STATS 3006 Mathematical Statistics III
Lecture notes

Lecturer: Gary Glonek

School of Mathematical Sciences, University of Adelaide

Semester 1, 2018

Course Outline

The purpose of this course is to present the theory of statistical inference

It consists of the following main sections:

- Random variables
  - Univariate distributions
  - Transformations of random variables
  - Multivariate distributions
  - Transformation of multiple random variables
  - Limit Theorems
- Statistical Inference
  - Estimation
  - Testing

Learning Objectives

This is essentially a theoretical course and the main emphasis is on understanding the mathematical concepts, their proofs and their implications. When you complete this course you should:

- Have a thorough knowledge of the commonly used distributions in statistics;
- Have a sound understanding of the theory of random variables;
- Have a sound understanding of the principles and theory of statistical inference;
- Be able to understand mathematical proofs of results concerning random variables and statistical inference;
- Be able to construct and present mathematical proofs of results concerning random variables and statistical inference.

Why is this important?

- The material in this course provides a framework and the building blocks you need to understand the methods introduced in the other statistics courses.
- If you work as a statistician, you will need to learn about statistical methods not discussed in our program. This course provides a foundation for you to be able to understand those methods.
- If you proceed to honours or higher level study in statistics, this course provides the foundation for a deeper understanding of the subject.

How to study this course

- The most important part of your learning will be through completing the tutorial and assignment questions.
- These will be mostly theoretical questions designed:
  - To be do-able by yourself based on the lecture material;
  - To be appropriately challenging;
  - To complement or reinforce the lecture material.
- The single most important thing is to work through the questions yourself.
  - You can easily find solutions on the internet.
  - You may get good assignment marks but will short circuit your learning.
  - Allow enough time: about 10 hours for an assignment and 5 hours for a tutorial.

Random Variables

We consider *discrete* and *continuous* random variables.

We use upper case, $X,\ Y,\ Z$, to denote random variables and lower case $x,\ y\ z$, to denote particular values.

A discrete random variable (RV) is described by its *probability function*

$$p(x) = P(\{X = x\})$$

and is represented by a probability histogram.

A continuous RV is described by its *probability density function* (PDF), $f(x) \geq 0$, for which

$$\begin{aligned} & P(\{a \leq X \leq b\}) \\ &= \int_a^b f(x)\,dx, \qquad \text{for all } a \leq b. \end{aligned}$$

Probability Density Function

In order to be a valid PDF, $f(x)$ must satisfy the following:

- $f(x) \geq 0$ for all x;
- $f(x)$ must be an integrable function;
- $$\int_{-\infty}^{\infty} f(x)dx = 1.$$

For a continuous random variable, X, and any constant, a,

$$P(X = a) = \int_{a}^{a} f(x)\, dx = 0.$$

Not all RVs are discrete or continuous

Example Daily rainfall is neither discrete nor continuous:

- On days when it rains, it is reasonable to model rain as a continuous variable.
- But on many days it will not rain at all, so the probability of exactly zero rainfall is non-zero.

Cumulative Distribution Function

The cumulative distribution function (CDF) is defined for any type of random variable by

$$F(x) = P(\{X \leq x\}).$$

It is computed by

$$F(x) = \begin{cases} \sum_{t:t\leq x} p(t) & \text{for } X \text{ discrete} \\ \int_{-\infty}^{x} f(t)dt & \text{for } X \text{ continuous.} \end{cases}$$

Expected Value

The *expected value* $E(X)$ of a *discrete* RV is given by

$$E(X) = \sum_x xp(x),$$

provided that $\sum_x |x|p(x) < \infty$ Otherwise the expectation is *not* defined.

The *expected value* of a *continuous* RV is given by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx,$$

provided that $\int_{-\infty}^{\infty} |x|f(x)dx < \infty$, otherwise it is *not* defined.

Expected Value of a Function

If X is a random variable and h is a function then

$$E\{h(X)\} = \begin{cases} \sum_x h(x)p(x) & \text{for } X \text{ discrete} \\ \int_{-\infty}^{\infty} h(x)f(x)\,dx & \text{for } X \text{ continuous} \end{cases}$$

Notes:

- As previously, the expectation is defined only when the sum or integral is absolutely convergent.
- We will say later that this statement is a theorem rather than a definition.

Moment Generating Function

The *moment generating function* (MGF) of a RV X is defined to be:

$$M_X(t) = E\left[e^{tX}\right] = \begin{cases} \sum_x e^{tx} p(x) & \text{for } X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{tx} f(x) dx & \text{for } X \text{ continuous} \end{cases}$$

$M_X(0) = 1$ for any distribution.

The mgf may or may not be defined for other values of t.

Moment Generating Function (continued)

If $M_X(t)$ defined for all t in some open interval containing 0, then:

- Moments of all orders exist;
- $E[X^r] = M_X^{(r)}(0)$

$$M'(0) = E(X)$$
$$M''(0) = E(X^2), \qquad \text{and so on.}$$

- $M_X(t)$ uniquely determines the distribution of X.

Discrete Distributions
The Bernoulli Distribution

Parameter $0 \leq p \leq 1$.

Notation $X \sim \text{Bern}(p)$.

Possible values $\{0, 1\}$.

Probability function

$$p(x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases} \qquad = p^x (1-p)^{1-x}$$

Moments $E(X) = p$ and $\text{var}(X) = p(1-p)$.

MGF $M(t) = 1 + p(e^t - 1)$.

The Binomial Distribution

Parameters $0 \leq p \leq 1$, $n \in \mathbb{N}$.

Notation $X \sim B(n, p)$.

Possible values $\{0, 1, \ldots, n\}$

Probability function $p(x) = \binom{n}{x} p^x (1-p)^{n-x}$.

Moments $E(X) = np$ and $\text{var}(X) = np(1-p)$.

MGF $M(t) = (1 + p(e^t - 1))^n$.

Genesis Consider set of n independent Bernoulli trials each with success probability p. The total number of successes, X, is $B(n, p)$.

Note The binomial distribution is also the distribution of the sum of independent Bernoulli variables.

The Geometric Distribution

Parameter $0 < p < 1$.

Notation $X \sim \text{Geom}(p)$.

Possible values $\{0, 1, 2, \ldots\}$

Probability function $p(x) = (1-p)^x p$.

Moments $E(X) = \frac{1-p}{p}$ and $\text{var}(X) = \frac{1-p}{p^2}$.

MGF $M(t) = \frac{p}{1-(1-p)e^t}$ for $t < -\log(1-p)$.

Genesis Consider sequence of independent Bernoulli trials each with success probability p. The total number of failures, X, preceding the first success is Geom(p).

The Geometric Distribution (2)

In some treatments, the geometric variable Y is defined to be the position of the first success in the sequence. The distribution of Y is obtained by modifying that of X.

Possible values $\{1, 2, 3, \ldots\}$

Probability function $p(x) = (1-p)^{y-1}p$.

Moments $E(Y) = \frac{1}{p}$ and $\text{var}(Y) = \frac{1-p}{p^2}$.

MGF $M_Y(t) = \frac{pe^t}{1-(1-p)e^t}$

The Negative Binomial Distribution

Parameter $0 < p < 1$, $n \in \mathbb{N}$.

Notation $X \sim NB(n, p)$.

Possible values $\{0, 1, 2, \ldots\}$

Probability function $p(x) = \binom{n+x-1}{n-1}(1-p)^x p^n$.

Moments $E(X) = \frac{n(1-p)}{p}$ and $\text{var}(X) = \frac{n(1-p)}{p^2}$.

MGF $M(t) = \left(\frac{p}{1-(1-p)e^t}\right)^n$

Genesis Consider sequence of independent Bernoulli trials each with success probability p. The total number of failures, X, preceding the n^{th} success is $NB(n, p)$.

Note The negative binomial distribution is the distribution of the sum of independent geometric variables. When $n = 1$ the negative binomial distribution reduces to the geometric distribution.

The Poisson Distribution

Parameter $\mu > 0$.

Notation $X \sim \text{Po}(\mu)$.

Possible values $\{0, 1, 2, \ldots\}$

Probability function $p(x) = \frac{e^{-\mu}\mu^x}{x!}$.

Moments $E(X) = \mu$ and $\text{var}(X) = \mu$.

MGF $M(t) = e^{\mu(e^t - 1)}$.

Genesis

1. The Poisson distribution arises as the distribution for the number of "point events" observed in a Poisson process.
2. The Poisson distribution also arises as the limit of the binomial distribution, $B(n, p)$ as

$$n \to \infty; \; p \to 0 \text{ such that } np \to \mu.$$

The Poisson Process

The Poisson Process with rate $\lambda > 0$ is a point process on $[0, \infty)$ that satisfies the following axioms.

1. The numbers of occurrences in disjoint intervals are independent.
2. The probability of 1 or more occurrences in any interval $[t, t + h)$ is $\lambda h + o(h)$ as $h \to 0$.
3. The probability of more than one occurrence in any interval $[t, t + h)$ is $o(h)$ as $h \to 0$.

$o(h)$ notation

$o(h)$ pronounced "small order h" is standard notation for any function $r(h)$ with the property

$$\lim_{h \to 0} \frac{r(h)}{h} = 0.$$

Examples

- $r(h) = h^2$ is $o(h)$ because

$$\lim_{h \to 0} \frac{r(h)}{h} = \lim_{h \to 0} \frac{h^2}{h} = \lim_{h \to 0} h = 0.$$

- $r(h) = h/2$ is not $o(h)$ because

$$\lim_{h \to 0} \frac{r(h)}{h} = \lim_{h \to 0} \frac{h/2}{h} = \lim_{h \to 0} 1/2 = 1/2 \neq 0.$$

- $r(h) = 100h^3$ is $o(h^2)$.

Derivation of Poisson Dist from Poisson Process

Consider a Poisson Process with rate $\lambda > 0$ and let X be the number of occurrences in the fixed interval $[0, t)$.

Then X has the Poisson distribution $X \sim \text{Po}(\lambda t)$.

A derivation via the Poisson limit for the binomial distribution is as follows.

- Divide the interval $[0, t)$ into n sub-intervals

$$[0, t/n),\ [t/n, 2t/n), \ldots, [(n-1)t/n, t).$$

- Define Y_i for $i = 1, 2, \ldots, n$

$$Y_i = \begin{cases} 1 & \text{if there is at least one occurrence in } [(i-1)t/n, it/n) \\ 0 & \text{otherwise.} \end{cases}$$

Derivation Poisson Dist from Poisson Process (2)

- Then $Y_1, Y_2, \ldots, Y_n$ are independent Bernoulli-p_n variables.
- So $X_n = \sum_{i=1}^{n} Y_i \sim B(n, p_n)$.
- By axiom 2 of the Poisson Process,

$$\lim_{n\to\infty} p_n = 0, \quad \lim_{n\to\infty} np_n = \lambda t.$$

- Hence, as $n \to \infty$ the distribution of X_n is $\text{Po}(\lambda t)$.
- For finite n, X_n may not be the total number of occurrences, as a sub-interval may have multiple occurrences.
- By axiom 3 of the Poisson process, the probability of multiple occurrences in any sub-interval becomes negligible as $n \to \infty$.

The Hypergeometric Distribution

Parameters $M, N \in \mathbb{N}$, $0 < n < M + N$.

Notation $X \sim \text{hyper}(M, N, n)$.

Possible values $\max(0, n - N) \leq x \leq \min(n, M)$.

Probability function $p(x) = \frac{\binom{M}{x}\binom{N}{n-x}}{\binom{M+N}{n}}$.

Moments $E(X) = \frac{M}{M+N}$ and $\text{var}(X) = \frac{M+N-n}{M+N-1} \times \frac{nMN}{(M+N)^2}$.

MGF Exists, but no useful expression.

Genesis Consider an urn containing M black and N white balls. Suppose n balls are sampled randomly *without* replacement and let X be the number of black balls chosen. Then X has a hypergeometric distribution.

The Hypergeometric Distribution (remarks)

- The possible values for X arise as follows.
  - The number of black balls in the sample cannot exceed the number of balls in the sample n and also number of black balls in the urn. Hence $X \leq \min(n, M)$.
  - For the total number of white balls in the sample, we obtain $n - X \leq n$ and also $n - X \leq N$ whereby $\max(0, n - N) \leq X$.
- When M, N are large compared to n, the hypergeometric distribution is approximated by the $B(n, p)$ distribution with $p = M/(M + N)$.
- Note the moment of the hypergeometric distribution are

$$E(X) = np \text{ and } \text{var}(X) = \frac{M + N - n}{M + N - 1} \times np(1 - p).$$

Continuous Distributions

The Uniform Distribution

Parameters $a < b \in \mathbb{R}$.

Notation $X \sim U(a, b)$.

Possible values $a < x < b$.

Probability density function (PDF) $f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{otherwise.} \end{cases}$

Cumulative Distribution Function (CDF)

$$f(x) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } a < x < b \\ 1 & \text{for } b \leq x \end{cases}$$

The Uniform Distribution

Moments $E(X) = \frac{a+b}{2}$ and $\text{var}(X) = \frac{(b-a)^2}{12}$.

MGF $M(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$.

Genesis Consider suppose a Poisson process is observed and it is known that there is exactly one occurrence in the the interval (a, b). Then the exact time of the occurence is $U(a, b)$.

Special case $U(0, 1)$ is the uniform distribution on the unit interval with PDF $f(x) = \begin{cases} 1 & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$

Exponential Distribution

Parameters $\lambda > 0$

Notation $X \sim \text{Exp}(\lambda)$

Possible values $x > 0$.

PDF $f(x) = \lambda e^{-\lambda x}$ for $x > 0$.

CDF $F(x) = 1 - e^{-\lambda x}$ for $x > 0$.

Moments $E(X) = 1/\lambda$ and $\text{var}(X) = 1/\lambda^2$.

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

Genesis The exponential distribution arises as the waiting time until the first occurrence in a Poisson process with rate parameter λ.

Note The exponential distribution has the memoryless property,

$$P(X > s + t | X > s) = P(X > t).$$

The Gamma Distribution

Parameters $\alpha > 0,\ \lambda > 0$

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

Possible values $x > 0$.

PDF $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ for $x > 0$.

CDF No simple expression.

Moments $E(X) = \alpha/\lambda$ and $\text{var}(X) = \alpha/\lambda^2$.

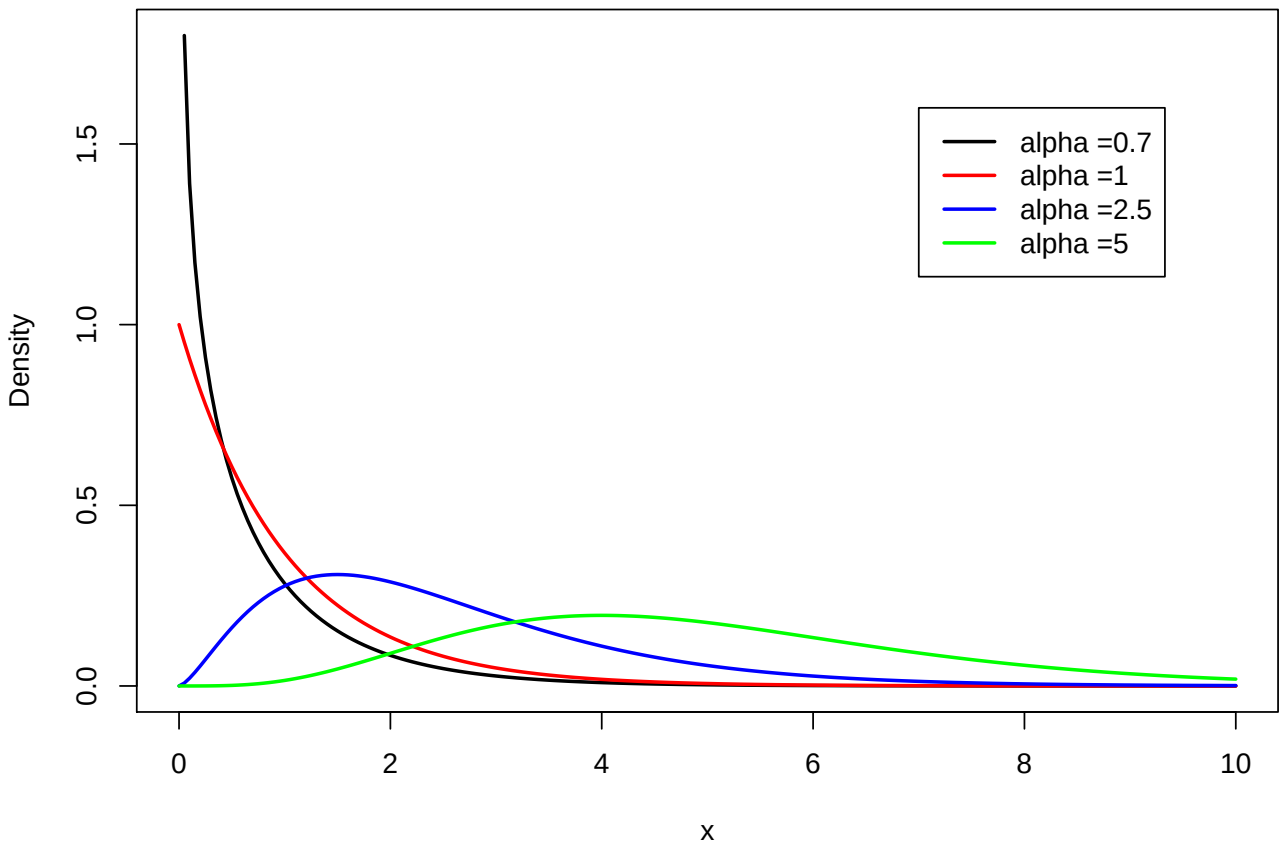
MGF $M(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha$.

Genesis If $Y_1, Y_2, \ldots, Y_r$ are indepedendent $\text{Exp}(\lambda)$ random variables and $X = Y_1 + Y_2 + \ldots + Y_r$ then $X \sim \text{Gamma}(r, \lambda)$. That is, X is the waiting time until the r^{th} occurrence in a Poisson process with rate λ.

The Gamma Distribution Remarks

- The Gamma $\left(\frac{r}{2}, \frac{1}{2}\right)$ distribution is also the χ^2_r distribution.
- The parameters α and λ are respectively called the shape parameter and the scale parameter.
- If $X \sim \text{Gamma}(\alpha, 1)$ and $Y = X/\lambda$ then $Y \sim \text{Gamma}(\alpha, \lambda)$.
- Examples of Gamma distributions with different shape parameters are shown below.

Gamma Densities



The Beta Distribution

Parameters $\alpha > 0,\ \beta > 0$

Notation $X \sim \text{Beta}(\alpha, \beta)$

Possible values $0 < x < 1.$

PDF $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ for $0 < x < 1.$

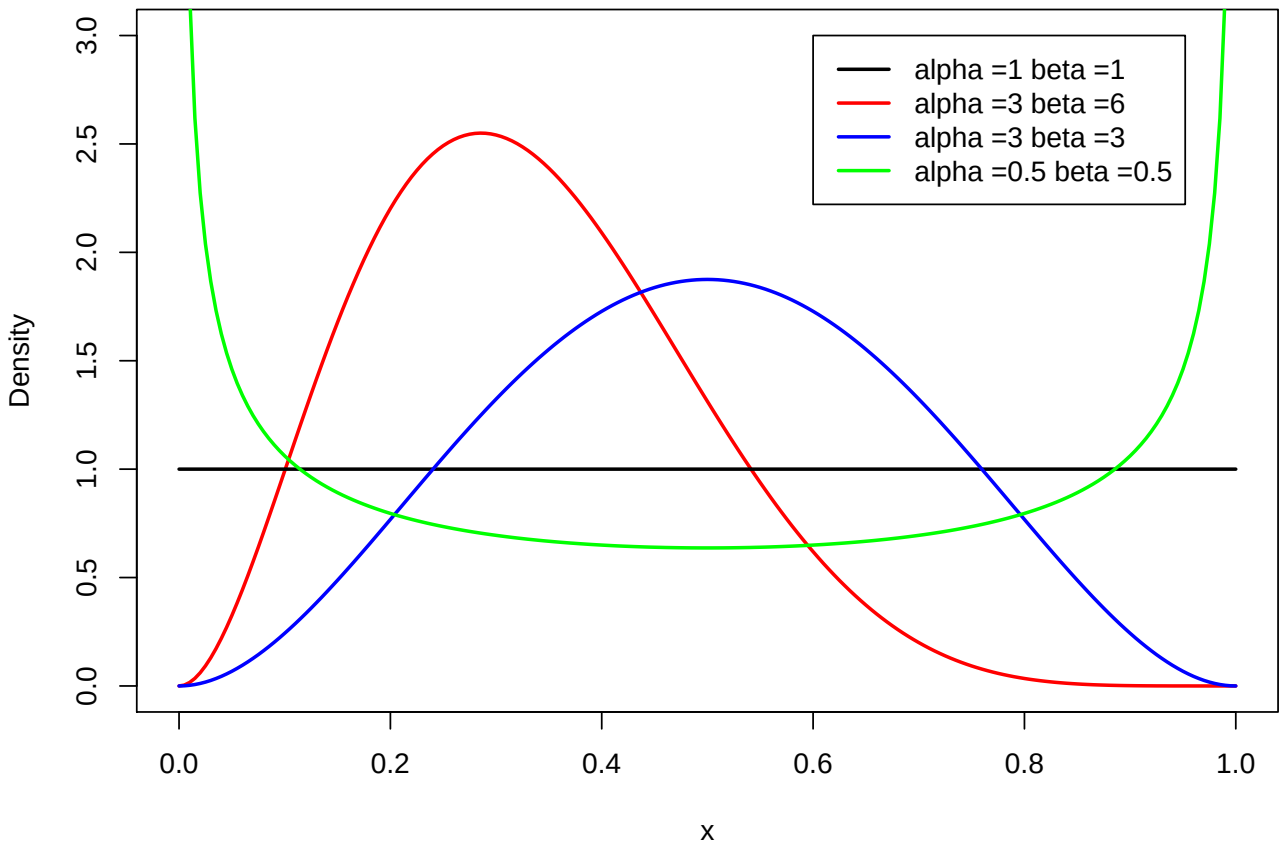
CDF No simple expression.

Moments $E(X) = \frac{\alpha}{\alpha+\beta}$ and $\text{var}(X) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$

MGF Exists but no simple expression.

Genesis If $Y_1 \sim \text{Gamma}(\alpha, \lambda)$ and $Y_2 \sim \text{Gamma}(\beta, \lambda)$ independently then $\frac{Y_1}{Y_1+Y_2} \sim \text{Beta}(\alpha, \beta)$.

Beta Densities



The Normal Distribution

Parameters $\mu \in \mathbb{R},\ \sigma^2 > 0$

Notation $X \sim N(\mu, \sigma^2))$

Possible values $x \in \mathbb{R}$.

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

CDF No simple expression.

Moments $E(X) = \mu$ and $\text{var}(X) = \sigma^2$.

MGF $M(t) = e^{\mu t + \sigma^2 t^2/2}$.

The Standard Normal Distribution

- When the $\mu = 0$ and $\sigma^2 = 1$ we obtain the standard normal distribution, $N(0, 1)$.
- The following notation is commonly used:
  - The symbol Z is used to denote an $N(0, 1)$ variable.
  - The standard normal PDF is denoted $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$.
  - The standard normal CDF is denoted

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt.$$

The Cauchy Distribution

Possible values $x \in \mathbb{R}$.

