MATH4065 University of Nottingham

Statistics

S1 Statistical modelling overview

S1.1 What is statistics?

One view: "Statistics is the science of identifying and quantifying structure in a population, apportioning sources of uncertainty, from information contained in a sample".

Another view: "Statistics is the science of collecting, organising, analysing and presenting data".

S1.2 The statistical paradigm

Statistical analysis consists of a number of steps:

Specification of population
Sampling design of a protocol
Data collection
Exploratory analysis
Model specification
Diagnostics (Model verification)
Inference
Interpretation
Communication

The procedure is not usually linear - some iteration may be necessary.

S1.3 Population and sample

From a sample of 52 university students, four individuals were found to be left-handed. We can easily summarise the sample information as the proportion $\frac{4}{52} = \frac{1}{13}$. However, what are we able to say about the population? What proportion of all university students are left-handed? Is $\frac{1}{13}$ a good estimate and what do we mean by 'good'?

We identify the important features of statistical inference in this example. Here, the *population* are all university students. The population has some *parameter* or characteritic, θ, which we wish to estimate. In this example, θ is the probability of an individual being left-handed.

From the population we take a *random sample*, which means each member of the population has an equal chance of being chosen. The sample gives rise to *data* $x_1, x_2, \ldots, x_n$. We estimate the parameter θ by means of a *statistic* $T(x_1, x_2, \ldots, x_n)$.

S1.4 Modelling assumptions

1. *Identically distributed assumption:* Every sample observation (data point) x is the outcome of a random variable X which has an identical distribution (either discrete or continuous) for every member of the population.

2. *Independence assumption:* The random variables $X_1, X_2, \ldots, X_n$ which give rise to the data points $x_1, x_2, \ldots, x_n$ are independent.

Note that we defined a random sample to be a set of i.i.d. random variables.

The subtle point here is that we are treating the observed data as just one possible outcome from the many different outcomes that could occur.

S1.5 Parametric models

In the parametric approach to statistics (inference), we assume that the random sample that we collect was generated by some specific probability distribution which is completely known, except for a small number of parameters. For example,

(a) we could assume that the annual income in the U.K. is normally distributed but we don't know its mean, μ, or its variance, σ^2;

(b) in studying the effectiveness of a certain drug's ability to decrease the size of tumours in laboratory rats, we assume that the outcome of the tumour size being decreased (or not) has a Binomial distribution where n is the known sample size and p is not.

Approaches to determining the underlying model:
1. Physical argument, e.g. counts of events from a Poisson process follow a Poisson distribution.
2. Mathematical argument, e.g. central limit theorem leading to a normal distribution.
3. Flexible model which fairly arbitrarily covers a wide range of possibilities.

S2 Parameter Estimation

S2.1 Preliminaries

Definition A *statistic*, $T(\boldsymbol{X})$, is any function of the random sample.

Note that since $T(\boldsymbol{X})$ is a function of random variables, it is also a random variable. Hence it will also have all the properties of a random variable. Most importantly, it has a distribution associated with it.

Definition A statistic that is used for the purpose of estimating an unknown population parameter is called an *estimator*.

Definition A realised value of an estimator, $T(\boldsymbol{x})$, (i.e. the value of $T(\boldsymbol{X})$ evaluated at a particular outcome of the random sample) is called an *estimate*.

Suppose we want to estimate the average annual income in the U.K. Let $X_1, X_2, \ldots, X_n$ be a random sample of annual incomes. Possible estimators might include:

1. $T_1(\boldsymbol{X}) = \frac{X_1+X_2+\cdots+X_n}{n}$,
2. $T_2(\boldsymbol{X}) = \text{min}(X_1, X_2, \ldots, X_n)$,
3. $T_3(\boldsymbol{X}) = X_1$.

S2.2 Judging estimators

Let θ be a population parameter we wish to estimate. Since any function of the sample data is a potential estimator of θ, how should we determine whether an estimator is good or not? What qualities should our estimator have?

Quality 1: Unbiasedness

Definition The estimator $T(\boldsymbol{X})$ is an *unbiased* estimate of θ if

$$E\left(T(\boldsymbol{X})\right) = \theta.$$

Otherwise, we say that the estimator $T(\boldsymbol{X})$ is biased and we define $B(T) = E\left(T(\boldsymbol{X})\right) - \theta$ to be the *bias* of T.

Definition If $B(T) \to 0$ as the sample size $n \to \infty$, then we say that $T(\boldsymbol{X})$ is *asymptotically unbiased* for θ.

Quality 2: Small variance

Definition If two estimators $T_1(\boldsymbol{X})$ and $T_2(\boldsymbol{X})$ are both unbiased for θ, then $T_1(\boldsymbol{X})$ is said to be *more efficient* than $T_2(\boldsymbol{X})$ if $\text{var}\left(T_1(\boldsymbol{X})\right) < \text{var}\left(T_2(\boldsymbol{X})\right)$.

We would ideally like an estimator that is unbiased with a small variance. So given two unbiased estimators, we choose the most efficient estimator (the estimator with the smallest variance). For comparing an estimator with a biased estimator, we can use the mean-square error to quantify the trade-off between bias and variance:

Definition The *mean-square error* of an estimator is defined by

$$\text{MSE}(T) = E\left[\left(T(\boldsymbol{X}) - \theta\right)^2\right].$$

Exercise: Prove $\text{MSE}(T) = \text{var}(T) + \left(B(T)\right)^2$.

Quality 3: Consistency

Definition An estimator $T(\boldsymbol{X})$ is said to be a *consistent* estimator for θ if

$$T(\boldsymbol{X}) \longrightarrow \theta \text{ as } n \to \infty.$$

This third desirable property can sometimes be established using the following theorem:

Theorem If $E\left[T(\boldsymbol{X})\right] \to \theta$ and Var $(T(\boldsymbol{X})) \to 0$ as $n \to \infty$, then $T(\boldsymbol{X})$ is a consistent estimator for θ.

Note that these are sufficient but not necessary conditions for consistency. Since MSE(T) = var$(T) + (B(T))^2$, then the theorem implies that if MSE$(T) \to 0$ then $T(\boldsymbol{X})$ is a consistent estimator for θ.

S2.3 Example: The sample mean

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from any population with mean μ and variance σ^2. The sample mean is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and is an estimator of μ. What are the properties of $\bar{X}$?

1. Unbiasedness

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n}(X_1 + X_2 + \ldots + X_n)\right] \\ &= \frac{1}{n}\{E[X_1] + E[X_2] + \ldots + E[X_n]\} \\ &= \frac{1}{n}\{\mu + \mu + \ldots + \mu\} = \frac{1}{n}n\mu = \mu. \end{aligned}$$

2. Variance

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

3. Mean-square error

$$\mathsf{MSE}(\bar{X}) = \mathsf{Var}(\bar{X}) + B(\bar{X})^2 = \frac{\sigma^2}{n}.$$

4. Consistency
Since $E[\bar{X}] \to \mu$ and $\mathsf{Var}(\bar{X}) \to 0$ as $n \to \infty$, then $\bar{X}$ is a consistent estimator for μ.

S2.4 Example: The sample variance

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from any population with mean μ and variance σ^2. Consider the estimator

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

Note that

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

Unbiasedness:

$$\begin{aligned}
E[\hat{\sigma}^2] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{1}{n}\sum_{i=1}^{n}(\bar{X} - \mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - \frac{1}{n}\sum_{i=1}^{n} E\left[(\bar{X} - \mu)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathsf{Var}(X_i) - \frac{1}{n}\sum_{i=1}^{n} \mathsf{Var}(\bar{X}) \\
&= \frac{1}{n}n\sigma^2 - \frac{1}{n}n\frac{\sigma^2}{n} = \frac{(n-1)\sigma^2}{n} \neq \sigma^2.
\end{aligned}$$

Hence $\hat{\sigma}^2$ is a biased, although asymptotically unbiased, estimator for σ^2. Therefore,

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

is an unbiased estimator of σ^2.

Useful formula

It can be shown (exercise) that

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

Notation

Given observed data $x_1, x_2, \ldots, x_n$ then we define

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right).$$

Similarly, if we have data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ we define

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(\bar{x} - x_i)(\bar{y} - y_i).$$

s_x^2 is called the **sample variance**;
s_{xy} is called the **sample covariance**.

NOTE that some text books define these quantities without the $1/(n-1)$ term, i.e. they are just the sums.

S3 Techniques for deriving estimators

S3.1 Method of Moments

Definition If $E[X^k]$ exists, then $E[X^k]$ is said to be the kth *moment* of the random variable X.

For example,

$E[X] = \mu$ is the first moment of X.

$E[X^2]$ is the second moment of X.

Var$(X) = E[X^2] - (E[X])^2$ is a function of the first and second moments.

Definition The kth *sample moment* is

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

The idea: Since,

$$E[\hat{\mu}_k] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^k\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i^k\right] = E\left[X_i^k\right],$$

then the kth sample moment is an unbiased estimator of the kth moment of a distribution. Therefore, if one wants to estimate the parameters from a particular distribution, one can write the parameters as a function of the moments of the distribution and then estimate them by their corresponding sample moments.

Example S3.1 Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution with mean μ and variance σ^2. Find the method of moments estimators for μ and σ^2.

$$\begin{aligned}
\hat{\mu} &= \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \\
\hat{\sigma}^2 &= \hat{\mu}_2 - (\hat{\mu}_1)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.
\end{aligned}$$

Example S3.2 Let $X_1, X_2, \ldots, X_n \sim \text{Bin}(m, \theta)$ (m known). Find the method of moments estimator for θ.

The first moment of the Binomial distribution is $m\theta$. Therefore,

$$\hat{\theta} = \frac{\hat{\mu}_1}{m} = \frac{\bar{X}}{m}.$$

Example S3.3 Let $X_1, X_2, \ldots, X_n \sim \text{Exp}(\theta)$. Find the method of moments estimator for θ.

