MATHEMATICAL STATISTICS III

Lecture Notes

Lecturer: Professor Patty Solomon

THE UNIVERSITY OF ADELAIDE AUSTRALIA

Contents

1 Distribution Theory

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

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

The PDF is a piecewise continuous function which integrates to 1 over the range of the RV. Note that

$$P(X = a) = \int_a^a f(x)\,dx = 0 \qquad \text{for any continuous RV.}$$

Example: 'Precipitation' is neither a discrete nor a continuous RV, since there is zero precipitation on some days; it is a mixture of both.

The cumulative distribution function (CDF) is defined by:

$$F(x) = P(\{X \leq x\}) \qquad \text{(area to left of } x\text{)}$$

completed by:
$$F(x) = \begin{cases} \displaystyle\sum_{t;t\leq x} p(t) \\ \\ \displaystyle\int_{-\infty}^{x} f(t)\,dt. \end{cases}$$

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

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

provided that $\displaystyle\sum_x |x|p(x) < \infty$ (*i.e.*, must converge). 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 $\displaystyle\int_{-\infty}^{\infty} |x|f(x) < \infty$, otherwise it is *not* defined.

(For example, the Cauchy distribution is momentless.)

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

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

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

$\downarrow$
moment
generating fct
of RV X.

$M_X(0) = 1$ always; the mgf may or may not be defined for other values of t.

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

1. Moments of all orders exist;
2. $E[X^r] \;=\; M_X^{(r)}(0)$ (rth order derivative) ;
3. $M_X(t)$ uniquely determines the distribution of X:

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

1.1 Discrete distributions

1.1.1 Bernoulli distribution

Parameter: $0 \le p \le 1$

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

Prob. function:

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

$$\begin{aligned} E(X) &= p \\ \text{Var } (X) &= p(1-p) \\ M_X(t) &= 1 + p(e^t - 1). \end{aligned}$$

1.1.2 Binomial distribution

Parameter: $0 \leq p \leq 1; \quad n > 0;$

MGF: $M_X(t) = \{1 + p(e^t - 1)\}^n$.

Consider a sequence of n independent Bern(p) trials. If X = total number of successes, then $X \sim B(n, p)$.

Probability function:

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

Note: If $Y_1, \ldots, Y_n$ are underlying Bernoulli RV's then $X = Y_1 + Y_2 + \cdots + Y_n$. (same as counting the number of successes).

Probability function if:

$$\begin{aligned} &p(x) \geq 0, \qquad x = 0, 1, \ldots, n, \\ &\sum_{x=0}^{n} p(x) = 1. \end{aligned}$$

1.1.3 Geometric distribution

This is a discrete waiting time distribution. Suppose a sequence of independent Bernoulli trials is performed and let X be the number of failures preceding the first success. Then $X \sim Geom(p)$, with

$$p(x) = p(1-p)^x, \quad x = 0, 1, 2, \ldots$$

1.1.4 Negative Binomial distribution

Suppose a sequence of independent Bernoulli trials is conducted. If X is the number of failures preceding the n^{th} success, then X has a negative binomial distribution.

Probability function:

$$p(x) = \underbrace{\binom{n+x-1}{n-1}}_{\substack{\text{ways of allocating} \\ \text{failures and successes}}} p^n(1-p)^x,$$

$$E(X) = \frac{n(1-p)}{p},$$
$$\text{Var } (X) = \frac{n(1-p)}{p^2},$$
$$M_X(t) = \left[\frac{p}{1-e^t(1-p)}\right]^n.$$

1. If $n = 1$, we obtain the geometric distribution.
2. Also seen to arise as sum of n independent geometric variables.

1.1.5 Poisson distribution

Parameter: rate $\lambda > 0$

MGF: $M_X(t) = e^{\lambda(e^t-1)}$

Probability function:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

1. The Poisson distribution arises as the distribution for the number of "point events" observed from a Poisson process.

Examples:

Figure 1: Poisson Example

Number of incoming calls to certain exchange in a given hour.

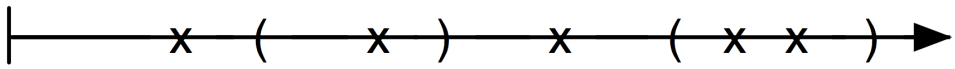
2. The Poisson distribution also arises as the limiting form of the binomial distribution:

$$n \to \infty, \qquad np \to \lambda$$
$$p \to 0$$

The derivation of the Poisson distribution (via the binomial) is underpinned by a Poisson process *i.e.*, a point process on $[0, \infty)$; see Figure 1.

AXIOMS for a Poisson process of rate $\lambda > 0$ are:

(A) The number of occurrences in disjoint intervals are independent.



(B) Probability of 1 or more occurrences in any sub-interval $[t, t+h)$ is $\lambda h + o(h) \quad (h \to 0)$ (approx prob. is equal to length of interval $x\lambda$).

(C) Probability of more than one occurrence in $[t, t+h)$ is $o(h) \quad (h \to 0)$ (i.e. prob is small, negligible).

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

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

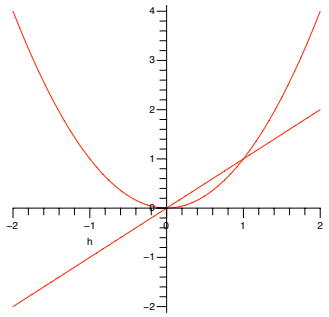


Figure 2: Small order h: functions h^4 (yes) and h (no)

1.1.6 Hypergeometric distribution

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.

Parameters: $M, N > 0, \qquad 0 < n \leq M + N$

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

Prob function:

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

$$E(X) = n\frac{M}{M+N}, \qquad \text{Var}(X) = \frac{M+N-n}{M+N-1}\frac{nMN}{(M+N)^2}.$$

The mgf exists, but there is no useful expression available.

1. The hypergeometric distribution is simply

$$\frac{\#\text{ samples with } x \text{ black balls}}{\#\text{ possible samples}},$$

$$= \frac{\binom{M}{x}\binom{N}{n-x}}{\binom{M+N}{n}}.$$

2. To see how the limits arise, observe we must have $x \leq n$ (i.e., no more than sample size of black balls in the sample.) Also, $x \leq M$, i.e., $x \leq \min(n, M)$.

   Similarly, we must have $x \geq 0$ (i.e., cannot have < 0 black balls in sample), and $n - x \leq N$ (i.e., cannot have more white balls than number in urn).

   i.e. $x \geq n - N$
   i.e. $x \geq \max(0, n - N)$.

3. If we sample with replacement, we would get $X \sim B\left(n, p = \frac{M}{M+N}\right)$. It is interesting to compare moments:

finite population correction
↑

hypergeometric	$E(X) = np$	Var $(X) = \frac{M+N-n}{M+N-1}\,[np(1-p)]$
binomial	$E(x) = np$	Var $(X) = np(1-p)$ ↓

when sample all balls in urn Var$(X) \sim 0$

4. When $M, N >> n$, the difference between sampling with and without replacement should be small.

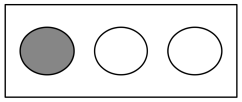


Figure 3: $p = \dfrac{1}{3}$

If white ball out →

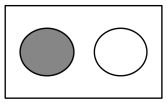


Figure 4: $p = \dfrac{1}{2}$ (without replacement)

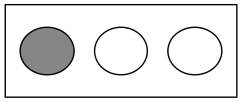


Figure 5: $p = \dfrac{1}{3}$ (with replacement)

Intuitively, this implies that for $M, N >> n$, the hypergeometric and binomial probabilities should be very similar, and this can be verified for fixed, n, x:

$$\lim_{\substack{M,N\to\infty \\ \frac{M}{M+N}\to p}} \frac{\binom{N}{x}\binom{M}{n-x}}{\binom{M+N}{n}} = \binom{n}{x} p^x (1-p)^{n-x}.$$

1.2 Continuous Distributions

1.2.1 Uniform Distribution

CDF, for $a < x < b$:

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{a}^{x} \frac{1}{b-a}\,dx = \frac{x-a}{b-a},$$

that is,

$$F(x) = \begin{cases} 0 & x \leq a \\ \dfrac{x-a}{b-a} & a < x < b \\ 1 & b \leq x \end{cases}$$

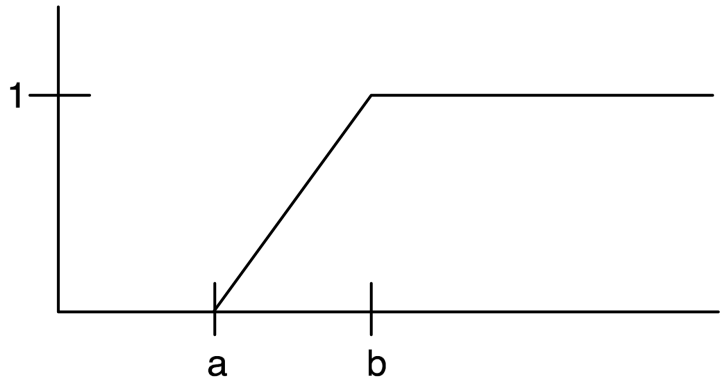


Figure 6: Uniform distribution CDF

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

A special case is the $U(0,1)$ distribution:

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$F(x) = x \qquad \text{for } 0 < x < 1,$$

$$E(X) = \frac{1}{2}, \qquad \text{Var}(X) = \frac{1}{12}, \qquad M(t) = \frac{e^t - 1}{t}.$$

1.2.2 Exponential Distribution

CDF:

$$\begin{aligned} F(x) &= 1 - e^{-\lambda x}, \\ M_X(t) &= \frac{\lambda}{\lambda - t}, \quad \lambda > 0, x \geq 0. \end{aligned}$$

This is the distribution for the waiting time until the first occurrence in a Poisson process with rate parameter $\lambda > 0$.

1. If $X \sim \text{Exp}(\lambda)$ then,

$$P(X \geq t + x | X \geq t) = P(X \geq x)$$
(memoryless property)

2. It can be obtained as limiting form of geometric distribution.

1.2.3 Gamma distribution

$$f(x) = \frac{\lambda^{alpha}}{\Gamma(\alpha} x^{\alpha-1} e^{-\lambda x}, \quad \alpha > 0, \lambda > 0, x \geq 0$$

with

$$\text{gamma} \quad \text{function} : \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$$

$$\text{mgf} : M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha, t < \lambda.$$

Suppose $Y_1, \ldots, Y_K$ are independent $\text{Exp}(\lambda)$ random variables and let $X = Y_1 + \cdots + Y_K$. Then $X \sim \text{Gamma}(K, \lambda)$, for K integer. In general, $X \sim \text{Gamma}(\alpha, \lambda)$, $\alpha > 0$.

1. α is the shape parameter,
   λ is the scale parameter

   Note: if $Y \sim$ Gamma $(\alpha, 1)$ and $X = \frac{Y}{\lambda}$, then $X \sim$ Gamma (α, λ). That is, λ is scale parameter.

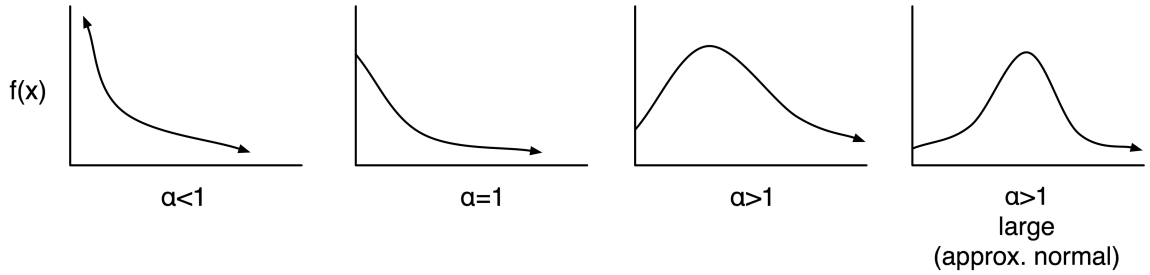


Figure 7: Gamma Distribution

2. Gamma $\left(\frac{p}{2}, \frac{1}{2}\right)$ distribution is also called χ^2_p (chi-square with p df) distribution if p is integer;
   $\chi^2_2 =$ exponential distribution (for 2 df).
3. Gamma (K, λ) distribution can be interpreted as the waiting time until the K^{th} occurrence in a Poisson process.

1.2.4 Beta density function

Suppose $Y_1 \sim$ Gamma $(\alpha, \lambda), Y_2 \sim$ Gamma (β, λ) independently, then,

$$X = \frac{Y_1}{Y_1 + Y_2} \sim \mathcal{B}(\alpha, \beta), \quad 0 \leq x \leq 1.$$

Remark: see soon for derivation!

1.2.5 Normal distribution

$X \sim N(\mu, \sigma^2)$; $M_X(t) = e^{t\mu} e^{t^2\sigma^2/2}$.

1.2.6 Standard Cauchy distribution

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

PDF: $f(x) = \frac{1}{\pi}\left(\frac{1}{1+x^2}\right)$; (location parameter $\theta = 0$)

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

$E(X), \mathrm{Var}(X), M_X(t)$ do **not** exist.

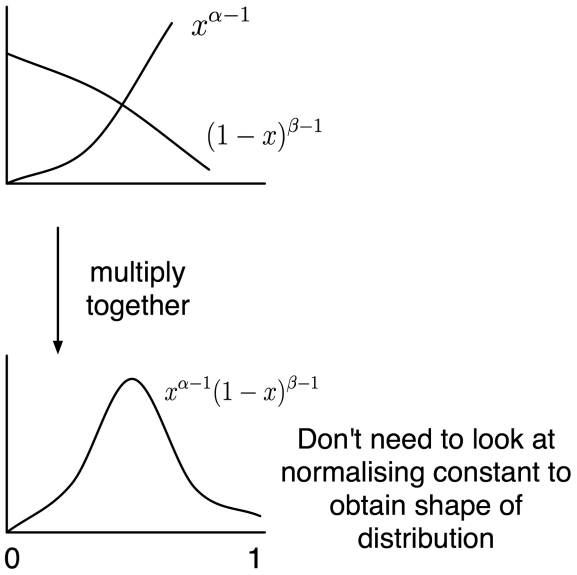


Figure 8: Beta Distribution

$\rightarrow$ the Cauchy is a bell-shaped distribution symmetric about zero for which **no** moments are defined.

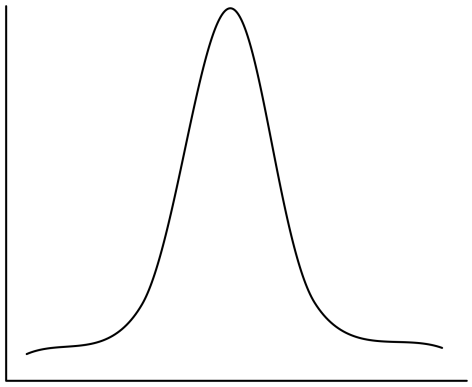


Figure 9: Cauchy Distribution

(Pointier than normal distribution and tails go to zero much slower than normal distribution.)

If $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$ independently, then $X = \dfrac{Z_1}{Z_2} \sim$ Cauchy distribution.