PDF $f(x) = \frac{1}{\pi} \times \frac{1}{1+x^2}$.

CDF $F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$.

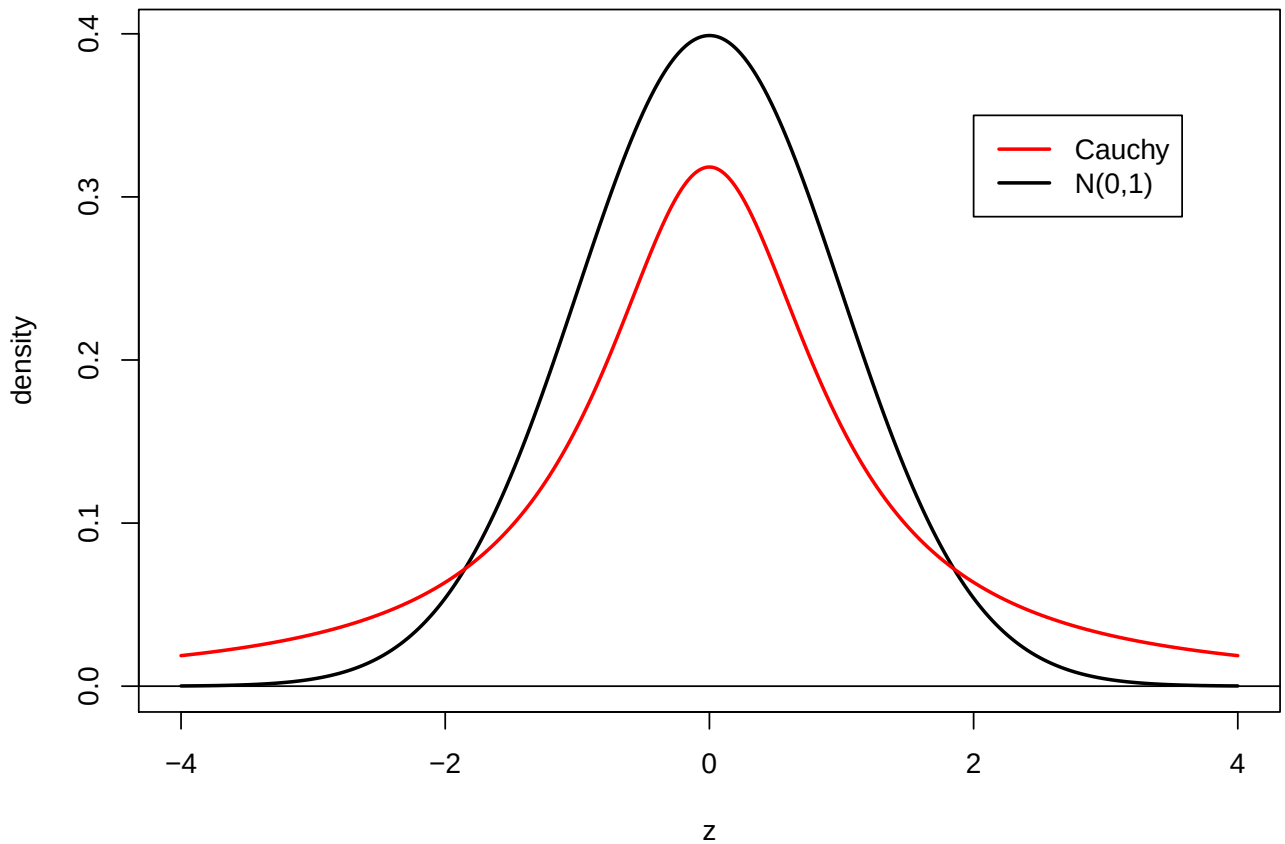
Moments Do not exist.

MGF Does not exist.

Genesis If $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,2)$ independently then Z_1/Z_2 has the Cauchy distribution.

Remark The Cauchy density appears superficially like the normal density but is pointier at the centre and has much heavier tails.

Cauchy and Normal Densities



Transformations of a Random Variable

If X is a RV and $Y = h(X)$, then Y is also a RV. If we know the distribution of X, for a given function $h(x)$, we should be able to find the distribution of $Y = h(X)$.

Theorem 1

Suppose X is a discrete RV with prob. function $p_X(x)$ and let $Y = h(X)$, where h is any function then:

$$p_Y(y) = \sum_{x:h(x)=y} p_X(x)$$

Note the sum is over all values x for which $h(x) = y$.

Proof of Theorem 1

Proof.

$$\begin{aligned} p_Y(y) &= P(Y = y) = P\{h(X) = y\} \\ &= \sum_{x:h(x)=y} P(X = x) \\ &= \sum_{x:h(x)=y} p_X(x) \end{aligned}$$

□

Monotonic transformations for continuous RVs

Theorem 2

Suppose X is a continuous RV with PDF $f_X(x)$ and let $Y = h(X)$, where $h(x)$ is differentiable and monotonic, i.e., either strictly increasing **or** *strictly decreasing.*

Then the PDF of Y is given by:

$$f_Y(y) = f_X\left(h^{-1}(y)\right)|h^{-1}(y)'|$$

Proof of Theorem 2

Assume first that h is increasing. Then

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(h(X) \leq y) \\ &= P(X \leq h^{-1}(y)) \\ &= F_X\left(h^{-1}(y)\right). \end{aligned}$$

Differentiating with respect to Y, we obtain

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(h^{-1}(y)\right) = f_X\left(h^{-1}(y)\right) h^{-1}(y)' \quad (1)$$

Proof of Theorem 2 (continued)

Now consider the case of $h(x)$ decreasing.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(h(X) \leq y) \\ &= P(X \geq h^{-1}(y)) \\ &= 1 - F_X\left(h^{-1}(y)\right) \end{aligned}$$

and hence

$$f_Y(y) = -f_X\left(h^{-1}(y)\right) h^{-1}(y)'. \qquad (2)$$

Proof of Theorem 2 (continued)

Finally, observe that if h is increasing then $h'(x)$ and hence $h^{-1}(y)'$ must be positive. Similarly for h decreasing, $h^{-1}(y)' < 0$. Hence (1) and (2) can be combined to give:

$$f_Y(y) = f_X\left(h^{-1}(y)\right)|h^{-1}(y)'|$$

□

Example: linear transformations

Suppose X is a continuous random variable and let $Y = h(X) = aX + b$.

Then $h^{-1}(y) = \frac{y-b}{a}$ and $h^{-1}(y)' = \frac{1}{a}$.

Hence, by Theorem 2,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Examples of linear transformations

- If $Z \sim N(0,1)$ and $X = \mu + \sigma Z$ with $\sigma > 0$ then $X \sim N(\mu, \sigma^2)$.
- If $X \sim N(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$ then $Z \sim N(0,1)$.
- If $X \sim \text{Gamma}(\alpha, 1)$ and $Y = X/\lambda$ with $\lambda > 0$ then $Y \sim \text{Gamma}(\alpha, \lambda)$.

The CDF transformation

Suppose X is a continuous RV with CDF $F_X(x)$, which is increasing over the range of X.

If $U = F_X(x)$, then $U \sim U(0,1)$.

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P\{F_X(X) \leq u\} \\ &= P\{X \leq F_X^{-1}(u)\} \\ &= F_X\{F_X^{-1}(u)\} \\ &= u, \qquad \text{for } 0 < u < 1. \end{aligned}$$

This is simply the CDF of $U(0,1)$, so the result is proved.

□

The inverse CDF transformation

The converse also applies. If $U \sim U(0,1)$ and F is any strictly increasing "on its range" CDF, then $X = F^{-1}(U)$ has CDF $F(x)$, i.e.,

$$\begin{aligned} F_X(x) &= P(X \leq x) = P\{F^{-1}(U) \leq x\} \\ &= P(U \leq F(x)) \\ &= F(x), \quad \text{as required.} \end{aligned}$$

□

This result is used in the generation of pseudo-random numbers.

If we can obtain $U(0,1)$ random numbers, we can use this result to transfrom them to random numbers with CDF $F(x)$.

Non-monotonic transformations

Theorem 2 applies only to monotonic transformations of a continuous RV.

If $h(x)$ is *not* monotonic, then $h(x)$ may not even be continuous.

For example, if $h(x) = [x]$ (the integer part of x) then the possible values for $Y = h(X)$ are integers.

However, if $h(x)$ is piecewise monotonic and X is a continuous random variable, then $h(X)$ is also a continuous random variable.

Example

Suppose $Z \sim N(0,1)$ and let $X = Z^2$.

This transformation is not monotonic.

Can find the PDF of X as follow.

Observe

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(Z^2 \leq x) \\ &= P(-\sqrt{x} \leq Z \leq \sqrt{x}) \\ &= \Phi(\sqrt{x}) - \Phi(-\sqrt{x}) \end{aligned}$$

where Φ is the $N(0,1)$ CDF.

Example (continued)

Hence

$$f_X(x) = F'_X(x) = \frac{1}{2}x^{-\frac{1}{2}}\phi(\sqrt{x}) + \frac{1}{2}x^{-\frac{1}{2}}\phi(-\sqrt{x})$$

where ϕ is the $N(0,1)$ PDF,

$$\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}.$$

Substituting for ϕ we obtain

$$f_X(x) = \frac{(\frac{1}{2})^{\frac{1}{2}}}{\sqrt{\pi}}x^{-\frac{1}{2}}e^{-\frac{1}{2}x}$$

which is the Gamma$(\frac{1}{2}, \frac{1}{2}) \equiv \chi^2_1$ distribution.

Note $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Moments of transformed variables

Suppose X is a RV and let $Y = h(X)$.

To find $E(Y)$,

1. Find the distribution of $Y = h(X)$ using preceding methods.
2. Find $E(Y) = \begin{cases} \sum_y yp(y) & Y \text{ discrete} \\ \int_{-\infty}^{\infty} yf(y)\,dy & Y \text{ continuous} \end{cases}$

Moments of transformed variables

Alternatively can use

Theorem 3

If X is a RV of either discrete or continuous type and $h(x)$ is any transformation (not necessarily monotonic), then $E(h(X))$ (provided it exists) is given by:

$$E(h(X)) = \begin{cases} \sum_x h(x)p(x) & X \quad \textit{discrete} \\ \int_{-\infty}^{\infty} h(x)f(x)dx & X \quad \textit{continuous} \end{cases}$$

□

The proof of this theorem is omitted.

Corollaries

1. If $E(X) = \mu$ and $\text{var}(X) = \sigma^2$, and $Y = aX + b$ for constants a, b, then $E(Y) = a\mu + b$ and $\text{var}(Y) = a^2\sigma^2$.
2. If $Y = h(X)$ then the MGF of Y, provided it exists, is

$$M_Y(t) = \begin{cases} \sum_x e^{t.h(x)} p(x) & \text{for } X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{t.h(x)} f(x) dx & \text{for } X \text{ continuous} \end{cases}$$

Example the CDF transformation

Suppose X is continuous with CDF $F(x)$, and $F(a) = 0,\ F(b) = 1$; (a, b can be $\pm\infty$ respectively).

Let $U = F(X)$. Observe that

$$\begin{aligned} M_U(t) &= \int_a^b e^{tF(x)} f(x)\, dx \\ &= \frac{1}{t} e^{tF(x)} \Big|_a^b \\ &= \frac{e^{tF(b)} - e^{tF(a)}}{t} \\ &= \frac{e^t - 1}{t}, \end{aligned}$$

which is the $U(0,1)$ MGF.

Multivariate distributions

Consider random variables $X_1, X_2, \ldots, X_r$.

When all variables are discrete, the *joint* distribution is described by the *joint probability function*

$$p(\boldsymbol{x}) = p(x_1, x_2, \ldots, x_r) = P(X_1 = x_1, X_2 = x_2, \ldots, X_r = x_r).$$

To be a valid probability function, $p(\boldsymbol{x})$ must satisfy

$$p(\boldsymbol{x}) \geq 0 \text{ for all } \boldsymbol{x} \text{ and } \sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1.$$

Continuous multivariate distributions

When all variables are discrete, the joint distribution is described by the *joint probability density function*

$$f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_r)$$

with the property

$$P(\boldsymbol{X} \in \mathcal{A}) = \int_{\mathcal{A}} f(x_1, x_2, \ldots, x_r) dx_1 dx_2 \ldots dx_r$$

for any *measurable set* $\mathcal{A}$.

To be a valid PDF, $f(\boldsymbol{x})$ must satisfy $f(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_r) dx_1 dx_2 \ldots dx_r = 1.$$

The joint CDF and independence

For random variables of all types, the *joint CDF* is defined by

$$F(\boldsymbol{x}) = F(x_1, x_2, \ldots, x_r) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_r \leq x_r).$$

Definition 4

The variables $X_1, X_2, \ldots, X_r$ are said to be mutually independent if

$$F(x_1, x_2, \ldots, x_r) = F_{X_1}(x_1) F_{X_2}(x_2) \cdots F_{X_r}(x_r)$$

for all $x_1, x_2, \ldots, x_r$.

Marginal distributions

Consider a random vector, $\boldsymbol{X} = (X_1, X_2, \ldots, X_r)$ and let

$$\boldsymbol{X}_1 = (X_1, X_2, \ldots, X_{r_1}) \text{ and } \boldsymbol{X}_2 = (X_{r_1+1}, X_{r_1+2}, \ldots, X_r).$$

Definition 5

Suppose $\boldsymbol{X}$ is discrete with joint probability function

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_1, \boldsymbol{x}_2) = p(x_1, x_2, \ldots, x_{r_1}, x_{r_1+1}, \ldots, x_r).$$

The marginal probability function of $\boldsymbol{X}_1$ is

$$\begin{aligned} p_{X_1}(\boldsymbol{x}_1) &= \sum_{\boldsymbol{x}_2} p(\boldsymbol{x}_1, \boldsymbol{x}_2) \\ &= \sum_{x_{r_1+1}} \sum_{x_{r_1+2}} \cdots \sum_{x_r} p(x_1, x_2, \ldots, x_{r_1}, x_{r_1+1}, \ldots, x_r). \end{aligned}$$

Marginal density

Definition 6

Suppose $\boldsymbol{X}$ is continuous with probability density function

$$f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_{r_1}, x_{r_1+1}, \ldots, x_r).$$

The marginal PDF of $\boldsymbol{X}_1$ is

$$f_{X_1}(\boldsymbol{x}_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_{r_1}, x_{r_1+1}, \ldots, x_r) dx_{r_1+1} \ldots dx_r.$$

Conditional distributions

Definition 7

Suppose $\boldsymbol{X}$ is discrete with joint probability function

$$p(\boldsymbol{x}) = p(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

The conditional probability function of $\boldsymbol{X}_2 | \boldsymbol{X}_1 = \boldsymbol{x}_1$ is

$$p_{X_2|X_1}(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2)}{p_{X_1}(\boldsymbol{x}_1)}.$$

provided $p_{X_1}(\boldsymbol{x}_1) > 0$.

The interpretation of the conditional probability function is

$$p_{X_2|X_1}(\boldsymbol{x}_2, \boldsymbol{x}_1) = P(\boldsymbol{X}_2 = \boldsymbol{x}_2 | \boldsymbol{X}_1 = \boldsymbol{x}_1).$$

Independence for discrete variables

Theorem 8

The discrete random variables $X_1, X_2, \ldots, X_r$ are independent if and only if their joint probability function satisfies either of the following two conditions:

- $$p(x_1, x_2, \ldots, x_r) = p_{X_1}(x_1) p_{X_2}(x_2) \cdots p_{X_r}(x_r)$$

  for all $x_1, x_2, \ldots, x_r$;

- $$p(x_1, x_2, \ldots, x_r) = \psi_1(x_1)\psi_2(x_2) \cdots \psi_r(x_r)$$

  for all $x_1, x_2, \ldots, x_r$ and some functions $\psi_1, \psi_2, \ldots, \psi_r$.

Conditional densities

Definition 9

Suppose $\boldsymbol{X}$ is continuous with joint PDF

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

The conditional density function of $\boldsymbol{X}_2|\boldsymbol{X}_1$ is

$$f_{X_2|X_1}(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{X_1}(\boldsymbol{x}_1)}.$$

provided $f_{X_1}(\boldsymbol{x}_1) > 0$.

Interpretation of conditional density

The conditional density does not, strictly speaking, produce conditional probabilities.

For example,

$$\int_a^b f_{X_2|X_2}(x_2|x_1)dx_2$$

is not

$$P(a < X_2 < b|X_1 = x_1)$$

because

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

is defined only for $P(B > 0)$.

But, for a continuous random variable, $P(X_1 = x_1) = 0$.

Interpretation of conditional density

However, if $f(x_1, x_2)$ is a continuous function of x_1, x_2, it can be checked that

$$\int_a^b f_{X_2|X_1}(x_2|x_1)dx_2 = \lim_{\varepsilon \to 0} P(a \leq X_2 \leq b | x_1 - \varepsilon < X_1 < x_1 + \varepsilon).$$

Independence for continuous variables

Theorem 10

The continuous random variables $X_1, X_2, \ldots, X_r$ are independent if and only if their joint PDF satisfies either of the following two conditions:

- $$f(x_1, x_2, \ldots, x_r) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_r}(x_r)$$

  for all $x_1, x_2, \ldots, x_r$;

- $$f(x_1, x_2, \ldots, x_r) = \psi_1(x_1)\psi_2(x_2) \cdots \psi_r(x_r)$$

  for all $x_1, x_2, \ldots, x_r$ and some functions $\psi_1, \psi_2, \ldots, \psi_r$.

Moments for multivariate distributions

Suppose $\boldsymbol{X} = (X_1, X_2, \ldots, X_r)$ be random vector and let $h(x_1, x_2, \ldots, x_r)$ be a real valued function.

Theorem 11

Provided it exists, the expected value can be computed by

$$E(h(X_1, X_2, \ldots, X_r)$$

$$= \begin{cases} \sum_{x_1} \sum_{x_2} \cdots \sum_{x_r} h(x_1, x_2, \ldots, x_r) p(x_1, x_2, \ldots, x_r) \\ \textit{in the discrete case} \\ \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, x_2, \ldots, x_r) f(x_1, x_2, \ldots, x_r) dx_1 dx_2 \cdots dx_r \\ \textit{in the continuous case} \end{cases}$$

□

Linear combinations

Theorem 12

Suppose $\boldsymbol{X}$ *is a random vector and let* $h_1,\ h_2, \ldots,\ h_k$ *be real valued functions and* $a_1,\ a_2, \ldots, a_k$ *be real constants. Then*

$$\begin{aligned} & E(a_1 h_1(\boldsymbol{X}) + a_2 h_2(\boldsymbol{X}) + \ldots + a_k h_k(\boldsymbol{X})) \\ = \quad & a_1 E(h_1(\boldsymbol{X})) + a_2 E(h_2(\boldsymbol{X})) + \ldots a_k E(h_k(\boldsymbol{X})) \end{aligned}$$

Proof (continuous case)

$$\begin{aligned} & E(a_1 h_1(\boldsymbol{X}) + a_2 h_2(\boldsymbol{X}) + \ldots + a_k h_k(\boldsymbol{X})) \\ = \quad & \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (a_1 h_1(\boldsymbol{x}) + \ldots + a_k h_k(\boldsymbol{x})) f(\boldsymbol{x}) dx_1 dx_2 \cdots dx_r \end{aligned}$$

Proof (continued)

$$
\begin{aligned}
= \quad & a_1 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_1(\boldsymbol{x}) f(\boldsymbol{x}) dx_1 dx_2 \cdots dx_r \\
& + a_2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_2(\boldsymbol{x}) f(\boldsymbol{x}) dx_1 dx_2 \cdots dx_r \\
& \vdots \\
& + a_k \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_k(\boldsymbol{x}) f(\boldsymbol{x}) dx_1 dx_2 \cdots dx_r \\
= \quad & a_1 E(h_1(\boldsymbol{X})) + a_2 E(h_2(\boldsymbol{X})) + \ldots + a_k E(h_k(\boldsymbol{X})).
\end{aligned}
$$

□

Linear combinations

Corollary 13

If $a_1,\ a_2, \ldots,\ a_k$ *are constants then*

$$E(a_1X_1 + a_2X_2 + \ldots + a_kX_k) = a_1E(X_1) + a_2E(X_2) + \ldots + a_kE(X_k).$$

□

Remark

This result does not require the assumption of independence.

Covariance and correlation

Definition 14

If X_1 and X_2 are random variables with $E(X_i) = \mu_i$ and $\text{var}(X_i) = \sigma_i^2$ then the covariance is

$$\text{cov}(X_1, X_2) = \sigma_{12} = E((X_1 - \mu_1)(X_2 - \mu_2))$$

and the correlation coefficient is

$$\text{cor}(X_1, X_2) = \rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}.$$

Correlation and independence

Theorem 15

If X_1 and X_2 are independent, then $\text{cov}(X_1, X_2) = \text{cor}(X_1, X_2) = 0$.

Proof (continuous case)

$$
\begin{aligned}
& \text{cov}(X_1, X_2) \\
= & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_1)(x_2-\mu_2)f(x_1,x_2)dx_1dx_2 \\
= & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_1)(x_2-\mu_2)f_{X_1}(x_1)f_{X_2}(x_2)dx_1dx_2 \\
= & \left(\int_{-\infty}^{\infty}(x_1-\mu_1)f_{X_1}(x_1)dx_1\right)\times\left(\int_{-\infty}^{\infty}(x_2-\mu_2)f_{X_2}(x_2)dx_2\right) \\
= & \ 0\times 0 = 0
\end{aligned}
$$

□

Correlation and independence

Remark

The converse does not apply. In particular, $\text{cor}(X_1, X_2) = 0$ does not imply that X_1 and X_2 are independent.

Covariance of linear combinations

Theorem 16

Suppose $X_1, X_2, \ldots, X_r$ *are random variables such that*

$$E(X_i) = \mu_i, \ \operatorname{cov}(X_i, X_j) = \sigma_{ij} \ (\operatorname{var}(X_i) = \sigma_{ii})$$

and let $a, a_2, \ldots, a_r, b_1, b_2, \ldots, b_r$ *be constants. Then*

$$\operatorname{cov}\left(\sum_{i=1}^{r} a_i X_i, \sum_{j=1}^{r} b_j X_j\right) = \sum_{i=1}^{r}\sum_{j=1}^{r} a_i b_j \sigma_{ij}.$$

Proof

$$
\begin{aligned}
\operatorname{cov}\left(\sum_{i=1}^{r} a_i X_i, \sum_{j=1}^{r} b_j X_j\right) &= E\left(\left\{\sum_{i=1}^{r} a_i (X_i - \mu_i)\right\} \times \left\{\sum_{j=1}^{r} b_j (X_j - \mu_j)\right\}\right) \\
&= E\left(\sum_{i=1}^{r}\sum_{j=1}^{r} a_i b_j (X_i - \mu_i)(X_j - \mu_j)\right) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r} a_i b_j E((X_i - \mu_i)(X_j - \mu_j)) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{r} a_i b_j \sigma_{ij}.
\end{aligned}
$$

□

Linear combinations of random variables

Corollary 17

If $a_1, a_2, \ldots, a_r$ are constants then

$$\text{var}\left(\sum_{i=1}^{r} a_i X_i\right) = \sum_{i=1}^{r}\sum_{j=1}^{r} a_i a_j \sigma_{ij}.$$

If it is also assumed that $X_1, X_2, \ldots, X_r$ are independent then

$$\text{var}\left(\sum_{i=1}^{r} a_i X_i\right) = \sum_{i=1}^{r} a_i^2 \sigma_{ii}.$$

□

Properties of the correlation coefficient

Corollary 18

For random variables X_1 and X_2, the correlation coefficient satisfies

$$|\rho_{12}| \leq 1$$

with equality if and only if X_1 satisfy a linear relationship with probability 1.

