept and examined for fertile eggs. The insemination rate declined from approximately 1 per day at age one week to about 0.6 per day at age eight weeks.

The data was as follows

Groups of 25 males kept with	mean life (days)	s.e.
1 uninterested female	64.80	15.6525
1 interested female	56.76	14.9284

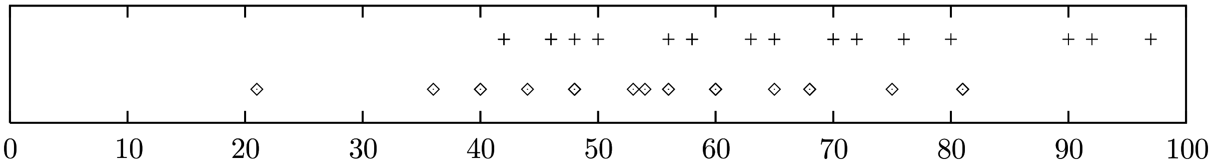
Here s.e. is an abbreviation for **standard error**, i.e. the value of $\hat{\sigma} = \sqrt{S_{xx}/(n-1)}$. Here $n = 25$. The mean life, $\bar{x}$ and the s.e., $\hat{\sigma}$, are sufficient statistics for (μ, σ^2), so there is nothing else we need to know about the individual values of the the longevities of these 50 flies in order to compute confidence intervals or test statistics.

From these summary statistics we can compute 95% confidence intervals for the mean lives of the control and test groups to be

$$[64.80 - 2.06(15.6526)/\sqrt{25}, 64.80 + 2.06(15.6526)/\sqrt{25}] = [58.35, 71.25]$$
$$[56.76 - 2.06(14.9284)/\sqrt{25}, 56.76 + 2.06(14.9284)/\sqrt{25}] = [50.61, 62.91]$$

It is interesting to look at the data, and doing so helps us check that lifespan is normally distributed about a mean. The longevities for control and test groups were

42 42 46 46 46 48 50 56 58 58 63 65 65 70 70 70 70 72 72 76 76 80 90 92 97
21 36 40 40 44 48 48 48 48 53 54 56 56 60 60 60 60 65 68 68 68 75 81 81 81



11.2 Single sample: testing a given mean, unknown variance (t-test)

Suppose that with the same assumptions as above it is required to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

Adopting the paradigm of the generalized likelihood ratio test we consider

$$\begin{aligned}
L_x(H_0, H_1) &= \frac{\max_{\mu,\sigma^2} f\bigl(x \mid \mu, \sigma^2\bigr)}{\max_{\sigma^2} f\bigl(x \mid \mu_0, \sigma^2\bigr)} \\
&= \frac{\bigl[2\pi \sum_i (x_i - \bar{x})^2/n\bigr]^{-n/2} \exp\left[-n/2\right]}{\bigl[2\pi \sum_i (x_i - \mu_0)^2/n\bigr]^{-n/2} \exp\left[-n/2\right]} \\
&= \left[\frac{\sum_i (x_i - \mu_0)^2}{\sum_i (x_i - \bar{x})^2}\right]^{n/2} \\
&= \left[\frac{\sum_i (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2}{\sum_i (x_i - \bar{x})^2}\right]^{n/2} \\
&= \left[1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_i (x_i - \bar{x})^2}\right]^{n/2} .
\end{aligned}$$

This is large when $T^2 := n(n-1)(\bar{x} - \mu_0)^2 \,/\, \sum_i (x_i - \bar{x})^2$ is large, equivalently when $|T|$ is large. Under H_0 we have $T \sim t_{n-1}$. So a size α test is the two-tailed test which rejects H_0 if $t > t_{\alpha/2}^{(n-1)}$ or if $t < -t_{\alpha/2}^{(n-1)}$.

Example 11.2 *Does jogging lead to a reduction in pulse rate? Eight non-jogging volunteers engaged in a one-month jogging programme. Their pulses were taken before and after the programme.*

pulse rate before	74	86	98	102	78	84	79	70
pulse rate after	70	85	90	110	71	80	69	74
decrease	4	1	8	-8	7	4	10	-4

Although there are two sets of data it is really just the changes that matter. Let the decreases in pulse rates be $x_1, \dots, x_8$ and assume these are samples from $N(\mu, \sigma^2)$ for some unknown σ^2. To test $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ we compute

$$\sum x_i = 22, \quad \bar{x} = 2.75, \quad \sum x_i^2 = 326, \quad S_{xx} = \sum x_i^2 - 8\bar{x}^2 = 265.5.$$

Hence the test statistic is

$$t = \frac{\sqrt{8}(2.75 - 0)}{\sqrt{265.5/(8-1)}} = 1.263,$$

which is to be compared to $t_{0.025}^{(7)} = 2.365$. Hence the data is not sufficient to reject H_0 at the 5% level. This may surprise you since 6 of the 8 subjects had lowered pulse rates. This sort of test is called a **paired samples t-test**.

11.3 Two samples: testing equality of means, unknown common variance (t-test)

We have the same samples as in 8.5, i.e., $X_1, \dots, X_m$ are IID $N(\mu_1, \sigma^2)$ and $Y_1, \dots, Y_n$ are IID $N(\mu_2, \sigma^2)$. These two samples are independent. It is required to test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, but now σ^2 is unknown. Note how this differs from the paired t-test above: the samples are not paired, and can be of unequal sizes.

As above, a maximum likelihood approach could convince us that the test should be of the form ‘reject H_0 if $(\bar{x} - \bar{y})^2/(S_{xx} + S_{yy})$ is large.

As in 8.5 we have that under H_0,

$$(\bar{X} - \bar{Y}) \left(\frac{1}{m} + \frac{1}{n} \right)^{-\frac{1}{2}} \frac{1}{\sigma} \sim N(0,1) .$$

If $S_{XX} = \sum_{i=1}^m (X_i - \bar{X})^2$, and $S_{YY} = \sum_{j=1}^n (Y_j - \bar{Y})^2$, then

$$(S_{XX} + S_{YY})/\sigma^2 \sim \chi^2_{m+n-2}$$

so that (since $\bar{X}$ is independent of S_{XX}, $\bar{Y}$ is independent of S_{YY} and the two samples are independent)

$$T = (\bar{X} - \bar{Y}) \Bigg/ \sqrt{\left(\frac{1}{m} + \frac{1}{n} \right) \left(\frac{S_{XX} + S_{YY}}{m+n-2} \right)} \sim t_{m+n-2}.$$

A test of size α rejects H_0 if $t > t_{\alpha/2}^{(m+n-2)}$ or if $t < -t_{\alpha/2}^{(m+n-2)}$.

Example 11.3 *For the fruitfly data we might test H_0 : that mean longevity is the same for males living with 8 interested females as with 8 uninterested females.* The test statistic is

$$t = (64.80 - 56.76)\Bigg/\sqrt{\left(\frac{1}{25}+\frac{1}{25}\right)\left(\frac{24(15.6525)+24(14.9284)}{25+25-2}\right)} = 1.859$$

which can be compared to $t_{0.025}^{(48)} = 2.01$, and therefore is not significant at the 0.05% level. H_0 is not rejected. (It is however, significant at the 10% level, since $t_{0.05}^{(48)} = 1.68$).

Similarly, we can give a 95% confidence interval for the difference of the means. This has endpoints

$$\begin{aligned}(64.80 - 56.76) \pm 2.01\sqrt{\left(\frac{1}{25}+\frac{1}{25}\right)\left(\frac{24(15.6525)+24(14.9284)}{25+25-2}\right)}&\\ = 8.04 \pm 8.695.&\end{aligned}$$

I.e., a 95% confidence interval for the extra longevity of celibate males is $[-0.655, 16.735]$ days. Notice again that finding we cannot reject $\mu_1 - \mu_2 = 0$ at the 5% level is equivalent to finding that the 95% confidence interval for the difference of the means contains 0.

In making the above test we have assumed that the variances for the two populations are the same. In the next lecture we will see how we might test that hypothesis.

11.4 Single sample: testing a given variance, unknown mean (χ^2-test)

Let $X_1, \dots, X_n$ be IID $N(\mu, \sigma^2)$, and suppose we wish to test $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$, where μ is unknown, and therefore a '**nuisance parameter**'.