1.3 Transformations of a single random variable

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

Theorem. *1.3.1 (Discrete case)*

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

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

(sum over all values x for which $h(x) = y$)

Proof.

$$p_Y(y) = P(Y = y) = P\{h(X) = y\}$$

$$= \sum_{x:h(x)=y} P(X = x)$$

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

□

Theorem. *1.3.2*

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:

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

Proof. Assume h is increasing. Then

$$F_Y(y) = P(Y \leq y) = P\{h(X) \leq y\}$$

$$= P\{X \leq h^{-1}(y)\}$$

$$= F_X\{h^{-1}(y)\}$$

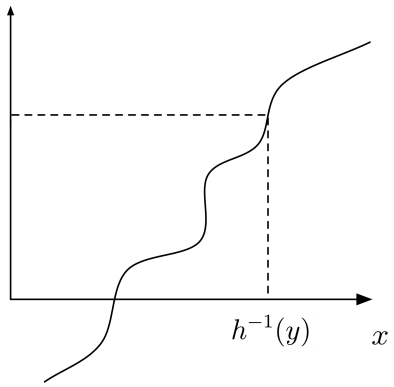


Figure 10: h increasing.

$$\Rightarrow f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X\{h^{-1}(y)\} \quad \text{(use Chain Rule)}$$

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

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

$$F_Y(y) = P(Y \leq y) = P\{h(X) \leq y\}$$

$$= P\{X \geq h^{-1}(y)\}$$

$$= 1 - F_X\{h^{-1}(y)\}$$

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

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:

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

□

Examples:

1. **Discrete transformation of single RV**

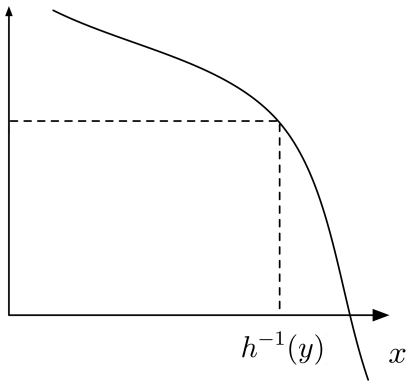


Figure 11: h decreasing.

$X \sim Po(\lambda)$ and Y is X rounded to the nearest multiple of 10.

Possible values of Y are: $0, 10, 20, \ldots$

$$P(Y = 0) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

$$= e^{-\lambda} + e^{-\lambda}\lambda + \frac{e^{-\lambda}\lambda^2}{2!} + \frac{e^{-\lambda}\lambda^3}{3!} + \frac{e^{-\lambda}\lambda^4}{4!};$$

$$P(Y = 10) = P(X = 5) + P(X = 6) + P(X = 7) + P(X = 8) + \ldots + P(X = 14),$$

and so on.

2. **Continuous transformation of single RV**

Let $Y = h(X) = aX + b, \quad a \neq 0.$

To find $h^{-1}(y)$ we solve for x in the equation $y = h(x)$, i.e.,

$$y = ax + b$$

$$\Rightarrow x = \frac{y - b}{a} \Rightarrow h^{-1}(y) = \frac{y - b}{a} = \frac{y}{a} - \frac{b}{a}$$

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

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

Specifically:

1. Suppose $Z \sim N(0,1)$ and let $X = \mu + \sigma Z,\ \sigma > 0$. Recall that Z has PDF:

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

Hence, $f_X(x) = \frac{1}{|\sigma|}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}(\sigma)}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, which is N$(\mu, \sigma^2)$ PDF.

2. Suppose $X \sim N(\mu, \sigma^2)$ and let $Z = \dfrac{X-\mu}{\sigma}$.

Find the PDF of Z.

Solution. Observe that $Z = h(X) = \dfrac{1}{\sigma}X - \dfrac{\mu}{\sigma}$,

$$\Rightarrow f_Z(z) = \sigma f_X\left(\frac{z + \frac{\mu}{\sigma}}{\frac{1}{\sigma}}\right) = \sigma f_X(\mu + \sigma z)$$

$$= \sigma \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{(-1)}{(2\sigma^2)}(\mu+\sigma z-\mu)^2}$$

$$= \frac{1}{\sqrt{2\pi}}\, e^{\frac{(-1)}{(2\sigma^2)}\sigma^2 z^2}$$

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

i.e., $Z \sim$ N$(0,1)$.

3. Suppose $X \sim$ Gamma $(\alpha, 1)$ and let $Y = \dfrac{X}{\lambda}, \lambda > 0$.

Find the PDF of Y.

Solution. Since $Y = \dfrac{1}{\lambda}\, X$, is a linear function, we have

$$f_Y(y) = \lambda f_X(\lambda y)$$

$$= \lambda \frac{1}{\Gamma(\alpha)}(\lambda y)^{\alpha-1}\ e^{-\lambda y} = \frac{\lambda^\alpha}{\Gamma(\alpha)}\ y^{\alpha-1}\ e^{-\lambda y},$$

which is the Gamma(α, λ) PDF.

1.4 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 $U \sim U(0,1)$.

Proof.

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

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

The 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.,

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

1.5 Non-monotonic transformations

Theorem 1.3.2 applies to monotonic (either strictly increasing or decreasing) transformations of a continuous RV.

In general, if $h(x)$ is <u>not</u> monotonic, then $h(x)$ may not even be continuous.

For example,

$$\text{if } h(x) = [x], \quad \nearrow \text{integer part of } x$$

then possible values for $Y = h(X)$ are the integers

$$\Rightarrow Y \text{ is discrete.}$$

However, it can be shown that $h(X)$ is continuous if X is continuous and $h(x)$ is piecewise monotonic.

1.6 Moments of transformed RVS

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

If we want to find $\boxed{E(Y)}$, we can proceed as follows:

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

(that is, forget X ever existed!)

Or use

Theorem. *1.6.1*

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:

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

Proof.

Not examinable.

□

Examples:

1. **CDF transformation**

   Suppose $U \sim U(0,1)$. How can we transform U to get an Exp(λ) RV?

Solution. Take $X = F^{-1}(U)$, where F is the Exp(λ) CDF. Recall $F(x) = 1 - e^{-\lambda x}$ for the Exp(λ) distribution.

To find $F^{-1}(u)$, we solve for x in $F(x) = u$,

i.e.,

$$u = 1 - e^{-\lambda x}$$

$$\Rightarrow 1 - u = e^{-\lambda x}$$

$$\Rightarrow \ln(1-u) = -\lambda x$$

$$\Rightarrow x = \frac{-\ln(1-u)}{\lambda}.$$

Hence if $U \sim U(0,1)$, it follows that $X = \dfrac{-\ln(1-U)}{\lambda} \sim \text{Exp}(\lambda)$.

Note: $Y = \dfrac{-\ln U}{\lambda} \sim \text{Exp}(\lambda)$ [both $(1-U)$ & U have $U(0,1)$ distribution]. m This type of result is used to generate random numbers. That is, there are good methods for producing $U(0,1)$ (pseudo-random) numbers. To obtain Exp(λ) random numbers, we can just get $U(0,1)$ numbers and then calculate $X = \dfrac{-\log U}{\lambda}$.

2. **Non-monotonic transformations**

Suppose $Z \sim N(0,1)$ and let $X = Z^2$; $h(Z) = Z^2$ is not monotonic, so Theorem 1.3.2 does not apply. However we can proceed as follows:

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

where

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

is the $N(0,1)$ CDF. That is,

$$\Rightarrow f_X(x) = \frac{d}{dx} F_X(x) = \phi\left(\sqrt{x}\right)\left(\frac{1}{2}x^{-1/2}\right) - \phi(-\sqrt{x})\left(-\frac{1}{2}x^{-1/2}\right),$$

where $\phi(a) = \Phi'(a) = \frac{1}{\sqrt{2\pi}} e^{-a^2/2}$ is the $N(0,1)$ PDF

$$= \frac{1}{2} x^{-1/2} \left[\frac{1}{\sqrt{2\pi}} e^{-x/2} + \frac{1}{\sqrt{2\pi}} e^{-x/2} \right]$$

$$= \frac{(1/2)^{1/2}}{\sqrt{\pi}} x^{1/2-1} e^{-1/2x,} \quad \text{which is the Gamma}\left(\frac{1}{2}, \frac{1}{2}\right) \quad \text{PDF.}$$

On the other hand, the distribution of Z^2 is also called the χ^2_1, distribution. We have proved that it is the same as the Gamma$\left(\frac{1}{2}, \frac{1}{2}\right)$ distribution.

3. **Moments of transformed RVS**:

   Example: if $U \sim U(0,1)$ and $Y = \dfrac{-\log U}{\lambda}$ then $Y \sim \text{Exp}(\lambda) \quad \Rightarrow \quad f(y) = \lambda e^{-\lambda y}, \quad y > 0.$

   Can check

$$E(Y) = \int_0^\infty \lambda y e^{-\lambda y} \, dy$$

$$= \frac{1}{\lambda}.$$

4. **Based on Theorem 1.6.1**:

   If $U \sim U(0,1)$ and $Y = -\dfrac{\log U}{\lambda}$, then according to Theorem 1.6.1,

$$E(Y) = \int_0^1 \frac{-\log u}{\lambda}(1) \, du$$

$$= -\frac{1}{\lambda} \int_0^1 \log u \, du = -\frac{1}{\lambda}(u \log u - u)\Big|_0^1$$

$$= -\frac{1}{\lambda} \left[u \log u \,\big|_0^1 - u\big|_0^1 \right]$$

$$= -\frac{1}{\lambda}[0 - 1]$$

$$= \frac{1}{\lambda}, \qquad \text{as required.}$$

There are some important consequences of Theorem 1.6.1:

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 $\text{Var}(Y) = a^2\sigma^2$.

   Proof. (Continuous case)

$$E(Y) = E(aX + b) = \int_{-\infty}^{\infty} (ax + b) f(x)\, dx$$

$$= \underbrace{a \int_{-\infty}^{\infty} x f(x)\, dx}_{E(X)} + \underbrace{b \int_{-\infty}^{\infty} f(x)\, dx}_{=1}$$

$$= a\mu + b.$$

$$\text{Var}(Y) = E\Big[\big\{Y - E(Y)\big\}^2\Big]$$

$$= E\Big[\big\{aX + b - (a\mu + b)\big\}^2\Big]$$

$$= E\big\{a^2(X - \mu)^2\big\}$$

$$= a^2 E\left\{(X - \mu)^2\right\}$$

$$= a^2 \,\text{Var}(X)$$

$$= a^2\sigma^2.$$

□

2. If X is a RV and $h(X)$ is any function, then the MGF of $Y = h(X)$, provided it exists is,

$$M_Y(t) = \begin{cases} \displaystyle\sum_x e^{th(x)} p(x) & X \quad \text{discrete} \\ \displaystyle\int_{-\infty}^{\infty} e^{th(x)} f(x)\, dx & X \quad \text{continuous} \end{cases}$$

This gives us another way to find the distribution of $Y = h(X)$.

i.e., Find $M_Y(t)$. If we recognise $M_Y(t)$, then by uniqueness we can conclude that Y has that distribution.

Examples

1. 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

$$M_U(t) = \int_a^b e^{tF(x)} f(x)\, dx$$

$$= \frac{1}{t} e^{tF(x)} \Big|_a^b = \frac{e^{tF(b)} - e^{tF(a)}}{t}$$

$$= \frac{e^t - 1}{t},$$

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

2. Suppose $X \sim U(0, 1)$, and let $Y = \dfrac{-\log X}{\lambda}$.

$$M_Y(t) = \int_0^1 e^{t(\frac{-\log x}{\lambda})}(1)\, dx$$

$$= \int_0^1 x^{-t/\lambda}\, dx$$

$$= \frac{1}{1 - t/\lambda} x^{1-t/\lambda} \Big|_0^1$$

$$= \frac{1}{1 - t/\lambda}$$

$$= \frac{\lambda}{\lambda - t},$$

   which is the MGF for $\text{Exp}(\lambda)$ distribution. Hence we can conclude that

$$Y = \frac{-\log X}{\lambda} \sim \text{ Exp}(\lambda).$$

#

1.7 Multivariate distributions

Definition. 1.7.1

If $X_1, X_2, \ldots, X_r$ are discrete RV's then, $\mathbf{X} = (X_1, X_2, \ldots, X_r)^T$ is called a discrete random vector.

The probability function $P(\mathbf{x})$ is:

$$P(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = P\left(\{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \cdots \cap \{X_r = x_r\}\right);$$

$$P(\mathbf{x}) = P(x_1, x_2, \ldots, x_r)$$

1.7.1 Trinomial distribution

$r = 2$

Consider a sequence of n independent trials where each trial produces:

Outcome 1:	with prob π_1
Outcome 2:	with prob π_2
Outcome 3:	with prob $1 - \pi_1 - \pi_2$

If X_1, X_2 are number of occurrences of outcomes 1 and 2 respectively then(X_1, X_2) have trinomial distribution.

Parameters: $\pi_1 > 0, \pi_2 > 0$ and $\pi_1 + \pi_2 < 1$; $n > 0$ fixed

Possible Values: integers (x_1, x_2) s.t. $x_1 \geq 0,\ x_2 \geq 0,\ x_1 + x_2 \leq n$

Probability function:

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

for

$$x_1, x_2 \geq 0, \quad x_1 + x_2 \leq n.$$

1.7.2 Multinomial distribution

Parameters: $n > 0,\ \boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_r)^T$ with $\pi_i > 0$ and $\sum_{i=1}^{r} \pi_i = 1$

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

Probability function:

$$P(\mathbf{x}) = \binom{n}{x_1, x_2, \ldots, x_r} \pi_1^{x_1} \pi_2^{x_2} \ldots \pi_r^{x_r} \quad \text{for}$$

$$x_i \geq 0,\ \sum_{i=1}^{r} x_i = n.$$

Remarks

1. Note $\binom{n}{x_1, x_2, \ldots, x_r} \overset{\text{def}}{=} \frac{n!}{x_1! x_2! \ldots x_r!}$ is the *multinomial coefficient*.
2. Multinomial distribution is the generalisation of the binomial distribution to r types of outcome.
3. Formulation differs from binomial and trinomial cases in that the redundant count $x_r = n - (x_1 + x_2 + \ldots x_{r-1})$ is included as an argument of $P(\mathbf{x})$.

1.7.3 Marginal and conditional distributions

Consider a discrete random vector

$$\begin{aligned} \mathbf{X} &= (X_1, X_2, \ldots, X_r)^T \quad \text{and let} \\ \mathbf{X}_1 &= (X_1, X_2, \ldots, X_{r_1})^T \& \\ \mathbf{X}_2 &= (X_{r_1+1}, X_{r_1+2}, \ldots, X_r)^T, \end{aligned}$$

so that $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$.

Definition. 1.7.2

If $\mathbf{X}$ has joint probability function $P_X(\mathbf{x}) = P_X(\mathbf{x}_1, \mathbf{x}_2)$ then the marginal probability function for $\mathbf{X}_1$ is :

$$P_{\mathbf{X}_1}(\mathbf{x}_1) = \sum_{\mathbf{x}_2} P_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2).$$

Observe that:

$$P_{X_1}(\mathbf{x}_1) = \sum_{\mathbf{x}_2} P_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{\mathbf{x}_2} P(\{\mathbf{X}_1 = \mathbf{x}_1\} \cap \{\mathbf{X}_2 = \mathbf{x}_2\})$$

$$= P(\mathbf{X}_1 = \mathbf{x}_1) \quad \text{by law of total probability.}$$

Hence the marginal probability function for $\mathbf{X}_1$ is just the probability function we would have if $\mathbf{X}_2$ was not observed.

The marginal probability function for $\mathbf{X}_2$ is:

$$P_{X_2}(\mathbf{x}_2) = \sum_{\mathbf{x}_1} P_X(\mathbf{x}_1, \mathbf{x}_2).$$

Definition. 1.7.3

Suppose $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$. If $P_{X_1}(\mathbf{x}_1) > 0$, we define the conditional probability function $\mathbf{X}_2|\mathbf{X}_1$ by:

$$P_{X_2|X_1}(\mathbf{x}_2|\mathbf{x}_1) = \frac{P_X(\mathbf{x}_1, \mathbf{x}_2)}{P_{X_1}(\mathbf{x}_1)}.$$

Remarks

1.

$$\begin{aligned} P_{X_2|X_1}(\mathbf{x}_2|\mathbf{x}_1) &= \frac{P(\{\mathbf{X}_1 = \mathbf{x}_1\} \cap \{\mathbf{X}_2 = \mathbf{x}_2\})}{P(\mathbf{X}_1 = \mathbf{x}_1)} \\ &= P(\mathbf{X}_2 = \mathbf{x}_2|\mathbf{X}_1 = \mathbf{x}_1). \end{aligned}$$

2. Easy to check $P_{X_2|X_1}(\mathbf{x}_2|\mathbf{x}_1)$ is a proper probability function with respect to $\mathbf{x}_2$ for each fixed $\mathbf{x}_1$ such that $P_{X_1}(\mathbf{x}_1) > 0$.