Proof

For any fixed, t,

$$0 \leq \text{var}(X_1 - tX_2) = \sigma_1^2 + t^2\sigma_2^2 - 2\sigma_{12}t.$$

Hence the quadratic $p(t) = \sigma_1^2 + t^2\sigma_2^2 - 2\sigma_{12}t$ has either one real root or no real roots.

Proof (continued)

No real roots:

$$\Delta < 0 \Leftrightarrow 4\sigma_{12}^2 - 4\sigma_1^2\sigma_2^2 < 0 \Leftrightarrow \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2} < 1 \Leftrightarrow \rho_{12}^2 < 1.$$

One real root: $\Delta = 0 \Rightarrow \rho_{12} = \pm 1$..

- There exists t_0 such that $p(t_0) = 0$
- That is, $\text{var}(X_1 - t_0 X_2) = 0$.
- In this case $X_1 - t_0 X_2 = c$ with probability 1, for some constant, c.
- That is, $X_1 = t_0 X_2 + c$ with probability 1.

□

Discrete multivariate distributions

The trinomial distribution

Dimension $r = 2$

Parameters $n \in \mathbb{N}$, π_1, π_2 such that $\pi_1 > 0, \pi_2 > 0, \pi_1 + \pi_2 < 1$.

Possible values $\{(x_1, x_2) : x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq n\}$.

Probability function

$$p(x_1, x_2) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} \pi_1^{x_1} \pi_2^{x_2} (1 - \pi_1 - \pi_2)^{n - x_1 - x_2}.$$

Genesis n independent trials, each trial results in:

Outcome 1, with probability π_1;

Outcome 2, with probability π_2;

Outcome 3, with probability $1 - \pi_1 - \pi_2$.

X_1 is the number of trials with Outcome 1 and X_2 is the number of trials with Outcome 2.

The trinomial distribution (continued)

Moments

- $E(X_i) = n\pi_i$,
- $\text{var}(X_i) = n\pi_i(1 - \pi_i)$,
- $\text{cov}(X_1, X_2) = -n\pi_1\pi_2$.

Marginal distribution $X_1 \sim B(n, \pi_1)$.

Conditional distribution $X_2|X_1 = x_1 \sim B\left(n - x_1, \frac{\pi_2}{1-\pi_1}\right)$.

The multinomial distribution

Dimension $r \geq 3$.

Parameters $n \in \mathbb{N}$, $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_r)$ s.t. $\pi_i > 0$, $\sum_{i=1}^{r} \pi_i = 1$.

Possible values $\boldsymbol{x} = (x_1, x_2, \ldots, x_r)$ s.t. $x_i \geq 0$, $\sum_{i=1}^{r} x_i = n$

Probability Function

$$p(x_1, x_2, \ldots, x_r) = \frac{n!}{x_1! x_2! \cdots x_r!} \pi_1^{x_1} \pi_2^{x_2} \ldots \pi_r^{x_r}$$

Genesis n independent trials such that each trial results in Outcome i with probability π_i, for $i = 1, 2, \ldots, r$.

X_i is the number of trials with Outcome i, for $i = 1, 2, \ldots, r$.

The multinomial distribution (continued)

Moments

- $E(X_i) = n\pi_i$,
- $\text{var}(X_i) = n\pi_i(1 - \pi_i)$,
- $\text{cov}(X_i, X_j) = -n\pi_i\pi_j$.

Remark The multinomial distribution is the extension of the trinomial distribution to a general number of outcomes.

In the trinomial formulation, the final category (Outcome 3) is obtained by subtraction $(n - x_1 - x_2)$.

In the multinomial formulation, the final category is represented by x_r but the constraint $\sum_{i=1}^{r} x_i = n$ is imposed.

Vector notation

Consider the random vector

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_r)^T,$$

with $E[X_i] = \mu_i$, $\text{var}(X_i) = \sigma_i^2 = \sigma_{ii}$, $\text{cov}(X_i, X_j) = \sigma_{ij}$.

Define the mean vector by:

$$\boldsymbol{\mu} = E(\boldsymbol{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_r \end{pmatrix}$$

and the variance matrix (covariance matrix) by:

$$\Sigma = \text{Var}(\boldsymbol{X}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1r} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2r} \\ \vdots & & \ddots & \vdots \\ \sigma_{r1} & \sigma_{n2} & \ldots & \sigma_{rr} \end{pmatrix}$$

The correlation matrix

The correlation matrix is defined by:

$$R = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1r} \\ \rho_{21} & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \rho_{r-1,r} \\ \rho_{r1} & & \cdots & 1 \end{pmatrix}$$

where

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}.$$

Linear combinations

Suppose $\boldsymbol{a} = (a_1, \ldots, a_r)^T$ is a vector of constants and observe that:

$$\boldsymbol{a}^T\boldsymbol{x} = a_1x_1 + a_2x_2 + \cdots + a_rx_r = \sum_{i=1}^{r} a_ix_i$$

Theorem 19

Suppose $\boldsymbol{X}$ *has* $E(\boldsymbol{X}) = \boldsymbol{\mu}$ *and* $\text{Var}(\boldsymbol{X}) = \Sigma$*, and let* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^r$ *be fixed vectors. Then,*

1. $E(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T\boldsymbol{\mu}$,
2. $\text{var}(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T\Sigma\boldsymbol{a}$,
3. $\text{cov}(\boldsymbol{a}^T\boldsymbol{X}, \boldsymbol{b}^T\boldsymbol{X}) = \boldsymbol{a}^T\Sigma\boldsymbol{b}$.

□

This theorem is just the re-statement of Corollary 13, Theorem 16 and Corollary 17 in matrix notation.

Linear transformations

Theorem 20

Suppose $\boldsymbol{X}$ *is a random vector with* $E(\boldsymbol{X}) = \boldsymbol{\mu}, \text{Var}(\boldsymbol{X}) = \Sigma$, *and let* $A_{p\times r}$ *and* $\boldsymbol{b} \in \mathbb{R}^p$ *be fixed.*

If $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$, *then*

$$E(\boldsymbol{Y}) = A\boldsymbol{\mu} + \boldsymbol{b} \text{ and } \text{Var}(\boldsymbol{Y}) = A\Sigma A^T.$$

□

This is also a re-statement of previously established results. Observe that if $\boldsymbol{a}_i^T$ is the i^{th} row of A, then $Y_i = \boldsymbol{a}_i^T\boldsymbol{X}$ and, moreover, the $(i,j)^{th}$ element of $A\Sigma A^T$ is

$$\boldsymbol{a}_i^T\Sigma\boldsymbol{a}_j = \text{cov}\left(\boldsymbol{a}_i^T\boldsymbol{X}, \boldsymbol{a}_j^T\boldsymbol{X}\right) = \text{cov}(Y_i, Y_j).$$

Properties of variance matrices

If $\Sigma = \text{Var}(\boldsymbol{X})$ for some random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_r)^T$, then it must satisfy certain properties.

Since $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$, it follows that Σ is a square $(r \times r)$, symmetric matrix.

Definition 21

The square, symmetric matrix M is said to be positive definite if $\boldsymbol{a}^T M \boldsymbol{a} > 0$ for every vector $\boldsymbol{a} \in \mathbb{R}^r$ such that $\boldsymbol{a} \neq \boldsymbol{0}$.

M is said to be non-negative definite if $\boldsymbol{a}^T M \boldsymbol{a} \geq 0$ for every vector $\boldsymbol{a} \in \mathbb{R}^r$.

□

Necessary and sufficient conditions

Theorem 22

It is necessary and sufficient that Σ be non-negative definite in order that it can be a variance matrix.

Proof

To demonstrate the necessity of this condition, consider the linear combination $\boldsymbol{a}^T\boldsymbol{X}$. By Theorem 19

$$0 \leq \text{var}(\boldsymbol{a}^T\boldsymbol{X}) = \boldsymbol{a}^T\Sigma\boldsymbol{a} \qquad \text{for every } \boldsymbol{a},$$

and hence Σ must be non-negative definite.

Proof (continued)

To demonstrate sufficiency, observe that every symmetric, non-negative definite matrix Σ is expressible in the form

$$\Sigma = AA^T.$$

Now let $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_r)^T$ be independent random variables with $\text{var}(Z_i) = 1$ so that

$$\text{Var}(\boldsymbol{Z}) = I_{r\times r}.$$

If $\boldsymbol{X} = A\boldsymbol{Z}$ then $\text{Var}(\boldsymbol{X}) = AA^T$ by Theorem 20, so every matrix of the form $\Sigma = AA^T$ is a variance matrix.

□

Degenerate distributions

Consider a random variable $\boldsymbol{X}$ with $\Sigma = \text{Var}(X)$ non-negative definite but not positive definite.

Then there exists a vector $\boldsymbol{a} \in \mathbb{R}^r$ such that

$$\boldsymbol{a}^T \Sigma \boldsymbol{a} = 0.$$

That is,

$$\text{var}(\boldsymbol{a}^T \boldsymbol{X}) = 0$$

for this $\boldsymbol{a}$.

The distribution of $\boldsymbol{X}$ is degenerate in the sense that either one of the X_i's is constant or else a linear combination of the other components.

Example

Consider the multinomial distribution with parameters n and $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_r)^T$ with $\pi_i > 0$ and $\sum_{i=1}^{r} \pi_i = 1$.

It can be checked that

$$\text{Var}(\boldsymbol{X}) = n(\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T).$$

and also that

$$\mathbf{1}^T(\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T)\mathbf{1} = 0.$$

We expect this to be the case because for the multimomial distribution, $\boldsymbol{X}$ is constrained so that

$$\mathbf{1}^T\boldsymbol{X} = \sum_{i=1}^{n} X_i = n.$$

Moment generating functions

Definition 23

If $X_1, X_2 \ldots, X_r$ are RVs, then the joint MGF (provided it exists) is given by

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = M_{\boldsymbol{X}}(t_1, t_2, \ldots, t_r) = E\left[e^{t_1 X_1 + t_2 X_2 + \cdots + t_r X_r}\right].$$

Theorem 24

If $X_1, X_2, \ldots, X_r$ are mutually independent, then the joint MGF satisfies

$$M_{\boldsymbol{X}}(t_1, \ldots, t_r) = M_{X_1}(t_1) M_{X_2}(t_2) \ldots M_{X_r}(t_r),$$

provided it exists.

Proof

$$
\begin{aligned}
& M_{\boldsymbol{X}}(t_1, \ldots, t_r) \\
= & \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t_1 x_1 + t_2 x_2 + \cdots + t_r x_r} f_{\boldsymbol{X}}(x_1, \ldots, x_r)\, dx_1 \ldots dx_r \\
= & \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t_1 x_1} e^{t_2 x_2} \ldots e^{t_r x_r} f_{X_1}(x_1) f_{X_2}(x_2) \ldots f_{X_r}(x_r)\ dx_1 \ldots dx_r \\
= & \left(\int_{-\infty}^{\infty} e^{t_1 x_1} f_{X_1}(x_1)\, dx_1 \right) \ldots \left(\int_{-\infty}^{\infty} e^{t_r x_r} f_{X_r}(x_r)\, dx_r \right) \\
= & M_{X_1}(t_1) M_{X_2}(t_2) \ldots M_{X_r}(t_r), \qquad \text{as required.}
\end{aligned}
$$

□

Continuous distributions

The uniform distribution on the unit square

Dimension: $r = 2$

Possible values: $0 < x_1 < 1,\ 0 < x_2 < 1.$

PDF:

$$f(x_1, x_2) = \begin{cases} 1 & \text{for } 0 < x_1 < 1,\ 0 < x_2 < 1; \\ 0 & \text{otherwise} \end{cases}$$

The uniform distribution on the unit disk

Dimension: $r = 2$

Possible values: $x_1^2 + x_2^2 < 1$.

PDF:

$$f(x_1, x_2) = \begin{cases} \frac{1}{\pi} & \text{for } x_1^2 + x_2^2 < 1; \\ 0 & \text{otherwise} \end{cases}$$

The Dirichlet distribution

Dimension: $r > 1$

Possible values: $x_i > 0$ and $\sum_{i=1}^{r} x_i < 1$.

Parameters: $\alpha_1 > 0,\ \alpha_2 > 0, \ldots, \alpha_{r+1} > 0$.

PDF:

$$\begin{aligned} & f(x_1, x_2, \ldots, x_r) \\ = \quad & \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_r + \alpha_{r+1})}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_r)\Gamma(\alpha_{r+1})} \\ \times \quad & x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \ldots x_r^{\alpha_r - 1} (1 - x_1 - \cdots - x_r)^{\alpha_{r+1}} \end{aligned}$$

The Dirichlet distribution (continued)

Moments:

$$\begin{aligned} E(X_i) &= \frac{\alpha_i}{\alpha_0}, \\ \text{var}(X_i) &= \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \\ \text{cov}(X_i, X_j) &= -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \end{aligned}$$

$$\text{where } \alpha_0 = \sum_{i=1}^{r+1} \alpha_i.$$

Marginal Distribution: $X_i \sim \text{Beta}(\alpha_i, \alpha_0 - \alpha_i)$.

The multivariate normal distribution

Dimension: $r > 1$.

Possible values: $\boldsymbol{x} \in \mathbb{R}^r$.

Parameters: $\boldsymbol{\mu} \in \mathbb{R}^r$, Σ $r \times r$, symmetric, positive definite.

Notation: $\boldsymbol{X} \sim N_r(\boldsymbol{\mu}, \Sigma)$.

PDF:

$$f(\boldsymbol{x}) = \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{r}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Moments: $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{X}) = \Sigma$.

MGF:

$$M(\boldsymbol{t}) = e^{\boldsymbol{t}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T \Sigma \boldsymbol{t}}$$

The bivariate normal distribution

Let $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$

The bivariate normal PDF is

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\left(\frac{x_1-\mu_1}{\sigma_1} \right)^2 \right.\right. \\ &\quad \left.\left. + \left(\frac{x_2-\mu_2}{\sigma_2} \right)^2 - 2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) \right] \right\}. \end{aligned}$$

Marginal and conditional distributions

Theorem 25

Suppose that

$$\begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \sim N_{r_1+r_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

Then, the marginal distribution of $\boldsymbol{X}_2$ *is* $\boldsymbol{X}_2 \sim N_{r_2}(\boldsymbol{\mu}_2, \Sigma_{22})$ *and the conditional distribution of* $\boldsymbol{X}_1|\boldsymbol{X}_2$ *is*

$$N_{r_1}\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

For convenience, let

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2) \text{ and } \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Proof (1)

This result can be proved by exhibiting the joint PDF as

$$f(\boldsymbol{x}_1, \boldsymbol{x}_2) = g(\boldsymbol{x}_2) h(\boldsymbol{x}_1, \boldsymbol{x}_2),$$

where

- $g(\boldsymbol{x}_2)$ is the $N_{r_2}(\boldsymbol{\mu}_2, \Sigma_{22})$ PDF,
- for each $\boldsymbol{x}_2$, $h(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is the $N_{r_1}(\boldsymbol{\mu}_{1|2}, \Sigma_{1|2})$ PDF.

  That is, $N_{r_1}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$ PDF.

Proof (2) The normalising constant

We need to show that

$$(2\pi)^{\frac{r_1+r_2}{2}}|\Sigma|^{\frac{1}{2}} = (2\pi)^{\frac{r_2}{2}}|\Sigma_{22}|^{\frac{1}{2}} \times (2\pi)^{\frac{r_1}{2}}|\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|^{\frac{1}{2}}$$

Let C be the partitioned matrix

$$C = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & \Sigma_{22}^{-1} \end{pmatrix},$$

recall $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ and observe

$$C\Sigma = \begin{pmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ \Sigma_{22}^{-1}\Sigma_{21} & I \end{pmatrix}.$$

Proof (3) The normalising constant

- Using the fact that $|AB| = |A||B|$ for square matrices A and B it follows that $|\Sigma| = |C\Sigma|/|C|$.
  - $|C| = |\Sigma_{22}^{-1}| = 1/|\Sigma_{22}|$.
  - $|C\Sigma| = |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|$.
- Hence,

$$|\Sigma| = |\Sigma_{22}| \times |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|$$

as required.

Proof (4) The exponent

We need to show

$$\begin{aligned}
& (\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \\
= \quad & (\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T \Sigma_{22}^{-1} (\boldsymbol{x}_2-\boldsymbol{\mu}_2) \\
& + (\boldsymbol{x}_1-\boldsymbol{\mu}_{1|2})^T \Sigma_{1|2}^{-1} (\boldsymbol{x}_1-\boldsymbol{\mu}_{1|2})
\end{aligned}$$

where

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2).$$

and

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Proof (5) The exponent

Using the formula for the inverse of a partitioned matrix from Tutorial 2, Question 6, we have

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{1|2}^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{1|2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{pmatrix}$$

where

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Proof (6) The exponent

$$
\begin{aligned}
& ((\boldsymbol{x}_1-\boldsymbol{\mu}_1)^T(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T)\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{x}_1-\boldsymbol{\mu}_1\\ \boldsymbol{x}_2-\boldsymbol{\mu}_2\end{pmatrix} \\
= \quad & (\boldsymbol{x}_1-\boldsymbol{\mu}_1)^T\Sigma_{1|2}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) \\
& -(\boldsymbol{x}_1-\boldsymbol{\mu}_1)^T\Sigma_{1|2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) \\
& -(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{1|2}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1) \\
& +(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{1|2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) \\
& +(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2) \\
= \quad & (\boldsymbol{x}_1-\boldsymbol{\mu}_{1|2})^T\Sigma_{1|2}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_{1|2})+(\boldsymbol{x}_2-\boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}(\boldsymbol{x}_2-\boldsymbol{\mu}_2)
\end{aligned}
$$

as required. $\square$

Transformations of several variables
Sums of discrete random variables

Theorem 26

If X_1, X_2 are discrete with joint probability function $p(x_1, x_2)$, and $Y = X_1 + X_2$, then:

1. *Y has probability function*

$$p_Y(y) = \sum_x p(x, y - x).$$

2. *If X_1, X_2 are independent,*

$$p_Y(y) = \sum_x p_{X_1}(x) p_{X_2}(y - x)$$

Proof

1. $$\begin{aligned} p_Y(y) &= P(\{Y = y\}) \\ &= \sum_x P(\{Y = y\} \cap \{X_1 = x\}) \text{ (law of total probability)} \\ &= \sum_x P(\{X_1 + X_2 = y\} \cap \{X_1 = x\}) \\ &= \sum_x P(\{X_2 = y - x\} \cap \{X_1 = x\}) \\ &= \sum_x p(x, y - x) \end{aligned}$$

2. Substitute $p(x, y - x) = p_{X_1}(x)p_{X_2}(y - x)$ in part (1) under the assumption of independence. □

Example

Suppose $X_1 \sim B(n_1, p)$ and $X_2 \sim B(n_2, p)$ independently. Then $Y = X_1 + X_2 \sim B(n_1 + n_2, p)$.

Proof.

$$\begin{aligned} p(y) &= \sum_x \binom{n_1}{x} p^x (1-p)^{n_1 - x} \binom{n_2}{y-x} p^{y-x} (1-p)^{n_2 - y + x} \\ &= p^y (1-p)^{n_1 + n_2 - y} \sum_{x = \max(0, y - n_2)}^{\min(n_1, y)} \binom{n_1}{x} \binom{n_2}{y-x} \\ &= \binom{n_1 + n_2}{y} p^y (1-p)^{n_1 + n_2 - y} \end{aligned}$$

as required. □

Sums of continuous random variables

Theorem 27

Suppose X_1, X_2 are continuous with PDF, $f(x_1, x_2)$, and let $Y = X_1 + X_2$. Then

1. $f_Y(y) = \int_{-\infty}^{\infty} f(x, y - x)\, dx.$
2. *If X_1, X_2 are independent, then*

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(y - x)\, dx.$$

Proof

1. $$\begin{aligned} F_Y(y) &= P(Y \le y) = P(X_1 + X_2 \le y) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y-x_1} f(x_1, x_2) dx_2 dx_1 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{y} f(x_1, t - x_1) dt dx_1 \qquad \text{(substitute } t = x_1 + x_2\text{)} \\ &= \int_{-\infty}^{y} \int_{-\infty}^{\infty} f(x_1, t - x_1) dx_1 dt \end{aligned}$$

   Differentiate with respect to y to obtain

$$f_y(y) = F'(y) = \int_{-\infty}^{\infty} f(x_1, y - x_1) dx_1.$$

2. Substitute $f(x, y - x) = f_{X_1}(x) f_{X_2}(y - x)$ in part (1) under the assumption of independence. □

Example

Suppose $X_1 \sim \text{Exp}(\lambda)$ and $X_2 \sim \text{Exp}(\lambda$ independently. Then $Y = X_1 + X_2 \sim \Gamma(2, \lambda)$.

Proof.

$$\begin{aligned} f_y(y) &= \int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(y-x) dx \\ &= \int_0^y \lambda e^{-\lambda x} \lambda e^{-\lambda(y-x)} dx \\ &= \lambda^2 e^{-\lambda y} \int_0^y 1 dx \\ &= \lambda^2 y e^{-\lambda y} = \frac{\lambda^2}{\Gamma(2)} y e^{-\lambda y} \text{ for } y > 0, \end{aligned}$$

as required. □

Ratio of continuous random variables

Theorem 28

Suppose X_1, X_2 are continuous with joint PDF $f(x_1, x_2)$, and let $Y = \frac{X_2}{X_1}$.

Then Y has PDF

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

If X_1, X_2 independent, we obtain

$$f_Y(y) = \int_{-\infty}^{\infty} |x| f_{X_1}(x) f_{X_2}(yx)\, dx.$$

Proof

1.

$$
\begin{aligned}
F_Y(y) &= P(\{Y \le y\}) \\
&= P(\{Y \le y\} \cap \{X_1 < 0\}) + P(\{Y \le y\} \cap \{X_1 > 0\}) \\
&= P(\{X_2 \ge yx_1\} \cap \{X_1 < 0\}) + P(\{X_2 \le yx_1\} \cap \{X_1 > 0\} \\
&= \int_{-\infty}^{0} \int_{x_1 y}^{\infty} f(x_1, x_2)\, dx_2 dx_1 + \int_{0}^{\infty} \int_{-\infty}^{x_1 y} f(x_1, x_2)\, dx_2 dx_1.
\end{aligned}
$$

(substitute $x_2 = tx_1 \Rightarrow dx_2 = x_1 dt$ in both inner integrals)

$$
\begin{aligned}
&= \int_{-\infty}^{0} \int_{y}^{-\infty} x_1 f(x_1, tx_1) dt\, dx_1 + \int_{0}^{\infty} \int_{-\infty}^{y} x_1 f(x_1, tx_1)\, dt\, dx_1 \\
&= \int_{-\infty}^{0} \int_{-\infty}^{y} (-x_1) f(x_1, tx_1)\, dt\, dx_1 + \int_{0}^{\infty} \int_{-\infty}^{y} x_1 f(x_1, tx_1)\, dt\, dx_1
\end{aligned}
$$

Proof (2)

Hence,

$$\begin{aligned} F_Y(y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{y} |x_1| f(x_1, tx_1) dt dx_1 \\ &= \int_{-\infty}^{y}\int_{-\infty}^{\infty} |x_1| f(x_1, tx_1) dx_1 dt \end{aligned}$$

Differentiate with respect to y to obtain

$$f_Y(y) = F'_Y(y) = \int_{-\infty}^{\infty} |x_1| f(x_1, yx_1) dx_1.$$

2. Substitute $f(x, yx) = f_{X_1}(x) f_{X_2}(yx)$ in part (1) under the assumption of independence. □

Example

Suppose $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$ independently. Then

$$Y = \frac{Z_2}{Z_1} \sim \text{Cauchy}$$

Proof

$$\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} |z|\phi(z)\phi(yz)dz \\
&= 2\int_0^{\infty} z\phi(z)\phi(yz)dz \\
&= 2\int_0^{\infty} \frac{z}{2\pi} e^{-\frac{1}{2}z^2(1+y^2)}dz \\
&= \frac{1}{\pi}\int_0^{\infty} e^{-u(1+y^2)}du \qquad \text{(subsitute } u = z^2/2 \Rightarrow du = zdz\text{)}
\end{aligned}$$

Example (continued)

Hence,

$$f_Y(y) = \frac{1}{\pi}\frac{1}{1+y^2}e^{-u(1+y^2)}\Big|_0^\infty = \frac{1}{\pi}\frac{1}{1+y^2}$$

as required.