Following Theorem 8.1, the likelihood ratio is

$$L_x(H_0, H_1) = \frac{\max_{\mu,\sigma^2} f\left(x \mid \mu, \sigma^2\right)}{\max_\mu f\left(x \mid \mu, \sigma_0^2\right)} = \frac{\left[2\pi\frac{\sum_i (x_i-\bar{x})^2}{n}\right]^{-n/2} \exp\left[-n/2\right]}{\left(2\pi\sigma_0^2\right)^{-n/2} \exp\left[-(1/2\sigma_0^2)\sum_i (x_i - \bar{x})^2\right]}$$

As in Section 8.4 this is large when $\sum_i (x_i - \bar{x})/n\sigma_0^2$ $(= S_{xx}/n\sigma_0^2)$ differs substantially from 1.

Under H_0, $S_{XX}/\sigma_0^2 \sim \chi_{n-1}^2$. Given the required size of test α, let a_1, a_2 be such that

$$\mathbb{P}(S_{XX}/\sigma_0^2 < a_1) + \mathbb{P}(S_{XX}/\sigma_0^2 > a_2) = \alpha$$

under H_0. Then a size α test is to reject H_0 if $S_{xx}/\sigma_0^2 < a_1$ or if $S_{xx}/\sigma_0^2 > a_2$.

Usually we would take $a_1 = F_{n-1}^{-1}(\alpha/2), a_2 = F_{n-1}^{-1}(1 - \alpha/2)$, where F_{n-1} is the distribution function of a χ_{n-1}^2 random variable.

12 The F-test and analysis of variance

The statistician's attitude to variation is like that of the evangelist to sin; he sees it everywhere to a greater or lesser extent.

12.1 F-distribution

If $X \sim \chi^2_m$ and $Y \sim \chi^2_n$, independently of X, then

$$Z = (X/m)/(Y/n) \sim F_{m,n},$$

has the **F-distribution** with (or on) m and n degrees of freedom.

Note that if $T \sim F_{m,n}$ then $1/T \sim F_{n,m}$. Tables for the F-distribution usually only give the upper percentage points. If we want to know x such that $\mathbb{P}(T < x) = 0.05$ we can use $F_{n,m}$ tables to find $1/x$ such that $\mathbb{P}(1/T > 1/x) = 0.05$.

Note that if $X \sim t_n$ then $X^2 \sim F_{1,n}$. It is always nice to recognise connections between distributions.

12.2 Two samples: comparison of variances (F-test)

Suppose $X_1, \dots, X_m$ are IID $N(\mu_1, \sigma_1^2)$ and $Y_1, \dots, Y_n$ are IID $N(\mu_2, \sigma_2^2)$, with the two samples independent. It is required to test $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_1 : \sigma_1^2 > \sigma_2^2$, with μ_1, μ_2 unknown nuisance parameters.

Now, by either the generalized likelihood ratio test, or common sense, we are led to consider the statistic

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{S_{XX}/(m-1)}{S_{YY}/(n-1)} \sim \frac{\sigma_1^2 \chi^2_{m-1}/(m-1)}{\sigma_2^2 \chi^2_{n-1}/(n-1)} = \frac{\sigma_1^2}{\sigma_2^2} F_{m-1,n-1}.$$

Thus, under H_0, $F \sim F_{m-1,n-1}$.

If H_1 is true, F will tend to be greater than when H_0 is true, so we reject H_0 if this ratio is large. A size α test is to reject H_0 if $f > F_\alpha^{(m-1,n-1)}$, the upper α point of $F_{m-1,n-1}$.

Example 12.1 *Suppose we wish to test the hypothesis that the variance of longevity is the same for male fruitflies kept with 1 interested or 1 uninterested female, i.e., $H_0 : \sigma_1^2 = \sigma_2^2$ against $H_0 : \sigma_1^2 \neq \sigma_2^2$.*The test statistic is

$$f = (15.6525)^2/(14.9284)^2 = 1.099,$$

which, as $F_{0.05}^{(24,24)} = 1.98$, is not significant at the 10% level (the test is two-tailed). Notice that in order to use F tables we put the larger of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ in the numerator.

12.3 Non-central χ^2

If $X_1, \dots, X_k$ are independent $N(\mu_i, 1)$ then $Z = \sum_{i=1}^k X_i^2$ has the **non-central chi-squared** distribution, $\chi_k^2(\lambda)$, with non-centrality parameter $\lambda = \sum_{i=1}^k \mu_i^2$. Note that $EW = k + \lambda$; thus a non-central χ_k^2 tends to be larger than a central χ_k^2.

To see that it is only the value of λ matters, let A be an orthogonal matrix such that $A\mu = (\lambda^{1/2}, 0, \dots, 0)^\top$, so $(A\mu)^\top(A\mu) = \mu^\top\mu = \lambda$. Let $Y = AX$; then $\sum_{i=1}^k X_i^2 = \sum_{i=1}^k Y_i^2$, with $Y_1^2 = \chi_1^2(\lambda)$ and $\sum_{i=2}^k Y_i^2 = \chi_{k-1}^2$.

12.4 One way analysis of variance

Analysis of variance (ANOVA) is a technique for testing hypotheses about means by looking at sample variances. We consider here the question of testing equality of the means of $k > 2$ groups. The mathematical model is:

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, k.$$

Thus there are n_i observations in the ith group. Let $\sum_{i=1}^k n_i = N$.

It is assumed that the ϵ_{ij} are IID $N(0, \sigma^2)$, and that our data consists of observations x_{ij} which are realisations of random variables X_{ij} satisfying the model.

One-way ANOVA is used to test the null hypothesis $H_0 : \mu_1 = \dots = \mu_k$. The alternative hypothesis H_1 is 'H_0 is not true'. Application of the generalized likelihood ratio test gives

$$\begin{aligned} L_x(H_0, H_1) &= \frac{\max_{\mu_1,\cdots,\mu_k,\sigma^2} (2\pi\sigma^2)^{-N/2} \exp\left[-\sum_{ij}(x_{ij} - \mu_i)^2/2\sigma^2\right]}{\max_{\mu,\sigma^2} (2\pi\sigma^2)^{-N/2} \exp\left[-\sum_{ij}(x_{ij} - \mu)^2/2\sigma^2\right]} \\ &= \left[\frac{s_0}{s_1}\right]^{N/2}, \quad \text{where} \quad s_0 := \sum_{ij}(x_{ij} - \bar{x}_{\cdot\cdot})^2 \quad \text{and} \quad s_1 := \sum_{ij}(x_{ij} - \bar{x}_{i\cdot})^2. \end{aligned}$$

Here, $\bar{x}_{\cdot\cdot} = \sum_{ij} x_{ij}/N = \sum_i n_i \bar{x}_{i\cdot}/N$ is the overall mean (and the MLE of μ under H_0). Similarly, $\bar{x}_{i\cdot} = \sum_{j=1}^{n_i} x_{ij}/n_i$ is the mean within the ith group (and the MLE of μ_i under H_1).

Thus we are led to consider rejecting H_0 when s_0/s_1 is large. Now

$$\begin{aligned} s_0 &= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot} + \bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 \\ &= \sum_i \sum_j \left[(x_{ij} - \bar{x}_{i\cdot})^2 + 2(x_{ij} - \bar{x}_{i\cdot})(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot}) + (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2\right] \\ &= \sum_{ij}(x_{ij} - \bar{x}_{i\cdot})^2 + \sum_i n_i(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 \\ &= s_1 + s_2, \end{aligned}$$

where $s_2 := \sum_i n_i(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$, and thus s_0/s_1 is large when s_2/s_1 is large.

s_1 is called the **within samples sum of squares** and s_2 is called the **between samples sum of squares**.

Now, whether or not H_0 is true, $\sum_j (X_{ij} - \bar{X}_{i\cdot})^2 \sim \sigma^2 \chi^2_{n_i - 1}$, since $E(X_{ij})$ depends only on i. Hence, $S_1 \sim \sigma^2 \chi^2_{N-k}$, since samples for different i are independent.

Also, $\sum_j (X_{ij} - \bar{X}_{i\cdot})^2$ is independent of $\bar{X}_{i\cdot}$, so that S_1 is independent of S_2. If H_0 is true $S_2 \sim \sigma^2 \chi^2_{k-1}$, and if H_0 is not true, $S_2 \sim \sigma^2 \chi^2_{k-1}(\lambda)$, where

$$E(S_2) = (k-1)\sigma^2 + \lambda, \quad \lambda = \textstyle\sum_{i=1}^k n_i(\mu_i - \bar{\mu})^2, \quad \bar{\mu} = \sum_i n_i \mu_i / N.$$

Intuitively, if H_0 is not true S_2 tends to be inflated.