3. $P_{X_1|X_2}(\mathbf{x}_1|\mathbf{x}_2)$ is defined by:

$$P_{X_1|X_2}(\mathbf{x}_1|\mathbf{x}_2) = \frac{P_X(\mathbf{x}_1, \mathbf{x}_2)}{P_{X_2}(x_2)}.$$

Definition. 1.7.4 (Independence)

Discrete RV's $X_1, X_2, \ldots, X_r$ are said to be independent if their joint probability function satisfies

$$P(x_1, x_2, \ldots, x_r) = p_1(x_1)p_2(x_2)\ldots p_r(x_r)$$

for some functions $p_1, p_2, \ldots, p_r$ and all $(x_1, x_2, \ldots, x_r)$.

Remarks

1. Observe that:

$$\begin{aligned} P_{X_1}(x_1) &= \sum_{x_2}\sum_{x_3}\cdots\sum_{x_r} P(x_1, \ldots, x_r) \\ &= p_1(x_1)\sum_{x_2}\sum_{x_3}\cdots\sum_{x_r} p_2(x_2)\ldots p_r(x_r) \\ &= c_1 p_1(x_1). \end{aligned}$$

Hence $\underset{\substack{\downarrow \\ \text{sum of probs.}}}{p_1(x_1)} \propto \underbrace{P_{X_1}(x_1)}_{\text{marginal prob.}}$ also $p_i(x_i) \propto P_{X_i}(x_i)$

$$\begin{aligned} P_{X_1|X_2\ldots X_r}(x_1|x_2, \ldots, x_r) &= \frac{P(x_1, \ldots, x_r)}{P_{x_2\ldots x_r}(x_1, \ldots, x_r)} \\ &= \frac{p_1(x_1)p_2(x_2)\ldots p_r(x_r)}{\sum_{x_1} p_1(x_1)p_2(x_2)\ldots p_r(x_r)} \\ &= \frac{p_1(x_1)p_2(x_2)\ldots p_r(x_r)}{p_2(x_2)p_3(x_3)\ldots p_r(x_r)\sum_{x_1} p_1(x_1)} \\ &= \frac{p_1(x_1)}{\sum_{x_1} p_1(x_1)} \\ &= \frac{\frac{1}{c}P_{X_1}(x_1)}{\frac{1}{c}\sum_{x_1} P_{X_1}(x_1)} \\ &= P_{X_1}(x_1). \end{aligned}$$

That is, $P_{X_1|X_2\ldots X_r}(x_1|x_2, \ldots, x_r) = P_{X_1}(x_1)$.

2. Clearly independence $\implies$

$$P_{X_i|X_1\ldots X_{i-1},X_{i+1}\ldots X_r}(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_r) = P_{X_i}(x_i).$$

Moreover, we have:

$P_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2) = P_{X_1}(\mathbf{x}_1)$ for any partitioning of $\mathbf{X} = \begin{pmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{pmatrix}$ if $X_1,\ldots,X_r$ are independent.

1.7.4 Continuous multivariate distributions

Definition. 1.7.5

The random vector $(X_1,\ldots,X_r)^T$ is said to have a continuous multivariate distribution with PDF $f(\mathbf{x})$ if

$$P(\mathbf{X}\in A) = \int\ldots\int_A f(x_1,\ldots,x_r)\,dx_1\ldots dx_r$$

for any measurable set A.

Examples

1. Suppose (X_1,X_2) have the trinomial distribution with parameters $n,\ \pi_1,\ \pi_2$. Then the marginal distribution of X_1 can be seen to be $B(n,\pi_1)$ and the conditional distribution of $X_2|X_1 = x_1$ is $B\big(n-x_1, \frac{\pi_2}{1-\pi_1}\big)$.

   Outcome 1 π_1
   Outcome 2 π_2
   Outcome 3 $1-\pi_1-\pi_2$

Examples of Definition 1.7.5

1. If X_1, X_2 have PDF

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

   The distribution is called uniform on $(0,1)\times(0,1)$.

   It follows that $P(X\in A) =$ Area (A) for any $A\in(0,1)\times(0,1)$:

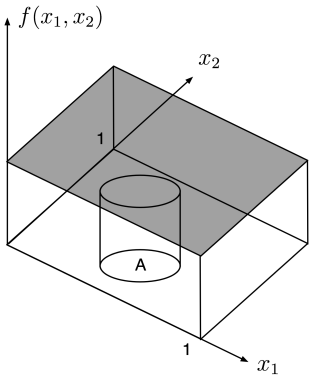


Figure 12: Area A

2. Uniform distribution on unit disk:

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

3. Dirichlet distribution is defined by PDF

$$f(x_1, x_2, \ldots, x_r) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_r + \alpha_{r+1})}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_r)\Gamma(\alpha_{r+1})} \times$$

$$x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \ldots x_r^{\alpha_r - 1} (1 - x_1 - \cdots - x_r)^{\alpha_{r+1}}$$

for $x_1, x_2, \ldots, x_r > 0$, $\sum x_i < 1$, and parameters $\alpha_1, \alpha_2, \ldots, \alpha_{r+1} > 0$.

#

Recall joint PDF:

$$P(\mathbf{X} \in A) = \int \ldots \int_A f(x_1, \ldots, x_r)\, dx_1 \ldots dx_r.$$

Note: the joint PDF must satisfy:

1. $f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$;
2. $\displaystyle\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_r)\, dx_1 \ldots dx_r = 1.$

Definition. 1.7.6

If $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ has joint PDF $f(\mathbf{x}) = f(\mathbf{x}_1, \mathbf{x}_2)$, then the marginal PDF of $\mathbf{X}_1$ is given by:

$$f_{X_1}(\mathbf{x}_1) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_{r_1}, x_{r_1+1}, \dots x_r)\ dx_{r_1+1} dx_{r_1+2} \dots dx_r,$$

and for $f_{X_1}(\mathbf{x}_1) > 0$, the conditional PDF is given by

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

Remarks:

1. $f_{X_2|X_1}(\mathbf{x}_2|\mathbf{x}_1)$ <u>cannot</u> be interpreted as the conditional PDF of $\mathbf{X}_2|\{\mathbf{X}_1 = \mathbf{x}_1\}$ because $P(\mathbf{X}_1 = \mathbf{x}_1) = 0$ for any continuous distribution.

   Proper interpretation is the limit as $\delta \to 0$ in $\mathbf{X}_2|\mathbf{X}_1 \in \mathrm{B}(\mathbf{x}_1, \delta)$.

2. $f_{X_2}(\mathbf{x}_2),\ f_{X_1|X_2}(\mathbf{x}_1|\mathbf{x}_2)$ are defined analogously.

Definition. 1.7.7 (Independence)

Continuous RV's $X_1, X_2, \dots, X_r$ are said to be (mutually) independent if their joint PDF satisfies:

$$f(x_1, x_2, \dots, x_r) = f_1(x_1) f_2(x_2) \dots f_r(x_r),$$

for some functions $f_1, f_2, \dots, f_r$ and all $x_1, \dots, x_r$

Remarks

1. Easy to check that if $X_1, \dots, X_r$ are independent then each $f_i(x_i) = c_i f_{X_i}(x_i)$.

   Moreover $c_1, c_2, \dots, c_r = 1$.

2. If $X_1, X_2, \dots, X_r$ are independent, then it can be checked that:

   $$f_{\mathbf{x}_1|\mathbf{x}_2}(\mathbf{x}_1|\mathbf{x}_2) = f_{X_1}(\mathbf{x}_1)$$

   for any partitioning of $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$.

Examples

1. If (X_1, X_2) has the uniform distribution on the unit disk, find:

   (a) the marginal PDF $f_{X_1}(x_1)$

   (b) the conditional PDF $f_{X_2|X_1}(x_2|x_1)$.

   Solution. Recall

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

$$\begin{aligned} \text{(a)} \quad f_{X_1}(x_1) &= \int_{-\infty}^{\infty} f(x_1, x_2)\, dx_2 \\ &= \int_{-\infty}^{-\sqrt{1-x_1^2}} 0\, dx_2 + \int_{-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}} \frac{1}{\pi}\, dx_2 + \int_{\sqrt{1-x_1^2}}^{\infty} 0\, dx_2 \\ &= 0 + \frac{x_2}{\pi}\Big|_{-\sqrt{1-x_1^2}}^{\sqrt{1-x_1^2}} + 0 \\ &= \begin{cases} \dfrac{2\sqrt{1-x_1^2}}{\pi} & -1 < x_1 < 1 \\ \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

   i.e., a semi-circular distribution (ours has some distortion).

   (b) The conditional density for $X_2|X_1$ is:

$$\begin{aligned} f_{X_2|X_1}(x_2|x_1) &= \frac{f(x_1, x_2)}{f_{X_1}(x_1)} \\ &= \begin{cases} \dfrac{1}{2\sqrt{1-x_1^2}} & \text{for } -\sqrt{1-x_1^2} < x_2 < \sqrt{1-x_1^2} \\ \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

   which is uniform $U(-\sqrt{1-x_1^2}, \sqrt{1-x_1^2})$.

2. If $X_1, \ldots, X_r$ are any independent continuous RV's then their joint PDF is:

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

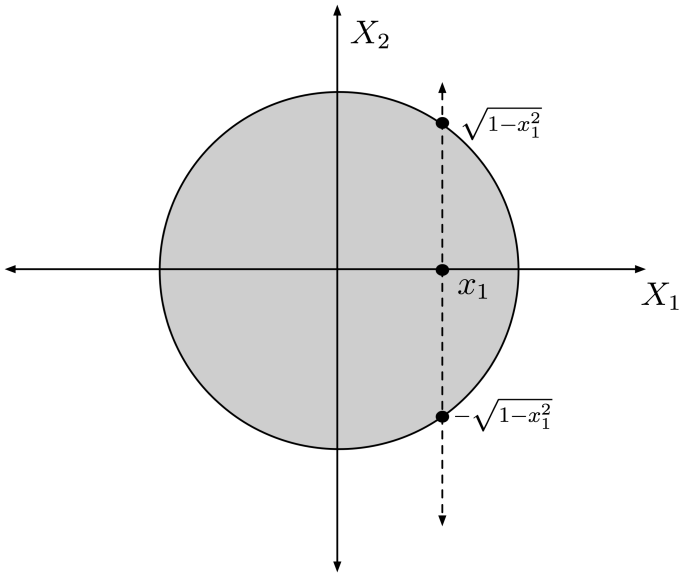


Figure 13: A graphic showing the conditional distribution and a semi-circular distribution.

E.g. If $X_1, X_2, \ldots, X_r$ are independent Exp(λ) RVs, then

$$f(x_1, \ldots, x_r) = \prod_{i=1}^{r} \lambda\, e^{-\lambda x_i}$$

$$= \lambda^r e^{-\lambda \sum_{i=1}^{r} x_i},$$

$$\text{for } x_i > 0 \quad i = 1, \ldots, n.$$

3. Suppose (X_1, X_2) are uniformly distributed on $(0,1) \times (0,1)$.
   That is,
$$f(x_1, x_2) = \begin{cases} 1 & 0 < x_1 < 1, 0 < x_2 < 1 \\ \\ 0 & \text{otherwise} \end{cases}$$
   Claim: X_1, X_2 are independent.

   Proof.
   Let
$$f_1(x_1) = \begin{cases} 1 & 0 < x_1 < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

$$f_2(x_2) = \begin{cases} 1 & 0 < x_2 < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

then $f(x_1, x_2) = f_1(x_1)f_2(x_2)$, and the two variables are independent. □

4. (X_1, X_2) is uniform on unit disk. Are X_1, X_2 independent? NO.

   Proof.
   We know:

$$f_{X_2|X_1}(x_2|x_1) = \begin{cases} \dfrac{1}{2\sqrt{1-x_1^2}} & -\sqrt{1-x_1^2} < x_2 < \sqrt{1-x_1^2} \\ \\ 0 & \text{otherwise,} \end{cases}.$$

   i.e., $U(-\sqrt{1-x_1^2}, \sqrt{1-x_1^2})$.
   On the other hand,

$$f_{X_2}(x_2) = \begin{cases} \dfrac{2}{\pi}\sqrt{1-x_2^2} & -1 < x_2 < 1 \\ \\ 0 & \text{otherwise,} \end{cases}$$

   i.e., a semicircular distribution.
   Hence $f_{X_2}(x_2) \neq f_{X_2|X_1}(x_2|x_1)$, so the variables cannot be independent. □

Definition. 1.7.8

The joint CDF of RVS's $X_1, X_2, \ldots, X_r$ is defined by:

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

Remarks:

1. Definition applies to RV's of any type, i.e., discrete, continuous, "hybrid".
2. Marginal CDF of $\mathbf{X}_1 = (X_1, \ldots, X_s)^T$ is $F_{X_1}(x_1, \ldots, x_s) = F_X(x_1, \ldots, x_s, \infty, \infty, \ldots, \infty)$, for $s < r$.
3. RV's $X_1, \ldots, X_r$ are defined to be independent if

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

4. The definitions above are completely general. However, in practice it is usually easier to work with the PDF/probability function for any given example.

1.8 Transformations of several RVs

Theorem. *1.8.1*

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)

$$P(\{Y = y\}) = \sum_x P(\{Y = y\} \cap \{X_1 = x\})$$

(law of total probability)

$$\downarrow$$

$$\left(\begin{array}{c} \text{If } A \text{ is an event } \& \ B_1, B_2 \ldots \text{ are events} \\ \\ \text{s.t. } \bigcup_i B_i = \mathcal{S} \ \& \ B_i \cap B_j = \phi \ \ (\text{for } j \neq i) \\ \\ \text{then } P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i) P(B_i) \end{array} \right)$$

$$= \sum_x P(\{X_1 + X_2 = y\} \cap \{X_1 = x\})$$

$$= \sum_x P(\{X_2 = y - x\} \cap \{X_1 = x\})$$

$$= \sum_x P(x, y - x).$$

□

Proof. (2)

Just substitute

$$P(x, y-x) = P_{X_1}(x)P_{X_2}(y-x).$$

□

Theorem. *1.8.2*

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.

$$\begin{aligned}
F_Y(y) &= P(Y \le y) \\
&= P(X_1 + X_2 \le y) \\
&= \int_{x_1=-\infty}^{\infty} \int_{x_2=-\infty}^{y-x_1} f(x_1, x_2)\,dx_2\,dx_1. \\
&\quad \text{Let } x_2 = t - x_1, \\
&\quad \Rightarrow \frac{dx_2}{dt} = 1 \\
&\quad \Rightarrow dx_2 = dt \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{y} f(x_1, t-x_1)\,dt\,dx_1 \\
&= \int_{-\infty}^{y} \left\{ \int_{-\infty}^{\infty} f(x_1, t-x_1)\,dx_1 \right\} dt \\
&\quad \Rightarrow f_Y(y) = F_Y'(y) = \int_{-\infty}^{\infty} f(x, y-x)\,dx.
\end{aligned}$$

□

Proof. (2)

Take $f(x, y-x) = f_{X_1}(x)\, f_{X_2}(y-x)$ if X_1, X_2 independent. □

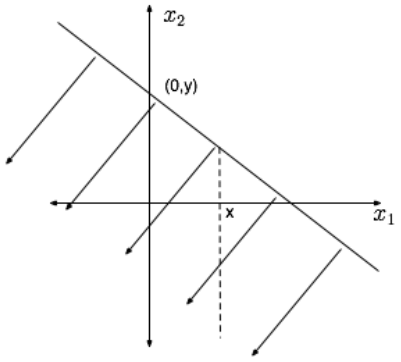


Figure 14: Theorem 1.8.2

Examples

1. Suppose $X_1 \sim B(n_1, p), X_2 \sim B(n_2, p)$ independently. Find the probability function for $Y = X_1 + X_2$.

   Solution.

$$P_Y(y) = \sum_x P_{X_1}(x) P_{X_2}(y - x)$$

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

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

$$= \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y} \underbrace{\sum_{x=\max\,(0,\,y-n_2)}^{\min\,(n_1,\,y)} \frac{\binom{n_1}{x}\binom{n_2}{y-x}}{\binom{n_1+n_2}{y}}}_{\substack{\text{sum of hypergeometric} \\ \text{probability function} = 1}}$$

$$= \binom{n_1+n_2}{y} p^y (1-p)^{n_1+n_2-y},$$

   i.e., $Y \sim B(n_1 + n_2, p)$.