□

Multivariate transformations

Suppose $h : \mathbb{R}^r \to \mathbb{R}^r$ is continuously differentiable.

That is,

$$h(x_1, x_2, \ldots, x_r) = \left(h_1(x_1, \ldots, x_r), h_2(x_1, \ldots, x_r), \ldots, h_r(x_1, \ldots, x_r)\right),$$

where each $h_i(x)$ is continuously differentiable. Let

$$H = \begin{pmatrix} \frac{\partial h_1}{\partial x_1} & \frac{\partial h_1}{\partial x_2} & \cdots & \frac{\partial h_1}{\partial x_r} \\ \frac{\partial h_2}{\partial x_1} & \frac{\partial h_2}{\partial x_2} & \cdots & \frac{\partial h_2}{\partial x_r} \\ \vdots & \vdots & \ldots & \vdots \\ \frac{\partial h_r}{\partial x_1} & \frac{\partial h_r}{\partial x_2} & \cdots & \frac{\partial h_r}{\partial x_r} \end{pmatrix} = \left(\frac{\partial h_i}{\partial x_j}\right)$$

The inverse transformation

If H is invertible for all $\boldsymbol{x}$, then there exists an inverse mapping:

$$g : \mathbb{R}^r \to \mathbb{R}^r$$

such that

$$g(h(\boldsymbol{x})) = \boldsymbol{x}.$$

It can be proved that the matrix of partial derivatives

$$G = \left(\frac{\partial g_i}{\partial y_j} \right) \text{ satisfies } G = H^{-1}.$$

The multivariate transformation rule

Theorem 29

Suppose $X_1, X_2, \ldots, X_r$ have joint PDF $f_X(x_1, \ldots, x_r)$, and let h, g, G be as above.

If $\boldsymbol{Y} = h(\boldsymbol{X})$, then $\boldsymbol{Y}$ has joint PDF

$$f_Y(y_1, y_2, \ldots, y_r) = f_X(g(\boldsymbol{y}))\,|\det G(\boldsymbol{y})|$$

Remarks

1. The proof of this result is the change of variables formula for multivariable integrals.
2. Can sometimes use H^{-1} instead of G, but need to be careful to evaluate $H^{-1}(\boldsymbol{x})$ at $\boldsymbol{x} = h^{-1}(\boldsymbol{y}) = g(\boldsymbol{y})$.

Example

Suppose $Z_1 \sim \mathrm{N}(0,1), Z_2 \sim \mathrm{N}(0,1)$ independently. Consider $h(z_1, z_2) = (r, \theta)^T$, where $r = h_1(z_1, z_2) = \sqrt{z_1^2 + z_2^2}$, and

$$\theta = h_2(z_1, z_2) = \begin{cases} \arctan\left(\dfrac{z_2}{z_1}\right) & \text{for } z_1 > 0 \\ \arctan\left(\dfrac{z_2}{z_1}\right) + \pi & \text{for } z_1 < 0 \\ \dfrac{\pi}{2} \operatorname{sgn} z_2 & \text{for } z_1 = 0, z_2 \neq 0 \\ 0 & \text{for } z_1 = z_2 = 0. \end{cases}$$

Then h maps $\mathbb{R}^2 \to [0, \infty) \times \left[-\dfrac{\pi}{2}, \dfrac{3\pi}{2}\right)$.

Example (continued)

The problem is to find the joint distribution of (R, θ).

Observe first the inverse mapping, g, can be seen to be:

$$g(r,\theta) = \begin{pmatrix} g_1(r,\theta) \\ g_2(r,\theta) \end{pmatrix} = \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix}$$

The derivative matrix is

$$G(r,\theta) = \begin{pmatrix} \frac{\partial g_1}{\partial r} & \frac{\partial g_1}{\partial \theta} \\ \frac{\partial g_2}{\partial r} & \frac{\partial g_2}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

so that

$$\det(G) = r\cos^2\theta + r\sin^2\theta = r.$$

Example (Continued)

Now apply Theorem 29.

Observed

$$f_Z(z_1, z_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_1^2} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_2^2} = \frac{1}{2\pi} e^{-\frac{1}{2}(z_1^2+z_2^2)}.$$

so that

$$f(r, \theta) = \begin{cases} \frac{r}{2\pi} e^{-\frac{r^2}{2}} & \text{for } r \geq 0, \frac{\pi}{2} \leq \theta < \frac{3\pi}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

Example (continued)

Hence, R and θ are independent such that

$$\theta \sim U\left(-\frac{\pi}{2}, \frac{3\pi}{2}\right)$$

and R has the Rayleigh distribution with PDF

$$f(r) = re^{-\frac{r^2}{2}} \text{ for } r \geq 0.$$

Note the distribution of R^2 is χ^2_2.

The method of regular transformations

Suppose $\boldsymbol{X} = (X_1, \ldots, X_r)^T$ and $h : \mathbb{R}^r \to \mathbb{R}^p$, where $p < r$ if

$$\boldsymbol{Y} = (Y_1, \ldots, Y_p)^T = h(\boldsymbol{X}).$$

One approach is as follows:

1. Choose a function, $d : \mathbb{R}^r \to \mathbb{R}^{r-p}$ and let $\boldsymbol{W} = d(\boldsymbol{X}) = (W_1, W_2, \ldots, W_{r-p})^T$.
2. Apply theorem 29 to $(Y_1, Y_2, \ldots, Y_p, W_1, W_2, \ldots, W_{r-p})^T$.
3. Integrate over $w_1, w_2, \ldots, w_{r-p}$ to get the marginal PDF for $\boldsymbol{Y}$.

Example

Suppose $X_1 \sim$ Gamma (α, λ) and $X_2 \sim$ Gamma (β, λ), independently. Find the distribution of $Y = X_1/(X_1 + X_2)$.

Solution

1. Try using $W = X_1 + X_2$.
2. If $h(x_1, x_2) = \begin{pmatrix} y = \frac{x_1}{x_1+x_2} \\ w = x_1 + x_2 \end{pmatrix}$, then $g(y, w) = \begin{pmatrix} yw \\ (1-y)w \end{pmatrix}$.

   Hence
   $$G = \begin{pmatrix} w & y \\ -w & 1-y \end{pmatrix}$$
   and
   $$\det(G) = w(1-y) - (-wy) = w > 0.$$

Example (continued)

3. Hence

$$\begin{aligned} f_{Y,W}(y,w) &= f_{X_1}(yw) f_{X_2}((1-y)w) w \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} (yw)^{\alpha-1} e^{-\lambda yw} \frac{\lambda^\beta}{\Gamma(\beta)} ((1-y)w)^{\beta-1} e^{-\lambda(1-y)w} w \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \lambda^{\alpha+\beta} y^{\alpha-1} (1-y)^{\beta-1} w^{\alpha+\beta-1} e^{-\lambda w}. \end{aligned}$$

Example (continued)

4.

$$
\begin{aligned}
f_Y(y) &= \int_0^\infty f_{Y,W}(y,w)dw \\
&= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \int_0^\infty \lambda^{\alpha+\beta} w^{\alpha+\beta-1} e^{-\lambda w} dw \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \int_0^\infty \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha+\beta)} w^{\alpha+\beta-1} e^{-\lambda w} dw \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \text{ for } 0 < y < 1.
\end{aligned}
$$

That is,

$$Y \sim \text{Beta}(\alpha, \beta).$$

Moment generating function methods

When it exists, the MGF can sometimes be used to find the distribution of transformed random variables.

The most important example is sums of random variables.

Theorem 30

Suppose $X_1, \ldots, X_r$ are RVs and let $Y = X_1 + X_2 + \cdots + X_r$. Then (assuming MGFs exist):

1. $M_Y(t) = M_{\boldsymbol{X}}(t, t, t, \ldots, t)$.
2. *If $X_1 \ldots X_r$ are independent then* $M_Y(t) = M_{X_1}(t) M_{X_2}(t) \ldots M_{X_r}(t)$.
3. *If $X_1, \ldots, X_r$ are IID with common MGF $M(t)$, then*

$$M_Y(t) = [M(t)]^r .$$

Proof

1. $$\begin{aligned} M_Y(t) &= E\left(e^{tY}\right) \\ &= E\left(e^{t(X_1+X_2+\cdots+X_r)}\right) \\ &= E\left(e^{tX_1+tX_2+\cdots+tX_r}\right) \\ &= M_X(t,t,t,\ldots,t) \end{aligned}$$

   as required.

2. Use the fact that when $X_1, X_2, \ldots, X_r$ are independent,

   $$M_X(t_1, t_2, \ldots, t_r) = M_{X_1}(t_1)M_{X_2}(t_2)\cdots M_{X_r}(t_r).$$

3. Substitute $M_{X_i}(t) = M(t)$ in the IID case. □

Example

Suppose $X \sim$ Bernoulli with parameter p. Then $M_X(t) = 1 + p(e^t - 1)$.

Now suppose $X_1, X_2, \ldots, X_n$ are IID Bernoulli with parameter p and

$$Y = X_1 + X_2 + \cdots + X_n.$$

Then $M_Y(t) = \left(1 + p(e^t - 1)\right)^n$, which agrees with the formula previously derived for the binomial distribution.

MGFs and linear transformations

The result for sums of random variables generalises to linear transformations.

Theorem 31

Suppose the r dimension random vector $\boldsymbol{X}$ has MGF $M_X(\boldsymbol{t})$ and let $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$, where $A_{p\times r}$ and $\boldsymbol{b} \in \mathbb{R}^p$ are constants. The MGF of $\boldsymbol{Y}$ is given by

$$M_Y(\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{b}} M_X(A^T\boldsymbol{t}).$$

Proof

$$\begin{aligned} M_Y(\boldsymbol{t}) &= E\left(e^{\boldsymbol{t}^T \boldsymbol{Y}}\right) \\ &= E\left(e^{\boldsymbol{t}^T (A\boldsymbol{X}+\boldsymbol{b})}\right) \\ &= e^{\boldsymbol{t}^T \boldsymbol{b}} E\left(e^{\boldsymbol{t}^T A\boldsymbol{X}}\right) \\ &= e^{\boldsymbol{t}^T \boldsymbol{b}} E\left(e^{(A^T \boldsymbol{t})^T \boldsymbol{X}}\right) \\ &= e^{\boldsymbol{t}^T \boldsymbol{b}} M_X(A^T \boldsymbol{t}) \end{aligned}$$

as required.

□

Linear transformations of normal variables

Theorem 32

Suppose $\boldsymbol{X} \sim N_r(\boldsymbol{\mu}, \Sigma)$ *and let*

$$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$$

where $A_{r\times r}$ *(invertible) and* $\boldsymbol{b} \in \mathbb{R}^r$ *are constants. Then*

$$\boldsymbol{Y} \sim N_r(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T).$$

Proof

Apply theorem 29 with $h(\boldsymbol{X}) = A\boldsymbol{X} + \boldsymbol{b}$. Since A is invertible,

$$g(\boldsymbol{y}) = h^{-1}(\boldsymbol{y}) = A^{-1}(\boldsymbol{y} - \boldsymbol{b}) \text{ and } G = A^{-1}.$$

Proof (continued)

Hence,

$$
\begin{aligned}
f_Y(\boldsymbol{y}) &= f_X(A^{-1}(\boldsymbol{y}-\boldsymbol{b}))|\det(A^{-1}| \\
&= \frac{|\det(A^{-1})|}{(2\pi)^{\frac{r}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(A^{-1}(\boldsymbol{y}-\boldsymbol{b})-\boldsymbol{\mu})^T\Sigma^{-1}(A^{-1}(\boldsymbol{y}-\boldsymbol{b})-\boldsymbol{\mu})} \\
&= \frac{1}{(2\pi)^{\frac{r}{2}}|A\Sigma A^T|^{\frac{1}{2}}} e^{-\frac{1}{2}(A^{-1}(\boldsymbol{y}-(A\boldsymbol{\mu}+\boldsymbol{b}))^T\Sigma^{-1}(A^{-1}(\boldsymbol{y}-(A\boldsymbol{\mu}+\boldsymbol{b}))} \\
&= \frac{1}{(2\pi)^{\frac{r}{2}}|A\Sigma A^T|^{\frac{1}{2}}} e^{-\frac{1}{2}((\boldsymbol{y}-(A\boldsymbol{\mu}+\boldsymbol{b}))^T(A\Sigma A^T)^{-1}(\boldsymbol{y}-(A\boldsymbol{\mu}+\boldsymbol{b}))}
\end{aligned}
$$

as required.

□

Linear transformations of normal variables (2)

Corollary 33

Suppose $\boldsymbol{X} \sim N_r(\boldsymbol{\mu}, \Sigma)$ *and let*

$$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$$

where $A_{p\times r}$ *of rank* $p < r$ *and* $\boldsymbol{b} \in \mathbb{R}^p$ *are constants. Then*

$$\boldsymbol{Y} \sim N_p(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T).$$

Proof

Apply the method of regular transformations. Find a matrix $C_{r-p\times r}$ such that the $r \times r$ matrix $\begin{pmatrix} A \\ C \end{pmatrix}$ is invertible.

Proof (continued)

Next consider the linear transformation,

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} = \begin{pmatrix} A \\ C \end{pmatrix} \boldsymbol{X} + \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{0}_{r-p\times 1} \end{pmatrix}.$$

Applying Theorem 32, it follows that

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix} \sim N_r \left(\begin{pmatrix} A\boldsymbol{\mu} + \boldsymbol{b} \\ C\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} A\Sigma A^T & A\Sigma C^T \\ C\Sigma A^T & C\Sigma C^T \end{pmatrix} \right).$$

Finally, observe that the marginal distribution is

$$\boldsymbol{Y}_1 \sim N_p(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T)$$

as required.

□

The one sample t statistic

Lemma 34

Suppose $X_1, X_2, \ldots, X_n$ *are i.i.d.* $N(\mu, \sigma^2)$ *RVs and let*

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

Then $\bar{X} \sim N(\mu, \sigma^2/n)$ *and* $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$ *independently.*

Proof (1)

Observe first that if $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ then the IID $N(\mu, \sigma^2)$ assumption may be written as:

$$\boldsymbol{X} \sim \mathrm{N}_n(\mu \mathbf{1}_n, \sigma^2 I_{n \times n})$$

1. $\bar{X} \sim \mathrm{N}(\mu, \sigma^2/n)$:

   Observe that $\bar{X} = A\boldsymbol{X}$, where

   $$A = \left(\frac{1}{n} \; \frac{1}{n} \; \ldots \; \frac{1}{n} \right)$$

   Hence, by Corollary 33

   $$\bar{X} \sim \mathrm{N}(A\mu\mathbf{1}, \sigma^2 A A^T) \equiv \mathrm{N}(\mu, \sigma^2/n), \qquad \text{as required.}$$

Proof (2)

2. Independence of $\bar{X}$ and S^2: Let

$$A = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} & \frac{1}{n} \\ 1-\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \cdots & -\frac{1}{n} & -\frac{1}{n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & 1-\frac{1}{n} & -\frac{1}{n} \end{pmatrix}$$

Proof (3)

Hence

$$A\boldsymbol{X} = \begin{pmatrix} \bar{X} \\ X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_{n-1} - \bar{X} \end{pmatrix}$$

It follows that $\mathrm{Var}(A\boldsymbol{X}) = \sigma^2 AA^T$ has the form

$$\begin{pmatrix} \frac{\sigma^2}{n} & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & \Sigma_{22} \end{pmatrix}.$$

Hence, $\bar{X}$ and $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_{n-1} - \bar{X})$ are independent, by multivariate normality.

Proof (4)

Finally, since

$$X_n - \bar{X} = -\Big((X_1 - \bar{X}) + (X_2 - \bar{X}) + \cdots + (X_{n-1} - \bar{X})\Big),$$

it follows that S^2 is a function of $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_{n-1} - \bar{X})$ and hence is independent of $\bar{X}$.

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

   Consider the identity:

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

   or, equivalently,

$$\sum_{i=1}^{n}\frac{(X_i - \mu)^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i - \bar{X})^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$$

Proof (5)

Let

$$\begin{aligned} R_1 &= \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \\ R_2 &= \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \\ R_3 &= \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \end{aligned}$$

If M_1, M_2, M_3 are the MGFs for R_1, R_2, R_3 respectively, then,

$$M_1(t) = M_2(t) M_3(t)$$

since $R_1 = R_2 + R_3$ with R_2 and R_3 independent.

Proof (6)

Hence, $M_2(t) = M_1(t)/M_3(t)$.

Now, observe

$$R_1 \sim \chi^2_n \quad \Rightarrow \quad M_1(t) = \left(\frac{1}{1-2t}\right)^{\frac{n}{2}}$$

$$R_3 \sim \chi^2_1 \quad \Rightarrow \quad M_3(t) = \frac{1}{1-2t}$$

and hence

$$M_2(t) = \left(\frac{1}{1-2t}\right)^{\frac{n-1}{2}}$$

so

$$\frac{(n-1)S^2}{\sigma^2} = R_2 \sim \chi^2_{n-1}$$

as required.

□

The one-sample t statistic

Theorem 35

If $X_1, \ldots, X_n$ are IID $N(\mu, \sigma^2)$, then

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Proof

Recall that t_k is the distribution of $\frac{Z}{\sqrt{V/k}}$, where $Z \sim \mathrm{N}(0, 1)$ and $V \sim \chi^2_k$ independently.

Now observe that:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \Bigg/ \sqrt{\frac{(n-1)S^2/\sigma^2}{(n-1)}}$$

Proof (continued)

Note that,

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

and

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

independently, from Lemma 34.

□

Limit Theorems

Let $X_1, X_2, X_3, \ldots,$ be an infinite sequence of RVs. We will consider 4 different types of convergence.

Definition 36

1. **Convergence in probability (weak convergence)**

   The sequence $\{X_n\}$ is said to converge to the constant $\alpha \in \mathbb{R}$, **in probability** if for every $\varepsilon > 0$

   $$\lim_{n\to\infty} P(|X_n - \alpha| > \varepsilon) = 0$$

2. **Convergence in quadratic mean** The sequence $\{X_n\}$ is said to converge to the constant $\alpha \in \mathbb{R}$, **in quadratic mean**

   $$\lim_{n\to\infty} E((X_n - \alpha)^2) = 0$$

3. **Almost sure convergence (strong convergence)** The sequence $\{X_n\}$ is said to converge **almost surely** to α if

$$P\left(\lim_{n\to\infty} X_n = \alpha\right) = 1.$$

4. **Convergence in distribution**

   The sequence of RVs $\{X_n\}$ with CDFs $\{F_n\}$ is said to converge in distribution to the RV X with CDF $F(x)$ if:

$$\lim_{n\to\infty} F_n(x) = F(x) \quad \text{for all continuity points of } F.$$

□

Remarks

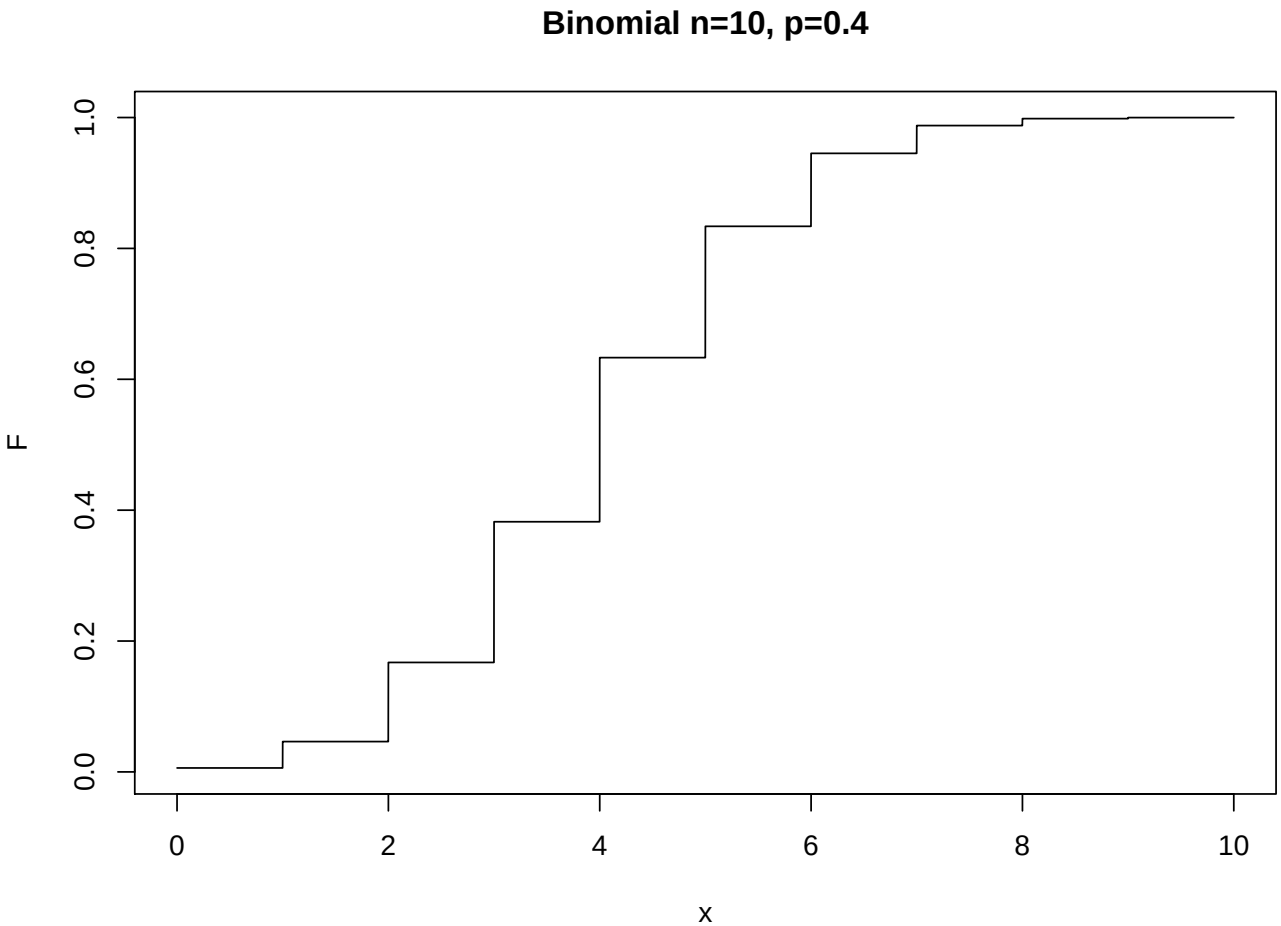
1. It can be checked that:

$$\begin{aligned} \text{Convergence in quadratic mean} &\Rightarrow \text{Convergence in probability} \\ \text{Almost sure convergence} &\Rightarrow \text{Convergence in probability} \end{aligned}$$

   but there is no simple relationship between convergence in quadratic mean and almost sure convergence.