So, if H_0 is true then $Q = \{S_2/(k-1)\}/\{S_1/(N-k)\} \sim F_{k-1,N-k}$, while if H_0 is not true, Q tends to be larger. So for a size α test we reject H_0 if $q > F_\alpha^{(k-1,N-k)}$.

An interpretation of this is that the variability in the total data set is $s_0 = \sum_{ij}(x_{ij} - \bar{x}_{\cdot\cdot})^2$. Under H_1 we expect x_{ij} to be about $\bar{x}_{i\cdot}$ and so a variability of $s_2 = \sum_{ij}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2$ is 'explained' by H_1. Statisticians say that H_1 'explains $(s_2/s_0)100\%$ of the variation in the data', (where since $s_0 = s_1 + s_2$, we must have $s_2/s_0 \leq 1$.) If s_2/s_0 is near 1, or equivalently if s_2/s_1 is large, then H_1 does much better than H_0 in explaining why the data has the variability it does.

Example 12.2 *Partridge and Farquhar did experiments with five different groups of 25 male fruitflies. In addition to the groups kept with 1 interested or 1 uninteresed female, 25 males were each kept with no companions, and groups of 25 were each kept with 8 uninterested or 8 interested females. The 'compliance' of the males who were supplied with 8 virgin females per day varied from 7 inseminations per day at age one week to just under 2 per day at age eight weeks.*

Groups of 25 males kept with	mean life (days)	s.e.
no companions	63.56	16.4522
1 uninterested female	64.80	15.6525
1 interested female	56.76	14.9284
8 uninterested females	63.36	14.5398
8 interested females	38.72	12.1021

Suppose we wish to test equality of means in the three control groups, i.e., those kept with either no companions, or 1 or 8 uninterested females (rows 1, 2 and 4).

First we reconstruct the sums of squares,

$$\textstyle\sum_{j=1}^{25}(x_{1j} - \bar{x}_1)^2 = 24(16.4522^2) = 6496.16$$

$$\textstyle\sum_{j=1}^{25}(x_{2j} - \bar{x}_2)^2 = 24(15.6525^2) = 5880.00$$

$$\textstyle\sum_{j=1}^{25}(x_{4j} - \bar{x}_4)^2 = 24(14.5398^2) = 5073.76$$

then we calculate the within and between sums of squares,

$$\bar{x} = (63.56 + 64.80 + 63.36)/3 = 63.91$$
$$s_1 = 6496.16 + 5880.00 + 5073.76 = 17449.92$$
$$s_2 = \sum_{i=1,2,4} 25(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^2 = \sum_{i=1,2,4} 25\bar{x}_{i\cdot}^2 - 75\bar{x}_{\cdot\cdot}^2 = 30.427$$

and finally we compute the test statistic,

$$q = \frac{30.427/(3-1)}{17449.92/(75-3)} = 0.0628.$$

It is usual to display this data in an ANOVA table of the following form.

Source of variation	Degrees of freedom		Sum of squares		Mean square		F statistic	
Between groups	$k-1$	2	s_2	30.427	$s_2/(k-1)$	15.213	$\frac{s_2/(k-1)}{s_1/(N-k)}$	0.0628
Within groups	$N-k$	72	s_1	17449.92	$s_1/(N-k)$	242.36		
Total	$N-1$	74	s_0	17480.35				

The value of 0.0628 is not significant compared to $F_{0.05}^{(2,72)} = 3.12$ and hence we do not reject the hypothesis of equal means.

A similar test for equality of all five group means gives a statistic with value 507.5, to be compared to $F_{0.05}^{(4,120)} = 2.45$. Clearly we reject the hypothesis of equal means. It does seem that sexual activity is associated with reduced longevity.

ANOVA can be carried out for many other experimental designs. We might want to investigate more than one treatment possibility, or combinations of treatments. (E.g., in the fruitfly experiments each male fly was kept separate from other males; we might want to do experiments in which males are kept with different numbers of interested females and/or competing males.) If there are k possible treatments which can be applied or not applied, then 2^k different combinations are possible and this may be more than is realistic. The subject of 'experimental design' has to do with deciding how to arrange the treatments so as to gather as much information as possible from the fewest observations. The data is to be analysed to compare treatment effects and this typically involves some sort of ANOVA. The methodology is the same as for the one-way ANOVA considered above; we consider a normalised quotient, such as q above, between the reduction in the residual sums of squares that is obtained when moving from H_0 to H_1 (e.g., $s_0 - s_1$) and the value of the residual sum of squares under H_1 (e.g., s_1). In subsequent lectures we will see further examples of this idea in the context of regression models.

13 Linear regression and least squares

Numbers are like people; torture them enough and they'll tell you anything.

13.1 Regression models

One of the most widely used examples of estimation in statistics is provided by linear regression. For example, if Y_i is the number of unemployed in UK in the ith month after some date, we might make the hypothesis that

$$Y_i = a + \beta i + \epsilon_i, \quad i = 1, \dots, n,$$

where $\epsilon_1, \dots, \epsilon_n$ are IID $N(0, \sigma^2)$ and a, β are some unknown constants. The business of estimating β is to do with detecting a trend in unemployment. A related problem is that of testing $H_0 : \beta = 0$, a test of whether there is any trend in unemployment.

The model above is a special case of the **simple linear regression model** in which, with the same assumptions on $\{\epsilon_i\}$, a, β, σ^2,

$$Y_i = a + \beta x_i + \epsilon_i, \quad i = 1, \dots, n,$$

where the x_i are known constants. In the case above $x_i = i$.

A **multiple regression** model has more that one explanatory variable on the right hand side, e.g.,

$$Y_i = a + \beta_1 \log i + \beta_2 z_{i-5} + \epsilon_i, \quad i = 1, \dots, n,$$

where perhaps z_{i-5} is the number of unemployed people who were in training programmes five months earlier. The estimation of a, β_1 and β_2 and associated tests are similar to what we find for simple linear regression.

13.2 Least squares/MLE

Suppose $Y_1, \dots, Y_n$ are independent and $Y_i = \alpha + \beta w_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, or equivalently that $Y_i \sim N(\alpha + \beta w_i, \sigma^2)$, and where α, β and σ^2 are unknown parameters and the w_i are known constants such that $\sum_i w_i = 0$.

Theorem 13.1 *The MLEs of α and β are obtained by minimizing*

$$S = \sum_{i=1}^{n} \big(Y_i - \mathbb{E}(Y_i)\big)^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta w_i)^2$$

with respect to α and β. These are called the **least squares estimators** *and are given by:*

$$\boxed{\hat{\alpha} = \bar{Y} \qquad \textit{and} \qquad \hat{\beta} = S_{wY}/S_{ww}}$$

where $S_{ww} = \sum_i w_i^2$, and $S_{wY} = \sum_i w_i Y_i$.

Proof. Since $Y_i \sim N(\alpha + \beta w_i, \sigma^2)$ the likelihood of of $y_1, \dots, y_n$ is

$$f_Y(y \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta w_i)^2\right) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-S/2\sigma^2}.$$

The maximum likelihood estimator minimizes S, and so at a minimum,

$$\left.\frac{\partial S}{\partial \alpha}\right|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} = -2\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}w_i) = 0\ , \qquad \left.\frac{\partial S}{\partial \beta}\right|_{\substack{\alpha=\hat{\alpha}\\ \beta=\hat{\beta}}} = -2\sum_{i=1}^{n} w_i(y_i - \hat{\alpha} - \hat{\beta}w_i) = 0.$$

Hence

$$\sum_{i=1}^{n} Y_i - n\hat{\alpha} = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} w_i Y_i - \hat{\beta}\sum_{i=1}^{n} w_i^2 = 0,$$

from which the answers follow. ■

13.3 Practical usage

Given a linear regression model

$$Y_i = a + \beta x_i + \epsilon_i,$$

in which $\sum_i x_i \neq 0$ we make the transformation $w_i = x_i - \bar{x}$ and consider

$$Y_i = \alpha + \beta w_i + \epsilon_i,$$

where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\alpha = a + \beta\bar{x}$. This gives the situation described in 13.1, and we can use results of 13.2 to estimate the regression and the results in 14.1 to perform tests. Making the necessary transformations we have

$$\boxed{\hat{a} = \bar{Y} - \hat{\beta}\bar{x} \qquad \text{and} \qquad \hat{\beta} = S_{xY}/S_{xx}}$$

where $S_{xx} = \mathbf{w}^\top\mathbf{w} = \sum_i (x_i - \bar{x})^2$ and $S_{xY} = \sum_i (x_i - \bar{x})(Y_i - \bar{Y})$.