2. Suppose $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$ independently. Let $X = Z_1 + Z_2$. Find PDF of X.

   Solution.

$$f_X(x) = \int_{-\infty}^{\infty} \phi(z)\phi(x-z)\,dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \frac{1}{\sqrt{2\pi}} e^{-(x-z)^2/2} dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{1}{2}(z^2+(x-z)^2)}\,dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{1}{2}\left(2(z-\frac{x}{2})^2\right)} e^{-x^2/4}\,dz$$

$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/4} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(z-\frac{x}{2})^2}{2\times\frac{1}{2}}}\,dz}_{\substack{=1 \\ \mathrm{N}\left(\frac{x}{2},\frac{1}{2}\right)\ \text{PDF}}}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-x^2/4},$$

   i.e., $X \sim N(0,2)$.

Theorem. *1.8.3 (Ratio of continuous RVs)*

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

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.

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

Substitute $x_2 = tx_1$ in both inner integrals;

$$\begin{aligned}
dx_2 &= x_1 dt; \\
&= \int_{-\infty}^{0} \int_{y}^{-\infty} x_1 f(x_1, tx_1) dt\, dx_1 + \int_{0}^{\infty} \int_{-\infty}^{y} x_1 f(x_1, tx_1)\, dt\, dx_1 \\
&= \int_{-\infty}^{0} \int_{-\infty}^{y} (-x_1) f(x_1, tx_1)\, dt\, dx_1 + \int_{0}^{\infty} \int_{-\infty}^{y} x_1 f(x_1, tx_1)\, dt\, dx_1 \\
&= \int_{-\infty}^{0} \int_{-\infty}^{y} |x_1| f(x_1, tx_1)\, dt\, dx_1 + \int_{0}^{\infty} \int_{-\infty}^{y} |x_1| f(x_1, tx_1)\, dt\, dx_1 \\
&= F_Y(y) \\
&\Rightarrow f_Y(y) = F_Y'(y) = \int_{-\infty}^{\infty} |x_1| f(x_1, yx_1)\, dx_1.
\end{aligned}$$

□

Example

1. If $Z \sim \mathrm{N}(0,1)$ and $X \sim \chi^2_k$ independently, then $T = \dfrac{Z}{\sqrt{X/k}}$ is said to have the t-distribution with k degrees of freedom. Derive the PDF of T.

Solution. Step 1:

Let $V = \sqrt{X/k}$. Need to find PDF of V. Recall χ^2_k is Gamma$\left(\frac{k}{2}, \frac{1}{2}\right)$ so

$$f_X(x) = \frac{(1/2)^{k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \qquad \text{for } x > 0.$$

Now,

$$v = h(x) = \sqrt{x/k}$$

$$\Rightarrow h^{-1}(v) = kv^2 \qquad \text{and}$$

$$h^{-1}(v)' = 2kv$$

$$\Rightarrow f_V(v) = f_X(h^{-1}(v))|h^{-1}(v)'|$$

$$= \frac{(1/2)^{k/2}}{\Gamma(k/2)} (kv^2)^{k/2-1} e^{-kv^2/2} (2kv)$$

$$= \frac{2(k/2)^{k/2}}{\Gamma(k/2)} v^{k-1} e^{-kv^2/2}, \qquad v > 0.$$

Step 2:

Apply Theorem 1.8.3 to find PDF of

$$T = \frac{Z}{V} = \int_{-\infty}^{\infty} |v| f_V(v) f_Z(vt)\, dv$$

$$= \int_0^{\infty} v \frac{2(k/2)^{k/2}}{\Gamma(k/2)} v^{k-1} e^{-kv^2/2} \frac{1}{\sqrt{2\pi}} e^{-t^2v^2/2}\, dv$$

$$= \frac{(k/2)^{k/2}}{\Gamma(k/2)\sqrt{2\pi}} \int_0^{\infty} v^{k-1} e^{-1/2(k+t^2)v^2} 2v\, dv;$$

substitute $u = v^2$

$$\Rightarrow du = 2v\,dv$$

$$= \frac{(k/2)^{k/2}}{\Gamma(k/2)\sqrt{2\pi}} \int_0^\infty u^{\frac{k-1}{2}} e^{-1/2(k+t^2)u}\,du$$

$$= \frac{(k/2)^{k/2}}{\Gamma(k/2)\sqrt{2\pi}} \left[\frac{\Gamma(\alpha)}{\lambda^\alpha}\right] = \frac{(k/2)^{k/2}}{\Gamma(k/2)\sqrt{2\pi}} \frac{\Gamma(\frac{k+1}{2})}{(1/2(k+t^2))^{(k+1)/2}}$$

$$= \frac{\Gamma(\frac{k+1}{2})}{\Gamma(k/2)\sqrt{2\pi}} \left(\frac{2}{k} \times \frac{1}{2}(k+t^2)\right)^{-\frac{k+1}{2}} \left(\frac{k}{2}\right)^{-1/2}$$

$$= \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})\sqrt{k\pi}} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \qquad -\infty < t < \infty$$

Remarks:

1. If we take $k = 1$, we obtain
$$f(t) \propto \frac{1}{1+t^2};$$
that is, t_1 = Cauchy distribution.
Can also check $\dfrac{\Gamma(1)}{\Gamma(\frac{1}{2})\sqrt{\pi}} = \dfrac{1}{\pi}$ as required, so that $f(t) = \dfrac{1}{\pi}\dfrac{1}{1+t^2}$.

2. As $k \to \infty, t_k \to \mathrm{N}(0,1)$. To see this directly consider limit of t-density as $k \to \infty$:

$$\text{i.e. } \lim_{k\to\infty} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}} \qquad \text{for } t \text{ fixed}$$

$$= \lim_{k\to\infty} \left(1 + \frac{t^2}{k}\right)^{\frac{-k}{2}}; \quad \text{let } \ell = \frac{k}{2}$$

$$= \lim_{\ell\to\infty} \left(1 + \frac{\frac{t^2}{2}}{\ell}\right)^{-\ell} = \frac{1}{e^{\frac{t^2}{2}}} = e^{\frac{-t^2}{2}}.$$

Recall standard limit:

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

$$\Rightarrow f_T(t) \to \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} \text{ as } k \to \infty.$$

1.8.1 Multivariate transformation rule

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

$$H = \begin{bmatrix} \dfrac{\partial h_1}{\partial x_1} & \dfrac{\partial h_1}{\partial x_2} & \cdots & \dfrac{\partial h_1}{\partial x_r} \\ \dfrac{\partial h_2}{\partial x_1} & \dfrac{\partial h_2}{\partial x_2} & \cdots & \dfrac{\partial h_2}{\partial x_r} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial h_r}{\partial x_1} & \dfrac{\partial h_r}{\partial x_2} & \cdots & \dfrac{\partial h_r}{\partial x_r} \end{bmatrix}$$

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

$$g : \mathbb{R}^r \to \mathbb{R}^r \quad \text{with property that:}$$
$$g(h(\mathbf{x})) = \mathbf{x}.$$

It can be proved that the matrix of partial derivatives

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

Theorem. *1.8.4*

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

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

Remark:

Can sometimes use H^{-1} instead of G, but need to be careful to evaluate $H^{-1}(\mathbf{x})$ at $\mathbf{x} = h^{-1}(\mathbf{y}) = g(\mathbf{y})$.

1.8.2 Method of regular transformations

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

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

One approach is as follows:

1. Choose a function, $d : \mathbb{R}^r \to \mathbb{R}^{r-s}$ and let $\mathbf{Z} = d(\mathbf{X}) = (Z_1, Z_2, \ldots, Z_{r-s})^T$.
2. Apply theorem 1.8.4 to $(Y_1, Y_2, \ldots, Y_s, Z_1, Z_2, \ldots, Z_{r-s})^T$.
3. Integrate over $Z_1, Z_2, \ldots, Z_{r-s}$ to get the marginal PDF for $\mathbf{Y}$.

Examples

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

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

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

The inverse mapping, g, can be seen to be:

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

To find distribution of (R, θ):

Step 1

$$G(r, \theta) = \begin{bmatrix} \dfrac{\partial}{\partial r} g_1(r, \theta) & \dfrac{\partial}{\partial \theta} g_1(r, \theta) \\ \dfrac{\partial}{\partial r} g_2(r, \theta) & \dfrac{\partial}{\partial \theta} g_2(r, \theta) \end{bmatrix}$$

$$\frac{\partial}{\partial r}g_1(r,\theta) = \frac{\partial}{\partial r}r\cos\theta = \cos\theta$$

$$\frac{\partial}{\partial \theta}(r\cos\theta) = -r\sin\theta$$

$$\frac{\partial}{\partial r}g_2(r,\theta) = \frac{\partial}{\partial r}r\sin\theta = \sin\theta$$

$$\frac{\partial}{\partial \theta}r\sin\theta = r\cos\theta$$

$$\Rightarrow G = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \\ \sin\theta & r\cos\theta \end{bmatrix}$$

$$\Rightarrow \det(G) = r\cos^2\theta - (-r\sin^2\theta)$$

$$= r\cos^2\theta + r\sin^2\theta = r.$$

Step 2

Now apply theorem 1.8.4.

Recall,

$$f_{\mathbf{Z}}(z_1,z_2) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_1^2} \times \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z_2^2}$$

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

$$
\begin{aligned}
f_{R,\theta}(r,\theta) &= f_{\mathbf{Z}}\big(g(r,\theta)\big)\,|\det G(r,\theta)| \\
&= f_{\mathbf{Z}}(r\cos\theta, r\sin\theta)|r| \\
&= \begin{cases} \dfrac{1}{2\pi} r\, e^{-\frac{1}{2}r^2} & \text{for } r \geq 0, -\frac{\pi}{2} \leq \theta < \frac{3\pi}{2} \\ 0 & \text{otherwise} \end{cases} \\
&= f_\theta(\theta) f_R(r),
\end{aligned}
$$

where

$$
f_\theta(\theta) = \begin{cases} \dfrac{1}{2\pi} & -\frac{\pi}{2} \leq \theta < \frac{3\pi}{2} \\ 0 & \text{otherwise} \end{cases}
$$

$$
f_R(r) = re^{-r^2/2} \quad \text{for } r \geq 0.
$$

Hence, if Z_1, Z_2 *i.i.d.* $\mathrm{N}(0,1)$, and we translate to polar coordinates, (R,θ) we find:

1. R, θ are independent.
2. $\theta \sim U(-\frac{\pi}{2}, \frac{3\pi}{2})$.
3. R has distribution called the Rayleigh distribution.

It is easy to check that $R^2 \sim \mathrm{Exp}(1/2) \equiv \mathrm{Gamma}(1, 1/2) \equiv \chi^2_2$.

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. Step 1: try using $Z = X_1 + X_2$.

Step 2: If $h(x_1, x_2) = \begin{pmatrix} y = \dfrac{x_1}{x_1 + x_2} \\ z = x_1 + x_2 \end{pmatrix}$, then $g(y,z) = \begin{pmatrix} yz \\ (1-y)z \end{pmatrix}$.

$$
G = \begin{pmatrix} z & y \\ -z & 1-y \end{pmatrix}
$$

$$\Rightarrow \det(G) = z(1-y) - (-zy)$$

$$= z(1-y) + zy = z,$$

where $z > 0$. Why?

$$f_{Y,Z}(y,z) = f_{X_1}(yz) f_{X_2}\big((1-y)z\big) z$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)}(yz)^{\alpha-1} e^{-\lambda yz} \frac{\lambda^\beta}{\Gamma(\beta)}\big((1-y)z\big)^{\beta-1} e^{-\lambda(1-y)z} z$$

$$= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \lambda^{\alpha+\beta} y^{\alpha-1}(1-y)^{\beta-1} z^{\alpha+\beta-1} e^{-\lambda z}.$$

Step 3:

$$f_Y(y) = \int_0^\infty f(y,z)\, dz$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \int_0^\infty \frac{\lambda^{\alpha+\beta} z^{\alpha+\beta-1}}{\Gamma(\alpha+\beta)} e^{-\lambda z}\, dz$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

$$= \text{Beta}(\alpha,\beta), \qquad \text{for } 0 < y < 1.$$

Exercise: Justify the range of values for y.

1.9 Moments

Suppose $X_1, X_2, \ldots, X_r$ are RVs. If $Y = h(\mathbf{X})$ is defined by a real-valued function h, then to find $E(Y)$ we can:

1. Find the distribution of Y.
2. Calculate $E(Y) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} y f_Y(y)\, dy & \text{continuous} \\ \displaystyle\sum_y y p(y) & \text{discrete} \end{cases}$

Theorem. *1.9.1*

If h and $X_1, \ldots, X_r$ are as above, then provided it exists,

$$E\big\{h(\mathbf{X})\big\} = \begin{cases} \displaystyle\sum_{x_1}\sum_{x_2}\cdots\sum_{x_r} h(x_1,\ldots,x_r)P(x_1,\ldots,x_r) & \text{if } X_1,\ldots,X_r \text{ discrete} \\ \displaystyle\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(x_1,\ldots,x_r)f(x_1,\ldots x_r)\,dx_1\ldots dx_r & \text{if } X_1,\ldots,X_r \text{ continuous} \end{cases}$$

Theorem. *1.9.2*

Suppose $X_1, \ldots, X_r$ are RVs, $h_1, \ldots, h_k$ are real-valued functions and $a_1, \ldots, a_k$ are constants. Then, provided it exists,

$$E\big\{a_1h_1(\mathbf{X}) + a_2h_2(\mathbf{X}) + \cdots + a_kh_k(\mathbf{X})\big\} = a_1E[h_1(\mathbf{X})] + a_2E[h_2(\mathbf{X})] + \cdots + a_kE\left[h_k(\mathbf{X})\right]$$

Proof. (Continuous case)

$$E\big\{a_1h_1(\mathbf{X}) + \cdots + a_kh_k(\mathbf{X})\big\} = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left\{a_1h_1(\mathbf{x}) + \cdots + a_kh_k(\mathbf{x})\right\}f_{\mathbf{x}}(\mathbf{x})\,dx_1dx_2\ldots dx_r$$

$$= a_1\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}h_1(\mathbf{x})f_X(\mathbf{x})\,dx_1\ldots dx_r$$

$$+ a_2\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}h_2(\mathbf{x})f_X(\mathbf{x})\,dx_1\ldots dx_r + \ldots$$

$$\ldots + a_k\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}h_k(\mathbf{x})f_X(\mathbf{x})\,dx_1\ldots dx_r$$

$$= \; a_1E[h_1(\mathbf{X})] + a_2E[h_2(\mathbf{X})] + \cdots + a_kE[h_k(\mathbf{X})],$$

as required.

□

Corollary.