2. In the definition of convergence in distribution, all values of x are continiuty points of F if X is a continuous random variable but not if X is discrete.

Remarks (continued)



Remarks (continued)

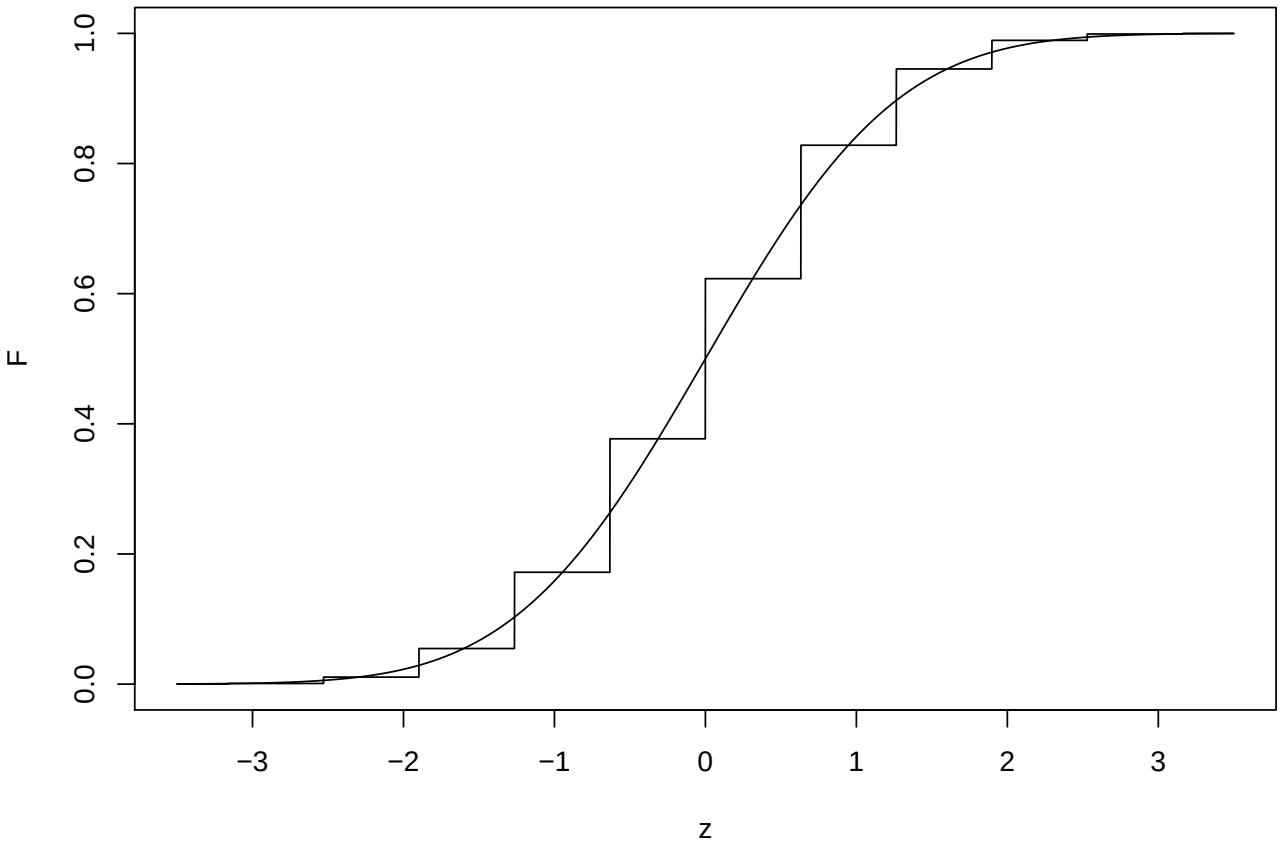
3. A sequence of discrete random variables can converge in distribution to a continuous random variable.

   For example, if $X_n \sim B(n, p)$ and

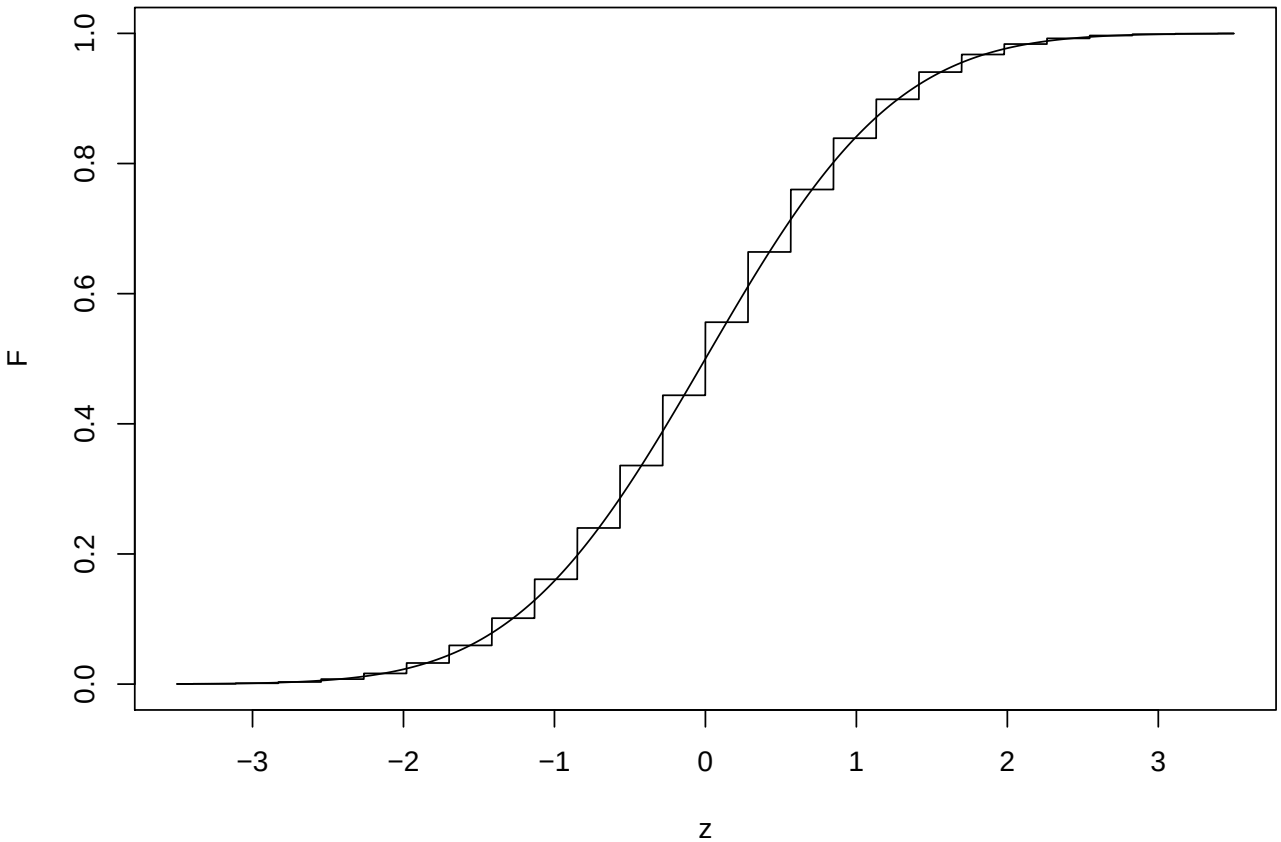
   $$Z_n = \frac{X_n - np}{\sqrt{np(1-p)}},$$

   it follows from the central limit theorem that the sequence $\{Z_n\}$ converges in distribution to $Z \sim N(0, 1)$.

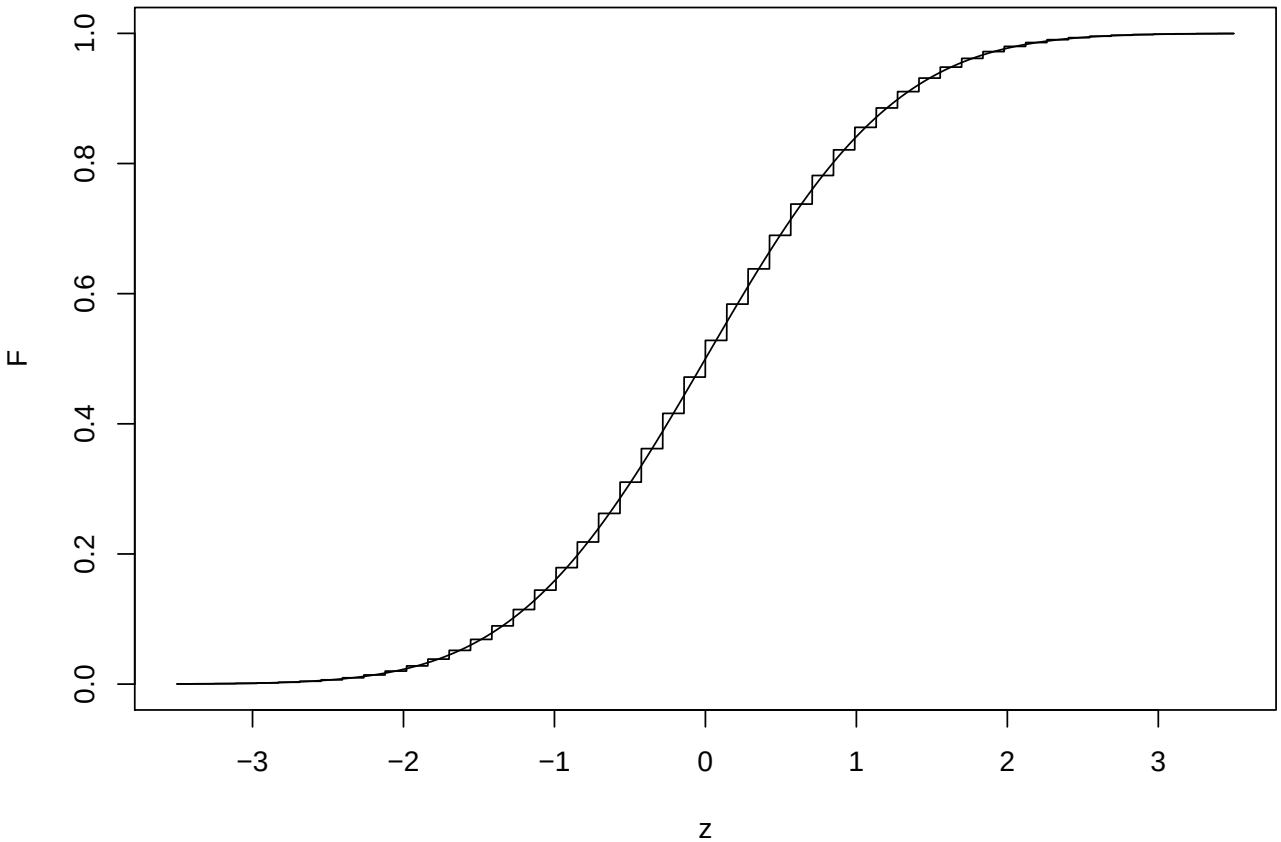
Binomial n=10, p=0.5



Binomial n=50, p=0.5



Binomial n=200, p=0.5



Remarks (continued)

4. If $F(x) = \begin{cases} 0 & x < \alpha \\ 1 & x \geq \alpha \end{cases}$

   then $X = \alpha$ with probability 1, and convergence in distribution to F is the same thing as convergence in probability to α.

5. Commonly used notation for convergence in distribution is either:

   (i) $\mathcal{L}[X_n] \to \mathcal{L}[X]$

   (ii) $X_n \xrightarrow[\mathcal{D}]{} \mathcal{L}[X]$ or e.g., $X_n \xrightarrow[\mathcal{D}]{} \mathrm{N}(0,1)$.

(Continuity Theorem)

Theorem 37

Let $M_n(t)$ be MGF of X_n and $M(t)$ be MGF of X. If

$$\lim_{n\to\infty} M_n(t) = M(t)$$

for each t in some open interval containing 0, then

$$\mathcal{L}[X_n] \to \mathcal{L}[X]$$

as $n \to \infty$.

□

The proof is omitted.

The weak law of large numbers

Theorem 38

Suppose $X_1, X_2, X_3, \ldots$ is a sequence of IID RVs with $E[X_i] = \mu, \text{var}(X_i) = \sigma^2$, and let

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Then $\bar{X}_n$ converges to μ in probability.

Proof of WLLN

We need to show for each $\epsilon > 0$ that

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Now observe that $E[\bar{X}_n] = \mu$ and $\text{var}(\bar{X}_n) = \frac{\sigma^2}{n}$. So by Chebyshev's inequality,

$$P(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0 \qquad \text{as } n \to \infty, \text{ for any fixed } \epsilon > 0.$$

□

Remarks

1. The proof given for Theorem 38 is really a corollary to the fact that $\bar{X}_n$ also converges to μ in quadratic mean.
2. There is also a version of this theorem involving almost sure convergence (strong law of large numbers). We will not discuss this.
3. The law of large numbers is one of the fundamental principles of statistical inference. That is, it is the formal justification for the claim that the "sample mean approaches the population mean for large n".

The central limit theorem

Lemma 39

Suppose a_n is a sequence of real numbers such that $\lim_{n\to\infty} na_n = a$ with $|a| < \infty$. Then,

$$\lim_{n\to\infty}(1 + a_n)^n = e^a.$$

The proof is omitted, but not difficult.

Remark

This is a simple generalisation of the standard limit,

$$\lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = e^x.$$

The central limit theorem

Consider a sequence of IID RVs $X_1, X_2, \ldots$ with $E(X_i) = \mu, \text{var}(X_i) = \sigma^2$ and such that the MGF, $M_X(t)$, is defined for all t in some open interval containing 0.

Let $S_n = \sum_{i=1}^{n} X_i$ and note that $E[S_n] = n\mu$ and $\text{var}(S_n) = n\sigma^2$.

Theorem 40

Let $X_1, X_2, \ldots, S_n$ be as above and let $Z_n = \dfrac{S_n - n\mu}{\sigma\sqrt{n}}$. Then

$$\mathcal{L}[Z_n] \to N(0,1) \qquad \text{as } n \to \infty.$$

Proof of the central limit theorem

Observe first that the moment generating function of the standard normal distribution is

$$M_Z(t) = e^{t^2/2}.$$

We will prove that

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}.$$

Let $U_i = \dfrac{X_i - \mu}{\sigma}$ and observe that $Z_n = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} U_i$.

Since the U_i are IID, it follows that

$$M_{Z_n}(t) = \left\{M_U(t/\sqrt{n})\right\}^n$$

Proof of the central limit theorem

Now consider the second order Taylor expansion,

$$M_U(t) = M_U(0) + M'_U(0)t + M''_U(0)\frac{t^2}{2} + r(t)$$

where

$$\lim_{s \to 0} \frac{r(s)}{s^2} = 0.$$

Since, $M'(U) = E(U) = 0$ and $\text{var}(U) = 1 \Rightarrow M''(U) = 1$, it follows that

$$M_U(t) = 1 + \frac{t^2}{2} + r(t).$$

Proof of the central limit theorem

Hence,

$$M_{Z_n}(t) = (1 + a_n)^n$$

where

$$a_n = \frac{t^2}{2n} + r(t/\sqrt{n}).$$

To complete the proof, need to show, for fixed t,

$$\lim_{n\to\infty} na_n = \frac{t^2}{2}$$

and then apply Lemma 39.

Proof of the central limit theorem

Since

$$\lim_{s \to 0} \frac{r(s)}{s^2} = 0,$$

it follows, for fixed t, that

$$\lim_{n \to \infty} nr(t/\sqrt{n}) = 0$$

and, hence,

$$\lim_{n \to \infty} na_n = \frac{t^2}{2}$$

so the proof is complete.

□

Remarks

1. The Central Limit Theorem can be stated equivalently for $\bar{X}_n$,

$$\text{i.e., } \mathcal{L}\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right) \to N(0,1) \quad \text{since} \quad \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

2. The Central Limit Theorem holds under conditions more general than those given above. In particular, with suitable assumptions,

   (i) $M_X(t)$ need not exist.

   (ii) $X_1, X_2, \ldots$ need not be IID.

Remarks

3. Theorems 38 and 40 are concerned with the asymptotic behaviour of $\bar{X}_n$.

   Theorem 38 states $\bar{X}_n \to \mu$ in prob as $n \to \infty$.

   Theorem 40 states $\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow[\mathcal{D}]{} \mathrm{N}(0,1)$ as $n \to \infty$.

   These results are not contradictory because $\mathrm{var}(\bar{X}_n) \to 0$, but the Central Limit Theorem is concerned with

$$\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\mathrm{var}(\bar{X}_n)}}.$$

Statistical Inference

Probability is concerned partly with the problem of predicting the behavior of the random variable X assuming we know its distribution.

Statistical inference is concerned with the inverse problem:

> *Given data $x_1, x_2, \ldots, x_n$ with unknown CDF $F(x)$, what can we conclude about $F(x)$?*

In this course, we are concerned with parametric inference. That is, we assume F belongs to a given family of distributions, indexed by the parameter $\boldsymbol{\theta}$:

$$\mathcal{F} = \{F(x; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$$

where Θ is the parameter space.

Examples

1. $X_1, X_2, \ldots, X_n$ IID $N(\mu, \sigma^2)$.
   - $\mathcal{F}$ is the family of normal distributions.
   - $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$.
   - $$\mathcal{F} = \left\{ F(x) : F(x) = \Phi\left(\frac{x-\mu}{\sigma^2}\right) \right\}.$$
2. $X_1, X_2, \ldots, X_n$ IID Bern(θ).
   - $\mathcal{F}$ is the family of Bernoulli distributions with success probability θ.
   - $\Theta = \{\theta \in [0, 1] \subset \mathbb{R}\}$.

Definitions

- A collection of IID random variables, $X_1, \ldots, X_n$, with common CDF $F(x; \boldsymbol{\theta})$, is said to be a random sample (from $F(x; \boldsymbol{\theta})$).
- Any function $T = T(x_1, x_2, \ldots, x_n)$ that can be calculated from the data (without knowledge of $\boldsymbol{\theta}$) is called a statistic.
- A statistic T with property $T(\boldsymbol{x}) \in \Theta$ for all $\boldsymbol{x}$ is called an estimator for $\boldsymbol{\theta}$.

Examples

1. Suppose $X_1, X_2, \ldots, X_n$ IID $N(\mu, \sigma^2)$.
   - $T = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a statistic.
   - The quantity $(\bar{X}, S^2)$ is a estimator for $\boldsymbol{\theta} = (\mu, \sigma^2)$, where

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

2. Suppose $X_1, X_2, \ldots, X_n$ IID Bern(θ) so that

$$Y = \sum_{i=1}^{n} X_i \sim B(n, \theta).$$

   - The quantity $\hat{\theta} = Y/n$ is an estimator for θ.

Mean squared error

An unsatisfactory aspect of the definition of an estimator is that it gives no guidance on how to recognise of construct good estimators.

We will consider the theory of estimation for a scalar parameter, θ.

Definition 41

The mean squared error of the estimator T of θ is defined by

$$MSE_T(\theta) = E\left((T-\theta)^2\right).$$

Example

Suppose $X_1, \ldots, X_n$ are IID Bern(θ) and $T = \hat{\theta} = Y/n$ where $Y = \sum_{i=1}^{n} X_i$. Since $Y \sim B(n, \theta)$ we have