We speak of 'regressing y on x'. A package such as MINITAB will return the estimated regression line in the form

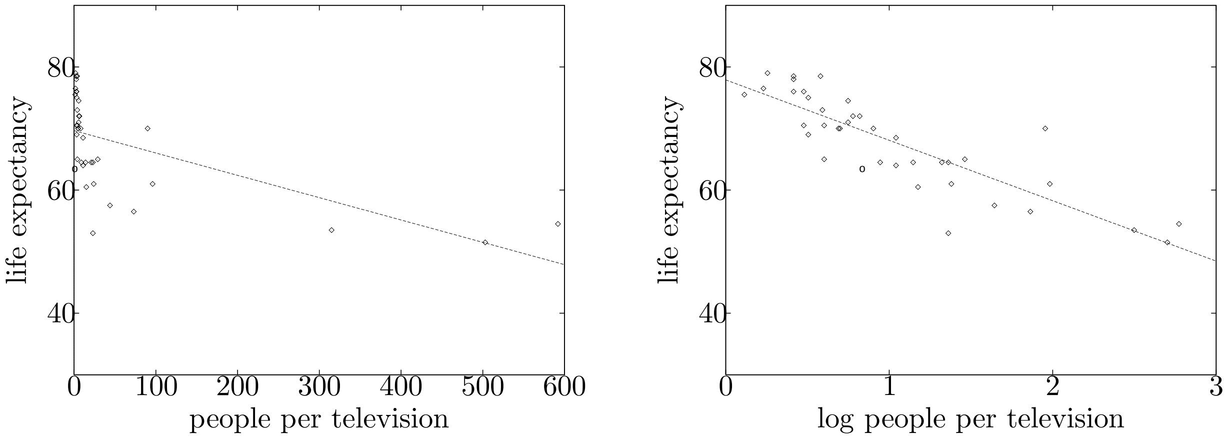
$$y = \hat{a} + \hat{\beta}x.$$

Note that the point $(\bar{x}, \bar{Y})$ always lies on the regression line, i.e., $\bar{Y} = \hat{a} + \hat{\beta}\bar{x}$.

Example 13.2 *The following data for 40 nations has been extracted from a 1993 almanac.*

The correlations of life expectancy with people/television and people/doctor are −0.606 and −0.666 respectively. Scatter plots suggests that a better fit might be obtained by a regression of life expectancy on either the logarithm of people/television or logarithm of people/doctor. When this is done the correlations are respectively −0.855 and −0.832.

country	mean life expectancy, y	people per television, u	people per doctor, v
Argentina	70.5	4.0	370
Bangladesh	53.5	315.0	6166
Brazil	65.0	4.0	684
⋮			⋮
United Kingdom	76.0	3.0	611
United States	75.5	1.3	404
Venezuela	74.5	5.6	576
Vietnam	65.0	29.0	3096
Zaire	54.0	*	23193



Let $x_i = \log_{10} u_i$ and consider fitting a regression of y against x. There is data for 38 countries (as television data for Zaire and Tanzania is missing). We compute the following summary statistics

$$\bar{y} = 67.76, \quad \bar{x} = 1.0322, \quad S_{yy} = 2252.37, \quad S_{xx} = 17.120, \quad S_{xy} = -167.917.$$

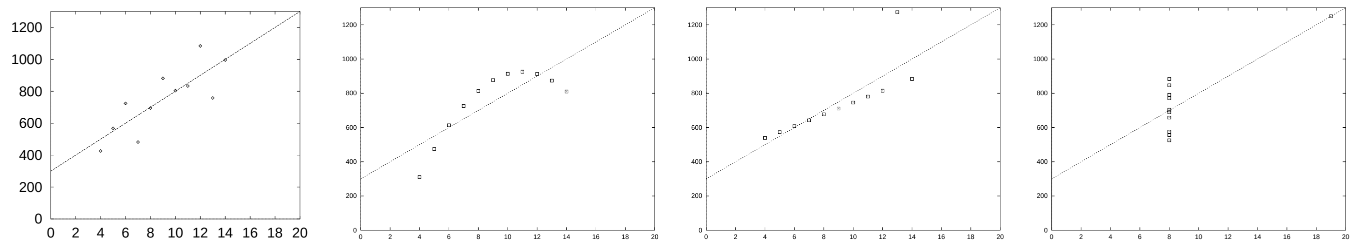
These give

$$\hat{\beta} = S_{xy}/S_{xx} = -9.808, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 77.887, \quad r = S_{xy}/(S_{xx}S_{yy})^{\frac{1}{2}} = -0.855.$$

Although people/television appears to be a useful predictor of a country's life expectancy, we don't really expect that sending shiploads of televisions to countries with short life expectancies would cause their people to live longer. This points up the obvious, but sometimes forgotten fact, that there may be correlation between two variables without causation.

13.4 Data sets with the same summary statistics

Tha application of regression analysis requires care. The following data sets have nearly the same value of the sufficient statistics $\bar{x}, \bar{y}, S_{xx}$ and S_{xy}. The regression line is about $y = 300 + 50\,x$ in each case. However, a simple linear regression is only appropriate in the first case. In the second case a quadratic would be more appropriate. The third case is affected by the presence of an **outlier** and the fourth case is really no more than a straight line fit through 2 points. The lesson is: plot the data!



10	804	8	695	13	758	9	881	11	833	14	996	6	724	4	426	12	1084	7	482	5	568
10	914	8	814	13	874	9	877	11	926	14	810	6	613	4	310	12	913	7	726	5	474
10	746	8	677	13	1274	9	711	11	781	14	884	6	608	4	539	12	815	7	642	5	573
8	658	8	576	8	771	8	884	8	847	8	704	8	525	19	1250	8	556	8	791	8	689

13.5 Other aspects of least squares

Least squares can be used to fit other models. For example, to fit a **regression through the origin** we would minimize

$$S = \textstyle\sum_i (Y_i - \beta x_i)^2$$

and get $\hat{\beta} = \sum_i x_i Y_i / \sum_i x_i^2$. To fit a multiple regrssion model, such as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \,,$$

in which y is predicted from p variables, $x_1, \ldots, x_p$, we would minimize

$$\textstyle\sum_i (Y_i - \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2$$

with respect to $\beta_0, \ldots, \beta_p$. Stationarity conditions give $p+1$ simultaneous linear equations in $\hat{\beta}_0, \ldots, \hat{\beta}_p$, which can be solved for these estimators.

Least squares estimators have many nice properties, one of which is that they are **best linear unbiased estimators**. Consider simple linear regression, in which $Y_i = a + \beta x_i + \epsilon_i$, where the ϵ_i are independent with common mean 0 and variance σ^2 (but are now not necessarily normal RVs). Suppose we want to estimate β by a linear function of the observations, i.e., $\hat{\beta} = \sum_i c_i Y_i$. If this estimator is to be unbiased then we need

$$\mathbb{E}\hat{\beta} = \textstyle\sum_i c_i (a + \beta x_i) = \beta, \quad \text{for all } a, \beta.$$

Hence we need $\sum_i c_i = 0$ and $\sum_i c_i x_i = 1$. Now the variance of the estimator is

$$\operatorname{var}(\hat{\beta}) = \operatorname{var}\left(\textstyle\sum_i c_i Y_i\right) = \sigma^2 \textstyle\sum_i c_i^2.$$

So we have the constrained optimization problem:

$$\text{minimize } \textstyle\sum_i c_i^2 \text{ subject to } \textstyle\sum_i c_i = 0 \text{ and } \textstyle\sum_i c_i x_i = 1.$$

Using Lagrangian methods it is easy to find the solution: $c_i = (x_i - \bar{x})/S_{xx}$. This gives the usual LSE $\hat{\beta} = S_{xy}/S_{xx}$. A similar analysis shows that the best linear unbiased estimator of a is also the usual LSE of a, i.e., $\hat{a} = \bar{Y} - \hat{\beta}\bar{x}$.

14 Hypothesis tests in regression models

Statisticians do it with a little deviance.

14.1 Distributions of the least squares estimators

Suppose as before that $Y_1, \dots, Y_n$ are independent and $Y_i = \alpha + \beta w_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$, equivalently $Y_i \sim N(\alpha + \beta w_i, \sigma^2)$, and where α, β and σ^2 are unknown parameters and the w_i are known constants such that $\sum_i w_i = 0$.