Provided it exists,

$$E\big(a_1X_1 + a_2X_2 + \cdots + a_rX_r\big) = a_1E[X_1] + a_2E[X_2] + \cdots + a_rE[X_r].$$

Definition. 1.9.1.

If X_1, X_2 are RVs with $E[X_i] = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$, $i = 1, 2$ we define

$$\sigma_{12} = \mathrm{Cov}(X_1, X_2) = E\big[(X_1 - \mu_1)(X_2 - \mu_2)\big]$$

$$= E\big[X_1 X_2\big] - \mu_1\mu_2,$$

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

Remark:

$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) = \big\{E[X^2] - (E[X])^2\big\}.$$

In some contexts, it is convenient to use the notation

$$\sigma_{ii} = \mathrm{Var}(X_i) \quad \text{instead of} \quad \sigma_i^2.$$

Theorem. *1.9.3*

Suppose $X_1, X_2, \ldots, X_r$ are RVs with $E[X_i] = \mu_i, \mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, and let $a_1, a_2, \ldots, a_r$, $b_1, b_2, \ldots, b_r$ be constants. Then

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

Proof.

$$\text{Cov}\left(\sum_{i=1}^{r} a_i X_i, \sum_{j=1}^{r} b_j X_j\right) = E\left[\left\{\sum_{i=1}^{r} a_i X_i - E\left(\sum_{i=1}^{r} a_i X_i\right)\right\}\left\{\sum_{j=1}^{r} b_j X_j - E\left(\sum_{j=1}^{r} b_j X_j\right)\right\}\right]$$

$$= E\left\{\left(\sum_{i=1}^{r} a_i X_i - \sum_{i=1}^{r} a_i \mu_i\right)\left(\sum_{j=1}^{r} b_j X_j - \sum_{j=1}^{r} b_j \mu_j\right)\right\}$$

$$= E\left[\left\{\sum_{i=1}^{r} a_i (X_i - \mu_i)\right\}\left\{\sum_{j=1}^{r} b_j (X_j - \mu_j)\right\}\right]$$

$$= E\left\{\sum_{i=1}^{r}\sum_{j=1}^{r} a_i b_j (X_i - \mu_i)(X_j - \mu_j)\right\}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{r} a_i b_j E\left\{(X_i - \mu_i)(X_j - \mu_j)\right\} \qquad \text{(Theorem 1.9.2)}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{r} a_i b_j \sigma_{ij}, \qquad \text{as required.}$$

□

Corollary. 1

Under the above assumptions,

$$\underbrace{\text{Var}\left(\sum_{i=1}^{r} a_i X_i\right)}_{\substack{\text{covariance} \\ \text{with itself} \\ = \text{ variance}}} = \sum_{i=1}^{r}\sum_{j=1}^{r} a_i a_j \sigma_{ij} = \sum_{i=1}^{r} a_i^2 \sigma_i^2 + 2\sum_{\substack{i,j \\ i<j}} a_i a_j \sigma_{ij}.$$

Corollary. 2

$\rho = \text{Corr}(X_1, X_2)$ satisfies $|\rho| \leq 1$; and $|\rho| = 1$ implies that X_2 is a linear function of X_1.

Proof. It follows from *Corollary 1* that

$$\text{Var}\left(\frac{X_1}{\sigma_1} + \frac{X_2}{\sigma_2}\right) = 2(1 + \rho)$$

Similarly, we can show that

$$\mathrm{Var}\left(\frac{X_1}{\sigma_1} - \frac{X_2}{\sigma_2}\right) = 2(1-\rho)$$

Then

$$2(1+\rho) \geq 0 \quad \Rightarrow \rho \geq -1$$

and

$$2(1-\rho) \geq 0 \quad \Rightarrow \rho \leq 1$$

Hence, $-1 \leq \rho \leq 1$, as required.

Now suppose $|\rho| = 1 \quad \Rightarrow \quad \Delta = 0$

$\Rightarrow$ there is single t_0 such that:

$$0 = q(t_0) = \mathrm{Var}(X_1 - t_0 X_2),$$

i.e., $X_1 = t_0 X_2 + c$ with probability 1. □

Theorem. *1.9.4*

If X_1, X_2 are independent, and $\mathrm{Var}(X_1), \mathrm{Var}(X_2)$ exist, then $\mathrm{Cov}(X_1, X_2) = 0$.

Proof. (Continuous)

$$\mathrm{Cov}(X_1, X_2) = E\big\{(X_1 - \mu_1)(X_2 - \mu_2)\big\}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_1)(x_2-\mu_2)f(x_1,x_2)\,dx_1dx_2$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_1-\mu_1)(x_2-\mu_2)f_{X_1}(x_1)f_{X_2}(x_2)\,dx_1\,dx_2 \qquad \text{(since } X_1, X_2 \text{ independent)}$$

$$= \left(\int_{-\infty}^{\infty}(x_1-\mu_1)f_{X_1}(x_1)\,dx_1\right)\left(\int_{-\infty}^{\infty}(x_2-\mu_2)f_{X_2}(x_2)\,dx_2\right)$$

$$= 0 \times 0$$

$$= 0.$$

□

Remark:

But the converse does NOT apply, in general!

That is, $\mathrm{Cov}(X_1, X_2) = 0 \not\Rightarrow X_1, X_2$ independent.

Definition. 1.9.2

If X_1, X_2 are RVs, we define the symbol $E[X_1|X_2]$ to be the expectation of X_1 calculated with respect to the conditional distribution of $X_1|X_2$,

i.e.,

$$E\big[X_1|X_2\big] = \begin{cases} \displaystyle\sum_{x_1} x_1 P_{X_1|X_2}(x_1|x_2) & X_1|X_2 \text{ discrete} \\ \\ \displaystyle\int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1|x_2)\, dx_1 & X_1|X_2 \text{ continuous} \end{cases}$$

Theorem. *1.9.5*

Provided the relevant moments exist,

$$E(X_1) = E_{X_2}\big\{E(X_1|X_2)\big\},$$

$$\mathrm{Var}(X_1) = E_{X_2}\big\{\mathrm{Var}(X_1|X_2)\big\} + \mathrm{Var}_{X_2}\big\{E(X_1|X_2)\big\}.$$

Proof. 1. (Continuous case)

$$\begin{aligned}
E_{X_2}\big\{E(X_1|X_2)\big\} &= \int_{-\infty}^{\infty} E(X_1|X_2) f_{X_2}(x_2)\,dx_2 \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x_1 f(x_1|x_2)\,dx_1 \right) f_{X_2}(x_2)\,dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f(x_1|x_2) f_{X_2}(x_2)\,dx_1\,dx_2 \\
&= \int_{-\infty}^{\infty} x_1 \underbrace{\int_{-\infty}^{\infty} f(x_1,x_2)\,dx_2}_{= f_{X_1}(x_1)}\,dx_1 \\
&= \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1)\,dx_1 \\
&= E(X_1).
\end{aligned}$$

□

1.9.1 Moment generating functions

Definition. 1.9.3

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

$$M_{\mathbf{X}}(\mathbf{t}) = M_{\mathbf{X}}(t_1, t_2, \ldots, t_r) = E\big[e^{t_1X_1 + t_2X_2 + \cdots + t_rX_r}\big].$$

Theorem. *1.9.6*

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

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

provided it exists.

Proof. (Theorem 1.9.6, continuous case)

$$M_{\mathbf{X}}(t_1,\dots,t_r) = \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} e^{t_1x_1+t_2x_2+\dots+t_rx_r} f_{\mathbf{X}}(x_1,\dots,x_r)\,dx_1\dots dx_r$$

$$= \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} e^{t_1x_1}e^{t_2x_2}\dots e^{t_rx_r} f_{X_1}(x_1)f_{X_2}(x_2)\dots f_{X_r}(x_r)\;dx_1\dots dx_r$$

$$= \left(\int_{-\infty}^{\infty} e^{t_1x_1} f_{X_1}(x_1)\,dx_1\right)\left(\int_{-\infty}^{\infty} e^{t_2x_2} f_{X_2}(x_2)\,dx_2\right)\dots\left(\int_{-\infty}^{\infty} e^{t_rx_r} f_{X_r}(x_r)\,dx_r\right)$$

$$= M_{X_1}(t_1)M_{X_2}(t_2)\dots M_{X_r}(t_r), \qquad \text{as required.}$$

□

We saw previously that if $Y = h(\mathbf{X})$, then we can find $M_Y(t) = E[e^{th(\mathbf{X})}]$ from the joint distribution of $X_1,\dots,X_r$ without calculating $f_Y(y)$ explicitly.

A simple, but important case is:

$$Y = X_1 + X_2 + \dots + X_r.$$

Theorem. *1.9.7*

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

1. *$M_Y(t) = M_{\mathbf{X}}(t,t,t,\dots,t)$.*
2. *If $X_1\dots X_r$ are independent then $M_Y(t) = M_{X_1}(t)M_{X_2}(t)\dots M_{X_r}(t)$.*
3. *If $X_1,\dots,X_r$ independent and identically distributed with common MGF $M_X(t)$, then:*
$$M_Y(t) = \big[M_X(t)\big]^r.$$

Proof.

$$\begin{aligned} M_X(t) &= E[e^{tY}] \\ &= E[e^{t(X_1+X_2+\cdots+X_r)}] \\ &= E[e^{tX_1+tX_2+\cdots+tX_r}] \\ &= M_{\mathbf{X}}(t,t,t,\ldots,t) \qquad \text{(using def 1.9.3).} \end{aligned}$$

For parts (2) and (3), substitute into (1) and use Theorem 1.9.6. □

Examples

Consider RVs X, V, defined by $V \sim \text{Gamma}(\alpha, \lambda)$ and the conditional distribution of $X|V \sim \text{Po}(V)$.

Find $E(X)$ and $\text{Var}(X)$.

Solution. Use

$$\begin{aligned} E(X) &= E\{E(X|V)\} \\ \text{Var}(X) &= E_V\{\text{Var}(X|V)\} + \text{Var}_V\{E(X|V)\} \\ E(X|V) &= V \\ \text{Var}(X|V) &= V \\ \text{so } E(X) &= E\{E(X|V)\} = E(V) = \frac{\alpha}{\lambda}. \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &= E_V\{\text{Var}(X|V)\} + \text{Var}_V\{E(X|V)\} \\ &= E_V(V) + \text{Var}_V(V) \\ &= \frac{\alpha}{\lambda} + \frac{\alpha}{\lambda^2} \\ &= \frac{\alpha}{\lambda}\left(1 + \frac{1}{\lambda}\right) = \alpha\left(\frac{1+\lambda}{\lambda^2}\right). \end{aligned}$$

Remark

The marginal distribution of X is sometimes called the negative binomial distribution. In particular, when α is an integer, it corresponds to the definition previously with $p = \dfrac{\lambda}{1+\lambda}$.

Examples

1. 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 *i.i.d.* Bernoulli with parameter p and

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

   then $M_Y(t) = \big(1 + p(e^t - 1)\big)^n$, which agrees with the formula previously given for the binomial distribution.

2. Suppose $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ independently. Find the MGF of $Y = X_1 + X_2$.

Solution. Recall that $M_{X_1}(t) = e^{t\mu_1} e^{t^2\sigma_1^2/2}$

$$M_{X_2}(t) = e^{t\mu_2} e^{t^2\sigma_2^2/2}$$

$$\Rightarrow M_Y(t) = M_{X_1}(t) M_{X_2}(t)$$

$$= e^{t(\mu_1+\mu_2)} e^{t^2(\sigma_1^2+\sigma_2^2)/2}$$

$$\Rightarrow Y \sim \mathrm{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right).$$

1.9.2 Marginal distributions and the MGF

To find the moment generating function of the marginal distribution of any set of components of $\mathbf{X}$, set to 0 the complementary elements of $\mathbf{t}$ in $M_{\mathbf{X}}(\mathbf{t})$.

Let $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)^T$, and $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2)^T$. Then

$$M_{\mathbf{X}_1}(\mathbf{t}_1) = M_{\mathbf{X}} \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{0} \end{pmatrix}.$$

To see this result: Note that if A is a constant matrix, and $\mathbf{b}$ is a constant vector, then

$$M_{A\mathbf{X}+\mathbf{b}}(\mathbf{t}) = e^{\mathbf{t}^T\mathbf{b}} M_{\mathbf{X}}(A^T\mathbf{t}).$$

Proof.

$$\begin{aligned} M_{A\mathbf{X}+\mathbf{b}}(\mathbf{t}) &= E[e^{\mathbf{t}^T(A\mathbf{X}+\mathbf{b})}] = e^{\mathbf{t}^T\mathbf{b}}E[e^{\mathbf{t}^TA\mathbf{X}}] \\ &= e^{\mathbf{t}^T\mathbf{b}}E[e^{(A^T\mathbf{t})^T\mathbf{X}}] = e^{\mathbf{t}^T\mathbf{b}}M_{\mathbf{X}}(A^T\mathbf{t}). \end{aligned}$$

□

Now partition

$$\mathbf{t}_{r\times 1} = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{t}_2 \end{pmatrix}$$

and

$$A_{l\times r} = (I_{l\times l} \quad \mathbf{0}_{l\times m}).$$

Note that

$$A\mathbf{X} = (I_{l\times l} \quad \mathbf{0}_{l\times m}) \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} = \mathbf{X}_1,$$

and

$$A^T\mathbf{t}_1 = \begin{pmatrix} I_{l\times l} \\ \mathbf{0}_{m\times l} \end{pmatrix} \mathbf{t}_1 = \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{0} \end{pmatrix}.$$

Hence,

$$\begin{aligned} M_{\mathbf{X}_1}(\mathbf{t}_1) = M_{A\mathbf{X}}(\mathbf{t}_1) &= M_{\mathbf{X}}(A^T\mathbf{t}_1) \\ &= M_{\mathbf{X}} \begin{pmatrix} \mathbf{t}_1 \\ \mathbf{0} \end{pmatrix}, \end{aligned}$$

as required.

Note that similar results hold for more than two random subvectors.

The major limitation of the MGF is that it may not exist. The *characteristic function* on the other hand is defined for all distributions. Its definition is similar to the MGF, with it replacing t, where $i = \sqrt{-1}$; the properties of the characteristic function are similar to those of the MGF, but using it requires some familiarity with complex analysis.

1.9.3 Vector notation

Consider the random vector

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

with $E[X_i] = \mu_i, \mathrm{Var}(X_i) = \sigma_i^2 = \sigma_{ii}, \mathrm{Cov}(X_i, X_j) = \sigma_{ij}$.

Define the mean vector by:

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

and the variance matrix (covariance matrix) by:

$$\Sigma = \mathrm{Var}(\mathbf{X}) = [\sigma_{ij}]_{\substack{i=1,\ldots,r \\ j=1,\ldots,r}}$$

Finally, the correlation matrix is defined by:

$$R = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1r} \\ & \ddots & \ddots & \\ & & \ddots & \rho_{r-1,r} \\ & & & 1 \end{pmatrix} \qquad \text{where } \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}.$$

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

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

Theorem. *1.9.8*

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

1. $E(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\boldsymbol{\mu}$
2. $\mathrm{Var}(\mathbf{a}^t\mathbf{X}) = \mathbf{a}^T\Sigma\mathbf{a}$
3. $\mathrm{Cov}(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{X}) = \mathbf{a}^T\Sigma\mathbf{b}$

Remark

It is easy to check that this is just a re-statement of Theorem 1.9.3 using matrix notation.

Theorem. *1.9.9*

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

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

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

Remark

This is also a re-statement of previously established results. To see this, observe that if $\mathbf{a}_i^T$ is the i^{th} row of A, we see that $Y_i = \mathbf{a}_i^T\mathbf{X}$ and, moreover, the $(i,j)^{th}$ element of

$$A\Sigma A^T = \mathbf{a}_i^T\Sigma\mathbf{a}_j = \mathrm{Cov}\big(\mathbf{a}_i^T\mathbf{X}, \mathbf{a}_j^T\mathbf{X}\big)$$

$$= \mathrm{Cov}(Y_i, Y_j).$$

1.9.4 Properties of variance matrices

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

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

Definition. 1.9.4

The square, symmetric matrix M is said to be positive definite [non-negative definite] if $\mathbf{a}^T M\mathbf{a} > 0$ $[\geq 0]$ for every vector $\mathbf{a} \in \mathbb{R}^r$ s.t. $\mathbf{a} \neq \mathbf{0}$.