For $x > 0$, $\theta > 0$, $f(x|\theta) = \theta e^{-\theta x}$. Then $E(X) = 1/\theta$, so $1/\hat{\theta} = \bar{X}$ and

$$\hat{\theta} = 1/\bar{X}.$$

Remarks

The sampling properties of the kth sample moment are fairly desirable:

1. $\hat{\mu}_k$ is an unbiased estimator of $E[X^k]$.
2. By the Central Limit Theorem, $\hat{\mu}_k$ is asymptotically normal.
3. $\hat{\mu}_k$ is a consistent estimator of $E[X^k]$.

If h is a continuous function, then $\hat{\theta} = h(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k)$ is a consistent estimator of $\theta = h(\mu_1, \mu_2, \ldots, \mu_k)$, but it may not be an unbiased or asymptotically normal estimator.

Finding theoretical moments as a function of θ is not always simple.

For some models, moments may not exist.

S3.2 Maximum likelihood estimation

In the study of probability, for random variables $X_1, X_2, \ldots, X_n$ we consider the joint p.m.f. or p.d.f. just a function of the random variables $X_1, X_2, \ldots, X_n$. We assume the parameter value(s) are completely known.

For example, if $X_1, X_2, \ldots, X_n$ is a random sample from a Poisson distribution with mean λ, then

$$p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

for $\lambda > 0$. See section P2.3.

However, in the study of statistics, we assume the parameter values are unknown. Therefore, if we are given a specific random sample $x_1, x_2, \ldots, x_n$, then $p(x_1, x_2, \ldots, x_n)$ will take on different values for each possible value of the parameters (λ in the Poisson example). Hence, we can consider $p(x_1, x_2, \ldots, x_n)$ to also be a function of the unknown parameter and write $p(x_1, x_2, \ldots, x_n|\lambda)$. We want to choose $\hat{\lambda}$ to be the value of λ which most likely produced the random sample $x_1, x_2, \ldots, x_n$, i.e. the value of λ which maximises $p(x_1, x_2, \ldots, x_n)$.

Example S3.4 Suppose we collect a random sample from a

Poisson distribution such that $X_1 = 1$, $X_2 = 2$, $X_3 = 3$ and $X_4 = 4$. Find the maximum likelihood estimator of λ.

The likelihood function is

$$L(\lambda) = p(x_1, x_2, x_3, x_4|\lambda) = p(1, 2, 3, 4|\lambda) = \frac{e^{-4\lambda}\lambda^{10}}{1!2!3!4!}.$$

Since $\log x$ is a monotonic increasing function, if we maximise $\log L(\lambda)$ this is equivalent to maximising $L(\lambda)$. Hence,

$$\log L(\lambda) = -4\lambda + 10 \log \lambda - \log(1!2!3!4!).$$

To maximise this function we solve

$$\frac{d \log L(\lambda)}{d\lambda} = 0.$$

Now, $\frac{d \log L(\lambda)}{d\lambda} = -4 + \frac{10}{\lambda} = 0$. Hence, $\hat{\lambda} = 5/2$.

Definition The *likelihood function* of the random variables $X_1, X_2, \ldots, X_n$ is the joint p.m.f. (discrete case) or joint p.d.f. (continuous case) of the observed data given the parameter θ. i.e. $L(\theta) = f(x_1, x_2, \ldots, x_n|\theta)$.

Note that if $X_1, X_2, \ldots, X_n$ are a random sample from a distribution with probability function $f(x|\theta)$ then

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

Definition The *maximum likelihood estimator* (m.l.e. or MLE) of θ is the value of θ which maximises $L(\theta)$.

Definition If $L(\theta)$ is the likelihood function of θ, then $l(\theta) = \log L(\theta)$ is called the *log likelihood function* of θ.

Example S3.5 Let $X \sim \text{Bin}(m, \theta)$. Find the MLE of θ.

$$L(\theta) = \binom{m}{x}\theta^x(1-\theta)^{m-x}, \quad 0 \leq \theta \leq 1.$$

Take the derivative of $L(\theta)$,

$$\frac{dL(\theta)}{d\theta} = \binom{m}{x}\theta^{x-1}(1-\theta)^{m-x-1}\left[x(1-\theta)-(m-x)\theta\right]$$

Setting $\frac{dL(\theta)}{d\theta} = 0$, we obtain

$$[x(1-\theta)-(m-x)\theta] = 0$$

Hence, $\hat{\theta} = \frac{x}{m}$ is a possible value for the MLE of θ.

Since $L(\theta)$ is a continuous function over $[0,1]$, the maximum must exist at either the stationary point or at one of the endpoints of the interval. Given, $L(0) = 0$, $L(1) = 0$, and $L\left(\frac{x}{m}\right) > 0$, $\hat{\theta} = \frac{x}{m}$ is the MLE of θ.

Example S3.6 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with mean λ. Find the MLE of λ.

$$L(\lambda) \;=\; p(x_1, x_2, \ldots, x_n|\lambda) = \frac{\mathrm{e}^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!},$$

where $\lambda > 0$. So,

$$l(\lambda) \;=\; -n\lambda + \sum_{i=1}^n x_i \log\lambda - \log\prod_{i=1}^n x_i!.$$

Now

$$\begin{aligned}\frac{dl(\lambda)}{d\lambda} &= -n + \frac{\sum_{i=1}^n x_i}{\lambda} = 0,\\ \text{so}\quad \hat{\lambda} &= \frac{\sum_{i=1}^n x_i}{n} = \bar{x}.\end{aligned}$$

Since $\frac{d^2l(\lambda)}{d\lambda^2} = \frac{-\sum_{i=1}^n x_i}{\lambda^2} < 0$, $\hat{\lambda} = \bar{X}$ is the MLE of λ.

S3.3 Maximum likelihood estimation - some comments

1. When finding the MLE, remember that you want to maximise the likelihood function. It may be more convenient to maximise the log likelihood function instead.

2. MLEs may not exist, and if they do, they may not be unique.

3. The likelihood function is NOT the probability distribution for θ. We assume θ is an unknown constant, not a random variable. In Bayesian statistics we will consider the parameter to be random.

4. The MLE has some nice large sample properties, including consistency, asymptotic normality and other optimality properties

5. The MLE can be used for non-independent data or non-identically distributed data as well.

6. Often the MLE cannot be found using calculus techniques and must be found numerically.

7. The MLE satisfies a useful invariance property. Namely, if $\phi = h(\theta)$, where $h(\theta)$ is a one-to-one function of θ, then the MLE of ϕ is given by $\hat{\phi} = h(\hat{\theta})$. For example, if $\phi = \frac{1}{\theta}$ and $\hat{\theta} = \bar{X}$ then $\hat{\phi} = \frac{1}{\hat{\theta}} = \frac{1}{\bar{X}}$.

S3.4 Further examples

Example S3.7 Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of $N(\theta, 1)$. Find the MLE of θ.

For each of the X_i

$$f(x_i|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x_i - \theta)^2\right\}.$$

Thus:

$$L(\theta) \propto \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}(x_i - \theta)^2\right\} = \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right\}$$

and

$$l(\theta) = \log L(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2 + \text{constant}.$$

Hence

$$\frac{dl(\theta)}{d\theta} = \sum_{i=1}^{n}(x_i - \theta) = 0$$

for a stationary point. Hence,

$$\hat{\theta} = \frac{\sum x_i}{n} = \bar{x},$$

which is verified as a maximum since

$$\frac{d^2 l(\theta)}{d\theta^2} = -n < 0.$$

Example S3.8 Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples of $U[0, \theta]$. Find the MLE of θ.

If $X_i \sim U[0, \theta]$, then its p.d.f. is given by

$$f(x|\theta) = \left\{\begin{array}{ll} \frac{1}{\theta} & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise} \end{array}\right. .$$

Therefore, if $0 \le x_i \le \theta$ for all $i = 1, \ldots, n$, then

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

Hence, $L(\theta)$ is a decreasing function of θ and its maximum must exist at the smallest value that θ can obtain. Since $\theta > \max\{x_1, x_2, \ldots, x_n\}$, the MLE of θ is $\hat{\theta} = \max\{x_1, x_2, \ldots, x_n\}$.

S4 Additional properties of estimators

S4.1 Sufficiency

Definition Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. A statistic $T(\boldsymbol{X}) = T$ is sufficient for θ if the conditional distribution of $\boldsymbol{X}|T$ does not depend on θ, i.e.

$$f(x_1, x_2, \ldots, x_n | T = t, \theta) = u(\boldsymbol{x}|t),$$

where u is a function of $\boldsymbol{x}$ only. Thus, T contains all the information about θ.

Example S4.1 Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from a Poisson distribution with parameter λ. Determine whether $\bar{X}$ is a sufficient statistic for λ.

We need to show $p(x_1, x_2, \ldots, x_n | T = t, \lambda)$ does not depend on λ.

$$\begin{aligned} p(x_1, \ldots, x_n | T = t, \lambda) &= \frac{P(X_1 = x_1, \ldots, X_n = x_n, \bar{X} = t)}{P(\bar{X} = t)} \qquad (1) \\ &= \frac{P(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^n X_i = nt)}{P(\sum_{i=1}^n X_i = nt)} \end{aligned}$$

Consider the denominator of (1). Recall $X_1, \ldots, X_n \sim$ i.i.d. Poi(λ), therefore $\sum_{i=1}^n X_i \sim$ Poi($n\lambda$). Consequently,

$$P\left(\sum_{i=1}^n X_i = nt\right) = \frac{e^{-n\lambda}(n\lambda)^{nt}}{(nt)!}. \qquad (2)$$

Now consider the numerator of (1). Since $\sum_{i=1}^n X_i = nt$ and $X_1 = x_1, \ldots, X_n = x_n$, then $\sum_{i=1}^n x_i = x_1 + \cdots + x_n = nt$

and $x_n = nt - \sum_{i=1}^{n-1} x_i$. Therefore,

$$\begin{aligned}
P&\left(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^{n} X_i = nt\right) \\
&= P\left(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = nt - \sum_{i=1}^{n-1} x_i\right) \\
&= \left(\prod_{i=1}^{n-1} P(X_i = x_i)\right) P\left(X_n = nt - \sum_{i=1}^{n-1} x_i\right) \\
&= \left(\prod_{i=1}^{n-1} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}\right) \frac{e^{-\lambda}\lambda^{\left(nt - \sum_{i=1}^{n-1} x_i\right)}}{\left(nt - \sum_{i=1}^{n-1} x_i\right)!} \\
&= \frac{e^{-(n-1)\lambda}\lambda^{\left(\sum_{i=1}^{n-1} x_i\right)}}{\prod_{i=1}^{n-1} x_i!} \frac{e^{-\lambda}\lambda^{\left(nt - \sum_{i=1}^{n-1} x_i\right)}}{\left(nt - \sum_{i=1}^{n-1} x_i\right)!} \\
&= \frac{e^{-n\lambda}\lambda^{(nt)}}{\left(nt - \sum_{i=1}^{n-1} x_i\right)! \prod_{i=1}^{n-1} x_i!} \qquad (3)
\end{aligned}$$

since $X_1, \ldots, X_n$ are independent.