Theorem 14.1

(i) $\hat{\alpha} = \bar{Y}$ *is distributed as* $N(\alpha, \sigma^2/n)$*;*

(ii) $\hat{\beta}$ *is distributed as* $N\big(\beta, (\mathbf{w}^\top \mathbf{w})^{-1}\sigma^2\big)$ *independently of* $\hat{\alpha}$*;*

(iii) the **residual sum of squares** R*, the minimised value of* S*, is distributed as* $\sigma^2 \chi^2_{n-2}$ *independently of* $\hat{\alpha}$ *and* $\hat{\beta}$*, and is equal to*

$$R = \sum Y_j^2 - n\bar{Y}^2 - (\mathbf{w}^\top \mathbf{w})\hat{\beta}^2;$$

(iv) $\hat{\sigma}^2 = R/(n-2)$ *is an unbiased estimator of* σ^2*.*

Proof. Let

$$A = \begin{pmatrix} 1/\sqrt{n} & \cdots & 1/\sqrt{n} \\ (\mathbf{w}^\top \mathbf{w})^{-1/2} w_1 & \cdots & (\mathbf{w}^\top \mathbf{w})^{-1/2} w_n \\ \vdots & & \vdots \\ \cdot & \cdots & \cdot \end{pmatrix}$$

be an orthogonal matrix by appropriate choice of rows $3, \dots, n$. Then $Z_1, \dots, Z_n$ are independent with $\mathbf{Z} = A\mathbf{Y} \sim N\big(A(\alpha\mathbf{1} + \beta\mathbf{w}), \sigma^2 \mathbf{I}\big)$, and

$$\begin{array}{ccccc} Z_1 & = & \sqrt{n}\hat{\alpha} & \sim & N\big(\sqrt{n}\alpha, \sigma^2\big) \\ Z_2 & = & (\mathbf{w}^\top \mathbf{w})^{1/2}\hat{\beta} & \sim & N\big((\mathbf{w}^\top \mathbf{w})^{1/2}\beta, \sigma^2\big) \\ Z_3 & = & \cdot & \sim & N\big(0, \sigma^2\big) \\ \vdots & & & & \vdots \\ Z_n & = & \cdot & \sim & N\big(0, \sigma^2\big) \end{array}$$

from which all the statements in (i) and (ii) follow.

(iii) and (iv) follow from

$$\sum_{i=1}^n Z_i^2 = n\bar{Y}^2 + (\mathbf{w}^\top \mathbf{w})\hat{\beta}^2 + \sum_{i=3}^n Z_i^2$$

and

$$\begin{aligned}\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} Y_i^2 & \\ &= \|(Y - \hat{\alpha}\mathbf{1} - \hat{\beta}\mathbf{w}) + \hat{\alpha}\mathbf{1} + \hat{\beta}\mathbf{w}\|^2 \\ &= \|Y - \hat{\alpha}\mathbf{1} - \hat{\beta}\mathbf{w}\|^2 + n\hat{\alpha}^2 + \hat{\beta}^2\|\mathbf{w}\|^2 \\ &\qquad \text{(since all cross-product terms vanish)} \\ &= R + n\bar{Y}^2 + (\mathbf{w}^\top \mathbf{w})\hat{\beta}^2\end{aligned}$$

So $R = \sum_3^n Z_i^2 \sim \sigma^2\chi^2_{n-2}$ and is independent of Z_1 and Z_2, i.e., of $\hat{\alpha}$ and $\hat{\beta}$. ■

14.2 Tests and confidence intervals

(a) A t-statistic may be constructed for testing the hypothesis that β takes a particular value β_0, since if β_0 is the true value of β:

$$T_0 = \frac{(\hat{\beta} - \beta_0)\sqrt{\mathbf{w}^\top \mathbf{w}}}{\sqrt{R/(n-2)}} = \frac{(\hat{\beta} - \beta_0)\sqrt{\mathbf{w}^\top \mathbf{w}}}{\hat{\sigma}} \sim t_{n-2}.$$

Therefore, to test $H_0 : \beta = \beta_0$ against $H_1 : \beta \neq \beta_0$, we compute t_0 and reject H_0 in a test of size α if $t_0 > t_{\alpha/2}^{(n-2)}$ or $t_0 < -t_{\alpha/2}^{(n-2)}$.

(b) A $(1-\alpha)100\%$ confidence interval may be found for β. Starting from the distributional result in (a) above, we find similarly as in Section 11.1,

$$P\left(\hat{\beta} - t_{\alpha/2}^{(n-2)}\hat{\sigma}/\sqrt{\mathbf{w}^\top \mathbf{w}} < \beta < \hat{\beta} + t_{\alpha/2}^{(n-2)}\hat{\sigma}/\sqrt{\mathbf{w}^\top \mathbf{w}}\right) = 1 - \alpha.$$

(c) We predict the value of Y that would be observed at a given w_0 by $\hat{Y} = \hat{\alpha} + \hat{\beta}w_0$. Then $Y - \hat{Y} \sim N\big(0, \sigma^2(1 + 1/n + w_0^2(\mathbf{w}^2\mathbf{w})^{-1})\big)$. Hence a $(1-\alpha)100\%$ **predictive confidence interval** for Y at w_0 is

$$\left[\hat{Y} - t_{\alpha/2}^{(n-2)}\hat{\sigma}\sqrt{1 + 1/n + w_0^2(\mathbf{w}^2\mathbf{w})^{-1}}\,, \,\hat{Y} + t_{\alpha/2}^{(n-2)}\hat{\sigma}\sqrt{1 + 1/n + w_0^2(\mathbf{w}^2\mathbf{w})^{-1}}\right]$$

14.3 The correlation coefficient

The **sample correlation coefficient** of x and y is defined as $r = S_{xy}/(S_{xx}S_{yy})^{\frac{1}{2}}$. Suppose $Y_i \sim N(a + \beta x_i, \sigma^2)$, independently for each i. The hypothesis that Y does not vary with x is $H_0 : \beta = 0$. The test statistic in (a) can be rewritten as follows:

$$t_0 = \frac{\hat{\beta}\sqrt{S_{xx}}}{\sqrt{R/(n-2)}} = \frac{(S_{xy}/S_{xx})\sqrt{S_{xx}}\sqrt{(n-2)}}{\sqrt{S_{yy} - S_{xy}^2/S_{xx}}} = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}}\,,$$

and so H_0 should be rejected if r^2 is near 1.

Note that the variation in the data is $S_{yy} = \sum_j (y_j - \bar{y})^2$. The regression model 'explains' variation of $\sum_j (\hat{y}_j - \bar{y})^2$ where $\hat{y}_i = \hat{a} + \hat{\beta} x_i$. One can check that $\sum_j (\hat{y}_j - \bar{y})^2 = S_{xy}^2 / S_{xx}$ and so the ratio of these is r^2. We say that 'the regression explains $100r^2$% of the variation in the data'.

14.4 Testing linearity

If we are fitting a linear regression, how can we know how good the fit of the estimated regression line is? In general we cannot be sure: a bad fit could quite well be caused by a large value of σ^2, which is unknown. We can test linearity if we are able to replicate the readings, so that, say, we take m readings at each value x_i, and get,

$$Y_{ij} = a + \beta x_i + \epsilon_{ij}, \quad j = 1, \dots, m,$$

for each $i = 1, \dots, n$. Then averaging over j for fixed i we have

$$\bar{Y}_i = a + \beta x_i + \eta_i = \alpha + \beta(x_i - \bar{x}) + \eta_i$$

where the η_i are IID $N(0, \sigma^2/m)$, independently of $\sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2$, which are IID $\sigma^2 \chi^2_{m-1}$. Now, if we do a linear regression of $\bar{Y}_i$ on x_i, the residual sum of squares is

$$\sum_{i=1}^n \big(\bar{Y}_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})\big)^2 \sim \frac{\sigma^2}{m} \chi^2_{n-2},$$

if the means are indeed linearly related. Thus to test linearity we consider

$$F = \frac{m \sum_{i=1}^n \big(\bar{Y}_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})\big)^2 / (n-2)}{\sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2 / n(m-1)} \sim F_{n-2, n(m-1)},$$

if the model of linearity holds. We reject the hypothesis if $f > F_\alpha^{(n-2), n(m-1)}$.