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

$\Longrightarrow$ To see the necessity of this condition, consider the linear combination $\mathbf{a}^T\mathbf{X}$.

By Theorem 1.9.8, we have

$$0 \leq \mathrm{Var}(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\Sigma\mathbf{a} \qquad \text{for every } \mathbf{a};$$

$\Longrightarrow \Sigma$ must be non-negative definite.

Suppose λ is an eigenvalue of Σ, and let $\mathbf{a}$ be the corresponding eigenvector. Then

$$\Sigma\mathbf{a} = \lambda\mathbf{a}$$
$$\Longrightarrow \mathbf{a}^T\Sigma\mathbf{a} = \lambda\mathbf{a}^T\mathbf{a} = \lambda||\mathbf{a}||^2.$$

Hence Σ is non-negative definite (positive definite) iff its eigenvalues are all non-negative (positive).

If Σ is non-negative definite <u>but not</u> positive definite, then there must be <u>at least</u> one zero eigenvalue. Let $\mathbf{a}$ be the corresponding eigenvector.

$\Longrightarrow \mathbf{a} \neq \mathbf{0}$ but $\mathbf{a}^T\Sigma\mathbf{a} = 0$. That is, $\mathrm{Var}(\mathbf{a}^T X) = 0$ for that $\mathbf{a}$.

$\Longrightarrow$ the distribution of 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.

Finally, recall that if $\lambda_1, \ldots, \lambda_r$ are eigenvalues of Σ, then $\det(\Sigma) = \prod_{i=1}^{r} \lambda_r$

$$\implies \det(\Sigma) > 0 \quad \text{for } \Sigma \text{ positive definite,}$$

and

$$\det(\Sigma) = 0 \quad \text{for } \Sigma \text{ non-negative definite but } \underline{\text{not}} \text{ positive definite.}$$

1.10 The multivariable normal distribution

Definition. 1.10.1

The random vector $\mathbf{X} = (X_1, \ldots, X_r)^T$ is said to have the r-dimensional multivariate normal distribution with parameters $\boldsymbol{\mu} \in \mathbb{R}^r$ and $\Sigma_{r\times r}$ positive definite, if it has PDF

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

We write $X \sim \mathrm{N}_r(\boldsymbol{\mu}, \Sigma)$.

Examples

1. $\underline{r = 2 \quad \text{The bivariate normal distribution}}$

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

   $$\implies \quad |\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2) \quad \text{and}$$

   $$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}$$

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

   $$\left.\left.\left. -2\rho \left(\frac{x_1-\mu_1}{\sigma_1} \right) \left(\frac{x_2-\mu_2}{\sigma_2} \right) \right] \right] \right\}.$$

2. If $Z_1, Z_2, \ldots, Z_r$ are *i.i.d.* $\mathrm{N}(0,1)$ and $\mathbf{Z} = (Z_1, \ldots, Z_r)^T$,

$$\Rightarrow f_{\mathbf{Z}}(z) = \prod_{i=1}^{r} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2}$$

$$= \frac{1}{(2\pi)^{r/2}}\, e^{-1/2\sum_{i=1}^{r} z_i^2} = \frac{1}{(2\pi)^{r/2}} e^{-\frac{1}{2}\mathbf{z}^T\mathbf{z}},$$

which is $\mathrm{N}_r(\mathbf{0}_r, I_{r\times r})$ PDF.

Theorem. *1.10.1*

Suppose $\mathbf{X} \sim N_r(\boldsymbol{\mu}, \Sigma)$ *and let* $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$*, where* $\mathbf{b} \in \mathbb{R}^r$ *and* $A_{r\times r}$ *invertible are fixed. Then* $\mathbf{Y} \sim N_r(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$.

Proof.

We use the transformation rule for PDFs:

If $\mathbf{X}$ has joint PDF $f_X(\mathbf{x})$ and $\mathbf{Y} = g(\mathbf{X})$ for $g : \mathbb{R}^r \to \mathbb{R}^r$ invertible, continuously differentiable, then $f_Y(\mathbf{y}) = f_X(h(\mathbf{y}))|H|$, where $h : \mathbb{R}^r \to \mathbb{R}^r$ such that $h = g^{-1}$ and $H = \left(\dfrac{\partial h_i}{\partial y_j}\right)$.

In this case, we have $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$

$$\Rightarrow h(\mathbf{y}) = A^{-1}(\mathbf{y} - \mathbf{b}) \left[\begin{array}{c}\text{solving for } \mathbf{x} \text{ in} \\ \mathbf{y} = A\mathbf{x} + \mathbf{b}\end{array}\right]$$

$$\Rightarrow H = A^{-1}.$$

Hence,

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{1}{(2\pi)^{r/2}|\Sigma|^{1/2}}\, e^{-1/2(A^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu})^T\Sigma^{-1}(A^{-1}(\mathbf{y}-\mathbf{b})-\boldsymbol{\mu})}|A^{-1}| \\ &= \frac{1}{(2\pi)^{r/2}|\Sigma|^{1/2}|AA^T|^{1/2}}\, e^{-1/2(\mathbf{y}-(A\boldsymbol{\mu}+\mathbf{b}))^T(A^{-1})^T\Sigma^{-1}(A^{-1})(\mathbf{y}-(A\boldsymbol{\mu}+\mathbf{b}))} \\ &= \frac{1}{(2\pi)^{r/2}|A\Sigma A^T|^{1/2}}\, e^{-1/2(\mathbf{y}-(A\boldsymbol{\mu}+\mathbf{b}))^T(A\Sigma A^T)^{-1}(\mathbf{y}-(A\boldsymbol{\mu}+\mathbf{b}))} \end{aligned}$$

which is exactly the PDF for $\mathrm{N}_r(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$. □

Now suppose $\Sigma_{r\times r}$ is any symmetric positive definite matrix and recall that we can write $\Sigma = E \wedge E^T$ where $\wedge = \text{ diag } (\lambda_1, \lambda_2, \ldots, \lambda_r)$ and $E_{r\times r}$ is such that $EE^T = E^T E = I$. Since Σ is positive definite, we must have $\lambda_i > 0, \quad i = 1, \ldots, r$ and we can define the symmetric square-root matrix by:

$$\Sigma^{1/2} = E \wedge^{1/2} E^T \quad \text{where } \wedge^{1/2} = \text{ diag } (\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_r}).$$

Note that $\Sigma^{1/2}$ is symmetric and satisfies:

$$\left(\Sigma^{1/2}\right)^2 = \Sigma^{1/2}\Sigma^{1/2^T} = \Sigma.$$

Now recall that if $Z_1, Z_2, \ldots, Z_r$ are *i.i.d.* $\mathrm{N}(0,1)$ then $\mathbf{Z} = (Z_1, \ldots, Z_r)^T \sim \mathrm{N}_r(\mathbf{0}, I)$.

Because of the *i.i.d.* $\mathrm{N}(0,1)$ assumption, we know in this case that $E(\mathbf{Z}) = \mathbf{0}, \mathrm{Var}(\mathbf{Z}) = I$.

Now, let $\mathbf{X} = \Sigma^{1/2}\mathbf{Z} + \boldsymbol{\mu}$:

1. From Theorem 1.9.9 we have

$$E(\mathbf{X}) = \Sigma^{1/2}E(\mathbf{Z}) + \boldsymbol{\mu} = \boldsymbol{\mu} \text{ and } \mathrm{Var}(\mathbf{X}) = \Sigma^{1/2}\,\mathrm{Var}(\mathbf{Z})\left(\Sigma^{1/2}\right)^T.$$

2. From Theorem 1.10.1,

$$\mathbf{X} \sim \mathrm{N}_r(\boldsymbol{\mu}, \Sigma).$$

Since this construction is valid for any symmetric positive definite Σ and any $\boldsymbol{\mu} \in \mathbb{R}^r$, we have proved,

Theorem. *1.10.2*

If $\mathbf{X} \sim N_r(\boldsymbol{\mu}, \Sigma)$ *then*

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

Theorem. *1.10.3*
If $\mathbf{X} \sim N_r(\boldsymbol{\mu}, \Sigma)$ *then* $\mathbf{Z} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu}) \sim N_r(0, I)$.

Proof.
Use Theorem 1.10.1. □

Suppose

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

where $\mathbf{X}_i, \boldsymbol{\mu}_i$ are of dimension r_i and Σ_{ij} is $r_i \times r_j$. In order that Σ be symmetric, we must have Σ_{11}, Σ_{22} symmetric and $\Sigma_{12} = \Sigma_{21}^T$.

We will derive the marginal distribution of $\mathbf{X}_2$ and the conditional distribution of $\mathbf{X}_1|\mathbf{X}_2$.

Lemma. 1.10.1

If $M = \begin{bmatrix} B & A \\ O & I \end{bmatrix}$ is a square, partitioned matrix, then $|M| = |B|$ (det).

Lemma. 1.10.2

If $M = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ then $|M| = |\Sigma_{22}||\Sigma_{11} - \Sigma_{12}\ \Sigma_{22}^{-1}\Sigma_{21}|$

Proof. (of Lemma 1.10.2)

Let

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

and observe

$$CM = \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ \Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix},$$

and

$$|C| = \frac{1}{|\Sigma_{22}|}, \qquad |CM| = |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|$$

Finally, observe $|CM| = |C||M|$

$$\Rightarrow \quad |M| = \frac{|CM|}{|C|} = |\Sigma_{22}||\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|.$$

□

Theorem. *1.10.4*

Suppose that

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

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

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

Proof.

We will exhibit $f(\mathbf{x}_1, \mathbf{x}_2)$ in the form

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

where $h(\mathbf{x}_1, \mathbf{x}_2)$ is a PDF with respect to $\mathbf{x}_1$ for each (given) $\mathbf{x}_2$.

It follows then that $f_{X_1|X_2}(\mathbf{x}_1|\mathbf{x}_2) = h(\mathbf{x}_1, \mathbf{x}_2)$ and $g(\mathbf{x}_2) = f_{X_2}(\mathbf{x}_2)$.

Now observe that:

$$f(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{(2\pi)^{(r_1+r_2)/2} \begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{vmatrix}^{1/2}} \quad \times$$

$$\exp\left[-\frac{1}{2}\left(\left(\mathbf{x_1} - \boldsymbol{\mu}_1\right)^T, \left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right)^T\right) \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix}\right]$$

and

1.

$$\begin{aligned} (2\pi)^{(r_1+r_2)/2} \begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix}^{1/2} &= (2\pi)^{r_1/2}(2\pi)^{r_2/2} \left|\Sigma_{22}\right|^{1/2} \left|\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right|^{1/2} \\ &= (2\pi)^{r_1/2} \left|\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right|^{1/2} (2\pi)^{r_2/2} \left|\Sigma_{22}\right|^{1/2} \end{aligned}$$

2.

$$\begin{aligned} \left((\mathbf{x}_1 - \boldsymbol{\mu}_1)^T, (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T\right) \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} &= \left((\mathbf{x}_1 - \boldsymbol{\mu}_1)^T, (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T\right) \\ \times \begin{bmatrix} \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & -\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{11}\right)^{-1} \\ & \times\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21} \\ \times\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} & \times\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} \\ & \times\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \end{aligned}$$

$$= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T V^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1) - (\mathbf{x}_1 - \boldsymbol{\mu}_1)^T V^{-1}\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

$$-(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}\Sigma_{21}V^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1) + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}\Sigma_{21}V^{-1}\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

$$+(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2).$$

How did this come about:

$$\begin{aligned}(x,y)\begin{bmatrix}a & b\\ c & d\end{bmatrix}\begin{pmatrix}x\\ y\end{pmatrix} &= (x,y)\begin{bmatrix}ax+by\\ cx+dy\end{bmatrix}\\ &= x(ax+by)+y(cx+dy)\\ &= ax^2+bxy+cyx+dy^2\end{aligned}$$

$$\begin{aligned}&= \big((\mathbf{x}_1-\boldsymbol{\mu}_1)-\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\big)^T V^{-1}\big((\mathbf{x}_1-\boldsymbol{\mu}_1)-\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x_2}-\boldsymbol{\mu}_2)\big)\\ &\quad+(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\end{aligned}$$

Note:

$$\begin{aligned}(\mathbf{x}-\mathbf{y})^T A(\mathbf{x}-\mathbf{y}) &= (\mathbf{x}-\mathbf{y})^T(A\mathbf{x}-A\mathbf{y})\\ &= \mathbf{x}^TA\mathbf{x}-\mathbf{x}^TA\mathbf{y}-\mathbf{y}^TA\mathbf{x}+\mathbf{y}^TA\mathbf{y}\end{aligned}$$

And: $\Sigma_{12}^T = \Sigma_{21}$

$$\begin{aligned}&= \big(\mathbf{x}_1-(\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))\big)^T\big(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big)^{-1}\times\\ &\quad\big(\mathbf{x}_1-(\boldsymbol{\mu}_1+\Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))\big)+(\mathbf{x}_2-\boldsymbol{\mu}_2)^T\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2).\end{aligned}$$

Combining (1), (2) we obtain:

$$f(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{(2\pi)^{(r_1+r_2)/2} \begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix}^{1/2}}$$

$$\times \exp\left\{ -\frac{1}{2}((\mathbf{x}_1 - \boldsymbol{\mu}_1)^T, (\mathbf{x}_2 - \boldsymbol{\mu}_2)^T) \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \right\}$$

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

$$\times \exp\left\{ -\frac{1}{2}(\mathbf{x}_1 - (\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)))^T \right.$$

$$\left. \times \left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1} (\mathbf{x}_1 - (\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2))) \right\}$$

$$\times \frac{1}{(2\pi)^{r_2/2} \left|\Sigma_{22}\right|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_2 - \boldsymbol{\mu}_2)^T \Sigma_{22}^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \right\}$$

$= h(\mathbf{x}_1, \mathbf{x}_2) g(\mathbf{x}_2)$, where for fixed $\mathbf{x}_2$, $h(\mathbf{x}_1, \mathbf{x}_2)$, is the

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

PDF, and $g(\mathbf{x}_2)$ is the $\mathrm{N}_{r_2}(\boldsymbol{\mu}_2, \Sigma_{22})$ PDF.

Hence, the result is proved. □

Theorem. *1.10.5*

Suppose that $\mathbf{X} \sim N_r(\boldsymbol{\mu}, \Sigma)$ *and* $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, *where* $A_{p\times r}$ *with linearly independent rows and* $\mathbf{b} \in \mathbb{R}^p$ *are fixed. Then* $\mathbf{Y} \sim N_p(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$.

[Note: $p \leq r$.*]*

Proof. Use method of regular transformations.

Since the rows of A are linearly independent, we can find $B_{(r-p)\times r}$ such that the $r \times r$ matrix $\begin{pmatrix} B \\ A \end{pmatrix}$ is invertible.

If we take $\mathbf{Z} = B\mathbf{X}$, we have

$$\begin{pmatrix}\mathbf{Z}\\ \mathbf{Y}\end{pmatrix} = \begin{pmatrix}B\\ A\end{pmatrix}\mathbf{X} + \begin{pmatrix}\mathbf{0}\\ \mathbf{b}\end{pmatrix} \sim N\left(\begin{bmatrix}B\boldsymbol{\mu}+\mathbf{0}\\ A\boldsymbol{\mu}+\mathbf{b}\end{bmatrix}, \begin{bmatrix}B\Sigma B^T & B\Sigma A^T\\ A\Sigma B^T & A\Sigma A^T\end{bmatrix}\right)$$

by Theorem 1.10.1.