Substituting (2) and (3) back into (1), we obtain

$$\begin{aligned}
p(x_1, \ldots, x_n | T = t, \lambda) &= \frac{\frac{e^{-n\lambda}\lambda^{(nt)}}{\left(nt - \sum_{i=1}^{n-1} x_i\right)! \prod_{i=1}^{n-1} x_i!}}{\frac{e^{-n\lambda}(n\lambda)^{nt}}{(nt)!}} \\
&= \frac{(nt)!}{n^{(nt)} \left(nt - \sum_{i=1}^{n-1} x_i\right)! \prod_{i=1}^{n-1} x_i!}
\end{aligned}$$

which does not involve λ. Therefore, we have shown that $\bar{X}$ is a sufficient statistic for λ.

Theorem Neyman-Fisher factorisation criterion.

The statistic $T(\boldsymbol{X})$ is sufficient for θ if and only if one can factor the likelihood function such that

$$L(\theta) = h(\boldsymbol{x})g(t, \theta),$$

where $h(\boldsymbol{x})$ does not depend on θ (whenever $L(\theta) > 0$) and g is *some* non-negative function of t and θ.

Alternatively, the log-likelihood must be expressible in the form: $l(\theta) = H(\boldsymbol{x}) + G(T(\boldsymbol{x}), \theta)$.

Example S4.2 Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\theta, 1)$. Show that $\bar{X}$ is sufficient for θ.

Consider

$$\begin{aligned}
L(\theta) &= f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}(x_i-\theta)^2} \\
&= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2} \\
&= (2\pi)^{-n/2} e^{-\frac{1}{2}\sum_{i=1}^{n}\left(x_i^2-2\theta x_i+\theta^2\right)} \\
&= (2\pi)^{-n/2} e^{-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2-2\theta\sum_{i=1}^{n} x_i+n\theta^2\right)} \\
&= (2\pi)^{-n/2} e^{-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2\right)} e^{-\frac{1}{2}\left(-2\theta n\bar{x}+n\theta^2\right)}.
\end{aligned}$$

Therefore, letting $h(\boldsymbol{x}) = (2\pi)^{-n/2} e^{-\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2\right)}$ and $g(\bar{X}, \theta) = e^{-\frac{1}{2}\left(-2\theta n\bar{x}+n\theta^2\right)}$ we can factor the likelihood function. So, by the Neyman-Fisher factorisation criterion, $\bar{X}$ is a sufficient statistic for θ.

Example S4.3 Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from a Poisson distribution with parameter λ. Show that $\bar{X}$ is a sufficient statistic for λ using the Neyman-Fisher factorisation criterion.

Consider

$$\begin{aligned}
L(\lambda) &= f(x_1, x_2, \ldots, x_n|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
&= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} = \frac{1}{\prod_{i=1}^{n} x_i!} e^{-n\lambda}\lambda^{n\bar{x}}.
\end{aligned}$$

If we let $h(\boldsymbol{x}) = \frac{1}{\prod_{i=1}^{n} x_i!}$ and $g(\bar{X}, \theta) = e^{-n\lambda}\lambda^{n\bar{x}}$, then we have been able to factor the likelihood function according to the criterion. So, $\bar{X}$ must be a sufficient statistic of λ.

Notes

1. One prefers to use a sufficient statistic as an estimator for θ since the sufficient statistic uses all of the sample information to estimate θ.

2. Sufficient statistics always exist, since $T(\boldsymbol{X}) = (X_1, X_2, \ldots, X_n)$ is itself a sufficient statistic. However, we would prefer a statistic that has as low a dimension as possible. A sufficient statistic with the lowest possible dimensionality is called a *minimal sufficient statistic*.

3. The MLE, if it exists, will always be a function of a sufficient statistic.

S4.2 Minimum variance estimators

Does there exist a best estimator?

Recall that in our previous discussions on qualities of estimators we said we would prefer an estimator with as small an MSE as possible. Unfortunately, if we consider the class of all estimators for a particular parameter, there does not exist such an optimality criterion. If we decide to limit ourselves to particular classes of estimators then there do exist certain optimality criterion.

Let's constrain ourselves to the class of unbiased estimators. Suppose that the random variables and their distributions satisfy the following regularity conditions:

1. The range of the random variables does not depend on θ. (e.g. $X \sim U(0, \theta)$ does not satisfy this condition.

2. The likelihood function is sufficiently smooth to allow us to interchange the operations of differentiation and integration.

3. The 2nd derivatives of the log-likelihood function exist.

Under the regularity conditions, the *Cramér-Rao inequality* states that if $T(\boldsymbol{X})$ is an unbiased estimator of θ, then

$$\mathsf{Var}(T(\boldsymbol{X})) \geq \frac{1}{I(\theta)},$$

where $I(\theta) = E\left[-\frac{d^2 l(\theta)}{d\theta^2}\right]$.

Definition $I(\theta)$ is called the *expected information* or *Fisher's information*.

Definition $\frac{1}{I(\theta)}$ is called the *Cramér-Rao lower bound*.

The Cramér-Rao inequality implies that the smallest the variance of any unbiased estimator can become is $1/I(\theta)$. If any unbiased estimator $T(\boldsymbol{X})$ is such that $\mathsf{Var}(T(\boldsymbol{X})) = 1/I(\theta)$, then we say that $T(\boldsymbol{X})$ is a *minimum variance unbiased estimator* (MVUE) as no other unbiased estimator will be able to obtain a smaller variance.

Example S4.4 Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from a Poisson distribution with parameter λ. Does $\hat{\lambda} = \bar{X}$ achieve the Cramér-Rao lower bound?

(i) $E[\bar{X}] = \frac{1}{n}\sum_{i=1}^n E[X_i] = \frac{1}{n}\sum_{i=1}^n \lambda = \lambda$. Therefore $\bar{X}$ is an unbiased estimator.
(ii)

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

This implies,

$$l(\lambda) = \log L(\lambda) = -n\lambda + \sum_{i=1}^{n} x_i \log \lambda - \log\left(\prod_{i=1}^{n} x_i!\right).$$

Therefore,

$$\frac{dl(\lambda)}{d\lambda} = -n + \frac{\sum_{i=1}^{n} x_i}{\lambda}, \quad \frac{d^2 l(\lambda)}{d\lambda^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2}$$

Computing Fisher's information,

$$\begin{aligned} I(\lambda) &= E\left[-\frac{d^2 l(\lambda)}{d\lambda^2}\right] = E\left[-\left(-\frac{\sum_{i=1}^{n} X_i}{\lambda^2}\right)\right] \\ &= \frac{\sum_{i=1}^{n} E[X_i]}{\lambda^2} = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}. \end{aligned}$$

Hence, according to the Cramér-Rao inequality, Var$(\bar{X}) \geq \frac{1}{I(\lambda)} = \frac{\lambda}{n}$.

Now, since $X_i \sim$ Poi(λ), Var$(\bar{X}) = \frac{\lambda}{n}$. Therefore, $\bar{X}$ is MVUE for λ.

S4.3 Asymptotic normality of the MLE

Theorem If $\hat{\theta}$ is the MLE of θ, then under certain regularity conditions it can be shown that

$$\sqrt{n}(\hat{\theta} - \theta) \to N\left(0, \frac{n}{I(\theta)}\right)$$

as $n \to \infty$.

Hence, approximately for sufficiently large sample sizes,

$$\hat{\theta} \sim N\left(\theta, \frac{1}{I(\theta)}\right).$$

Asymptotic properties of the MLE:

1. $\hat{\theta}$ is asymptotically unbiased.

2. $\hat{\theta}$ is asymptotically fully efficient. (i.e. the variance of $\hat{\theta}$ approaches the Cramér-Rao lower bound: Var$(\hat{\theta}) \to I(\theta)^{-1}$ as $n \to \infty$.

3. $\hat{\theta}$ is asymptotically normally distributed.

Although for large n, Var$(\hat{\theta}) \approx \frac{1}{I(\theta)}$, when θ is unknown then $I(\theta)$ (which is a function of θ) is also unknown. Consequently, if we need to know the variance we may need to estimate it as well. To do this it may be convenient to replace the expected information $I(\theta)$ with the observed information

$$I_0(\hat{\theta}) = -\frac{d^2 l(\theta)}{d\theta^2}\bigg|_{\theta=\hat{\theta}}.$$

Although the asymptotic properties of the MLE are quite good, the properties are true for large samples (i.e. as $n \to \infty$). The properties do not necessarily hold for small samples and for any finite sample they are approximations. The quality of the approximation will depend on the underlying distribution.

S4.4 Invariance property

If $\phi = g(\theta)$, where g is one-to-one monotonic function of θ, then $\hat{\phi} = g(\hat{\theta})$ is the MLE of ϕ for large n

$$\hat{\phi} \approx N\left(\phi, \frac{[g'(\theta)]^2}{I(\theta)}\right),$$

where $g'(\theta) = \frac{dg}{d\theta}$.

Note that for $\hat{\phi} = g(\hat{\theta})$ to be the MLE of ϕ it is not necessary for g to be strictly one-to-one. It is sufficient for the range of g to be an interval.

Example S4.5

Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. We have shown $\hat{\lambda} = \bar{X}$ is the MLE of λ.
(a) What is its asymptotic distribution?
(b) Compute $P(X_i = 0)$.
(c) Find the MLE for $P(X_i = 0)$ and its asymptotic distribution.