14.5 Analysis of variance in regression models

Example 14.2 *This tables shows other data for male fruitflies.*

Groups of 25 males kept with	mean life (days)	s.e.	length (mm)	s.e.	sleep (%/day)	s.e.
no companions	63.56	16.4522	0.8360	0.084261	21.56	12.4569
1 uninterested female	64.80	15.6525	0.8256	0.069886	24.08	16.6881
1 interested female	56.76	14.9284	0.8376	0.070550	25.76	18.4465
8 uninterested females	63.36	14.5398	0.8056	0.081552	25.16	19.8257
8 interested females	38.72	12.1021	0.8000	0.078316	20.76	10.7443

'Length' is the length of the fruitfly's thorax. It turns out that longevity (y) is positively correlated to thorax size (x) (as plots of the data show).

Suppose we consider only the data for rows 2 and 3 and adopt a model that for $i = 2, 3$,

$$y_{ij} = a_i + \beta x_{ij}, \quad j = 1 \ldots , 25.$$

Let $\bar{a} = \frac{1}{2}(a_2 + a_3)$. Our model 'explains' the observed variation in longevity within group i in terms of the sum of two effects: firstly, an effect due to thorax size, $\bar{a} + \beta x_{ij}$; secondly, an effect specific to group i, $a_i - \bar{a}$. We would like to test

$$H_0 : a_2 = a_3 \quad \text{against} \quad H_1 : a_2 \neq a_3 \,.$$

To do this we need to fit the appropriate regression models under the two hypotheses by minimizing the residual sum of squares

$$S = \sum_{j=1}^{25} (y_{2j} - a_2 - \beta x_{2j})^2 + \sum_{j=1}^{25} (y_{3j} - a_3 - \beta x_{3j})^2 \,.$$

Under H_1 we minimize freely over a_2, a_3, β and get $\hat{a}_2 = -46.04$, $\hat{a}_3 = -55.69$, $\hat{\beta} = 134.25$, with residual sum of squares $R_1 = 6962.90$.

Under H_0 we minimize subject to $a_2 = a_3$ and get $\hat{a}_2 = \hat{a}_3 = -45.82$, $\hat{\beta} = 128.18$, with residual sum of squares $R_0 = 8118.39$. We can write

$$R_0 = (R_0 - R_1) + R_1.$$

The degrees of freedom of H_0 and H_1 are 2 and 3 respectively. It can be shown that $R_1 \sim \sigma^2 \chi^2_{50-3}$, whether or not H_0 is true. Also R_1 and $R_0 - R_1$ are independent. If H_0 is true, then $R_0 - R_1 \sim \sigma^2 \chi^2_{3-2}$. If H_0 is not true then $R_0 - R_1$ is inflated.

As we have done previously for ANOVA in Section 12.4, we compute an F statistic

$$f = \frac{(R_0 - R_1)/(3-2)}{R_1/(50-3)} = 7.80,$$

which upon comparison to $F_{0.05}^{(1,47)} = 4.21$ leads us to reject H_0; there is indeed a significant difference between the longevities of the males in the two groups. This is the opposite to what we found with a t-test for equality of means in Example 11.3. The explanation is that the mean thorax size happens to be greater within the group of the males exposed to interested females. This is usually associated with greater longevity. When we take into account the fact that this group did not show the greater longevity that would be appropriate to its greater mean thorax size then we do find a difference in longevities between males in this group and those in the group that were kept with a nonreceptive female.

Thus we see that the analysis in Example 11.3 was deficient. There is a lesson in this example, which might be compared to that in Simpson's paradox.

15 Computational methods

Computers have freed statisticians from the grip of mathematical tractability.

15.1 Analysis of residuals from a regression

Lacking powerful computers, statisticians could once only analyse data in ways that were not computationally too difficult, or using pre-calculated tables. Modern computing power, high resolution screens and computational packages such as MINITAB and SPLUS make it easy to display and analyse data in useful ways.

In our regression models we hypothesised that errors are IID $N(0, \sigma^2)$. It is worth checking this by an analysis of the residuals. We estimate the errors by

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{a} - \hat{\beta} x_i\,.$$

It can be shown by a calculation, which we omit here, that

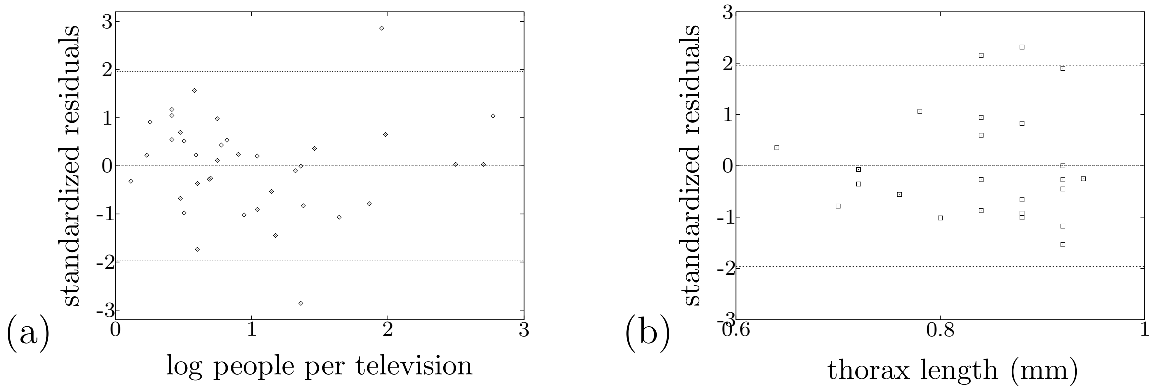
$$\mathrm{var}(Y_i - \hat{Y}_i) = \big(1 - 1/n - (x_i - \bar{x})^2/S_{xx}\big)\sigma^2.$$

Recall that an estimate of σ^2 is obtained from the residual sum of squares, R, as $\hat{\sigma}^2 = R/(n-2)$. So we can calculate **standardized residuals**,

$$\hat{\epsilon}_{\mathrm{s},i} = \frac{Y_i - \hat{Y}_i}{\hat{\sigma}\sqrt{\mathrm{var}(Y_i - \hat{Y}_i)/\sigma^2}} = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - 1/n - (x_i - \bar{x})^2/S_{xx}}}\,,$$

These should be distributed approximately as $N(0, 1)$.

Example 15.1 *Standardized residuals for (a) life expectancy on log people per television, and (b) fruitfly longevity on thorax length (for the 25 kept with no companions).*



We draw lines at ±1.96, the values between which samples from a $N(0, 1)$ will lie 95% of the time. In (a) the pattern of residuals is consistent with samples from $N(0, 1)$. In (b) it looks as though the magnitude of the errors might be increasing with thorax length. This is known as 'heteroscedasticity'. Perhaps a better model would be $\epsilon_i \sim N(0, \sigma^2 x_i)$. This would suggest we try fitting, with $\eta_i \sim N(0, \sigma^2)$:

$$y_i/\sqrt{x_i} = a/\sqrt{x_i} + \beta\sqrt{x_i} + \eta_i\,.$$

15.2 Discriminant analysis

A technique which would be impossible in practice without computer assistance is the credit-scoring used by banks and others to screen potential customers.

Suppose a set of individuals $\{1, 2, \ldots, n\}$ can be divided into two disjoint sets, A and B, of sizes n_A and n_B respectively. Those in set A are known good credit risks and those in set B are known bad credit risks. For each individual we have measured p variables which we believe to be related to credit risk. These might be years at present address, annual income, age, etc. For the ith individual these are $x_{i1}, \ldots, x_{ip}$. The question is: given measurements for a new individual, say $x_{01}, \ldots, x_{0p}$, is that individual more likely to be a good or bad credit risk? Is he more similar to the people in group A or to those in group B?

One approach to this problem is to use least squares to fit a model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

where y_i is defined to be 1 or -1 as $i \in A$ or $i \in B$. Then the 'discriminant function'

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p}$$

is used to classify the new individual as being in group A or group B as $\hat{y}_0$ is closer to $(1/n_A) \sum_{i \in A} \hat{y}_i$ or to $(1/n_B) \sum_{i \in B} \hat{y}_i$. We do not go any further with the theory here. The point is that this is a practically important application of statistics, but a lot of calculation is required to find the discriminant function. Of course a mail order company will experiment with building its discriminant function upon different variables and doing this research is also computer-intensive.

Other uses of discriminant analysis, (and related ideas of 'cluster analysis' when there are more than two groups), include algorithms used in speech recognition and in finance to pick investments for a portfolio.