$$E(T) = \theta, \ \text{var}(T) = \frac{\theta(1-\theta)}{n}$$

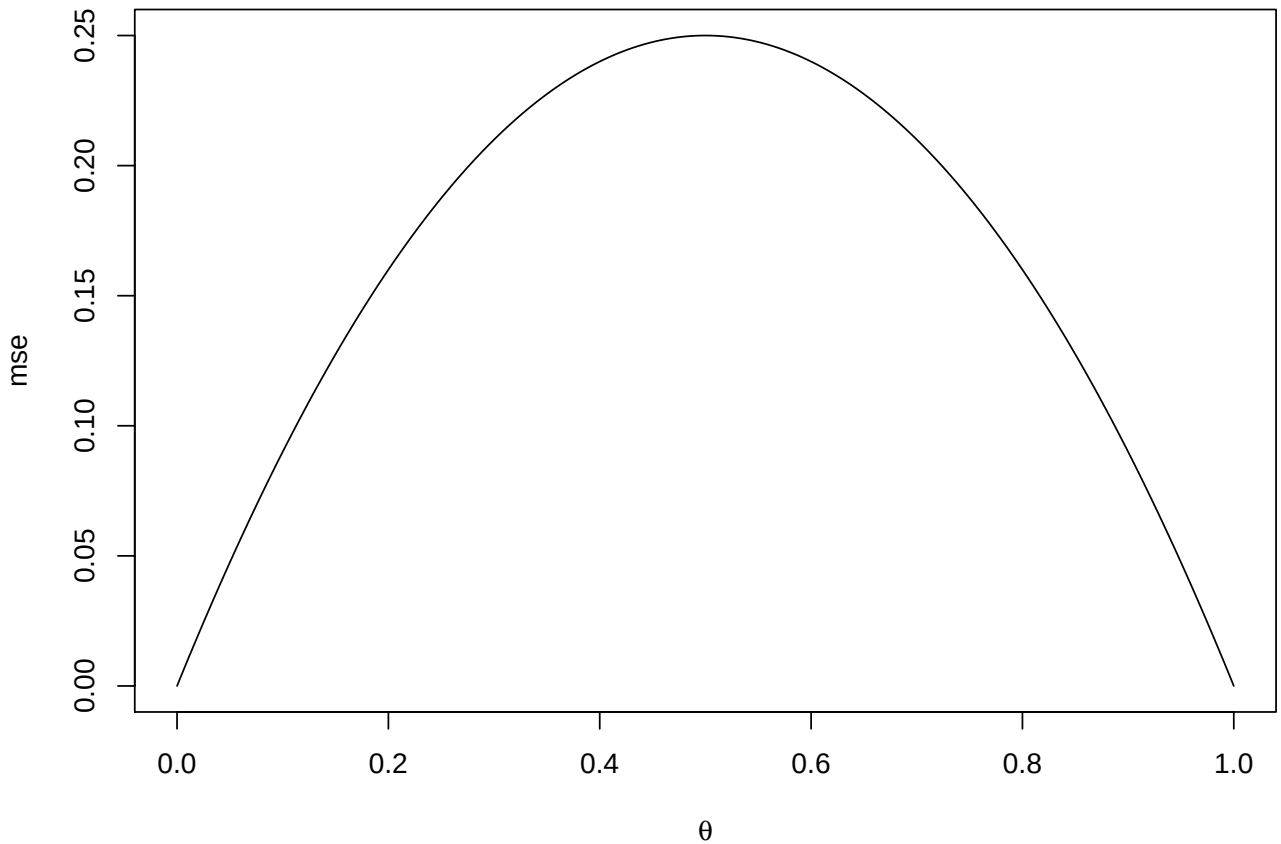
and hence

$$\text{MSE}_T(\theta) = \frac{\theta(1-\theta)}{n}.$$

Remarks

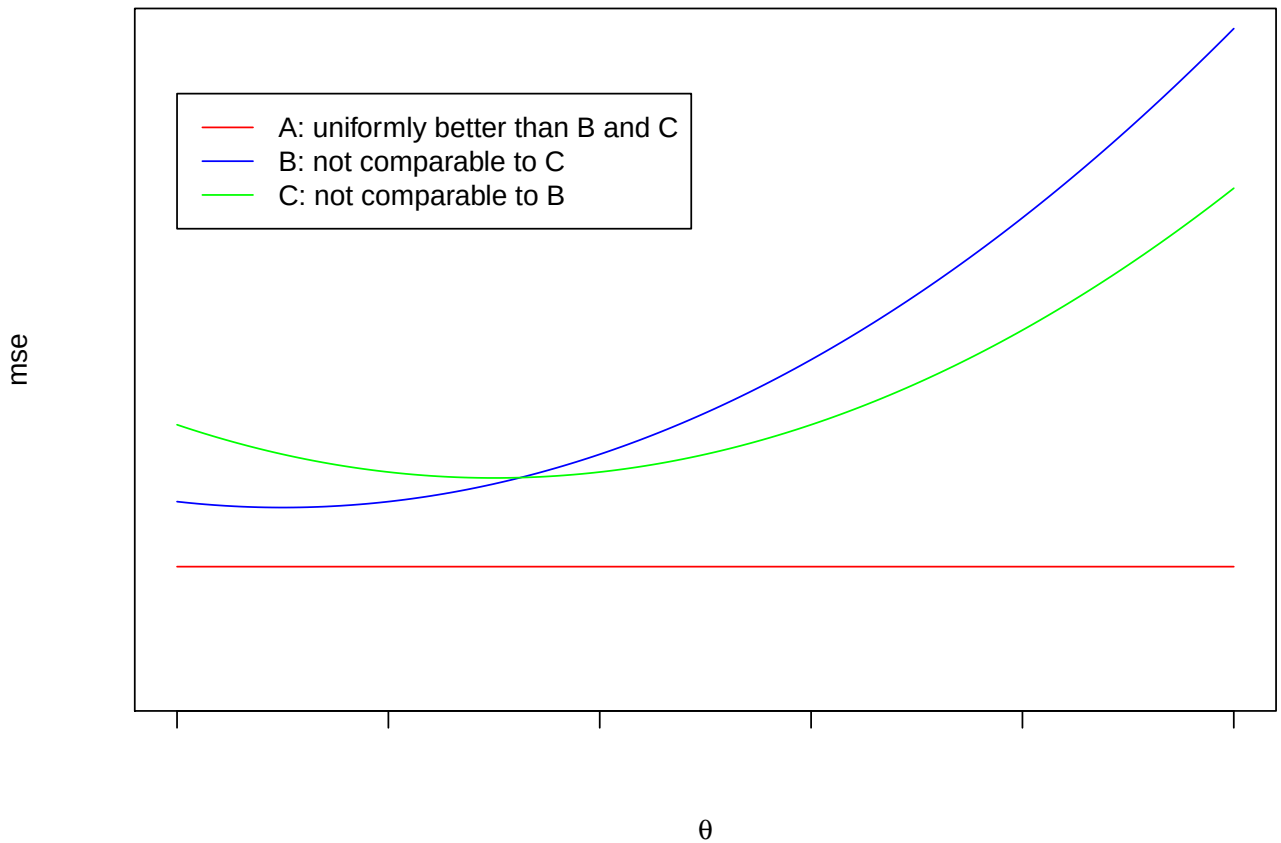
1. In this example it happens that $\text{MSE}_T = \text{var}(T)$ because $E(T) = \theta$.
2. This example also illustrates the fact that MSE_T is generally a function of θ rather than a single number.

MSE for a binomial proportion



Mean squared error

Intuitively, a good estimator is one for which MSE is as small as possible. However, quantifying this idea is complicated, because *MSE* is a function of θ, not just a number.



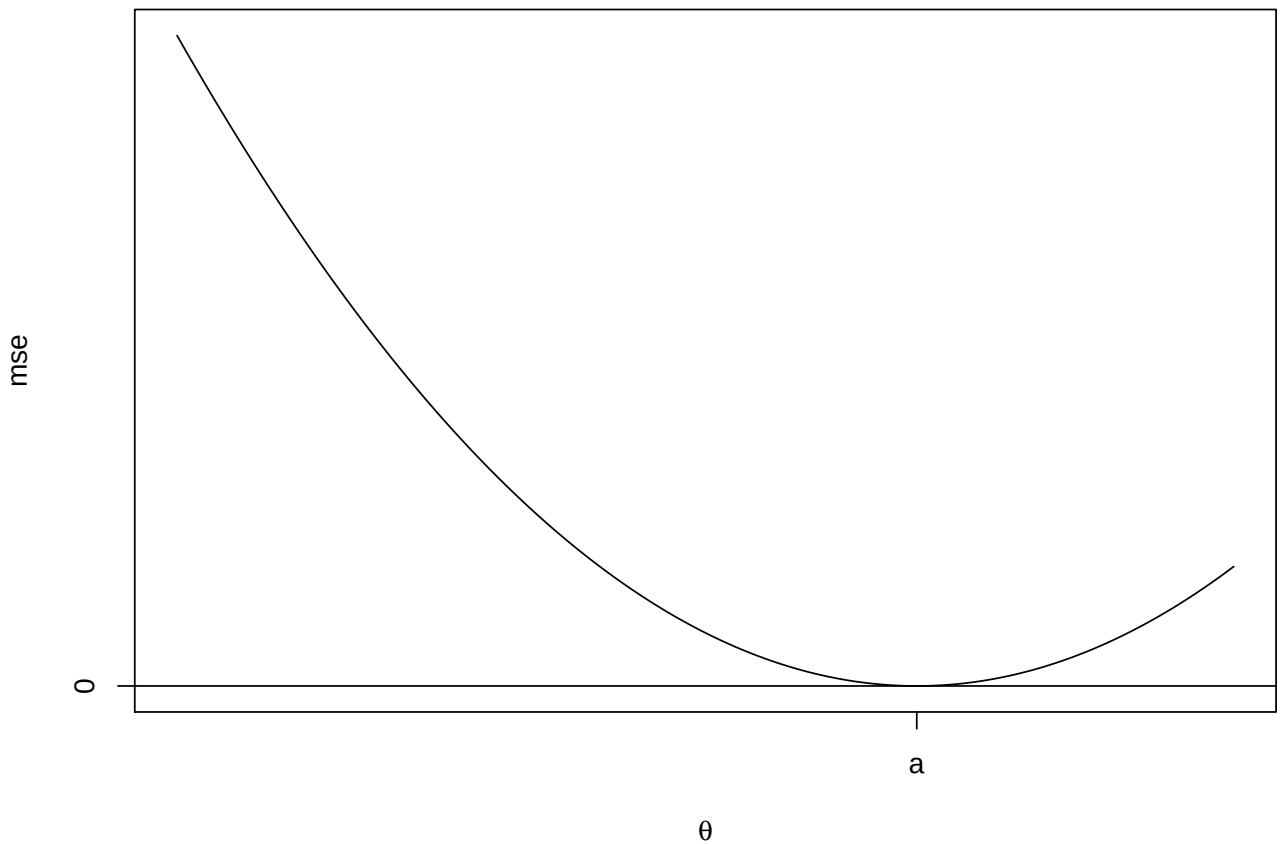
Non-existence of a minimum MSE estimator

It turns out that, except in trivial cases, there does not exist a uniformly minimum MSE estimator.

- Suppose T^* is a uniformly minimum MSE estimator for θ.
- Consider the trivial estimator $T_a = a$ for any constant a.
  - Note T_a is an estimator because it can be calculated from the data without knowledge of the parameter θ.
- The MSE of T_a is just $\text{MSE}_{T_a}(\theta) = (\theta - a)^2$ and hence

$$\text{MSE}_{T_a}(a) = 0.$$

Non-existence of a minimum MSE estimator



Non-existence of a minimum MSE estimator

- T^* would have to satisfy $\text{MSE}_{T^*}(\theta) \leq \text{MSE}_{T_a}(\theta)$ for all θ and, in particular,
$$\text{MSE}_{T^*}(a) = 0.$$
- Since a is arbitary, we must have
$$\text{MSE}_{T^*}(a) = 0$$
for all a.
- It follows that $T^* = \theta$ with probability 1.
- Hence we conclude that no such estimator exists except in trivial cases.

Unbiased estimators

Definition 42

- The bias of the estimator T is defined by:

$$b_T(\theta) = E(T) - \theta.$$

- An estimator T with $b_T(\theta) = 0$, so that

$$E(T) = \theta,$$

is said to be unbiased.

Remarks

- Although unbiasedness is an appealing property, not all commonly used estimates are unbiased.
  - The sample variance,

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

    is unbiased for σ^2.
  - However, the sample standard deviation $S = \sqrt{S^2}$ is a biased estimator for σ.
- In some situations it is impossible to construct unbiased estimators for the parameter of interest.
- Intuitively, having small bias would appear to be a pre-requisite for a good estimator.

Bias and MSE

Theorem 43

$$MSE_T(\theta) = \text{var}(T) + b_T(\theta)^2$$

Proof.

$$\begin{aligned}
\text{MSE}_T(\theta) &= E((T-\theta)^2) \\
&= E((T - E(T) + E(T) - \theta)^2) \\
&= E((T - E(T))^2) + (E(T) - \theta)^2 \\
&\quad +2(E(T) - \theta)E(T - E(T)) \\
&= \text{var}(T) + b_T(\theta).
\end{aligned}$$

□

The log-likelihood function

Consider data $x_1, x_2, \ldots, x_n$ assumed to be observations of random variables $X_1, X_2, \ldots, X_n$ with joint PDF $f_x(\boldsymbol{x}; \theta)$ or probability function $p_x(\boldsymbol{x}; \theta)$.

Definition 44

The likelihood function is defined by

$$L(\theta; \boldsymbol{x}) = \begin{cases} f_x(\boldsymbol{x}; \theta) & \text{for } X \text{ continuous} \\ p_x(\boldsymbol{x}; \theta) & \text{for } X \text{ discrete} \end{cases}$$

The log likelihood function is:

$$\ell(\theta; \boldsymbol{x}) = \log L(\theta; \boldsymbol{x}) \qquad \text{(log is the natural log i.e., ln).}$$

The log-likelihood function

If $X_1, X_2, \ldots, X_n$ are independent, the log likelihood function can be written as:

$$\ell(\theta; \boldsymbol{x}) = \sum_{i=1}^{n} \log f_i(x_i; \theta).$$

If $X_1, X_2, \ldots, X_n$ are IID, we have:

$$\ell(\theta; \boldsymbol{x}) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

Remark

Analogous formulae apply in the discrete case.

The score and the Fisher information

Definition 45

Consider a statistical problem with log-likelihood $\ell(\theta; \boldsymbol{x})$. The score is defined by:

$$\mathcal{U}(\theta; \boldsymbol{x}) = \frac{\partial \ell}{\partial \theta}$$

and the Fisher information is

$$\mathcal{I}(\theta) = E\left(\mathcal{U}(\theta; \boldsymbol{X})^2\right).$$

Remark

For a single observation with PDF $f(x, \theta)$, the information is

$$i(\theta) = E\left\{\left(\frac{\partial}{\partial \theta} \log f(X, \theta)\right)^2\right\}.$$

In the case of $x_1, x_2, \ldots, x_n$ IID , we have $\mathcal{I}(\theta) = ni(\theta)$. [STATS 3006-183]

The expected score

Consider a statistical problem with log-likelihood $\ell(\theta; \boldsymbol{x})$ and recall the score is defined by $\mathcal{U}(\theta; \boldsymbol{x}) = \frac{\partial \ell}{\partial \theta}$.

Theorem 46

Under suitable regularity conditions,

$$E\left(\mathcal{U}(\theta; \boldsymbol{X})\right) = 0.$$

Note

The regularity conditions are needed to allow the exchange of integration with respect to x and differentiation with respect to θ. The requirements are essentially:

- The PDF is differentiable with respect to θ;
- The domain of X does not depend on θ.

Proof

$$
\begin{aligned}
E\{\mathcal{U}(\theta;\boldsymbol{X})\} &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathcal{U}(\theta;\boldsymbol{x})f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial \log f(\boldsymbol{x};\theta)}{\partial\theta}f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial f(\boldsymbol{x};\theta)/\partial\theta}{f(\boldsymbol{x};\theta)}f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial f(\boldsymbol{x};\theta)}{\partial\theta}dx_1\ldots dx_n \\
\text{(regularity)} &= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
&= \frac{\partial}{\partial\theta}1 = 0, \text{ as required.}
\end{aligned}
$$

□

Variance of the score

Theorem 47

Under suitable regularity conditions,

$$\text{var}(\mathcal{U}(\theta; \boldsymbol{X})) = \mathcal{I}(\theta).$$

Proof.

Recall, by definition,

$$\mathcal{I}(\theta) = E\left(\mathcal{U}(\theta; \boldsymbol{X})^2\right).$$

From Theorem 46, we have $E(\mathcal{U}(\theta; \boldsymbol{X}) = 0$, whereby

$$\text{var}(\mathcal{U}(\theta; \boldsymbol{X})) = \mathcal{I}(\theta),$$

as required. □

Information and curvature of likelihood

Theorem 48

Under suitable regularity conditions,

$$\mathcal{I}(\theta) = E\left(-\frac{\partial^2 \ell(\theta; \boldsymbol{X})}{\partial \theta^2}\right).$$

Note

The regularity conditions are similar to those required for Theorem 46.

Proof

Observe first that

$$\begin{aligned}\frac{\partial^2 \ell(\theta;\boldsymbol{x})}{\partial\theta^2} &= \frac{\partial}{\partial\theta}\left(\frac{(\partial f(\boldsymbol{x};\theta)/\partial\theta)}{f(\boldsymbol{x};\theta)}\right)\\ &= \frac{(\partial^2 f(\boldsymbol{x};\theta)/\partial\theta^2)f(\boldsymbol{x};\theta) - (\partial f(\boldsymbol{x};\theta)/\partial\theta)^2}{f(\boldsymbol{x};\theta)^2}\\ &= \frac{(\partial^2 f(\boldsymbol{x};\theta)/\partial\theta^2)}{f(\boldsymbol{x};\theta)} - \mathcal{U}(\theta;\boldsymbol{x})^2.\end{aligned}$$

Taking expectation of both sides yields,

$$E\left(\frac{\partial^2 \ell(\theta;\boldsymbol{X})}{\partial\theta^2}\right) = E\left(\frac{\partial^2 f(\boldsymbol{X};\theta)/\partial\theta^2}{f(\boldsymbol{X};\theta)}\right) - \mathcal{I}(\theta). \qquad (3)$$

Proof (continued)

Next, observe,

$$
\begin{aligned}
& E\left(\frac{\partial^2 f(\boldsymbol{X};\theta)/\partial\theta^2}{f(\boldsymbol{X};\theta)}\right) \\
= & \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left(\frac{\partial^2 f(\boldsymbol{x};\theta)/\partial\theta^2}{f(\boldsymbol{x};\theta)}\right) f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
= & \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{\partial^2}{\partial\theta^2} f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
= & \frac{\partial^2}{\partial\theta^2}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
= & \frac{\partial^2}{\partial\theta^2}1 = 0,
\end{aligned}
$$

Proof (continued)

Substituting in (3), gives

$$E\left(\frac{\partial^2 \ell(\theta; \boldsymbol{x})}{\partial \theta^2}\right) = 0 - \mathcal{I}(\theta)$$

and, hence,

$$\mathcal{I}(\theta) = E\left(-\frac{\partial^2 \ell(\theta; \boldsymbol{X})}{\partial \theta^2}\right)$$

as required.

□

Example

Suppose $Y \sim B(n, \theta)$.

- $\ell(\theta) = \log \binom{n}{y} + y \log \theta + (n - y) \log(1 - \theta)$.
- $\mathcal{U}(\theta; y) = \frac{y - n\theta}{\theta(1-\theta)}$
- Can check $E(\mathcal{U}(\theta; Y)) = 0$ since $E(Y) = n\theta$.
- Also follows that $\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)}$.
- To check Theorem 48,

$$\frac{\partial^2 \ell}{\partial \theta^2} = \frac{-n\theta(1-\theta) - (Y - n\theta)(1 - 2\theta)}{(\theta(1-\theta))^2}$$

so

$$E\left(-\frac{\partial^2 \ell}{\partial \theta^2}\right) = \frac{n}{\theta(1-\theta)} = \mathcal{I}(\theta)$$

as required.

The Cramér Rao lower bound

Theorem 49

If T is an unbiased estimator for θ, then, under suitable regularity conditions,

$$\text{var}(T) \geq \frac{1}{\mathcal{I}(\theta)}.$$

Proof

Consider

$$\text{cov}(T(\boldsymbol{X}), \mathcal{U}(\theta; \boldsymbol{X})) = E\Big(T(\boldsymbol{X})\mathcal{U}(\theta; \boldsymbol{X})\Big)$$

since, by Theorem 46,

$$E(\mathcal{U}(\theta, \boldsymbol{X})) = 0.$$

Proof (continued)

Now,

$$
\begin{aligned}
E\Big(T(\boldsymbol{X})\mathcal{U}(\theta;\boldsymbol{X})\Big) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(\boldsymbol{x})\mathcal{U}(\theta;\boldsymbol{x})f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(\boldsymbol{x})\frac{\partial f(\boldsymbol{x};\theta)/\partial\theta}{f(\boldsymbol{x};\theta)}f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(\boldsymbol{x})\frac{\partial}{\partial\theta}f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
&= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(\boldsymbol{x})f(\boldsymbol{x};\theta)dx_1\ldots dx_n \\
&= \frac{\partial}{\partial\theta}E\Big(T(\boldsymbol{X})\Big) \\
&= \frac{\partial}{\partial\theta}\theta = 1.
\end{aligned}
$$

Proof (continued)

Hence,

$$\text{cov}\left(T(\boldsymbol{X}), \mathcal{U}(\theta; \boldsymbol{X})\right) = 1.$$

Finally, consider the inequality,

$$\text{cov}\left(\mathcal{U}(\theta; \boldsymbol{X}), T(\boldsymbol{X})\right)^2 \leq \text{var}\left(\mathcal{U}(\theta; \boldsymbol{X})\right) \text{var}\left(T(\boldsymbol{X})\right),$$

so that

$$\text{var}\left(T(\boldsymbol{X})\right) \geq \frac{1}{\text{var}\left(\mathcal{U}(\theta; \boldsymbol{X})\right)} = \frac{1}{\mathcal{I}(\theta)}$$

to complete the proof.

□

Example

Suppose $Y \sim B(n, \theta)$ and consider $T = \hat{\theta} = Y/n$.

- Then, $E(T) = \theta$ so T is an unbiased estimator for θ.
- Have seen previously that

$$\text{var}(T) = \frac{1}{n}\theta(1-\theta).$$

- Have also seen that

$$\mathcal{I}(\theta) = \frac{n}{\theta(1-\theta)}.$$

- Since $\text{var}(T) = \frac{1}{\mathcal{I}(\theta)}$ we can conclude that T is the minimum variance unbiased estimator for θ.

Estimators that achieve the CRLB

Theorem 50

The unbiased estimator $T(\boldsymbol{X})$ can achieve the Cramer-Rao Lower Bound only if the joint PDF (or probability function) has the form:

$$f(\boldsymbol{x};\theta) = exp\{A(\theta)T(\boldsymbol{x}) + B(\theta) + h(\boldsymbol{x})\},$$

where A, B are functions such that $\theta = \frac{-B'(\theta)}{A'(\theta)}$, and h is some function of $\boldsymbol{x}$.

Proof

Observe from Theorem 49 that var$(T(\boldsymbol{X})$ attains the CRLB when

$$\text{cor}\Big(T(\boldsymbol{X}), \mathcal{U}(\boldsymbol{X})\Big) = 1.$$

Proof (continued)

Recall that two random variables, U and T have $\text{cor}(U, T) = 1$ if and only if $U = aT + b$ for some constants $a > 0$ and b.

Hence we must have

$$U(\theta; \boldsymbol{X}) = a(\theta)T(\boldsymbol{X}) + b(\theta)$$

for some functions $a(\theta) > 0$ and $b(\theta)$

Note that $a(\theta)$ and $b(\theta)$ are constant functions (as opposed to random variables) in this context.

Now recall that $\mathcal{U}(\theta; \boldsymbol{x}) = \frac{\partial \ell}{\partial \theta}$, so

$$\begin{aligned} \ell(\theta; \boldsymbol{x}) &= \int a(\theta)T(\boldsymbol{x}) + b(\theta)d\theta \\ &= A(\theta)T(\boldsymbol{x}) + B(\theta) + h(\boldsymbol{x}) \end{aligned}$$

Proof (continued)

where $A'(\theta) = a(\theta)$, $B'(\theta) = b(\theta)$ and $h(\boldsymbol{x})$ is the combined constant (with respect to θ) of integration.

Hence,

$$f(\boldsymbol{x};\theta) = \exp(\ell(\theta;\boldsymbol{x})) = \exp\{A(\theta)T(\boldsymbol{x}) + B(\theta) + h(\boldsymbol{x})\}$$

as required.

Now, recall that $E\{\mathcal{U}(\theta;\boldsymbol{X})\} = 0$ and observe

$$E\Big(\mathcal{U}(\theta;\boldsymbol{X})\Big) = E\Big(A'(\theta)T(\boldsymbol{X}) + B'(\theta)\Big) = A'(\theta)E\Big(T(\boldsymbol{X})\Big) + B'(\theta)$$

Hence, $E\Big(T(\boldsymbol{X})\Big) = \frac{-B'(\theta)}{A'(\theta)}$, and, in order that $T(\boldsymbol{X})$ be unbiased for θ, it must be that

$$\frac{-B'(\theta)}{A'(\theta)} = \theta.$$

□

Example

Suppose $Y \sim B(n, \theta)$. The log-likelihood is

$$\begin{aligned}
\ell(\theta; Y) &= \exp\left(\log\binom{n}{y} + y\log\theta + (n-y)\log(1-\theta)\right) \\
&= \exp\left(\frac{y}{n} \times n\log\frac{\theta}{1-\theta} + n\log(1-\theta) + \log\binom{n}{y}\right) \\
&= \exp\Big(T(y)A(\theta) + B(\theta) + h(y)\Big)
\end{aligned}$$

where

$$A(\theta) = n\log\frac{\theta}{1-\theta},\ B(\theta) = n\log(1-\theta),\ \text{and}\ h(y) = \log\binom{m}{y}.$$

Example (continued)

Observe

$$A'(\theta) = \frac{n}{\theta} - \frac{n}{1-\theta} = \frac{n}{\theta(1-\theta)}$$

and

$$B'(\theta) = -\frac{n}{1-\theta}$$

so,

$$-\frac{B'(\theta)}{A'(\theta)} = \theta$$

as required.