(a) According to the previous theorem, since $\hat{\lambda}$ is the MLE of λ, then $\hat{\lambda} \to N\left(\lambda, \frac{1}{I(\lambda)}\right)$. We have shown that $I(\lambda) = \frac{n}{\lambda}$, therefore, $\hat{\lambda} \to N\left(\lambda, \frac{\lambda}{n}\right)$.

(b) $P(X_i = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$.

(c) If $p = P(X_i = 0) = e^{-\lambda}$, then since the range of p is $(0, \infty)$, the MLE is $\hat{p} = e^{-\hat{\lambda}} = e^{-\bar{X}}$. By the invariance property

$$\hat{p} \to N\left(p, \frac{[g'(\lambda)]^2}{I(\lambda)}\right),$$

where $g'(\lambda) = -e^{-\lambda}$. Therefore $\hat{p} \to N\left(p, \frac{e^{-2\lambda}}{n/\lambda}\right)$. Writing it now in terms of the parameter p, if $p = e^{-\lambda}$, then $\lambda = -\log(p)$ and

$$\hat{p} \to N\left(p, \frac{-p^2 \log(p)}{n}\right)$$

as $n \to \infty$.

S5 Interval estimation

S5.1 Confidence intervals

If we are interested in estimating a given parameter θ we can find some estimator $T(\boldsymbol{X})$ using some appropriate method, i.e. Method of Moments, Maximum Likelihood or Least Squares. $T(\boldsymbol{X})$ is called a point estimator since the estimate of θ that we report is one particular point in the parameter space.

For example, when we are interested in estimating the percentage of UK residents who are in favour of the Government's policies, we can collect a random sample of UK residents and compute the sample proportion of the people in favour of the policies. We then report that the Government has, say, a 54% approval rating.

The difficulty that arises, though, is what does 54% mean? How exact is our estimate? The point estimator does not give us that information. Instead it is helpful to also include information about the variability of the estimate given, and that will depend both upon the true underlying variance of the population and the sampling distribution of the estimator that we use.

We have 2 options:

1. Report the value of the estimate and the standard deviation of the estimate, which is often called the standard error of the estimate. For example, the Government has a 54% approval rating with a 2% standard error.

2. Construct an interval estimate for the parameter which incorporates both information about the point estimate, its standard error, and the sampling distribution of the estimator.

For example, a 95% confidence interval for the Government's approval rating is 52.4% to 55.6%.

Definition A $100(1-\alpha)\%$ *confidence interval* for the parameter θ is an interval constructed from a random sample such that if we were to repeat the experiment a large number of times the interval would contain the true value of θ in $100(1-\alpha)\%$ of the cases.

Note that the interval will depend on the value of the estimate and the sampling distribution of the estimator.

Example S5.1 Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean θ and known variance σ_0^2. Construct a $100(1-\alpha)\%$ confidence interval for θ.

First, we need a point estimator for θ, the mean of the normal distribution. Let $\hat{\theta} = \bar{x}$, the sample mean.

Next we need to determine the sampling distribution of the estimator $\hat{\theta}$. Since $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution, then $\hat{\theta} = \bar{x} \sim N\left(\theta, \frac{\sigma_0^2}{n}\right)$.

We want to find endpoints $\hat{\theta}_1$ and $\hat{\theta}_2$ such that

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha.$$

Note that $\hat{\theta}_1$ and $\hat{\theta}_2$ are random values which are determined by the random sample.

Note that there exist an infinite number of $100(1-\alpha)\%$ confidence intervals for θ. We would like to chose the one that is "best", i.e. the one for which the length of the interval $\hat{\theta}_2 - \hat{\theta}_1$ is the shortest, which will in general be the interval which is symmetric around θ if the distribution of $\hat{\theta}$ is symmetric.

Since $\hat{\theta} = \bar{x} \sim N\left(\theta, \frac{\sigma_0^2}{n}\right)$, if we standardise, then we get

$$P\left(-z_{\alpha/2} \leq \frac{\bar{x}-\theta}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

Solving for θ we get,

$$\begin{aligned} P\left(-z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \bar{x}-\theta \leq z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) &= 1-\alpha \\ P\left(-\bar{x}-z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq -\theta \leq -\bar{x}+z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) &= 1-\alpha \\ P\left(\bar{x}-z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \theta \leq \bar{x}+z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) &= 1-\alpha \end{aligned}$$

Therefore, a $100(1-\alpha)\%$ confidence interval for θ, where θ is the mean of a normal distribution with known variance σ_0^2 is

$$\left(\bar{x}-z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \bar{x}+z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right).$$

Example S5.2 Now suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with mean θ and unknown variance σ^2. Construct a $100(1-\alpha)\%$ confidence interval for θ.

Again we use $\hat{\theta} = \bar{x}$, since $\bar{x}$ is the MVUE of θ. We know that $\bar{x} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$, so

$$\frac{\bar{x}-\theta}{\sigma/\sqrt{n}} \sim N(0,1),$$

but now the variance σ^2 is unknown. Hence if we want to find the confidence interval for θ, we need to estimate σ^2.

Recall $\frac{\bar{x}-\theta}{s/\sqrt{n}} \sim t_{n-1}$. Therefore

$$P\left(-t_{n-1,\alpha/2} \leq \frac{\bar{x}-\theta}{s/\sqrt{n}} \leq t_{n-1,\alpha/2}\right) = 1-\alpha.$$

Isolating θ we get,

$$P\left(\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \leq \theta \leq \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha.$$

Therefore a $100(1-\alpha)\%$ confidence interval for θ, where θ is the mean of the normal distribution with unknown variance σ^2, is given by

$$\bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}.$$

S5.2 Asymptotic distribution of the MLE

Suppose $\hat{\theta}$ is the MLE of θ, then we know that $\hat{\theta} \to N\left(\theta, \frac{1}{I(\theta)}\right)$ as $n \to \infty$. Consequently we can construct an approximate $100(1-\alpha)\%$ confidence interval for θ. Since θ is unknown we will also need to approximate $\frac{1}{I(\theta)}$ with the observed information

$$I_0(\hat{\theta}) = E\left(-\frac{d^2 l(\theta)}{d\theta^2}\right)\bigg|_{\theta=\hat{\theta}}.$$

Consequently,

$$P\left(-z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sqrt{\frac{1}{I_0(\hat{\theta})}}} \leq z_{\alpha/2}\right) = 1 - \alpha,$$

and an approximate $100(1-\alpha)\%$ confidence interval for θ is given by

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{1}{I_0(\hat{\theta})}}.$$

This method is extremely useful since it is often quite straightforward to evaluate the MLE and the observed information. Nonetheless it is an approximation, and should only be trusted for large values of n (though the quality of the approximation will vary from model to model).

Example S5.3 Consider $Y_1, Y_2, \ldots, Y_n \sim$ Exp(θ^{-1}) independently. Construct an approximate 95% confidence interval for θ.

For each of the Y_i $(y_i > 0)$,

$$f(y_i|\theta) = \theta^{-1}e^{-y_i/\theta}.$$

Thus

$$L(\theta) = \theta^{-n}e^{-\sum_{i=1}^n y_i/\theta} \Rightarrow l(\theta) = -n\log\theta - \sum_{i=1}^n \frac{y_i}{\theta}$$

and

$$\hat{\theta} = \sum_{i=1}^n \frac{y_i}{n} = \bar{y}.$$

Now,

$$\frac{d^2l(\theta)}{d\theta^2} = \frac{n}{\theta^2} - \frac{2\sum_{i=1}^n y_i}{\theta^3}$$

so that

$$I_0(\hat{\theta}) = -\left(\frac{n}{\bar{y}^2} - \frac{2n\bar{y}}{\bar{y}^3}\right) = \frac{n}{\bar{y}^2}.$$

Hence, an approximate 95% confidence interval for θ is

$$\bar{y} \pm 1.96 \times \sqrt{\frac{\bar{y}^2}{n}}.$$

Example S5.4 Consider $Y_1, Y_2, \ldots, Y_n \sim N(\theta, 1)$ independently. Construct an approximate 95% confidence interval for θ.

In example S3.7 we showed that $\hat{\theta} = \bar{y}$ and

$$\frac{d^2l(\theta)}{d\theta^2} = -n.$$

Hence, $I_0(\hat{\theta}) = n$, and a 95% confidence interval for θ is

$$\bar{y} \pm 1.96 \times \sqrt{\frac{1}{n}}.$$

S6 Hypothesis testing

S6.1 Null and alternative hypotheses

In estimation, we are interested in asking ourselves the question what is the value of some particular parameter of interest in the population. For example, what is the average annual income of residents in the UK?

Often there are times in statistics when we are not interested in the specific value of the parameter, but rather are interested in asserting some statement about the parameter of interest. Some examples:

1. We want to claim that the average annual income of UK residents is more than $\pounds 35,000$.

2. We want to assess whether the average annual income of men in academia in the UK is the same as that of women at similar ranks.

3. We want to determine whether the number of cars crossing a certain intersection follows a Poisson distribution or whether it is more likely to come from a normal distribution.

To perform a statistical hypothesis test, one needs to specify two disjoint hypotheses in terms of the parameters of the distribution that are of interest. They are

H_0 : Null Hypothesis,
H_1 : Alternative Hypothesis.

Traditionally, we choose H_0 to be the hypothesis claiming equality and H_1 to be the claim that we would like to assert, unless we want to claim equality. Let's return to our examples:

1. We want to claim that the average annual income of UK residents is more than $£35,000$. We test

$$H_0 : \mu \leq 35,000 \quad \text{vs.} \quad H_1 : \mu > 35,000.$$

2. We want to assess whether the average annual income of men in academia in the UK is the same as that of women at similar ranks. We test

$$H_0 : \mu_{\text{men}} = \mu_{\text{women}} \quad \text{vs.} \quad H_1 : \mu_{\text{men}} \neq \mu_{\text{women}}.$$

3. We want to determine whether the number of cars crossing a certain intersection follows a Poisson distribution or whether it is more likely to come from a normal distribution. We test

$$H_0 : X \sim \text{Poisson}(2) \quad \text{vs.} \quad H_1 : X \sim \text{Normal}(2, 2).$$

Hypotheses where the distribution is completely specified are called simple hypotheses. For example, H_0 and H_1 in example 3 and H_0 in example 2 are all simple hypotheses.