15.3 Principal components / factor analysis

Suppose we have measured a large number of variables, say $x_{i1}, \ldots, x_{ip}$, for individuals, $i = 1, \ldots, n$. Maybe these are answers to p questions on a psychological test, such as the Myers–Briggs. The question is: can we find a much smaller number of variables (factors) which explain most of the variation? In the Myers–Briggs test subjects answer a large number of questions of the sort 'when the telephone rings, are you pleased?' and the answers are converted to scores on 4 factors measuring strengths of extroversion, intuition, thinking and judging. How might these four factors have been identified from the data?

To keep things simple, we explain an approach via 'principal components analysis'. True factor analysis involves some further ideas that we skip over here. We begin by finding that linear function of the variables with the greatest variance, i.e.,

$$\text{maximize } \sum_{i=1}^{n}\big[(\beta_1 x_{i1}+\cdots+\beta_p x_{ip})-(\beta_1\bar{x}_1+\cdots+\beta_p\bar{x}_p)\big]^2 \quad \text{subject to } \sum_{i=1}^{p}\beta_i^2=1$$

where $\bar{x}_i$ is the mean of the ith variable within the population. Equivalently,

$$\text{maximize } \beta^\top G\beta \quad \text{subject to } \beta^\top\beta=1,$$

where G is the $p\times p$ matrix with $G_{jk}=\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$. By Lagrangian methods we find that the maximum equals the largest eigenvalue of G, say λ_1, and is achieved when β is the corresponding right hand eigenvector, say $\beta^1=(\beta_1^1,\dots,\beta_p^1)^\top$. We call β^1 the 'first principal component'. Similarly, we can find the eigenvector β^2 of G corresponding to the second largest eigenvalue, λ_2. Continuing, we find an orthogonal set of eigenvectors $\beta^1,\dots,\beta^m$, $m<p$, such that the proportion of variance explained, i.e.,

$$\sum_{j=1}^{m}\sum_{i=1}^{n}\big[(\beta_1^j x_{i1}+\cdots+\beta_p^j x_{ip})-(\beta_1^j\bar{x}_1+\cdots+\beta_p^j\bar{x}_p)\big]^2 \Bigg/ \sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2$$

is near 1. This amounts to the same thing as $\sum_{j=1}^{m}\lambda_j/\sum_{j=1}^{p}\lambda_j$; indeed the denominator above is $\text{trace}(G)=\sum_{j=1}^{p}\lambda_j$. The above ratio is also the proportion of variation explained by using least squares to fit

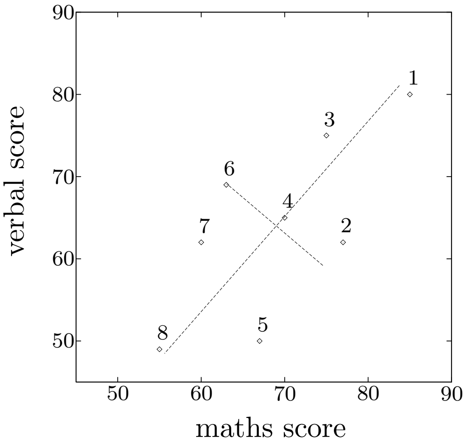
$$x_{ij}=\alpha_1^j z_{i1}+\cdots+\alpha_m^j z_{im}+\epsilon_{ij},$$

when we take $z_{ij}=\beta_1^j x_{i1}+\cdots\beta_p^j x_{ip}$. Here z_{ij} is the 'score of individual i on factor j'.

The final step is to try to give some natural interpretation to the factors, $z_1,\dots,z_m$. For example, if we observe that the components of β^1 which are large in magnitude seem to match up with components of x which have something to do with whether or not an individual is extroverted, and other components of β^1 are near 0, then we might interpret factor 1 as an 'extroversion factor'. Then if z_{i1}, the score of individual i on this factor, is large and positive we could say that i is extroverted, and if large and negative that i is introverted.

To be fair, we should say that things are rarely so simple in practice and that many statisticians are dubious about the value of factor analysis. For one thing, the factors depend on the relative units in which the variables are measured.

Nevertheless, here is a simple illustration for $p=2$, $m=1$, $n=8$. Suppose 8 students are scored on two tests, one consisting of verbal puzzles and the other of maths puzzles; the ith student scores (x_{i1},x_{i2}). The first principal component is a line through the data which minimizes the sum of squared differences between the data points and their orthogonal projections onto this line. A reasonable name for this component might be 'IQ'. The 'IQ' of student i is $z_{i1}=\beta_1^1 x_{i1}+\beta_2^1 x_{i2}$.



student	math score	verbal score	IQ factor	mathmo factor
1	85	80	116.1	12.1
2	77	62	97.2	17.8
3	75	75	105.8	7.8
4	70	65	94.9	10.5
5	67	50	81.6	18.1
6	63	69	93.4	2.6
7	60	62	86.1	4.9
8	55	49	73.0	9.6

15.4 Bootstrap estimators

Suppose students are scored on two tests, and we wish to reduce their scores to single 'IQ' scores. Let $\mathbf{x} = (x_1, \dots, x_8)$ be the vector of test scores, where $x_i = (x_{i1}, x_{i2})$. Define the statistic $t(\mathbf{x}) = \lambda_1(\mathbf{x}) / \sum_{j=1}^2 \lambda_j(\mathbf{x})$, i.e., the proportion of variation that is explained by a single factor corresponding to the first principal component of $G(\mathbf{x})$. Suppose we are interested in $\theta = \mathbb{E}\, t(\mathbf{X})$, a measure of how well we can do on average when using this procedure to summarise 8 pairs of test scores in 8 single 'IQ' scores.

We can estimate θ by $\hat{\theta} = t(\mathbf{x})$. But to assess the accuracy of $\hat{\theta}$ we need to know its variance. This depends on the distribution from which our IID samples $X_1, \dots, X_8$ have been drawn, say F. It is no surprise that there is not a nice formula for the variance of $t(\mathbf{X})$, nor that percentage points of the distribution of $t(\mathbf{X})$ have not been tabulated; (that would require some assumption about F, e.g., that it is bivariate normal).

A modern method of estimating the variance of $\hat{\theta}$ is the **bootstrap estimate**. The idea is to approximate F by the empirical distribution $\hat{F}$, a sample from which is equally likely to take any of the values $x_1, \dots, x_8$. We take a sample of 8 pairs of tests scores from $\hat{F}$; this corresponds to randomly choosing 8 out of the set $\{x_1, \dots, x_8\}$, *with replacement.* Perhaps we get $\mathbf{x}^* = (x_3, x_8, x_1, x_2, x_3, x_3, x_5, x_1)$. From this sample we calculate a value of the estimator, $\hat{\theta}^* = t(\mathbf{x}^*)$. We repeat this procedure B times, to get $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$. Of course we use a computer to do the random sampling, the calculation of G and of λ_1. The bootstrap estimate of the variance of $\hat{\theta} = t(\mathbf{X})$ under F is then the estimate of the variance of $t(\mathbf{X})$ under $\hat{F}$ given by

$$\hat{\sigma}_{\hat{\theta}}^2 = \frac{1}{B-1} \sum_{i=1}^{B} \left(\hat{\theta}_i^* - \frac{1}{B} \sum_{k=1}^{B} \hat{\theta}_k^* \right)^2 .$$

For the data above, $z_1 = 0.653x_1 + 0.757x_2$. The proportion of variation explained is $\hat{\theta} = t(\mathbf{x}) = 0.86$. A bootstrap estimate with $B = 240$ gives $\hat{\sigma}_{\hat{\theta}} = 0.094$.

Formalisation of the bootstrap method dates from 1979; the study of its use for constructing estimators, tests and confidence intervals is an active area of research.

16 Decision theory

To guess is cheap, to guess wrongly is expensive. (Old Chinese proverb)

16.1 The ideas of decision theory

We began the course with the following definition of Statistics:

> *a collection of procedures and principles for gaining and processing information in order to make decisions when faced with uncertainty.*

We have studied various ways to process data and to draw inferences from it: e.g., point estimation, interval estimation, hypothesis testing and regression modelling. There have been some key concepts, such as unbiasedness, the Neyman-Pearson lemma, and the fact that least squares estimators are the best linear unbiased estimators of regression parameters. But there are things that we have done which may seem to have been rather ad hoc, and which beg unanswered questions: e.g., do we always prefer unbiased estimators to biased ones? Do we care about estimators being linear? Sometimes we have have done things simply so that we can get an answer.

Decision theory attempts to provide Statistics with a satisfying foundation by placing everything within a unifying framework. In this framework the act of decision making is made central and ideas of optimality are introduced so that one can properly speak about making the 'best' inference. The conclusions are often the same as those reached by other means, but can also lead in new directions.