Exponential families

Definition 51

A probability density function (or probability function) is said to be a single parameter exponential family if it has the form:

$$f(x;\theta) = \exp\Big(A(\theta)t(x) + B(\theta) + h(x)\Big)$$

for all $x \in \mathcal{D} \subseteq \mathbb{R}$, where the $\mathcal{D}$ does not depend on θ.

□

If $X_1, \ldots, X_n$ is a random sample from an exponential family, the joint PDF (or probability function) becomes

$$\begin{aligned} f(\boldsymbol{x};\theta) &= \prod_{i=1}^{n} \exp\Big(A(\theta)t(x_i) + B(\theta) + h(x_i)\Big) \\ &= \exp\left(A(\theta)\sum_{i=1}^{n} t(x_i) + nB(\theta) + \sum_{i=1}^{n} h(x_i)\right). \end{aligned}$$

Minimum variance unbiased estimators

If $X_1, \ldots, X_n$ is a random sample from an exponential family, the quantity

$$T = \frac{1}{n} \sum_{i=1}^{n} t(x_i)$$

is the minimum variance unbiased estimator for the quantity

$$-\frac{B'(\theta)}{A'(\theta)}.$$

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from the Po(μ) distribution.

Observe now that

$$p(x) = \frac{e^{-\mu}\mu^x}{x!} = \exp\Big(x \log \mu - \mu - \log(x!)\Big).$$

so the Po(μ) distribution is an exponential family with

$$t(x) = x,\ A(\mu) = \log(\mu),\ B(\mu) = -\mu,\ \text{and}\ h(x) = -\log(x!).$$

It follows that

$$T = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X} \text{ is the MVUE for } -\frac{B'(\mu)}{A'(\mu)} = \mu.$$

Sufficient Statistics

Definition 52

Consider data with PDF (or probability function), $f(\boldsymbol{x};\theta)$. A statistic, $S(\boldsymbol{x})$, is called a sufficient statistic for θ if $f(\boldsymbol{x}|s;\theta)$ does not depend on θ for all s.

□

Remarks

1. We will see that sufficient statistics capture all of the information in the data $\boldsymbol{x}$ that is relevant to θ.
2. If we consider vector-valued statistics, then this definition admits trivial examples, such as $\boldsymbol{s} = \boldsymbol{x}$, since

$$P(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{S} = \boldsymbol{s}) = \begin{cases} 1 & \text{if } \boldsymbol{x} = \boldsymbol{s} \\ 0 & \text{otherwise.} \end{cases}$$

which does not depend on θ.

Example

Suppose $x_1, x_2, \ldots, x_n$ are IID Bernoulli-θ observations and let $S(\mathbf{x}) = \sum_{i=1}^{n} x_i$. Then S is sufficient for θ.

Proof.

$$P(\boldsymbol{x}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)} = \theta^{s}(1-\theta)^{n-s}.$$

Next, observe that

$$p(\boldsymbol{x}|s) = \frac{P(\{\boldsymbol{X} = \boldsymbol{x}\} \cap \{S = s\})}{P(S = s)} = \frac{P(\boldsymbol{X} = \boldsymbol{x})}{P(S = s)}$$

Finally, since $S \sim \mathrm{B}(n, \theta)$

$$p(\boldsymbol{x}|s) = \begin{cases} \frac{1}{\binom{n}{s}} & \text{if } \sum_{i=1}^{n} x_i = s \\ 0, & \text{otherwise.} \end{cases}$$

The factorisation theorem

Theorem 53

Suppose $X_1, \ldots, X_n$ have joint PDF (or probability function) $f(\boldsymbol{x}; \theta)$. Then $S = S(\boldsymbol{x})$ is a sufficient statistic for θ if and only if

$$f(\boldsymbol{x}; \theta) = g(S(\boldsymbol{x}); \theta) h(\boldsymbol{x})$$

for some functions g, h.

Proof (discrete case)

Suppose $\boldsymbol{X}$ has joint probability function

$$p(\boldsymbol{x}; \theta) = g(S(\boldsymbol{x}); \theta) h(\boldsymbol{x})$$

and observe that $S(\boldsymbol{X})$ has marginal probability function

$$p_S(s) = \sum_{\boldsymbol{y}: S(\boldsymbol{y}) = s} p(\boldsymbol{y}; \theta) = g(s; \theta) \sum_{\boldsymbol{y}: S(\boldsymbol{y}) = s} h(\boldsymbol{y})$$

Proof (continued)

Hence,

$$P(\boldsymbol{x}|s;\theta) = \begin{cases} \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y}:S(\boldsymbol{y})=s} h(\boldsymbol{y})} & \text{if } S(\boldsymbol{x}) = s \\ 0 & \text{otherwise,} \end{cases}$$

which does not depend on θ.

Conversely, suppose $S(\boldsymbol{X})$ is a sufficient statistic for θ, so that

$$p(\boldsymbol{x}|s;\theta) = p(\boldsymbol{x}|s)$$

does not depend on θ. We can always write

$$p(\boldsymbol{x};\theta) = p(s;\theta)p(\boldsymbol{x}|s;\theta)$$

and in this case take

$$g(s;\theta) = p(s;\theta) \text{ and } h(\boldsymbol{x}) = p(\boldsymbol{x}|s)$$

as required. □

Remark

The proof in the continuous case is more involved because $\boldsymbol{X}|s$ need not have a continuous distribution.

Example

Suppose $X_1, X_2, \ldots, X_n$ IID $\mathrm{N}(\mu, \sigma^2)$, σ^2 known. Then,

$$
\begin{aligned}
f(\boldsymbol{x}; \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/(2\sigma^2)} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(\frac{1}{2\sigma^2}\left(2\mu\sum_{i=1}^{n} x_i - n\mu^2\right)\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right)
\end{aligned}
$$

Hence, by the factorisation theorem,

$$S = \sum_{i=1}^{n} X_i$$

is a sufficient statistic for μ.

[STATS 3006-209]

Example

If $X_1, X_2, \ldots, X_n$ are IID with an exponential family distribution

$$f(x) = \exp\Big(A(\theta)t(x) + B(\theta) + h(x)\Big).$$

Then

$$f(\boldsymbol{x}; \theta) = \exp\left(A(\theta)\sum_{i=1}^{n} t(x_i) + nB(\theta)\right)\exp\left(\sum_{i=1}^{n} h(x_i)\right)$$

so

$$S = \sum_{i=1}^{n} t(X_i)$$

is sufficient for θ by the factorization theorem.

The Rao-Blackwell Theorem

Theorem 54

If T is an unbiased estimator for θ and S is a sufficient statistic for θ, then the quantity

$$T^* = E(T|S)$$

is also an unbiased estimator for θ with $\text{var}(T^*) \leq \text{var}(T)$. *Moreover,* $\text{var}(T^*) = \text{var}(T)$ *if and only if $T^* = T$ with probability 1.*

Proof

To establish unbiasedness, observe

$$E(T^*) = E(E(T|S)) = E(T) = \theta$$

since, by assumption, T is unbiased for θ.

Proof (continued)

To show $\text{var}(T^*) \leq \text{var}(T)$ observe,

$$\begin{aligned} \text{var}(T) &= \text{var}(E(T|S)) + E(\text{var}(T|S)) \\ &= var(T^*) + E(\text{var}(T|S)) \\ &\leq \text{var}(T^*) \end{aligned}$$

as required.

Finally, observe that

$$\text{var}(T) = \text{var}(T^*) \Leftrightarrow E(\text{var}(T|S)) = 0 \Leftrightarrow \text{var}(T|S) = 0$$

for almost all s, or equivalently $T = E(T|S) = T^*$ with probability 1.

□

Remark

In the proof of the Rao-Blackwell theorem, the role of sufficiency has not been made explicit.

The reason for this requirement is to ensure that T^* is a statistic.

That is, that T^* does not depend on θ.

Since S is assumed to be sufficient for θ,

$$T^* = \int_{-\infty}^{\infty} T(\boldsymbol{x}) f(\boldsymbol{x}|s) d\boldsymbol{x},$$

which does not depend on θ.

Example (Rao-Blackwell Theorem)

Suppose $X_1, \ldots, X_n$ IID $\mathrm{N}(\mu, \sigma^2)$ with σ^2 known.

Want to estimate μ.

Take $T = X_1$, as an unbiased estimator for μ.

Know $S = \sum_{i=1}^{n} X_i$ is a sufficient statistic for μ.

According to the Rao-Blackwell theorem, $T^* = E(T|S)$ will be an improved (unbiased) estimator for μ.

Example (continued)

If $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, then

$$\boldsymbol{X} \sim N_n(\mu \mathbf{1}, \sigma^2 I).$$

To find the joint distribution of (T, S), observe

$$\begin{pmatrix} T \\ S \end{pmatrix} = A\boldsymbol{X} \text{ where } A = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 1 \end{pmatrix}$$

and hence

$$\begin{pmatrix} T \\ S \end{pmatrix} \sim N_2\left(A(\mu \mathbf{1}), \sigma^2 A A^T\right) = N_2\left(\begin{pmatrix} \mu \\ n\mu \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & n\sigma^2 \end{pmatrix}\right).$$

Example (continued)

The conditional distribution of $T|S$ is given by

$$T|s \sim N\left(\mu + \frac{\sigma^2}{n\sigma^2}(s - n\mu), \sigma^2 - \frac{\sigma^4}{n\sigma^2}\right) = N\left(\frac{1}{n}s, \left(1 - \frac{1}{n}\right)\sigma^2\right).$$

Hence

$$T^* = E(T|S) = \frac{1}{n}\sum X_i = \bar{X}.$$

is a better (or equal) estimator for μ.

In this case T^* is the MVUE but this is not always guaranteed to be the case.

Remarks

- The example shows that the Rao-Blackwell theorem can be used to improve estimators.
- The more important consequence is the principle that estimators should be functions of a sufficient statistic.
- There is further theory of sufficient statistics that we do not discuss:
  - Minimal sufficiency;
  - Completeness.

Construction of estimators - method of moments

Consider a random sample $X_1, X_2, \ldots, X_n$ from $F(\theta)$ & let $\mu = \mu(\theta) = E(X)$.

Definition 55

The method of moments estimator $\tilde{\theta}$ is defined as the solution to the equation

$$\bar{x} = \mu(\tilde{\theta}).$$

Example

Suppose $X_1, \ldots, X_n \sim$ IID Exp(λ) and note $E(X) = \frac{1}{\lambda}$. The method of moments estimator is defined as the solution to the equation

$$\bar{x} = \frac{1}{\tilde{\lambda}}$$

and hence

$$\tilde{\lambda} = \frac{1}{\bar{x}}.$$

Vector parameters

The method of moments estimator can be adapted to a vector valued parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$.

- Let $\mu_k(\boldsymbol{\theta}) = E(X^k)$ for $k = 1, 2, \ldots, p$.
- Let $m_k = \dfrac{1}{n} \sum_{i=1}^{n} x_i^k$.
- The method of moments estimator is defined to be the solution to the system of equations

$$\mu_k(\boldsymbol{\theta}) = m_k \text{ for } k = 1, 2, \ldots, p.$$

Example

Suppose $X_1, X_2, \ldots, X_n$ are IID $\mathrm{N}(\mu, \sigma^2)$ and let $\boldsymbol{\theta} = (\mu, \sigma^2)$.

$p = 2 \Rightarrow$ two equations in two unknowns:

$$\mu_1(\boldsymbol{\theta}) = E(X) = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$\mu_2(\boldsymbol{\theta}) = E(X^2) = \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

The method of moments estimator is therefore

$$\tilde{\mu} = \bar{x} \text{ and } \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Remarks

1. The method of moments is appealing in its simplicity and is motivated by the fact that $\bar{X}$ is the BLUE for μ.
2. The method of moments has desirable statistical properties under fairly mild assumptions:
    1. Consistency:

$$\tilde{\theta} \to \theta \text{ in probability as } n \to \infty$$

    2. Asymptotic normality:

$$\mathcal{L}\left(\frac{\tilde{\theta} - \theta}{\sqrt{\text{var}(\tilde{\theta})}}\right) \to N(0,1) \text{ as } n \to \infty$$

Non-uniqueness of method of moments estimator

Suppose $X_1, \ldots, X_n$ is a random sample from F_{θ_X} and let $\tilde{\theta}_X$ be method of moments estimator.

Let $Y = h(X)$ for any invertible function $h(X)$.

Then $Y_1, \ldots, Y_n$ contains exactly the same information as $X_1, \ldots, X_n$.

Therefore we would like $\tilde{\theta}_Y = \tilde{\theta}_X$.

Unfortunately method of moments estimation does not have this property.

Example

Consider $X_1, \ldots, X_n$ IID Exp(λ).

We saw previously that

$$\tilde{\lambda}_X = \frac{1}{\bar{x}}.$$

Suppose $Y_i = X_i^2$ (which is invertible for $X_i > 0$).

To obtain $\tilde{\lambda}_Y$, observe

$$E(Y) = E(X^2) = \frac{2}{\lambda^2} \text{ and hence } \tilde{\lambda}_Y = \sqrt{\frac{2n}{\sum x_i^2}} \neq \frac{1}{\bar{x}}.$$

Construction of estimators - maximum likelihood

Consider a statistical problem with log-likelihood, $\ell(\theta; \boldsymbol{x})$.

Definition 56

The maximum likelihood estimate (MLE) $\hat{\theta}$ is the solution to the problem

$$\max_{\theta \in \Theta} \ell(\theta; \boldsymbol{x})$$

That is,

$$\hat{\theta} = \operatorname*{argmax}_{\theta \in \Theta} \ell(\theta; \boldsymbol{x}).$$

Remark

In practice, maximum likelihood estimates are obtained by solving the score equation

$$\frac{\partial}{\partial \theta} \ell(\theta; \boldsymbol{x}) = \mathcal{U}(\theta; \boldsymbol{x}) = 0$$

Example

Suupose $Y \sim B(n, \theta)$.

The log-likelihood function is

$$\ell(\theta; y) = \log\left\{\binom{n}{y}\theta^y(1-\theta)^{n-y}\right\} = \log\binom{n}{y} + y\log\theta + (n-y)\log(1-\theta).$$

The score function is

$$\mathcal{U}(\theta; y) = \frac{y}{\theta} - \frac{n-y}{1-\theta} = \frac{y - n\theta}{\theta(1-\theta)}.$$

The MLE is therefore

$$\hat{\theta} = \frac{y}{n}.$$

Invariance of the MLE

Suppose X has PDF $f(x;\theta)$ and $Y = h(X)$, where h is strictly monotonic. Then,

$$f_Y(y;\theta) = f_X(h^{-1}(y);\theta)|h^{-1}(y)'|.$$

Consider data $x_1, x_2, \ldots, x_n$ transformed to $y_1, y_2, \ldots, y_n$:

$$\begin{aligned}
\ell_Y(\theta;\boldsymbol{y}) &= \log\left\{\prod_{i=1}^{n} f_Y(y_i;\theta)\right\} \\
&= \log\left\{\prod_{i=1}^{n} f_X(h^{-1}(y_i);\theta)|h^{-1}(y_i)'|\right\} \\
&= \log\prod_{i=1}^{n} f_X(x_i;\theta) + \log\prod_{i=1}^{n}|h^{-1}(y_i)'| \\
&= \ell_X(\theta;\boldsymbol{x}) + \log\left(\prod_{i=1}^{n}|h^{-1}(y_i)'|\right).
\end{aligned}$$

Invariance (continued)

Since $\log\left(\prod |h^{-1}(y_i)'|\right)$ does not depend on θ, it follows that $\hat{\theta}$ maximizes $\ell_X(\theta; \boldsymbol{x})$ if and only if it maximizes $\ell_Y(\theta; \boldsymbol{y})$.

That is, the value of $\hat{\theta}$ doesn't change if the data are transformed.

Transformations of the parameter

Suppose $\phi = \phi(\theta)$ is a 1-1 transformation of θ and observe that

$$\ell_\phi(\phi(\theta); \boldsymbol{x}) = \ell_\theta(\theta, \boldsymbol{x}).$$

Hence,

$$\underset{\theta}{\operatorname{argmax}}\, \ell_\phi(\phi(\theta); \boldsymbol{x}) = \underset{\theta}{\operatorname{argmax}}\, \ell_\theta(\theta, \boldsymbol{x}).$$

and then the MLEs obey the transformation rule, $\hat{\phi} = \phi(\hat{\theta})$.

MLE is a function of sufficient statistic

Suppose $T(\boldsymbol{X})$ is a sufficient statistic for θ.

By the factorisation theorem,

$$f(\boldsymbol{x};\theta) = g(T(\boldsymbol{x});\theta)h(\boldsymbol{x})$$

so

$$\ell(\theta) = \log g(T(\boldsymbol{x});\theta) + \log h(\boldsymbol{x}).$$

Since $h(\boldsymbol{x})$ does not depend on θ,

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}} \log g(T(\boldsymbol{x});\theta),$$

so $\hat{\theta}$ depends on $\boldsymbol{x}$ only through the sufficient statistic $T(\boldsymbol{x})$.

Example

Suppose $X_1, X_2, \ldots, X_n$ are IID Exp(λ). Then

$$f_{\boldsymbol{X}}(\boldsymbol{x}; \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} = \lambda^n e^{-\lambda n \bar{x}}.$$

By the Factorization Theorem, $\bar{X}$ is sufficient for λ. To get the MLE,

$$\mathcal{U}(\lambda, \boldsymbol{x}) = \frac{\partial \ell}{\partial \lambda} = \frac{\partial}{\partial \lambda}(n \log \lambda - n\lambda \bar{x}) = \frac{n}{\lambda} - n\bar{x}.$$

Hence

$$\frac{\partial \ell}{\partial \lambda} = 0 \Rightarrow \frac{1}{\lambda} = \bar{x} \Rightarrow \hat{\lambda} = \frac{1}{\bar{x}}.$$

As required, $\hat{\lambda}$ is a function of the sufficient statistic $\bar{x}$.

Example (continued)

Now let $Y_i = \log X_i$ for $i = 1, 2, \ldots, n$.

If $X \sim$Exp(λ) and $Y = \log X$, then $f_Y(y) = \lambda e^{-\lambda e^y} e^y$.
Hence,

$$\ell(\lambda; \boldsymbol{y}) = n \log \lambda - \lambda \sum_{i=1}^{n} e^{y_i} + \sum_{i=1}^{n} y_i$$

and

$$\frac{\partial}{\partial \lambda} \ell(\lambda; \boldsymbol{y}) = \frac{n}{\lambda} - \sum_{i=1}^{n} e^{y_i}.$$

It follows that

$$\hat{\lambda} = n \Big/ \sum_{i=1}^{n} e^{y_i}.$$

Finally, note $y_i = \log x_i$, so that $e^{y_i} = x_i$ and

$$\hat{\lambda} = 1/\bar{x},$$

as required.

Example (continued)

Finally suppose $\theta = \log(\lambda)$. The log-likelihood becomes

$$\ell(\theta; \boldsymbol{x}) = n(\theta - e^{\theta}\bar{x})$$

so

$$\frac{\partial}{\partial\theta}\ell(\theta; \boldsymbol{x}) = n(1 - e^{\theta}\bar{x}).$$

Solving $\frac{\partial\ell}{\partial\theta} = 0$ to obtain the MLE yields,

$$\hat{\theta} = -\log\bar{x} = \log\hat{\lambda}$$

as required.

Estimating functions

Maximum likelihood estimation and method of moments can both be generated by the use of estimating functions.

An estimating function, $H(\boldsymbol{x};\theta)$, has the property
$E\Big(H(\boldsymbol{X};\theta)\Big) = 0.$

H can be used to define an estimator $\tilde{\theta}$ which is a solution to the equation

$$H(\boldsymbol{x};\theta) = 0.$$

The method of moments estimator is obtained by taking

$$H(\boldsymbol{(}x)) = \bar{x} - E(\bar{X}).$$

The maximum likelihood estimator is obtained by taking

$$H(\boldsymbol{x};\theta) = \mathcal{U}(\theta;\boldsymbol{x}).$$

Asymptotic theory

Definition 57

Suppose $X_1, X_2, X_3, \ldots$ is a sequence of IID random variables with common PDF (or probability function) $f(x;\theta)$ and let $T_n = T(X_1, X_2, \ldots, X_n)$ be an estimator for θ based on $X_1, X_2, \ldots, X_n$.

1. T is said to be a consistent estimator for θ if

$$T_n \to \theta \text{ in probability as } n \to \infty.$$

2. T is said to be asymptotically normal if it is consistent and

$$\mathcal{L}\left(\sqrt{nv}(T_n - \theta)\right) \to N(0,1)$$

for some v.

Asymptotic theory

Suppose $X_1, X_2, X_3, \ldots$ is a sequence of IID random variables with common PDF (or probability function) $f(x; \theta)$.

Assume:

1. θ_0 is the true value of the parameter θ;
2. f is s.t. $f(x; \theta_1) = f(x; \theta_2)$ for all $x \Rightarrow \theta_1 = \theta_2$.

We will show (in outline) that if $\hat{\theta}_n$ is the MLE based on $X_1, X_2, \ldots, X_n$, then:

1. Consistency:

$$\hat{\theta}_n \to \theta_0 \text{ in probability, as } n \to \infty.$$

2. Asymptotic Normality:

$$\mathcal{L}\left(\sqrt{ni(\theta_0)}(\hat{\theta}_n - \theta_0)\right) \to \mathrm{N}(0, 1) \text{ as } n \to \infty.$$

Consistency

Lemma 58

Suppose f is such that $f(x;\theta_1) = f(x;\theta_2)$ for all $x \Rightarrow \theta_1 = \theta_2$. Then, under suitable regularity conditions,

$$\ell^*(\theta) = E_{\theta_0}\Big(\log f(X;\theta)\Big)$$

is maximized uniquely by $\theta = \theta_0$.

Proof

$$\begin{aligned} & \ell^*(\theta) - \ell^*(\theta_0) \\ = & \int_{-\infty}^{\infty} (\log f(x;\theta)) f(x;\theta_0) dx - \int_{-\infty}^{\infty} (\log f(x;\theta_0)) f(x;\theta_0) dx \\ = & \int_{-\infty}^{\infty} \log\left(\frac{f(x;\theta)}{f(x;\theta_0)}\right) f(x;\theta_0) dx \end{aligned}$$

Proof of Lemma 58 (continued)

Since

$$\log(r) \leq r - 1$$

for all, $r > 0$ with equality if and only if $r = 1$, it follows that

$$\begin{aligned} \ell^*(\theta) - \ell^*(\theta_0) &\leq \int_{-\infty}^{\infty} \left(\frac{f(x;\theta)}{f(x;\theta_0)} - 1 \right) f(x;\theta_0) dx \\ &= \int_{-\infty}^{\infty} (f(x;\theta) - f(x;\theta_0)) dx \\ &= \int_{-\infty}^{\infty} f(x;\theta) dx - \int_{-\infty}^{\infty} f(x;\theta_0) dx = 0 \end{aligned}$$

Equality is achieved if and only if

$$\frac{f(x;\theta)}{f(x;\theta_0)} = 1 \text{ for all } x \Rightarrow f(x;\theta) = f(x;\theta_0) \text{ for all } x \Rightarrow \theta = \theta_0.$$

□

Consistency (continued)

Lemma 59

Let $\bar{\ell}_n(\theta; \boldsymbol{X}) = \frac{1}{n}\ell(\theta, X_1, \ldots, X_n)$. *Then, under suitable regularity conditions,*

$$\bar{\ell}_n(\theta; \boldsymbol{X}) \to \ell^*(\theta) \text{ in probability as } n \to \infty, \text{ for each } \theta \in \Theta.$$

Proof

Observe that

$$\begin{aligned}
\bar{\ell}_n(\theta; \boldsymbol{X}) &= \frac{1}{n} \log \prod_{i=1}^{n} f(X_i; \theta) \\
&= \frac{1}{n} \sum_{i=1}^{n} \log f(X_i; \theta) \\
&= \frac{1}{n} \sum_{i=1}^{n} L_i(\theta), \text{ where } L_i(\theta) = \log f(X_i; \theta).
\end{aligned}$$

Proof of Lemma 59 (continued)

Since $L_i(\theta)$ are IID with $E(L_i(\theta)) = \ell^*(\theta)$, it follows by the weak law of large numbers that

$$\bar{\ell}_n(\theta; \boldsymbol{X}) \to \ell^*(\theta) \text{ in probability as } n \to \infty.$$

□

Consistency

To summarise, we have proved:

1. $\bar{\ell}_n(\theta, \boldsymbol{x}) \to \ell^*(\theta)$ in probability as $n \to \infty$;
2. $\ell^*(\theta)$ is maximized when $\theta = \theta_0$.