Hypotheses where the distribution is not completely specified are called composite hypotheses. For example, H_0 and H_1 in example 1 and H_1 in example 2 are all composite hypotheses.

The conclusion of a hypothesis test

We will **reject** H_0 if there is sufficient information from our sample that indicates that the null hypothesis cannot be true, i.e. the alternative hypothesis is true.

We will **not reject** H_0 if there is not sufficient sample information to refute our claim.

S6.2 Type I and Type II errors

Decision	Truth	
	H_0 True	H_1 True
Reject H_0	Type I error	Correct
Do Not Reject H_0	Correct	Type II error

Definition The *significance level* or *size* of the test is

$$\begin{aligned}\alpha &= P(\text{Type I error}) \\ &= P(\text{Reject } H_0 | H_0 \text{ true}).\end{aligned}$$

Typical choices for α are 0.01, 0.05 and 0.10.

Definition The probability of a Type II error is

$$\begin{aligned}\beta &= P(\text{Type II error}) \\ &= P(\text{Do Not Reject } H_0 | H_1 \text{ true}).\end{aligned}$$

Notes

1. It can be shown that there is an inverse relationship between α and β i.e. as α increases, β decreases and vice versa. Therefore for any fixed sample size one can only choose to control one of the types of error, so we choose to control Type I error and select our hypotheses initially so the "worse" error is the Type I error.

2. The value of both α and β depend on the value of the underlying parameters. Consequently, we can control α by choosing H_0 to include equality of the parameter and can show that the largest the Type I error can become is at the point of equality and so choose that to be the size. In example 1 above, for example,

$$\begin{aligned}\alpha &= P(\text{rejecting } H_0 | \mu = 35,000) \\ &\geq P(\text{rejecting } H_0 | \mu \leq 35,000).\end{aligned}$$

Therefore $H_0 : \mu \leq 35,000$ is often just written as $H_0 : \mu = 35,000$.

3. Because H_0 contains the equality, H_1 is usually a composite hypotheses. Therefore $\beta = P(\text{Type II error})$ is a function of the parameter within the alternative parameter space.

Definition The *power* of the test is

$$\begin{aligned} 1 - \beta &= 1 - P(\text{Type II error}) \\ &= P(\text{Reject } H_0 | H_1 \text{ true}). \end{aligned}$$

The power can be thought of as the probability of making a correct decision.

S6.3 Tests for means, σ known

Test 1 $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$; σ^2 known

We assume either
(i) $X_1, X_2, \ldots, X_n$ are a random sample from a normal distribution with known variance σ^2, or
(ii) The sample size n is sufficiently large so that we can assume $\bar{X}$ is approximately normally distributed by the Central Limit Theorem and that either the variance is known or the sample variance $s^2 \approx \sigma^2$.

Step 1: Choose a *test statistic* based upon the random sample for the parameter we want to base our claim on. For example, we are interested in μ so we want to choose a good estimator of μ as our test statistic. i.e. $\hat{\mu} = \bar{X}$.

Step 2: Specify a *decision rule*. For example, we want to claim $\mu < \mu_0$. Therefore, our decision rule is to reject H_0 if $\bar{X} < c$, where c is called the *cut-off* value for the test.

Step 3: Based upon the sampling distribution of the test statistic and the specified significance level of the test, solve for the specific value of the cut-off value c. To find c,

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

Since $P(Z < -z_\alpha) = \alpha$, where z_α is found in Neave's Tables, then

$$-z_\alpha = \frac{c - \mu_0}{\sigma/\sqrt{n}}$$

and $c = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$.

So, the **decision rule** is to reject H_0 if $\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$ or, equivalently,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_\alpha.$$

Test 2 Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$; σ^2 known

This is similar to the previous test, except the **decision rule** is to reject H_0 if $\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$ or, equivalently,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha.$$

Note that both these tests are called *one-sided tests*, since the rejection region falls on only one side of the outcome space.

Test 3 Test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$; σ^2 known

The **decision rule** is now to reject H_0 if $\bar{X} < \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $\bar{X} > \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or, equivalently,

$$|Z| = \left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| > z_{\alpha/2}.$$

This is called a *two-sided test* because the decision rule partitions the outcome space into two disjoint intervals.

Example S6.1 Suppose a coffee machine is designed to dispense 6 ounces of coffee per cup with a (known) $\sigma = 0.2$ (where we assume the amount of coffee dispensed is normally distributed). A random sample of $n = 20$ cups gives $\bar{x} = 5.94$. Test whether the machine is correctly filling the cups.

We test $H_0 : \mu = 6.0$ vs. $H_1 : \mu \neq 6.0$ at $\alpha = 0.05$.

The decision rule is to reject H_0 if $|Z| = \left|\frac{\bar{x}-6.0}{0.2/\sqrt{20}}\right| > z_{0.05/2} = z_{0.025} = 1.96$. Now

$$|Z| = \left|\frac{5.94 - 6.0}{0.2/\sqrt{20}}\right| = |-1.34| < 1.96.$$

Therefore, we conclude that there is not enough statistical evidence to reject H_0 at $\alpha = 0.05$.

S6.4 p values

When our sample information determines a particular conclusion to our hypothesis test, we report that we either reject or do not reject H_0 at a particular significance level α.

For example, we would have reached the same decision in Example S6.1 whether $|Z| = 1.34$ or $|Z| = 1.95$. Whereas, if we had chosen $\alpha = 0.10$, we would have rejected H_0 if

$|Z| = 1.95 > z_{0.10/2} = 1.6449$, but we would not reject H_0 if $|Z| = 1.34 < z_{0.10/2} = 1.6449$.

Hence when we report our conclusion the reader doesn't know how sensitive our decision is to the choice of α.

Note that the choice of α should be made before the test is performed; otherwise, we run the risk of inducing experimenter bias!

Definition The p *value* of a test is the probability of rejecting H_0 with the value of the test statistic obtained from the data given H_0 is true.

If we report the conclusion of the test, as well as the p value then the reader can decide how sensitive our result was to our choice of α.

Example S6.2 Compute the p value for the test in Example S6.1.

In the coffee cup example above we were given $\bar{x} = 5.94$, $n = 20$ and $\sigma = 0.2$. Our decision rule was to reject H_0 if $|Z| = \left|\frac{\bar{x}-6.0}{0.2/\sqrt{20}}\right| > z_{0.025}$.

To compute the p-value for the test we want to find,

$$\begin{aligned} P\left(|Z| > \left|\frac{5.94 - 6.0}{0.2/\sqrt{20}}\right|\right) &= 2P(Z > 1.34) \\ &= 2 \times 0.0901 \\ &= 0.1802. \end{aligned}$$

Notes

1. We multiplied the probability by 2 since we are computing

the p value for a two-sided test, where there is an equal-sized rejection region at both tails of the distribution. For a one-tailed test we only need to compute the probability of rejecting in one direction.

2. The p value implies that if we had chosen an α of at least 0.1802 then we would have been able to reject H_0.

3. In applied statistics, the p value is interpreted as the sample providing:

$$\begin{cases} \text{strong evidence against } H_0 & \text{if } p \leq 0.01, \\ \text{evidence against } H_0 & \text{if } p \leq 0.05, \\ \text{slight evidence against } H_0 & \text{if } p \leq 0.10, \\ \text{no evidence against } H_0 & \text{if } p > 0.10. \end{cases}$$

S6.5 Tests concerning normal means (σ unknown)

Assume $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with unknown variance σ^2.

Test 4 $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$; σ^2 unknown

The decision rule is to reject H_0 if $\bar{X} < c$. As before,

$$\begin{aligned} \alpha &= P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ true}) \\ &= P(\bar{X} < c | \mu = \mu_0) \\ &= P\left(\bar{X} < c | \bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)\right), \end{aligned}$$

but recall σ^2 is unknown. Therefore, we will need to estimate it using s^2. Now,

$$\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}.$$

Hence,

$$\begin{aligned}\alpha &= P(\bar{X} < c|\mu = \mu_0) = P\left(\frac{\bar{X} - \mu_0}{s/\sqrt{n}} < \frac{c - \mu_0}{s/\sqrt{n}}\right) \\ &= P\left(T < \frac{c - \mu_0}{s/\sqrt{n}}\right).\end{aligned}$$

Now, $P(T < -t_{n-1,\alpha}) = \alpha$, where $t_{n-1,\alpha}$ is found in Neave's Tables, so

$$-t_{n-1,\alpha} = \frac{c - \mu_0}{s/\sqrt{n}}$$

and $c = \mu_0 - t_{n-1,\alpha}\frac{s}{\sqrt{n}}$.

Therefore, the **decision rule** is to reject H_0 if $\bar{X} < \mu_0 - t_{n-1,\alpha}\frac{s}{\sqrt{n}}$ or, equivalently,

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} < -t_{n-1,\alpha}.$$

Test 5 $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$; σ^2 unknown

This is similar to the previous test, except the **decision rule** is to reject H_0 if $\bar{X} > \mu_0 + t_{n-1,\alpha}\frac{s}{\sqrt{n}}$ or, equivalently,

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} > t_{n-1,\alpha}.$$

Test 6 $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$; σ^2 unknown

The **decision rule** is now to reject H_0 if

$$|T| = \left|\frac{\bar{X} - \mu_0}{s/\sqrt{n}}\right| > t_{n-1,\alpha/2}.$$

Example S6.3 Suppose that σ is unknown in Example S6.1. (We still assume the amount of coffee dispensed is normally distributed). A random sample of $n = 20$ cups gives $\bar{x} = 5.94$ and $s^2 = 0.1501^2$. Test whether the machine is correctly filling the cups.

We test $H_0 : \mu = 6.0$ vs. $H_1 : \mu \neq 6.0$ at $\alpha = 0.05$.

The decision rule is to reject H_0 if $|T| = \left|\frac{\bar{x}-6.0}{0.1501/\sqrt{20}}\right| > t_{20-1,0.05/2} = t_{19,0.025} = 2.093$.