Hence, from Theorem 1.10.4, the marginal distribution for $\mathbf{Y}$ is

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

□

1.10.1 The multivariate normal MGF

The multivariate normal moment generating function for a random vector $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ is given by

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

Prove this result as an exercise!

The *characteristic function* of $\mathbf{X}$ is

$$E[\exp(i\mathbf{t}^T\mathbf{X})] = \exp\left(i\mathbf{t}^T\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right)$$

The marginal distribution of $\mathbf{X}_1$ (or $\mathbf{X}_2$) is easy to derive using the multivariate normal MGF.

Let

$$\mathbf{t} = \begin{pmatrix}\mathbf{t}_1\\ \mathbf{t}_2\end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{pmatrix}.$$

Then the marginal distribution of $\mathbf{X}_1$ is obtained by setting $\mathbf{t}_2 = \mathbf{0}$ in the expression for the MGF of $\mathbf{X}$.

Proof:

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= \exp\left(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}\right)\\ &= \exp\left(\mathbf{t_1}^T\boldsymbol{\mu_1} + \mathbf{t_2}^T\boldsymbol{\mu_2} + \frac{1}{2}\mathbf{t_1}^T\Sigma_{11}\mathbf{t}_1 + \mathbf{t_1}^T\Sigma_{12}\mathbf{t}_2 + \frac{1}{2}\mathbf{t_2}^T\Sigma_{22}\mathbf{t}_2\right). \end{aligned}$$

Now,

$$M_{\mathbf{X}_1}(\mathbf{t}_1) = M_{\mathbf{X}}\begin{pmatrix}\mathbf{t}_1\\ \mathbf{0}\end{pmatrix} = \exp\left(\mathbf{t_1}^T\boldsymbol{\mu_1} + \frac{1}{2}\mathbf{t_1}^T\Sigma_{11}\mathbf{t}_1\right),$$

which is the MGF of $\mathbf{X}_1 \sim N_{r_1}(\boldsymbol{\mu_1}, \Sigma_{11})$.

Similarly for $\mathbf{X}_2$. This means that all marginal distributions of a multivariate normal distribution are multivariate normal themselves. Note though, that in general the opposite implication is not true: there are examples of non-normal multivariate distributions whose marginal distributions are normal.

1.10.2 Independence and normality

We have seen previously that X, Y independent $\Rightarrow \mathrm{Cov}(X, Y) = 0$, but not vice-versa.

An exception is when the data are (jointly) normally distributed.

In particular, if (X, Y) have the bivariate normal distribution, then $\mathrm{Cov}(X, Y) = 0 \Longleftrightarrow$ X, Y are independent.

Theorem. *1.10.6*

Suppose $X_1, X_2, \ldots, X_r$ have a multivariate normal distribution. Then $X_1, X_2, \ldots, X_r$ are independent if and only if $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$.

Proof.

$(\Longrightarrow) X_1, \ldots, X_r$ independent

$\Longrightarrow \mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$, has already been proved.

$(\Longleftarrow)$ Suppose $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$

$$\begin{aligned}
\Longrightarrow \mathrm{Var}(\mathbf{X}) = \Sigma &= \begin{bmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \sigma_{rr} \end{bmatrix} \\
\Longrightarrow \Sigma^{-1} &= \begin{bmatrix} \sigma_{11}^{-1} & 0 & \dots & 0 \\ 0 & \sigma_{22}^{-1} & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \sigma_{rr}^{-1} \end{bmatrix} \qquad \text{and } |\Sigma| = \sigma_{11}\sigma_{22}\dots\sigma_{rr} \\
\Longrightarrow f_{\mathbf{X}}(\mathbf{x}) &= \frac{1}{(2\pi)^{r/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})} \\
&= \frac{1}{(2\pi)^{r/2}(\sigma_{11}\sigma_{22}\dots\sigma_{rr})^{1/2}} e^{-\frac{1}{2}\sum_{i=1}^{r}\frac{(x_i-\mu_i)^2}{\sigma_{ii}}} \\
&= \left(\frac{1}{\sqrt{2\pi}\sqrt{\sigma_{11}}} e^{-\frac{1}{2}\frac{(x_1-\mu_1)^2}{\sigma_{11}}}\right)\left(\frac{1}{\sqrt{2\pi}\sqrt{\sigma_{22}}} e^{-\frac{1}{2}\frac{(x_2-\mu_2)^2}{\sigma_{22}}}\right)\dots \\
&\quad \dots\left(\frac{1}{\sqrt{2\pi}\sqrt{\sigma_{rr}}} e^{-\frac{1}{2}\frac{(x_r-\mu_r)^2}{\sigma_{rr}}}\right) \\
&= f_1(x_1)f_2(x_2)\dots f_r(x_r)
\end{aligned}$$

$$\Longrightarrow X_1, \dots, X_r \qquad \text{are independent.}$$

□

Note:

$$\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) &= (x_1 - \mu_1 x_2 - \mu_2 \dots x_r - \mu_r) \\
&\quad \times \begin{bmatrix} \sigma_{11}^{-1} & 0 & \dots & 0 \\ 0 & \sigma_{22}^{-1} & \dots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \sigma_{rr}^{-1} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_r - \mu_r \end{bmatrix} \\
&= \sum_{i=1}^{r} \frac{(x_i - \mu_i)^2}{\sigma_{ii}}.
\end{aligned}$$

Remark

The same methods can be used to establish a similar result for block diagonal matrices. The simplest case is the following, which is most easily proved using moment generating functions:

$$\mathbf{X}_1, \mathbf{X}_2 \qquad \text{are independent}$$

if and only if $\Sigma_{12} = \mathbf{0}$, i.e.,

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim \mathrm{N}_{r_1+r_2}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right).$$

Theorem. *1.10.7*

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

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

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

(*Note: S^2 here is a random variable, and is not to be confused with the sample covariance matrix, which is also often denoted by S^2. Hopefully, the meaning of the notation will be clear in the context in which it is used.*)

Proof. Observe first that if $\mathbf{X} = (X_1, \ldots, X_n)^T$ then the *i.i.d.* assumption may be written as:

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

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

   Observe that $\bar{X} = B\mathbf{X}$, where

   $$B = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{bmatrix}$$

   Hence, by Theorem 1.10.5,

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

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

$$A = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} & \frac{1}{n} \\ 1-\frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} & -\frac{1}{n} \\ -\frac{1}{n} & 1-\frac{1}{n} & \dots & -\frac{1}{n} & -\frac{1}{n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1-\frac{1}{n} & -\frac{1}{n} \end{bmatrix}$$

and observe

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

We can check that $\mathrm{Var}(A\mathbf{X}) = \sigma^2 AA^T$ has the form:

$$\sigma^2 \begin{bmatrix} \frac{1}{n} & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \Sigma_{22} & \\ 0 & & & \end{bmatrix}$$

$\Rightarrow \bar{X}, (X_1 - \bar{X}, X_2 - \bar{X}, \dots, X_{n-1} - \bar{X})$ are independent.

(By multivariate normality.)

Finally, since

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

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

3. Prove:

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

Consider the identity:

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

(subtract and add $\bar{X}$ to first term)

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

$$\text{and let } R_1=\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma^2}$$

$$R_2=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}=\frac{(n-1)S^2}{\sigma^2}$$

$$R_3=\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2$$

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

$$M_1(t)=M_2(t)M_3(t) \qquad \text{since} \quad \begin{array}{l} R_1=R_2+R_3 \\ \text{with } R_2 \text{ and } R_3 \text{ independent} \\ \quad\downarrow \qquad\quad \downarrow \\ \text{depends} \quad \text{depends only} \\ \text{only on} \quad\ \text{on } \bar{X} \\ S^2 \end{array}$$

$$\Rightarrow M_2(t)=\frac{M_1(t)}{M_3(t)}.$$

Next, observe that $R_3=\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2\sim\chi^2_1$

$$\Rightarrow M_3(t)=\frac{1}{(1-2t)^{1/2}},$$

and

$$(R_1) \qquad \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2_n$$

$$\Rightarrow \qquad M_1(t) = \frac{1}{(1-2t)^{n/2}}.$$

$$\therefore \qquad M_2(t) = \frac{1}{(1-2t)^{(n-1)/2}},$$

which is the mgf for χ^2_{n-1}.

Hence,

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

□

Corollary.

If $X_1, \ldots, X_n$ are *i.i.d.* $N(\mu, \sigma^2)$, then:

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

Proof.

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

Now observe that:

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

and 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. □

1.11 Limit Theorems

1.11.1 Convergence of random variables

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

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**

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

3. **Almost sure convergence (Strong convergence)**

   The sequence $\{X_n\}$ is said to converge almost surely to α if for each $\varepsilon > 0$,

   $$|X_n - \alpha| > \varepsilon \quad \text{for only finite number of } n \geq 1.$$

   Remarks

   (1),(2),(3) are related by:

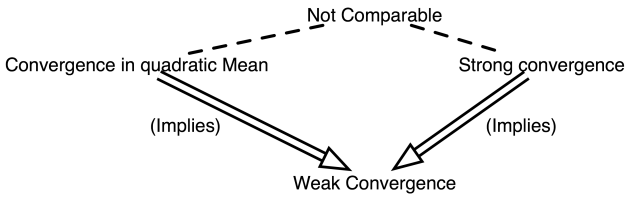


Figure 15: Relationships of convergence

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:

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

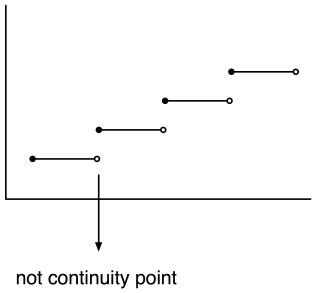


Figure 16: Discrete case

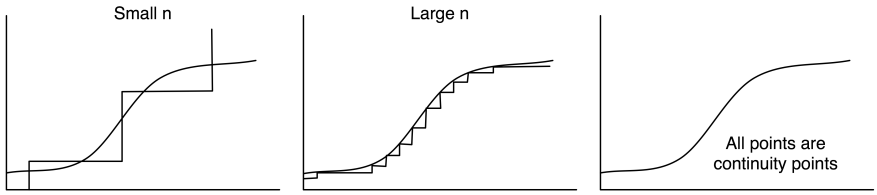


Figure 17: Normal approximation to binomial (an application of the Central Limit Theorem) continuous case

Remarks

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

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

2. 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., $X_n \xrightarrow[\mathcal{D}]{} \mathrm{N}(0,1)$.

3. An important result that we will use without proof is as follows:

   Let $M_n(t)$ be MGF of X_n and $M(t)$ be MGF of X. Then if $M_n(t) \to M(t)$ for each t in some open interval containing 0, as $n \to \infty$, then $\mathcal{L}[X_n] \xrightarrow[\mathcal{D}]{} \mathcal{L}[X]$.

   (Sometimes called the Continuity Theorem.)

Theorem. *1.11.1 (Weak law of large numbers)*

Suppose $X_1, X_2, X_3, \ldots$ is a sequence of i.i.d. RVs with $E[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2$, and let

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

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

Proof. 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 $\mathrm{Var}(\bar{X}_n) = \dfrac{\sigma^2}{n}$. 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 1.11.1 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".

Lemma. 1.11.1

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

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

Proof.

Omitted (but not difficult).

□

Remarks

This is a simple generalisation of the standard limit,

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

Consider a sequence of *i.i.d.* RVs $X_1, X_2, \ldots$ with $E(X_i) = \mu, \mathrm{Var}(X_i) = \sigma^2$ 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 $\mathrm{Var}(S_n) = n\sigma^2$.

Theorem. *1.11.2 (Central Limit Theorem)*

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.

We will use the fact that it is sufficient to prove that

$$\boxed{M_{Z_n}(t) \to e^{t^2/2} \quad \text{for each fixed } t}$$

[Note: if $Z \sim \mathrm{N}(0,1)$ then $M_Z(t) = e^{t^2/2}$.]

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

Since the U_i are independent we have:

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

$$\left[\begin{array}{ll} M_{aX}(t) & = E[e^{taX}] \\ & = M_X(at) \end{array}\right]$$

Now,

$$E(U_i) = 0 \Rightarrow M_U'(0) = 0,$$

$$\mathrm{Var}(U_i) = 1 \Rightarrow M_U''(0) = 1.$$

Consider the second-order Taylor expansion,

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

$$M_U(t) = 1 + 0 + t^2/2 + r(t), \quad \text{where } \lim_{s\to 0}\frac{r(s)}{s^2} = 0$$

$$\downarrow$$

$$(M_U(0))(\text{as } M_U'(0) = 0)$$

$$= 1 + t^2/2 + r(t).$$

$$\Rightarrow$$

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

$$= \{1 + t^2/2n + r(t/\sqrt{n})\}^n$$

$$= (1 + a_n)^n,$$

where

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

Next observe that $\lim_{n\to\infty} na_n = \frac{t^2}{2}$ for fixed t.

[To check this observe that:

$$\lim_{n\to\infty}\frac{nt^2}{2n} = \frac{t^2}{2} \qquad \text{and } \lim_{n\to\infty} nr(t/\sqrt{n})$$

$$= \lim_{n\to\infty}\frac{t^2 r(t/\sqrt{n})}{(t/\sqrt{n})^2}$$

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

Note: $s = t/\sqrt{n} \to 0$ as $n \to \infty$.]

Hence we have shown $M_{Z_n}(t) = (1 + a_n)^n$, where

$$\lim_{n\to\infty} na_n = t^2/2, \text{ so by Lemma 1.11.1,}$$

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

□

Remarks

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

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

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 *i.i.d.*.

3. Theorems 1.11.1 and 1.11.2 are concerned with the asymptotic behaviour of $\bar{X}_n$.

Theorem 1.11.1 states $\bar{X}_n \to \mu$ in prob as $n \to \infty$.

Theorem 1.11.2 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 $\dfrac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\mathrm{Var}(\bar{X}_n)}}$.

2 Statistical Inference

2.1 Basic definitions and terminology

Probability is concerned partly with the problem of predicting the behavior of the RV 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 $\boldsymbol{\theta}$:

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

where Θ is the parameter space.

Examples

(1) $\Im$ is the family of normal distributions:
$\boldsymbol{\theta} = (\mu, \sigma^2)$
$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\}$
then $\Im = \left\{F(x) : F(x) = \Phi\left(\frac{x-\mu}{\sigma^2}\right)\right\}$.

(2) $\Im$ is the family of Bernoulli distributions with success probability θ:
$\Theta = \{\theta \in [0,1] \subset \mathbb{R}\}$.

In this framework, the problem is then to use the data $x_1, \ldots, x_n$ to draw conclusions about $\boldsymbol{\theta}$.

Definition. 2.1.1

A collection of *i.i.d.* RVs, $X_1, \ldots, X_n$, with common CDF $F(x; \boldsymbol{\theta})$, is said to be a random sample (from $F(x; \boldsymbol{\theta})$).

Definition. 2.1.2

Any function $T = T(x_1, x_2, \ldots, x_n)$ that can be calculated from the data (without knowledge of θ) is called a statistic.

Example

The sample mean $\bar{x}$ is a statistic ($\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$).

Definition. 2.1.3

A statistic T with property $T(\mathbf{x}) \in \Theta \quad \forall \mathbf{x}$ is called an estimator for θ.

Example
For $x_1, x_2, \ldots, x_n$ *i.i.d.* $\mathrm{N}(\mu, \sigma^2)$, we have $\boldsymbol{\theta} = (\mu, \sigma^2)$. The quantity $(\bar{x}, s^2)$ is an estimator for $\boldsymbol{\theta}$, where $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$, $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

There are two important concepts here: the first is that *estimators* are random variables; the second is that you need to be able to distinguish between random variables and their realisations. In particular, an estimate is a realisation of a random variable.