Since $\hat{\theta}_n$ maximizes $\bar{\ell}_n(\theta, \boldsymbol{X})$, it can be proved, with suitable regularity conditions, that $\hat{\theta}_n \to \theta_0$ in probability as $n \to \infty$.

□

Asymptotic normality

Theorem 60

Let $\mathcal{U}_n(\theta; \boldsymbol{X})$ denote the score based on $X_1, \ldots, X_n$.

Then, under suitable regularity conditions,

$$\mathcal{L}\left(\frac{\mathcal{U}_n(\theta_0; \boldsymbol{X})}{\sqrt{ni(\theta_0)}}\right) \to N(0,1) \text{ as } n \to \infty.$$

Proof

$$\begin{aligned}
\mathcal{U}_n(\theta_0; \boldsymbol{X}) &= \frac{\partial}{\partial\theta} \log \prod_{i=1}^{n} f(X_i; \theta)|_{\theta=\theta_0} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i; \theta)|_{\theta=\theta_0} \\
&= \sum_{i=1}^{n} \mathcal{U}_i \text{ where } \mathcal{U}_i = \frac{\partial}{\partial\theta} \log f(X_i; \theta)|_{\theta=\theta_0}.
\end{aligned}$$

Proof (continued)

Since $\mathcal{U}_1, \mathcal{U}_2, \ldots$ are IID with $E(\mathcal{U}_i) = 0$ and $\text{var}(\mathcal{U}_i) = i(\theta)$, and

$$\left(\sum_{i=1}^{n} \mathcal{U}_i - nE(\mathcal{U}) \right) \Big/ \sqrt{n \, \text{var}(\mathcal{U})} = \frac{\mathcal{U}_n(\theta_0; \boldsymbol{x})}{\sqrt{ni(\theta_0)}}$$

it follows from the Central Limit Theorem that

$$\mathcal{L}\left(\frac{\mathcal{U}_n(\theta_0; \boldsymbol{x})}{\sqrt{ni(\theta_0)}} \right) \to \text{N}(0, 1).$$

as $n \to \infty$.

□

Asymptotic normality of the MLE

Theorem 61

Under suitable regularity conditions,

$$\mathcal{L}\left(\sqrt{ni(\theta_0)}\left(\hat{\theta}_n - \theta_0\right)\right) \to N(0,1) \text{ as } n \to \infty.$$

Outline Proof

Recall the mean value theorem,

$$\frac{f(b) - f(a)}{b - a} = f'(c) \text{ for some } a < b < c$$

or equivalently,

$$f(b) = f(a) + (b - a)f'(c).$$

Proof (continued)

Applying the mean value theorem to $\mathcal{U}_n$, yields

$$\mathcal{U}_n(\hat{\theta}_n) = \mathcal{U}_n(\theta_0) + \mathcal{U}_n'(\theta_1)(\hat{\theta}_n - \theta_0)$$

for some θ_1 between θ_0 and $\hat{\theta}_n$.

Since $\mathcal{U}_n(\hat{\theta}_n) = 0$ by definition, it follows that

$$\mathcal{U}_n(\theta_0) = -\mathcal{U}_n'(\theta_1)(\hat{\theta}_n - \theta_0)$$

and, hence,

$$\sqrt{ni(\theta_0)}(\hat{\theta}_n - \theta_0) = \frac{\mathcal{U}_n(\theta_0)}{\sqrt{ni(\theta_0)}} \times \frac{ni(\theta_0)}{-\mathcal{U}_n'(\theta_1)}.$$

Proof (continued)

Next observe by Theorem 60

$$\mathcal{L}\left(\frac{\mathcal{U}_n(\theta_0)}{\sqrt{ni(\theta_0)}}\right) \to N(0,1)$$

as $n \to \infty$.

Now consider, the quantity $-\frac{1}{n}\mathcal{U}_n'(\theta)$. Since

$$\mathcal{U}_n'(\theta) = \sum_{i=1}^{n} \frac{\partial^2 \ell(\theta; X_i)}{\partial \theta^2},$$

it follows from the weak law of large numbers and Theorem 46, that for each fixed θ

$$-\frac{1}{n}\mathcal{U}_n'(\theta) \to i(\theta) \text{ in probability, as } n \to \infty.$$

Proof (continued)

Since θ_1 lies between θ_0 and $\hat{\theta}_n$, and

$$\hat{\theta}_n \to \theta_0 \text{ in probability, as } n \to \infty$$

it can be shown that

$$-\frac{1}{n}\mathcal{U}_n'(\hat{\theta}_n) \to i(\theta_0) \text{ in probability, as } n \to \infty.$$

Finally, it can be conluded that

$$\frac{ni(\theta_0)}{-\mathcal{U}_n'(\theta_1)} \to 1 \text{ in probability, as } n \to \infty$$

and then that

$$\mathcal{L}\left(\sqrt{ni(\theta)}\left(\hat{\theta} - \theta_0\right)\right) \to N(0,1) \text{ as } n \to \infty.$$

□

Remarks

We have shown, under regularity conditions that maximum likelihood estimates are:

1. Invariant under transformation of the data;
2. Invariant under transformation of the parameter;
3. Always a function of a sufficient statistic;
4. Consistent;
5. Asymptotically normal;
6. Attains the Cramér-Rao lower bound asymptotically.

Tests of Hypotheses and confidence intervals

Motivating example

Suppose $X_1, X_2, \ldots, X_n$ are IID $N(\mu, \sigma^2)$, and consider $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$.

If σ^2 is known, then the test of H_0 with significance level α is defined by the test statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

and the rule, reject H_0 if $|z| \geq z(\alpha/2)$.

A $100(1-\alpha)\%$ CI for μ is given by

$$\left(\bar{X} - z(\alpha/2)\frac{\sigma}{\sqrt{n}}, \bar{X} + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right).$$

It is easy to check that the confidence interval contains all values of μ_0 that are acceptable null hypotheses.

Tests of Hypotheses

Consider a statistical problem with parameter

$$\theta \in \Theta = \Theta_0 \cup \Theta_A, \qquad \text{where } \Theta_0 \cap \Theta_A = \phi.$$

We consider the null hypothesis,

$$H_0 : \theta \in \Theta_0$$

and the alternative hypothesis

$$H_A : \theta \in \Theta_A.$$

The hypothesis testing set up can be represented as:

Test Result	Actual Status: H_0 true	Actual Status: H_A true
Accept H_0	$\surd$	type II error
Reject H_0	type I error	$\surd$

Neyman-Pearson theory

It is desirable to use a test that makes both the type I and type II error rates as small as possible.

However, these requirements conflict with each other.

- Reducing the type I error makes it harder to reject H_0.
- Reducing the type II error makes it easier to reject H_0.

The Neyman-Pearson approach to hypothesis testing is to control the type I error rate at a suitably small value of α and then choose a test that makes the type II error rate as small as possible.

Simple Hypotheses

Consider simple null and alternative hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_a : \theta = \theta_a$$

Recall that the type I error rate is defined by

$$\alpha = P(\text{reject } H_0 | H_0 \text{ true}).$$

and the type II error rate by

$$\beta = P(\text{accept } H_0 | H_0 \text{ false}).$$

The power is defined by $1 - \beta = P(\text{reject } H_0 | H_0 \text{ false})$ so minimising the type II error rate is equivalent to maximising power.

The Neyman-Pearson Lemma

Theorem 62

Consider the test $H_0 : \theta = \theta_0$ *vs.* $H_a : \theta = \theta_a$ *defined by the rule*

$$\textit{reject } H_0 \textit{ for } \frac{f(\boldsymbol{x}; \theta_0)}{f(\boldsymbol{x}; \theta_a)} \leq k,$$

for some constant $k > 0$.

Let α^* *be the type I error rate and* $1 - \beta^*$ *be the power for this test. Then any other test with* $\alpha \leq \alpha^*$ *will have* $(1 - \beta) \leq (1 - \beta^*)$.

Proof

Let

$$Ra^*(\boldsymbol{x}) = \begin{cases} 1 & \text{if } f(\boldsymbol{x};\theta_0)/f(\boldsymbol{x};\theta_a) \leq k \\ 0 & \text{otherwise} \end{cases}$$

so that $R^*(\boldsymbol{x})$ indicates whether H_0 is rejected or accepted by the likelihood ratio test.

Let $R(\boldsymbol{x})$ be the indicator for another test with $\alpha \leq \alpha^*$.

Observe

$$\alpha^* = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} R^*(\boldsymbol{x}) f(\boldsymbol{x};\theta_0) dx_1 \ldots dx_n$$

and

$$\alpha = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} R(\boldsymbol{x}) f(\boldsymbol{x};\theta_0) dx_1 \ldots dx_n.$$

Proof (continued)

Similarly

$$1-\beta^* = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} R^*(\boldsymbol{x}) f(\boldsymbol{x};\theta_a) dx_1 \ldots dx_n$$

and

$$1-\beta = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} R(\boldsymbol{x}) f(\boldsymbol{x};\theta_a) dx_1 \ldots dx_n.$$

Now observe

$$R^*(\boldsymbol{x})\{kf(\boldsymbol{x};\theta_a) - f(\boldsymbol{x};\theta_0)\} \geq R(\boldsymbol{x})\{kf(\boldsymbol{x};\theta_a) - f(\boldsymbol{x};\theta_0)\}$$

for all $\boldsymbol{x}$.

Proof (continued)

. Integrating both sides with respect to $\boldsymbol{x}$, we obtain

$$
\begin{aligned}
& k(1-\beta^*) - \alpha^* \geq k(1-\beta) - \alpha \\
\Rightarrow \quad & k(1-\beta^*) - (\alpha^* - \alpha) \geq k(1-\beta) \\
\Rightarrow \quad & k(1-\beta^*) \geq k(1-\beta) \text{ since } \alpha \leq \alpha^* \\
\Rightarrow \quad & (1-\beta^*) \geq (1-\beta) \text{ since } k > 0
\end{aligned}
$$

as required.

□

Size of the test

In hypothesis testing the size of the test is defined to be

$$\alpha = P(\text{Reject } H_0 | H_0).$$

For a composite null hypothesis, the size of the test is defined to be

$$\alpha = \sup_{\theta \in \Theta_0} P(\text{Reject } H_0 | \theta).$$

A test is said to have significance level α if the size of the test is less than or equal to α.

Ideally, the size and the significance level would be the same but this is not always possible, for example when the data are discrete.

Example (Neyman Pearson Lemma)

Suppose $X_1, X_2, \ldots, X_n$ are IID $N(\mu, \sigma^2)$, with σ^2 given, and consider

$$H_0 : \mu = \mu_0 \text{ vs } H_A : \mu = \mu_a, \qquad \mu_a > \mu_0$$

Then,

$$\begin{aligned}
\frac{f(\boldsymbol{x}; \mu_0)}{f(\boldsymbol{x}; \mu_a)} &= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu_0)^2\right\}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu_a)^2\right\}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\mu_0 + n\mu_0^2\right)\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\mu_a + n\mu_a^2\right)\right\}} \\
&= \exp\left\{\frac{1}{2\sigma^2}(2n\bar{x}(\mu_0 - \mu_a) - n\mu_0^2 + n\mu_a^2)\right\}
\end{aligned}$$

Example (continued)

For a constant k,

$$\frac{f(\boldsymbol{x};\mu_0)}{f(\boldsymbol{x};\mu_a)} \leq k \quad \Leftrightarrow \quad (\mu_0 - \mu_a)\bar{x} \leq k^* \quad \Leftrightarrow \quad \bar{x} \geq c,$$

for a suitably chosen c.

That is, H_0 is rejected when $\bar{x}$ is too large.

To choose c, we use the fact that

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1) \quad \text{under } H_0.$$

Hence, the usual z-test,

$$\text{reject } H_0 \text{ if } \bar{z} \geq z(\alpha)$$

is the Neyman-Pearson LR test in this case.

Remarks

1. This example shows that the one-sided z test is also uniformly most powerful for

$$H_0 : \mu = \mu_0 \text{ vs. } H_A : \mu > \mu_0$$

2. The result can be extended to the case of

$$H_0 : \mu \leq \mu_0 \text{ vs. } H_A : \mu > \mu_0$$

   In this case $\alpha = \max_{\mu \leq \mu_0} P(\text{Reject } H_0 \mid \mu)$.

3. The construction fails for two-sided alternatives, such as

$$H_0 : \mu = \mu_0 \text{ vs. } H_A : \mu \neq \mu_0$$

   and no uniformly most powerful test exists.

Large sample tests

Consider data $x_1, x_2, \ldots, x_n$ with log-likelihood function $\ell(\theta; \boldsymbol{x})$ where $\theta \in \mathbb{R}$ is a scalar parameter and the hypotheses

$$H_0 : \theta = \theta_0 \text{ vs } H_A : \theta \neq \theta_0.$$

The Wald test

The Wald test statistic is

$$W = \sqrt{\mathcal{I}(\hat{\theta})}\Big(\hat{\theta} - \theta_0\Big).$$

A test with approximate significance level α is defined by the critical region:

Reject H_0: for $|W| \geq z_{\alpha/2}$,

Accept H_0: for $|W| < z_{\alpha/2}$.

The Wald test (continued)

The approximate significance level derives from Theorem 59, the asymptotic normality of the maximum likelihood estimator,

$$\mathcal{L}\left(\sqrt{\mathcal{I}(\theta)}\left(\hat{\theta}-\theta\right)\right) \to N(0,1) \text{ as } n \to \infty.$$

The Score test

The score test statistic is defined by

$$U = \frac{\mathcal{U}(\theta_0, \boldsymbol{x})}{\sqrt{\mathcal{I}(\theta_0)}}.$$

A test with approximate significance level α is defined by the critical region:

Reject H_0: for $|U| \geq z_{\alpha/2}$,

Accept H_0: for $|U| < z_{\alpha/2}$.

The approximate significance level derives from Theorem 58, the asymptotic normality of the score,

$$\mathcal{L}\left(\mathcal{U}(\theta; \boldsymbol{X}) \Big/ \sqrt{\mathcal{I}(\theta)}\right) \to N(0,1) \text{ as } n \to \infty.$$

The likelihood ratio test

The likelihood ratio test statistic is defined by

$$G^2 = 2\Big(\ell(\hat{\theta}; \boldsymbol{x}) - \ell(\theta_0; \boldsymbol{x})\Big)$$

A test with approximate significance level α is defined by the critical region:

Reject H_0: for $G^2 \geq \chi^2_{1,\alpha}$,

Accept H_0: for $G^2 < \chi^2_{1,\alpha}$.

Example

Suppose $X_1, X_2, \ldots, X_n$ are IID (Po(λ)), and consider

$$H_0 : \lambda = \lambda_0 \text{ vs } H_A : \lambda \neq \lambda_0.$$

Recall

$$\begin{aligned} \ell(\lambda; \boldsymbol{x}) &= n(\bar{x} \log \lambda - \lambda) - \log \prod_{i=1}^{n} x_i!, \\ \mathcal{U}(\lambda; \boldsymbol{x}) &= \frac{n(\bar{x} - \lambda)}{\lambda}, \\ \mathcal{I}(\lambda) &= n/\lambda \text{ and} \\ \hat{\lambda} &= \bar{x}. \end{aligned}$$

Example (continued)

- The Wald test statistic is

$$W = \sqrt{\mathcal{I}(\hat{\lambda})}(\hat{\lambda} - \lambda_0) = \frac{\hat{\lambda} - \lambda_0}{\sqrt{\hat{\lambda}/n}} = \frac{\bar{x} - \lambda_0}{\sqrt{\hat{\lambda}/n}}.$$

- The Score test statistic is

$$U = \frac{\mathcal{U}(\lambda_0; \boldsymbol{x})}{\sqrt{\mathcal{I}(\lambda_0)}} = \frac{n(\bar{x} - \lambda_0)}{\lambda_0(\sqrt{n/\lambda_0})} = \frac{\bar{x} - \lambda_0}{\sqrt{\lambda_0/n}}.$$

- The likelihood ratio test statistic is

$$G^2 = 2(\ell(\hat{\lambda}) - \ell(\lambda_0)) = 2n\left(\bar{x}\log\frac{\hat{\lambda}}{\lambda_0} - (\hat{\lambda} - \lambda_0)\right).$$

Asymptotic equivalence of Wald and score tests

The Wald test and the score test are asymptotically equivalent under the null hypothesis.

Recall in the proof of theorem 59, we obtained

$$\sqrt{ni(\theta_0)}(\hat{\theta}_n - \theta_0) = \frac{\mathcal{U}_n(\theta_0)}{\sqrt{ni(\theta_0)}} \times \frac{ni(\theta_0)}{-\mathcal{U}_n'(\theta_1)}$$

and argued for large n that

$$\frac{ni(\theta_0)}{-\mathcal{U}_n'(\theta_1)} \to 1$$

in probability as $n \to \infty$.

Asymptotic equivalence (continued)

When H_0 is true, it can also be shown (under regularity conditions) that

$$\frac{\sqrt{ni(\theta_0)}}{\sqrt{ni(\hat{\theta})}} \to 1$$

in probability as $n \to \infty$. Since

$$\frac{\sqrt{ni(\theta_0)}}{\sqrt{ni(\hat{\theta})}} W = \frac{ni(\theta_0)}{-\mathcal{U}_n'(\theta_1)} U,$$

the two tests are asymptotically equivalent under H_0.

Asymptotic distribution of likelihood ratio test

An outline proof of the asymptotic χ^2_1 of for G^2 is obtained by considering the second order Taylor expansion of $\ell(\theta_0)$ about $\hat{\theta}$

$$\begin{aligned}\ell(\theta_0) &= \ell(\hat{\theta}) + \ell'(\hat{\theta})(\theta_0 - \hat{\theta}) \\ &\quad + \ell''(\hat{\theta})(\theta_0 - \hat{\theta})^2/2 + r(\theta_0 - \hat{\theta})\end{aligned}$$

where $\lim_{s \to 0} r(s)/s^2 = 0$.

But by definition, $\ell'(\hat{\theta}) = 0$, so that

$$2\Big(\ell(\hat{\theta}) - \ell(\theta_0)\Big) = -\ell''(\hat{\theta})(\hat{\theta} - \theta_0)^2 - r(\theta_0 - \hat{\theta})$$

and hence

$$G^2 = W^2 \frac{-\ell''(\hat{\theta})}{ni(\hat{\theta})} - r(\theta_0 - \hat{\theta}).$$

Asymptotic distribution of likelihood ratio test

It can be shown (under regularity conditions) that

$$\frac{-\ell''(\hat{\theta})}{ni(\hat{\theta})} \to 1$$

in probability as $n \to \infty$ and when H_0 is true,

$$r(\theta_0 - \hat{\theta}) \to 0$$

in probability.

Hence, G^2 is asymptotically equivalent to W^2 and since $W \sim N(0,1)$ under H_0, it follows that $G^2 \sim \chi^2_1$.

Remarks

1. In the formulation of Theorems 58 and 59, the symbol θ_0 was used to denote the true value of the parameter. In the present discussion, we are using θ_0 to represent the null hypothesis value.

2. The outline proofs have been given in the simple case of IID data so that
$$\mathcal{I}(\theta) = ni(\theta).$$
However, the theory applies more generally when the data are not IID.

Confidence intervals

A $100(1-\alpha)\%$ CI for θ is a random interval, (L, U) with the property

$$P((L, U) \ni \theta) = 1 - \alpha.$$

It is easy to check that the test defined by rule:

"*Accept* $H_0 : \theta = \theta_0$ *iff* $\theta_0 \in (L, U)$"

has significance level α.

Conversely, given a hypothesis test $H_0 : \theta = \theta_0$ with significance level α, it can be proved that the set

$$R = \{\theta_0 : H_0 : \theta = \theta_0 \text{ is accepted}\}$$

is a $100(1-\alpha)\%$ confidence region for θ.

Note we, describe the set as a confidence region as it may not be an interval.

Large sample confidence intervals

Each of the Wald, score and likelihood ratio tests can be inverted to obtain an approximate $100(1-\alpha)\%$ confidence interval.

Wald test: Solve for θ_0 in $|W| < z_{\alpha/2}$. The solution is the interval

$$\left(\hat{\theta} - z_{\alpha/2}\Big/\sqrt{\mathcal{I}(\hat{\theta})}, \hat{\theta} + z_{\alpha/2}\Big/\sqrt{\mathcal{I}(\hat{\theta})}\right).$$

Score test: Solve for θ_0 in $|U| < z_{\alpha/2}$.

Likelihood ratio test: Solve for θ_0 in $G^2 < \chi^2_{1,\alpha}$.

Example

Suppose $X_1, X_2, \ldots, X_n$ are IID (Po(λ)).

The Wald test is inverted by solving λ_0 in the equation

$$|W| < z_{\alpha/2}.$$

The result is the confidence interval

$$\left(\hat{\lambda} - z_{\alpha/2}\sqrt{\hat{\lambda}/n},\ \hat{\lambda} + z_{\alpha/2}\sqrt{\hat{\lambda}/n}\right).$$

Example (continued)

To invert the score test, we need to solve for λ_0 in the equation

$$
\begin{aligned}
& |U| < z_{\alpha/2} \\
\Rightarrow \quad & (\bar{x} - \lambda_0)/\sqrt{\lambda_0/n} < z_{\alpha/2} \\
\Rightarrow \quad & (\bar{x} - \lambda_0)^2/(\lambda_0/n) < z^2_{\alpha/2} \\
\Rightarrow \quad & (\bar{x} - \lambda_0)^2 < z^2_{\alpha/2}\lambda_0/n
\end{aligned}
$$

This is just a quadratic equation in λ_0 than can be solved explicitly.

Example (continued)

The invert the likelihood ratio test, we need to solve for λ_0 in the equation,

$$G^2 = 2n\left\{\bar{x}\log\frac{\bar{x}}{\lambda_0} - (\bar{x} - \lambda_0)\right\} < \chi^2_{1,\alpha}.$$

This can be solved numerically.

The three confidence intervals are illustrated in the case $n = 20$ and $\bar{x} = 7.5$

Example (continued)