Now

$$|T| = \left|\frac{5.94 - 6.0}{0.1501/\sqrt{20}}\right| = |-1.7876| < 2.093.$$

Therefore, we do not reject H_0 at $\alpha = 0.05$. The p value is

$$p = 2P(t_{19} > |-1.7876|) \approx 2 \times 0.05 = 0.10.$$

S6.6 Confidence intervals and two-sided tests

Consider the two-sided t-test of size α. We reject H_0 if $|T| = \left|\frac{\bar{X}-\mu_0}{s/\sqrt{n}}\right| > t_{n-1,\alpha/2}$. This implies we do not reject H_0 if

$$\begin{aligned} & |T| = \left|\frac{\bar{X} - \mu_0}{s/\sqrt{n}}\right| \leq t_{n-1,\alpha/2} \\ \Leftrightarrow \quad & -t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \leq \bar{X} - \mu_0 \leq t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \\ \Leftrightarrow \quad & \bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}. \end{aligned}$$

But

$$\left(\bar{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for μ. Consequently, if μ_0, the value of μ under H_0, falls within the $100(1-\alpha)\%$ confidence interval for μ, then we will **not** reject H_0 at significance level α.

In general, therefore, there is a correspondence between the "acceptance region" of a statistical test of size α and the related $100(1-\alpha)\%$ confidence interval. Therefore, we will **not** reject $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ at level α if and only if θ_0 lies within the $100(1-\alpha)\%$ confidence interval for θ.

Example S6.4: For the coffee machine in Example S6.3 we wanted to test $H_0 : \mu = 6.0$ vs. $H_1 : \mu \neq 6.0$ at $\alpha = 0.05$. We were given a random sample of $n = 20$ cups with $\bar{x} = 5.94$ and $s^2 = 0.1501^2$. Construct a 95% confidence interval for μ.

The limits of a 95% confidence interval for μ are

$$\begin{aligned} \bar{x} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} &= 5.94 \pm t_{20-1,0.05/2}\frac{0.1501}{\sqrt{20}} \\ &= 5.94 \pm 2.093\frac{0.1501}{\sqrt{20}} \end{aligned}$$

so the 95% confidence interval for μ is

$$(5.8698, 6.0102).$$

If we use the confidence interval to perform our test, we see that

$$\mu_0 = 6.0 \in (5.8698, 6.0102),$$

so we will not reject H_0 at $\alpha = 0.05$.

S6.7 Other types of tests

Test 7 $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 \neq \sigma_2^2$

Let $X_1, X_2, \ldots, X_m \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma_2^2)$ be two independent random samples from normal populations.

The test statistic is $F = \frac{s_1^2}{s_2^2}$, where

$$\begin{aligned} s_1^2 &= \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 \\ s_2^2 &= \frac{1}{n-1}\sum_{i-1}^{n}(Y_i - \bar{Y})^2. \end{aligned}$$

Recall that

$$\begin{aligned} (m-1)\frac{s_1^2}{\sigma_1^2} &\sim \chi^2_{m-1}, \\ (n-1)\frac{s_2^2}{\sigma_2^2} &\sim \chi^2_{n-1}. \end{aligned}$$

Note that s_1^2 and s_2^2 are independent since the samples are independent. Therefore,

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim \frac{\chi^2_{m-1}/(m-1)}{\chi^2_{n-1}/(n-1)} \sim F_{m-1,n-1}.$$

Under $H_0 : \sigma_1^2 = \sigma_2^2$, therefore

$$F = \frac{s_1^2}{s_2^2} \sim F_{m-1,n-1}.$$

The **decision rule** is to reject H_0 if

$$F = \frac{s_1^2}{s_2^2} < F_{m-1,n-1,\alpha/2}$$

or if

$$F = \frac{s_1^2}{s_2^2} > F_{m-1,n-1,1-\alpha/2}.$$

When using Neave's Tables it may be helpful to remember that

$$F_{\nu_1,\nu_2,q} = \frac{1}{F_{\nu_2,\nu_1,1-q}}.$$

Test 8 $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$; σ^2 unknown

Assume $X_1, X_2, \ldots, X_m \sim N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_n \sim N(\mu_2, \sigma^2)$ are two independent random samples with unknown (common) variance σ^2.

The **decision rule** is to reject H_0 if

$$|T| = \left| \frac{\bar{X} - \bar{Y}}{\sqrt{s_p^2 \left(\frac{1}{m} + \frac{1}{n}\right)}} \right| > t_{m+n-2,\alpha/2},$$

where $s_p^2 = \frac{(m-1)s_X^2 + (n-1)s_Y^2}{m+n-2}$ is the pooled sample variance.

Note that

1. $(\bar{X} - \bar{Y}) \sim N\left((\mu_1 - \mu_2), \sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)\right)$ which implies
$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim N(0, 1).$$

2. $(m + n - 2)\frac{s_p^2}{\sigma^2} \sim \chi^2_{m+n-2}$

3. s_p^2 is independent of $\bar{X} - \bar{Y}$.

Therefore,

$$\frac{\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\sigma^2\left(\frac{1}{m}+\frac{1}{n}\right)}}}{\sqrt{\frac{(m+n-2)s_p^2}{(m+n-2)\sigma^2}}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim t_{m+n-2}.$$

Under H_0, $\mu_1 - \mu_2 = 0$, hence

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{s_p^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim t_{m+n-2}.$$

Example S6.5 Suppose one wants to test whether the time it takes to get from a blood bank to a hospital via two different routes is the same on average. Independent random samples are selected from each of the different routes and we obtain the following information:

Route X:	$m = 10$	$\bar{x} = 34$	$s_1^2 = 17.111$
Route Y:	$n = 12$	$\bar{y} = 30$	$s_2^2 = 9.454$

Test $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ at level $\alpha = 0.05$.

To perform the t-test we need the variances to be equal, so we test $H_0 : \sigma_1^2 = \sigma_2^2$ vs. $H_1 : \sigma_1^2 \neq \sigma_2^2$ at level $\alpha = 0.05$.

The decision rule is to reject H_0 if $F = \frac{s_1^2}{s_2^2} < F_{m-1,n-1,\alpha/2}$ or if $F = \frac{s_1^2}{s_2^2} > F_{m-1,n-1,1-\alpha/2}$.

Computing $F = \frac{s_1^2}{s_2^2} = \frac{17.111}{9.454} = 1.81$.

$F_{9,11,.025} = \frac{1}{F_{11,9,.975}} = \frac{1}{3.915} = 0.255$ and $F_{9,11,.975} = 3.59$.

Hence $F_{9,11,.025} < F < F_{9,11,.975}$, so we do not reject H_0 at $\alpha = 0.05$. Therefore we can assume the variances from the two samples are the same.

Now we test $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$ at level $\alpha = 0.05$

The decision rule is to reject H_0 if

$$|T| = \left| \frac{\bar{X} - \bar{Y}}{\sqrt{s_p^2 \left(\frac{1}{m} + \frac{1}{n}\right)}} \right| > t_{m+n-2,\alpha/2}.$$

Computing, $s_p^2 = \frac{9\times 17.111 + 11\times 9.454}{10+12-2} = 12.9$ so

$$T = \left| \frac{34 - 30}{\sqrt{12.9\left(\frac{1}{10} + \frac{1}{12}\right)}} \right| = 2.6 > t_{20,.025} = 2.086.$$

Therefore we reject H_0 that the journey times are the same on average at $\alpha = 0.05$. The p value is $P(|T| > 2.6) = 2P(T_{20} > 2.6) \approx 2 \times .009 = 0.018$.

Test 9 $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$; non-independent samples

Suppose we have two groups of observations $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ where there is an obvious pairing between the observations (e.g. before and after studies, different measuring devices, etc.) Note that the samples are no longer independent.

An equivalent set of hypotheses is $H_0 : \mu_d = \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_d = \mu_1 - \mu_2 \neq 0$.

Let $D_i = X_i - Y_i$ for $i = 1, \ldots, n$ and assume $D_1, D_2, \ldots, D_n \sim$ i.i.d. $N(\mu_d, \sigma_d^2)$.

The **decision rule** is to reject H_0 if

$$\left| \frac{\bar{D}}{s_d/\sqrt{n}} \right| > t_{n-1,\alpha/2}.$$

Example S6.6 In a medical study of patients given a drug and a placebo, sixteen patients were paired up with members of each pair having similar age and being the same sex. One of each pair received the drug and the other the placebo. The response score for each patient was found.

Pair number	1	2	3	4	5	6	7	8
Given drug	0.16	0.97	1.57	0.55	0.62	1.12	0.68	1.69
Given placebo	0.11	0.13	0.77	1.19	0.46	0.41	0.40	1.28

Are the responses for the drug and placebo significantly different?

This is a 'matched-pair' problem, since we expect a relation between the values of each pair. The difference within each pair is

Pair number	1	2	3	4	5	6	7	8
$d_i = y_i - x_i$	0.05	0.84	0.80	-0.64	0.16	0.71	0.28	0.41

We consider the d_i's to be a random sample from $N(\mu_d, \sigma_d^2)$. Now $\bar{d} = 0.326$, $s_d^2 = 0.241$ so $s_d = 0.490$.

We test $H_0 : \mu_d = 0$ vs $H_1 : \mu_d \neq 0$. and reject H_0 if

$$\left| \frac{\bar{d}}{s_d/\sqrt{n}} \right| = 1.88 > t_{n-1,\alpha/2}.$$

Now, $t_{7,0.05} = 1.89$, so we would not reject H_0 at the 10% level (just).

S7 Bayesian Inference

S7.1 What is Bayesian inference?

Classical (frequentist) approach With this approach, we assume that $f(\boldsymbol{x}|\theta) = f(x_1, x_2, \ldots, x_n|\theta)$ is the likelihood function of the random sample $X_1, X_2, \ldots, X_n$ where θ is an unknown but fixed parameter for the distribution which generated the random sample. The parameter, θ, is always assumed to be a *constant* which we want to estimate or test hypotheses about.