The decision theoretic approach begins with a careful definition of all the elements of a decision problem. It is imagined that there is a decision-maker who is to choose an *action* a from a set A. He is to do this based upon observation of a random variable, or *data* X. This X (typically a vector $X_1, \dots, X_n$) has a probability distribution which depends on an unknown parameter θ. Here θ denotes a *state of nature*. The set of all possible values of θ is the *parameter space* Θ.

The decision is to be made by a *statistical decision function* (or rule) d; this is a function which specifies $d(x)$ as the action to be taken when the observed data is $X = x$. On taking action $a = d(X)$ the decision-maker incurs a *loss* of $L(\theta, a)$. A good decision function is one that has a small value of the *risk function*

$$R(\theta, d) = \mathbb{E}\big[L(\theta, d(X))\big]\,,$$

where this expectation is taken over X.

Clearly if $R(\theta, d_1) \le R(\theta, d_2)$ for all θ and $R(\theta, d_1) < R(\theta, d_2)$ for some θ then we would never want to use rule d_2, since d_1 can always do as well and sometimes better. We say d_2 is *inadmissible*.

Decision theory requires several lectures or a whole course to cover fully. Here we just give the flavour of some of the ideas.

Example 16.1 *In* Nature *(29 August, 1996, p. 766) Matthews gives the following table for various outcomes of Meteorological Office forecasts of weather covering* 1000 *one-hour walks in London.*

	Rain	No rain	Sum
Forecast of rain	66	156	222
Forecast of no rain	14	764	778
Sum	80	920	1000

Should one pay any attention to weather forecasts when deciding whether or not to carry an umbrella?

To analyse this question in a decision-theoretic way, let W, F and U be respectively the events that it is going to rain (be wet), that rain has been forecast, and that we carry an umbrella. The possible states of nature are W and W^c. The data is $X = F$ or $X = F^c$. Possible actions are chosen from the set $A = \{U, U^c\}$. We might present the loss function as

	W^c	W
U^c	L_{00}	L_{01}
U	L_{10}	L_{11}

For example, we might take $L_{01} = 4$, $L_{11} = 2$, $L_{10} = 1$, $L_{00} = 0$. Of course these are subjective choices, but most people would probably rank the four outcomes this way.

One possible decision function is given by $d_1(X) = U^c$, i.e., never carry an umbrella. It's risk function is

$$R(W^c, d_1) = L_{00}; \quad R(W, d_1) = L_{01}\,.$$

Another possible decision function is given by $d_2(F) = U$ and $d_2(F^c) = U^c$, i.e., carry an umbrella if and only if rain is forecast. The risk function is

$$R(W^c, d_2) = (764/920)L_{00} + (156/920)L_{10}; \quad R(W, d_2) = (66/80)L_{11} + (14/80)L_{01}\,.$$

We see that if $\theta = W^c$ then d_1 is better, but if $\theta = W$ then d_2 is better. Thus neither rule is uniformly better for both states of nature. Both d_1 and d_2 are admissible. By averaging over the states of nature we have the so-called *Bayes risk*, defined as

$$B(d) = \mathbb{E}[R(\theta, d)],$$

where the expected value is now taken over θ. For example, in our problem, $\mathbb{P}(W) = 0.08$ and $\mathbb{P}(W^c) = 0.92$, so $B(d) = 0.08R(W, d) + 0.92R(W^c, d)$.

The *Bayes rule* is defined as the rule d which minimizes the Bayes risk. Thus to find the Bayes rule for our problem, we must compare

$$B(d_1) = .08L_{01} + .92L_{00}$$

to

$$\begin{aligned} B(d_2) &= .08\big[(66/80)L_{11} + (14/80)L_{01}\big] + .92\big[(764/920)L_{00} + (156/920)L_{10}\big] \\ &= .066L_{11} + .014L_{01} + .764L_{00} + .156L_{10}. \end{aligned}$$

It follows that it is better to ignore weather forecasts and simply go for walks without an umbrella, if

$$B(d_1) < B(d_2) \iff \Delta := \frac{L_{01} - L_{11}}{L_{10} - L_{00}} < \frac{.156}{.066} = 2.364,$$

which can hold for reasonable values of the loss function, such as those given above, for which $\Delta = 2$. It all depends how you feel about getting wet versus the inconvenience of carrying an umbrella. Similar analysis shows that the commonly followed rule of always carrying an umbrella is better than doing so only if rain is forecast only if one is very adverse to getting wet, i.e., if $\Delta > 764/14 \doteqdot 53$.

16.2 Posterior analysis

In Lecture 5 we considered a decision theoretic approach to the point estimation problem. We used a loss function $L(\theta, a)$ to measure the loss incurred by estimating the value of a parameter to be a when its true value is θ. Then $\hat{\theta}$ was chosen to minimize $\mathbb{E}[L(\theta, \hat{\theta})]$, where this expectation is over θ with respect to the posterior distribution $p(\theta \mid x)$.

Another way to think about the decision problem above is similar. We consider the expected loss under the posterior distribution. The posterior distribution for rain, given the data that there has been a forecast of rain, is $\mathbb{P}(W \mid F) = 66/222 \doteqdot 0.30$. (Note that this is less than 0.50!) Hence, given a forecast of rain, the expected loss if we carry an umbrella is

$$B(U \mid F) = (66/222)L_{11} + (156/222)L_{10},$$

whereas if we don't carry an umbrella the expected loss is

$$B(U^c \mid F) = (66/222)L_{01} + (156/222)L_{00}.$$

Not surprisingly, this leads to exactly the same criterion for choosing between d_1 and d_2 as we have already found above.

This is a general principle: the Bayes rule, d, can be determined as the action a which minimizes $\mathbb{E}_{\theta|X}[R(\theta, a)]$, this expectation being taken over θ with respect to the posterior distribution $p(\theta \mid x)$.

16.3 Hypothesis testing as decision making

We conclude by elucidating a decision theoretic approach to hypothesis testing. Consider the problem of testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$. On the basis of an observation X we must decide in favour of H_0 (i.e., take action a_0) or decide in favour of H_1 (i.e., take action a_1).

For the case of so-called *0–1 loss* we take $L(\theta_0, a_0) = L(\theta_1, a_1) = 0$ and $L(\theta_0, a_1) = L(\theta_1, a_0) = 1$. I.e., there is unit loss if and only if we make the wrong decision. The risk function is then simply the probability of making the wrong decision, so $R(\theta_0, d) = \mathbb{P}(d(X) = a_1 \mid H_0)$ and $R(\theta_1, d) = \mathbb{P}(d(X) = a_0 \mid H_1)$.

Suppose we have prior probabilities on H_0 and H_1 of p_0 and p_1 respectively. This gives Bayes risk of

$$B(d) = p_0 R(\theta_0, d) + p_1 R(\theta_1, d).$$

As we have seen in the previous section the Bayes rule minimizes the posterior losses, so we should choose $d(X)$ to be a_1 or a_0 as

$$\frac{B(a_0 \mid x)}{B(a_1 \mid x)} = \frac{\mathbb{P}(H_1 \mid x)}{\mathbb{P}(H_0 \mid x)} = \frac{p_1 \mathbb{P}(x \mid H_1)}{p_0 \mathbb{P}(x \mid H_0)} = \frac{p_1}{p_0} \frac{f(x \mid \theta_1)}{f(x \mid \theta_0)}$$

is greater or less than 1.

This is of course simply a likelihood ratio test. Observe, however, that we have reached this form of test by a rather different route than in Lecture 6.

16.4 The classical and subjective points of view

The decision theoretic approach to statistical inference is appealing for the way it directly addresses issues such as loss, risk, admissibility, etc. These have intuitive interpretations in terms of the economics of decision making.

Decision theory also has the philosophical merit or dismerit, depending on your point of view, that it incorporates the Bayesian notions of prior and posterior beliefs. In the analysis of the hypothesis test above, we had to introduce a prior distribution on H_0 and H_1, as given by the probabilities p_0 and p_1. Some statisticians argue that this is fine; people always come to decision problems armed with prior beliefs, if only an uninformed belief expressed as $p_0 = p_1 = 1/2$. Others take the 'classical' line that statistical procedures should not depend upon the introduction of subjective prior beliefs on the part of the person analysing the data. They argue that only the data should matter: two people should automatically come to the exactly the same conclusion when presented with the same data. Most practising statisticians are happy to take the best of both viewpoints, letting the actual question under consideration decide which concepts and procedures are most helpful.