The underlying assumption for the interpretation of our results comes from the frequency interpretation of probability, i.e. probability is the relative frequency that our desired result occurs in repeated trials.

Bayesian approach With this approach, we assume θ is a random variable for which we can assign a probability distribution based on previous knowledge or beliefs.

Therefore, probability is interpreted as reflecting a degree of reasonable belief.

The two approaches use different techniques and terminology:

Classical approach	Bayesian approach
maximum likelihood	prior distribution
method of moments	posterior distribution
sampling dist. of $\hat{\theta}$	risk (loss) function
unbiasedness, etc.	expected loss
confidence intervals	
type I and II errors	

S7.2 The Bayesian approach to inference

Let $f(\boldsymbol{x}|\theta) = f(x_1, x_2, \dots, x_n|\theta)$ be the p.d.f. or p.m.f. of some distribution with unknown parameter θ. Based on a random sample of observations $X_1, X_2, \dots, X_n$ from this distribution, we want to determine where θ lies in the parameter space, Ω.

The key steps to the Bayesian approach are:
1. Specify the likelihood model $f(x_1, x_2, \dots, x_n|\theta)$ for the random sample.
2. Determine a prior distribution $f(\theta)$ for the unknown parameter θ.
3. Calculate the posterior distribution of θ using Bayes' Theorem.
4. Draw inferences based on the posterior distribution.

The choice of likelihood model is similar to the frequentist approach. We can base the choice of likelihood on the model which produces the data (e.g. Bernoulli, Binomial, Poisson) or some "common knowledge" or previous work.

S7.3 The prior distribution

Before any observations are collected, the experimenter assigns a probability distribution to θ, $f(\theta)$, called the *prior distribution* of θ. The distribution is based upon previous knowledge and beliefs as to the relative likelihood that the true value of θ lies within each region of the parameter space Ω.

Note that a prior distribution, $f(\theta)$, is a probability distribution so it will have all the properties of a probability distribution.

For example, let p be the probability of obtaining a head when tossing a coin. Then $\Omega = [0, 1]$, and we might assign

$$f(p) = \begin{cases} 1.9 & \text{if } 0.45 \leq p \leq 0.55 \\ 0.9 & \text{if } 0 \leq p < 0.45 \text{ or } 0.55 < p \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Notes

1. Different prior beliefs will lead to a different posterior distribution. The analysis is subjective and therefore a different experimenter may draw a different conclusion.

2. In most cases the effect of the prior becomes less influential as more data becomes avaiable. The choice of prior is less important if there is enough data.

3. It may be possible to choose a prior which is consistent with our beliefs but which also makes the mathematics relatively straightforward.

4. If there is no prior information about a parameter we can choose a prior distribution which reflects our ignorance about the parameter.

There are some common choices for prior distributions:

1. Uniform distribution

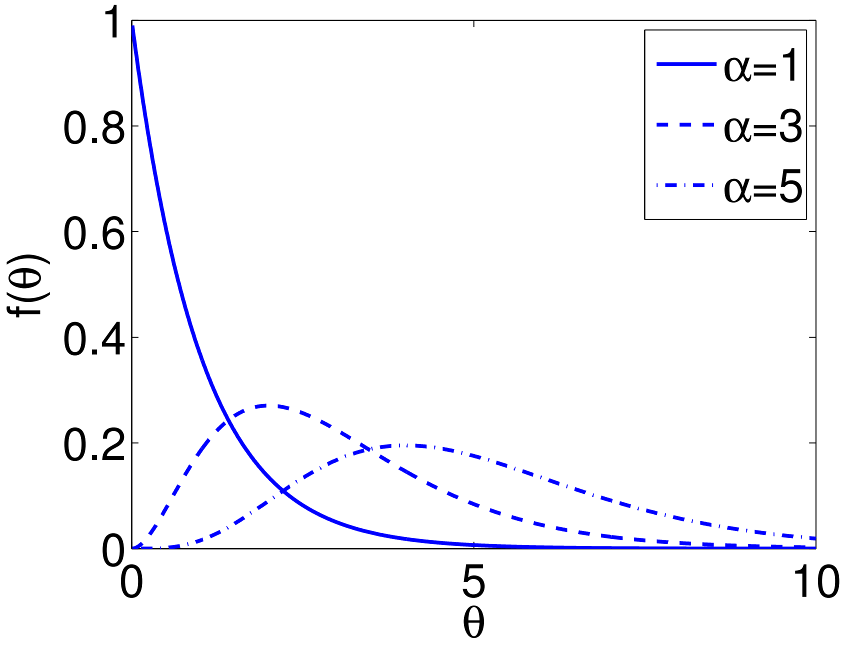
$$f(\theta) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq \theta \leq b \\ 0 & \text{otherwise} \end{cases}$$

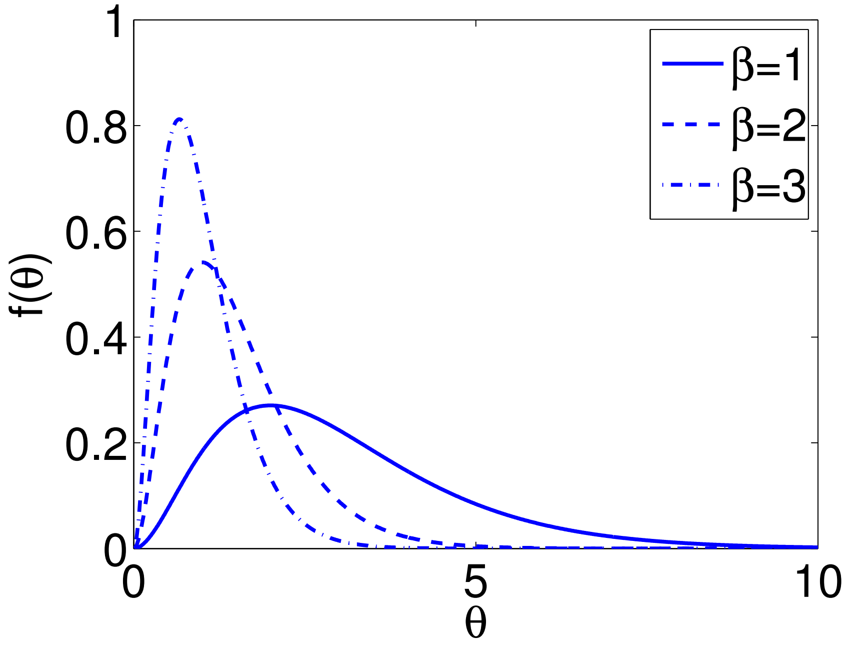
Note that any point in the interval $\Omega = [a, b]$ is equally likely to be a candidate for θ. It is used when the experimenter has no previous knowledge or belief where θ lies in Ω (and the parameter lies in an interval of finite length).

2. Gamma distribution

$$f(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} & \text{if } \theta > 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that $E[\theta] = \frac{\alpha}{\beta}$ and $\text{Var}(\theta) = \frac{\alpha}{\beta^2}$. It is used to model parameters which are positive. We can choose α and β either based on assumptions about the mean and variance of θ or to get the desired functional form.





The figures show the p.d.f. of the gamma distribution with $\beta = 1$, top, and $\alpha = 3$, bottom.

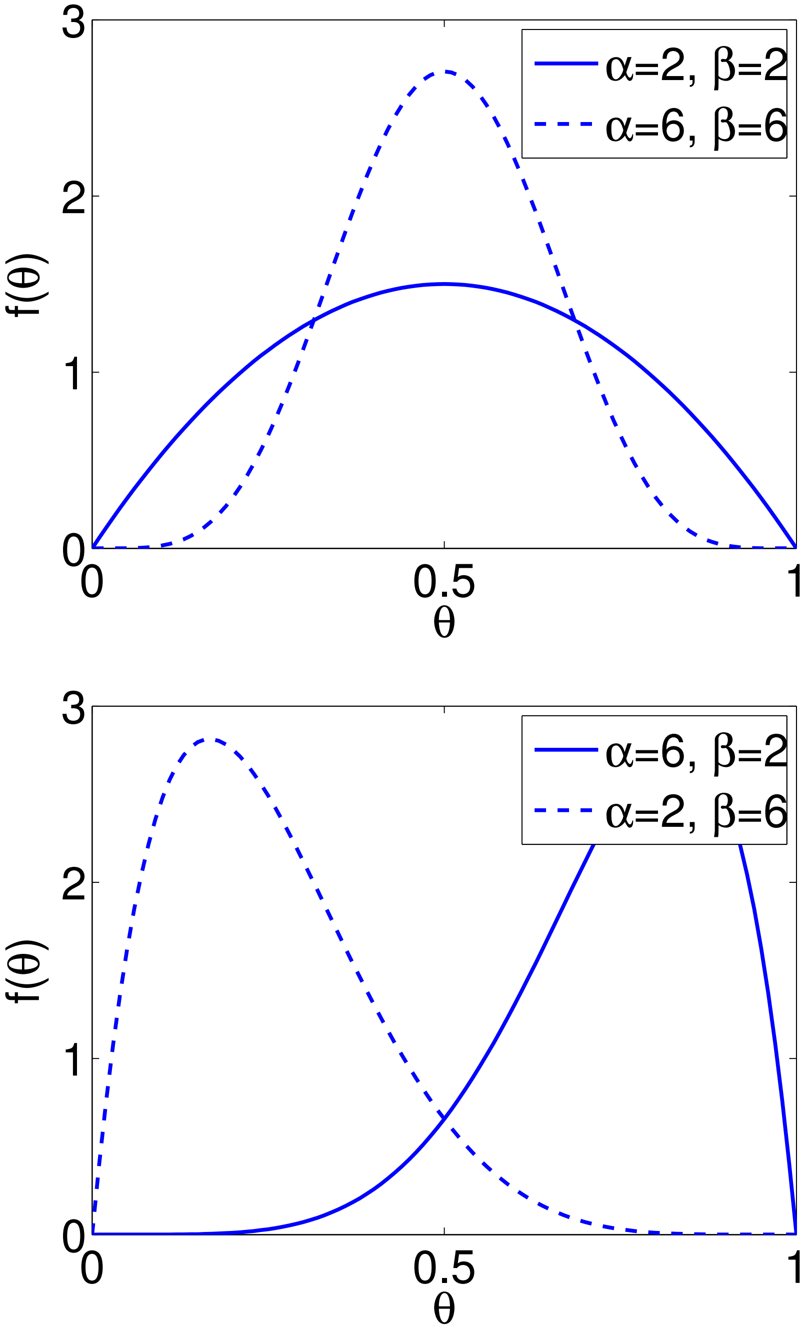
3. Beta distribution

$$f(\theta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} & \text{if } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that $E[\theta] = \frac{\alpha}{\alpha+\beta}$ and

$$\text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

It is used to model parameters that take on values only on (0,1). The choices of α and β may be based upon your knowledge of the mean and variance of θ or your belief on the shape of the underlying distribution.



The figures show the p.d.f. of the beta distribution for different values of α and β.

4. Normal distribution

$$f(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

if $-\infty < \theta < \infty$. It is used to model parameters that can take on any real value, but the probability is symmetric about some point μ and the variance is chosen to reflect how tightly you would expect the distribution of θ to lie around μ.

S7.4 Bayes' theorem

Once you have observed the random sample, $X_1, X_2, \ldots, X_n$, we want to revise the probability distribution of θ to incorporate the information you observe about $X_1, X_2, \ldots, X_n$.

Definition The p.d.f. (p.m.f.) of the *posterior distribution* of θ given $X_1, X_2, \ldots, X_n$, is defined to be

$$\begin{aligned} f(\theta|X_1, X_2, \ldots, X_n) &= \frac{f(x_1, x_2, \ldots, x_n, \theta)}{f(x_1, x_2, \ldots, x_n)} \\ &= \frac{f(x_1, x_2, \ldots, x_n|\theta)f(\theta)}{\int_\Omega f(x_1, x_2, \ldots, x_n|\theta)f(\theta)d\theta}. \end{aligned}$$

Note that:
1. $f(\theta)$ is the prior distribution of θ.
2. $f(x_1, x_2, \ldots, x_n|\theta)$ is the typical likelihood function of $X_1, X_2, \ldots, X_n$ with θ fixed.
3. $f(x_1, x_2, \ldots, x_n) = \int_\Omega f(x_1, x_2, \ldots, x_n|\theta)f(\theta)d\theta$ will be a constant with regard to the p.d.f. of θ. It is referred to as the normalising constant.

Therefore,

$$f(\theta|X_1, X_2, \ldots, X_n) \propto f(x_1, x_2, \ldots, x_n|\theta)f(\theta).$$

Instead of calculating $f(\theta|X_1, X_2, \ldots, X_n)$ explicitly, one instead calculates

$$f(x_1, x_2, \ldots, x_n|\theta)f(\theta).$$

Often we can specify the normalising constant afterwards, if needed.

S7.5 Examples

Example S7.1 $X_1, \ldots, X_n$ iid $N(\theta, \sigma^2)$ where σ^2 is known, and prior $\theta \sim N(b, d^2)$.

$$\begin{aligned} \pi(\theta|\boldsymbol{x}) &\propto \pi(\boldsymbol{x}|\theta)\pi(\theta) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\sum(x_i-\theta)^2\right\} \\ &\qquad \times \exp\left\{-\frac{1}{2d^2}(\theta-b)^2\right\} \end{aligned}$$

$$\begin{aligned} &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(n\theta^2 - 2\theta n\bar{x}\right) - \frac{1}{2d^2}\left(\theta^2 - 2\theta b\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{d^2}\right)\left[\theta^2 - 2\theta\frac{\left(\frac{n\bar{x}}{\sigma^2}+\frac{b}{d^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{d^2}\right)}\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{d^2}\right)\left[\theta - \frac{\left(\frac{n\bar{x}}{\sigma^2}+\frac{b}{d^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{d^2}\right)}\right]^2\right\}, \end{aligned}$$

and so $\pi(\theta|\boldsymbol{x}) \sim N(B, D^2)$ where

$$B = \frac{\left(\frac{n\bar{x}}{\sigma^2}+\frac{b}{d^2}\right)}{\left(\frac{n}{\sigma^2}+\frac{1}{d^2}\right)}, \quad D^2 = \frac{1}{\frac{n}{\sigma^2}+\frac{1}{d^2}}.$$

Example S7.2 Suppose we have a two-sided coin. We want to estimate, θ, the probability of obtaining a head. Based on our experiences we believe the coin should be fair with a high degree of certainty. To test our assertion, we flip the coin 10 times and observe the number of heads that results. Estimate θ using a Bayesian approach.

1. Firstly, we determine the prior distribution of θ. Since $0 < \theta < 1$, we assume $\theta \sim$ Beta(α, β). Furthermore, if we assume $E[\theta] = \frac{1}{2}$ and Var$(\theta) = \frac{1}{16}$, then we can solve for α and β as follows.

$$E[\theta] = \frac{\alpha}{\alpha+\beta} = \frac{1}{2} \quad \Rightarrow \quad 2\alpha = \alpha+\beta$$

So $\alpha = \beta$. Also,

$$\begin{aligned} \text{Var}(\theta) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha^2}{(2\alpha)^2(2\alpha+1)} \\ &= \frac{1}{4(2\alpha+1)} = \frac{1}{16}. \end{aligned}$$

So $2\alpha + 1 = 4$. Hence, $\alpha = \frac{3}{2}$ and $\beta = \frac{3}{2}$.

Consequently,

$$f(\theta) = \begin{cases} \frac{\Gamma(3)}{\Gamma(3/2)\Gamma(3/2)} \theta^{\frac{3}{2}-1}(1-\theta)^{\frac{3}{2}-1} & \text{if } 0 < \theta < 1 \\ 0 & \text{otherwise} \end{cases}$$

2. Next, we determine the likelihood function for X, the number of heads observed. $X \sim \text{Bin}(10, \theta)$ implies

$$f(x|\theta) = \left(\begin{array}{c} 10 \\ x \end{array} \right) \theta^x (1-\theta)^{10-x}$$

if $x = 0, 1, 2, \ldots, 10$.

3. Now we can calculate the posterior distribution.

$$\begin{aligned} f(\theta|x) &= \frac{f(x|\theta)f(\theta)}{\int_\Omega f(x|\theta)f(\theta)d\theta} \propto f(x|\theta)f(\theta) \\ &\propto \left(\begin{array}{c} 10 \\ x \end{array} \right) \theta^x (1-\theta)^{10-x} \frac{\Gamma(3)}{\Gamma(3/2)\Gamma(3/2)} \theta^{\frac{3}{2}-1}(1-\theta)^{\frac{3}{2}-1} \\ &\propto \theta^{x+\frac{3}{2}-1}(1-\theta)^{10-x+\frac{3}{2}-1} \end{aligned}$$

Therefore, $f(\theta|x)$ is proportional to a Beta distribution with $\alpha = x + \frac{3}{2}$ and $\beta = 10 - x + \frac{3}{2}$, so

$$f(\theta|x) = C\theta^{x+\frac{3}{2}-1}(1-\theta)^{10-x+\frac{3}{2}-1}$$

if $0 < \theta < 1$ and 0 otherwise, where

$$C = \frac{\Gamma(13)}{\Gamma\left(x + \frac{3}{2}\right)\Gamma\left(10 - x + \frac{3}{2}\right)}.$$

Definition If the prior and posterior distributions belong to the same family, G, the data have a distribution belonging to the family H, then we say that G is a family of *conjugate priors* to H.

In Example S7.2 we see that Beta is a family of conjugate priors to Binomial distribution. More families of conjugate priors are given in the following table:

Data	Parameter	Prior/Post.
Bernoulli (Binomial)	$0 < \theta < 1$	Beta
Normal (known var.)	$-\infty < \theta < \infty$	Normal
Poisson	$\lambda > 0$	Gamma
Exponential	$\theta > 0$	Gamma

S7.6 Estimation in a Bayesian context

Let θ be the parameter we are interested in estimating. Let a denote a possible estimate of θ.

In Bayesian statistics, one describes decisions as a set of possible actions. Estimation is considered to be making a decision about the value of θ, so each possible value that θ could take on can be thought of as an action, hence the notation, a.

Definition For each possible value of $\theta \in \Omega$ and each possible estimate $a \in \Omega$, there is a number $L(\theta, a)$ which measures the *loss* or *cost* to the experimenter when the true value of the parameter is θ and the estimate is a. $L(\theta, a)$ is a function of θ and a is called the *loss function*.

Note that in Bayesian statistics, since θ is a random variable, then $L(\theta, a)$ is also a random variable.

The idea of estimation in a Bayesian context consists of
1. Choosing a functional form for $L(\theta, a)$.
2. Choosing the value of a which minimises $E(L(\theta, a)|\boldsymbol{x})$.

The most common loss function is the squared-error loss function,

$$L(\theta, a) = (\theta - a)^2.$$

Theorem The value of a which minimises $E(L(\theta, a)|\boldsymbol{x})$ with the squared-error loss function is $E(\theta|\boldsymbol{x})$. Therefore, the Bayes' estimator of θ is $E(\theta|\boldsymbol{x})$.

Example S7.3 Consider example S7.2. If we observe 8 heads what is the Bayes' estimator of θ?

Our estimate of θ was $E(\theta) = \frac{1}{2}$ before we collected any data. After collecting our random sample, $\theta|x \sim$ Beta $\left(x + \frac{3}{2}, 10 - x + \frac{3}{2}\right)$. Therefore,

$$\begin{aligned} E(\theta|x) &= \frac{x + 3/2}{x + 3/2 + 10 - x + 3/2} \\ &= \frac{x + 3/2}{13} = \frac{2x + 3}{26}. \end{aligned}$$

If we observe 8 heads, then $E(\theta|x) = \frac{19}{26} = 0.731$.

Note that if we wanted to further investigate θ, the probability of obtaining a head, we could take another sample and

$$\theta|x \sim \text{Beta}\left(8 + \frac{3}{2}, 2 + \frac{3}{2}\right)$$

would become our prior distribution, and we would then compute the posterior again.

Note that if we choose a different loss function we will get a potentially different estimate for θ.