For example, strictly speaking, $x_1, x_2, \ldots, x_n$ are *realisations* of random variables $X_1, X_2, \ldots, X_n$, and $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ is a realisation of $\bar{X}$; $\bar{X}$ is an *estimator*, and $\bar{x}$ is an estimate.

We will find that from now on however, that it is often more convenient, if less rigorous, to use the same symbol for estimator and estimate. This arises especially in the use of $\hat{\theta}$ as both estimator and estimate, as we shall see.

#

An unsatisfactory aspect of Definition 2.1.3 is that it gives no guidance on how to recognize (or construct) good estimators.

Unless stated otherwise, we will assume that θ is a scalar parameter in the following.

2.1.1 Criteria for good estimators

In broad terms, we would like an estimator to be "as close as possible to" θ with high probability.

Definition. 2.1.4

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

$$MSE_T(\theta) = E\{(T - \theta)^2\}.$$

Example:
Suppose $X_1, \ldots, X_n$ are *i.i.d.* Bernoulli θ RV's, and $T = \bar{X}$ ='proportion of successes'.

Since $nT \sim B(n, \theta)$ we have

$$E(nT) = n\theta, \qquad \text{Var}(nT) = n\theta(1-\theta)$$

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

$$\implies MSE_T(\theta) = \text{Var}(T) = \frac{\theta(1-\theta)}{n}.$$

Remark:
This example shows that $MSE_T(\theta)$ must be thought of as a function of θ rather than just a number.

For example: see Figure 18

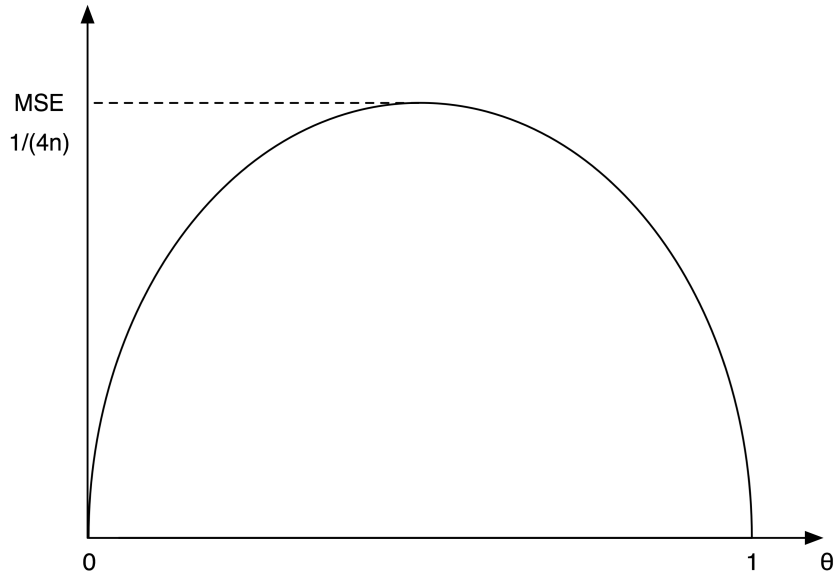


Figure 18: $MSE_T(\theta)$ as a function of θ

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. See Figure 19.

For this reason, it turns out it is not possible to construct a minimum MSE estimator in general.

To see why, suppose T^* is a minimum MSE estimator for θ. Now consider the estimator $T_a = a$, where $a \in \mathbb{R}$ is arbitrary. Then for $a = \theta$, $T_\theta = \theta$ with $MSE_{T_\theta}(\theta) = 0$.

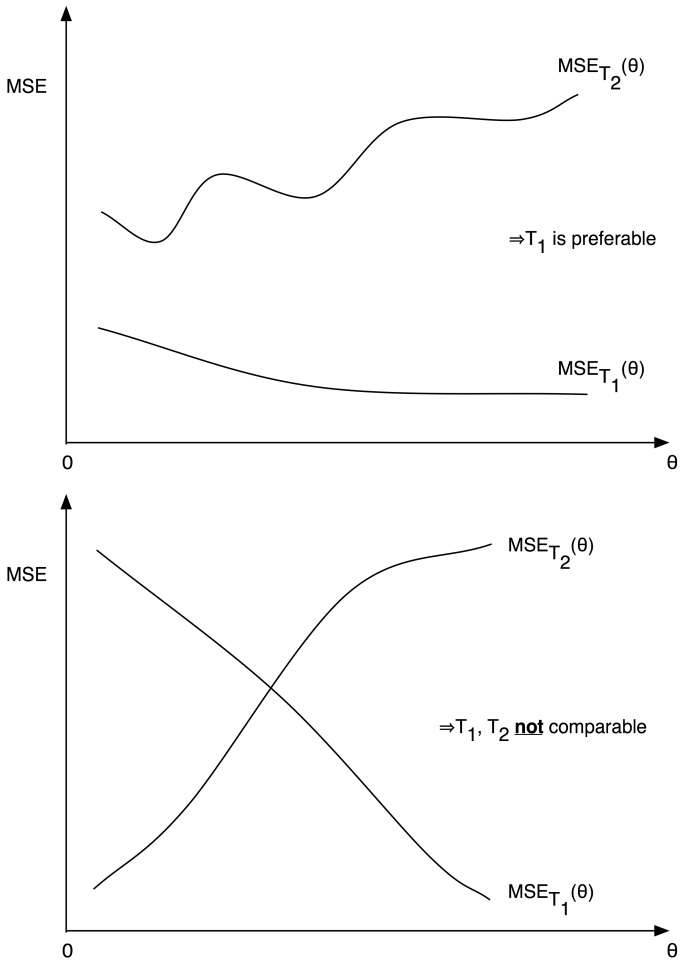


Figure 19: $MSE_T(\theta)$ as a function of θ

Observe $MSE_{T_a}(a) = 0$; hence if T^* exists, then we must have $MSE_{T^*}(a) = 0$. As a is arbitrary, we must have $MSE_{T^*}(\theta) = 0 \quad \forall \theta, \in \Theta$
$\implies T^* = \theta$ with probability 1.

Therefore we conclude that (excluding trivial situations) no minimum MSE estimator can exist.

Definition. 2.1.5 The bias of the estimator T is defined by:

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

An estimator T with $b_T(\theta) = 0$, i.e., $E(T) = \theta$, is said to be unbiased.

Remarks:

(1) Although unbiasedness is an appealing property, not all commonly used

estimates are unbiased and in some situations it is impossible to construct unbiased estimators for the parameter of interest.

Example: (Unbiasedness isn't everything)

$$E(s^2) = \sigma^2.$$

$$\text{If } E(s) = \sigma,$$

$$\text{then } \text{Var}(s) = E(s^2) - \{E(s)\}^2$$

$$= 0$$

which is <u>not</u> the case

$$\implies E(s) < \sigma.$$

(2) Intuitively, unbiasedness would seem to be a pre-requisite for a good estimator. This is to some extent formalized as follows:

Theorem. *2.1.1*

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

Remark:
Restricting attention to unbiased estimators excludes estimators of the form $T_a = a$. We will see that this permits the construction of Minimum Variance Unbiased Estimators (MVUE's) in some cases.

2.2 Minimum Variance Unbiased Estimation

2.2.1 Likelihood, score and Fisher Information

Consider data $x_1, x_2, \ldots, x_n$ assumed to be observations of RV's $X_1, X_2, \ldots, X_n$ with joint PDF (probability function)

$$f_x(\mathbf{x}; \theta) \qquad (P_x(\mathbf{x}; \theta)).$$

Definition. 2.2.1

The likelihood function is defined by

$$\mathcal{L}(\theta;\mathbf{x}) = \begin{cases} f_x(\mathbf{x};\theta) & X \text{ continuous} \\ P_x(\mathbf{x};\theta) & X \text{ discrete.} \end{cases}$$

The log likelihood function is:

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

Remark
If $x_1, x_2, \ldots, x_n$ are independent, the log likelihood function can be written as:

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

If $x_1, x_2, \ldots, x_n$ are *i.i.d.*, we have:

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

Definition. 2.2.2

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

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

and the Fisher information is

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

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

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

In the case of $x_1, x_2, \ldots, x_n$ *i.i.d.* , we have

$$\mathcal{I}(\theta) = ni(\theta).$$

Theorem. *2.2.1*

Under suitable regularity conditions,

$$E\{\mathcal{U}(\theta;\mathbf{X})\} = 0$$

$$\textit{and } \mathrm{Var}\{\mathcal{U}(\theta;\mathbf{X})\} = \mathcal{I}(\theta).$$

Proof.

$$\begin{aligned}
\text{Observe } E\{\mathcal{U}(\theta;\mathbf{X})\} &= \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\mathcal{U}(\theta;\mathbf{x})f(\mathbf{x};\theta)dx_1\dots dx_n \\
&= \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x};\theta)\right)f(\mathbf{x};\theta)dx_1\dots dx_n \\
&= \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\left(\frac{\frac{\partial}{\partial\theta}f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)}\right)f(\mathbf{x};\theta)dx_1\dots dx_n \\
&= \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(\mathbf{x};\theta)dx_1\dots dx_n \\
\text{regularity} \implies &= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}f(\mathbf{x};\theta)dx_1\dots dx_n \\
&= \frac{\partial}{\partial\theta}1 \\
&= 0, \text{ as required.}
\end{aligned}$$

To calculate $\text{Var}\{\mathcal{U}(\theta;\mathbf{X})\}$, observe

$$\begin{aligned}
\frac{\partial^2\ell(\theta;\mathbf{x})}{\partial\theta^2} &= \frac{\partial}{\partial\theta}\left\{\frac{\frac{\partial}{\partial\theta}f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)}\right\} \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(\mathbf{x};\theta)f(\mathbf{x};\theta)-\left(\frac{\partial}{\partial\theta}f(\mathbf{x};\theta)\right)^2}{[f(\mathbf{x};\theta)]^2} \\
&= \frac{\frac{\partial^2}{\partial\theta^2}f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)}-\mathcal{U}^2(\theta;\mathbf{x}) \\
\implies \mathcal{U}^2(\theta;\mathbf{x}) &= \frac{\frac{\partial^2}{\partial\theta^2}f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)}-\frac{\partial^2\ell}{\partial\theta^2} \\
\implies \text{Var}\{\mathcal{U}(\theta;\mathbf{X})\} &= E\{\mathcal{U}^2(\theta;\mathbf{X})\} \qquad (\text{as } E(\mathcal{U})=0) \\
&= E\left(\frac{\frac{\partial^2}{\partial\theta^2}f(\mathbf{X};\theta)}{f(\mathbf{X};\theta)}\right)+\mathcal{I}(\theta) \qquad (\text{by definition of } \mathcal{I}(\theta)).
\end{aligned}$$

Finally observe that:

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

Hence, we have proved $\mathrm{Var}\{\mathcal{U}(\theta;\mathbf{X})\} = \mathcal{I}(\theta)$.

□

2.2.2 Cramer-Rao Lower Bound

Theorem. *2.2.2 (Cramer-Rao Lower Bound)*

If T is an unbiased estimator for θ, then $\mathrm{Var}(T) \geq \dfrac{1}{\mathcal{I}(\theta)}$.

Proof.

$$\begin{aligned}
\text{Observe } \operatorname{Cov}\{T(\mathbf{X}),\mathcal{U}(\theta;\mathbf{X})\} &= E\{T(\mathbf{X})\mathcal{U}(\theta;\mathbf{X})\} \\
&= \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} T(\mathbf{x})\mathcal{U}(\theta;\mathbf{x})f(\mathbf{x};\theta)dx_1\dots dx_n \\
&= \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} T(\mathbf{x})\frac{\frac{\partial}{\partial\theta}f(\mathbf{x};\theta)}{f(\mathbf{x};\theta)}f(\mathbf{x};\theta)dx_1\dots dx_n \\
&= \frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} T(\mathbf{x})f(\mathbf{x};\theta)dx_1\dots dx_n \\
&= \frac{\partial}{\partial\theta}E\{T(\mathbf{X})\} \\
&= \frac{\partial}{\partial\theta}\theta \\
&= 1.
\end{aligned}$$

To summarize, $\operatorname{Cov}(T,\mathcal{U}) = 1$.

Recall that $\operatorname{Cov}^2(T,\mathcal{U}) \leq \operatorname{Var}(T)\operatorname{Var}(\mathcal{U})$ [i.e. $|\rho| \leq 1$ and divide both sides by RHS]

$$\implies \operatorname{Var}(T) \geq \frac{\operatorname{Cov}^2(T,\mathcal{U})}{\operatorname{Var}(\mathcal{U})} = \frac{1}{\mathcal{I}(\theta)}, \text{ as required.}$$

□

Example

Suppose $X_1, X_2, \dots, X_n$ are *i.i.d.* Po(λ) RV's, and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We will prove that $\bar{X}$ is a MVUE for λ.

Proof.

(1) Recall if $X \sim$Po(λ), then $P(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$, $E(X) = \lambda$ and $\text{Var}(X) = \lambda$.
Hence, $E(\bar{X}) = \lambda$ and $\text{Var}(\bar{X}) = \dfrac{\lambda}{n}$. Hence $\bar{X}$ is unbiased for λ.

(2) To show that $\bar{X}$ is MVUE, we will show that $\text{Var}(\bar{X}) = \dfrac{1}{\mathcal{I}(\lambda)}$.
Step 1:
Log-likelihood is

$$\begin{aligned} \ell(\lambda; \mathbf{x}) &= \log P(\mathbf{x}; \lambda) \\ &= \log \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \\ \text{so } \ell(\lambda; \mathbf{x}) &= \log e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i} \prod_{i=1}^{n} \frac{1}{x_i!} \\ &= -n\lambda + \sum_{i=1}^{n} x_i \log\lambda - \log \prod_{i=1}^{n} x_i! \\ &= -n\lambda + n\bar{x}\log\lambda - \log \prod_{i=1}^{n} x_i!. \end{aligned}$$

Step 2:
Find $\dfrac{\partial^2 \ell}{\partial \lambda^2}$;

$$\frac{\partial \ell}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda} \implies \frac{\partial^2 \ell}{\partial \lambda^2} = \frac{-n\bar{x}}{\lambda^2}.$$

Step 3:

$$\begin{aligned} \mathcal{I}(\lambda) &= -E\left(\frac{\partial^2 \ell}{\partial \lambda^2}\right) \\ \implies \mathcal{I}(\lambda) &= E\left(\frac{n\bar{X}}{\lambda^2}\right) \\ &= \frac{n}{\lambda^2} E(\bar{X}) \\ &= \frac{n\lambda}{\lambda^2} \\ &= \frac{n}{\lambda}. \end{aligned}$$

(3) Finally, observe that $\text{Var}(\bar{X}) = \dfrac{\lambda}{n} = \dfrac{1}{\mathcal{I}(\lambda)}$.

By Theorem 2.2.2, any unbiased estimator T for λ must have

$$\begin{aligned} \text{Var}(T) &\geq \frac{1}{\mathcal{I}(\lambda)} = \frac{\lambda}{n} \\ \implies \bar{X} &\text{ is a MVUE.} \end{aligned}$$

□

Theorem. *2.2.3*

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

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

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

Proof. Recall from the proof of Theorem 2.2.2 that the bound arises from the inequality

$$\text{Cor}^2(T(\mathbf{X}), \mathcal{U}(\theta;\mathbf{X})) \leq 1,$$

where $\mathcal{U}(\theta;\mathbf{x}) = \dfrac{\partial \ell}{\partial \theta}$.

Moreover, it is easy to see that the Cramer-Rao Lower Bound (CRLB) can be achieved only when

$$\text{Cor}^2\{T(\mathbf{X}), \mathcal{U}(\theta;\mathbf{X})\} = 1$$

Hence the CRLB is achieved only if $\mathcal{U}$ is a linear function of T with probability 1 (correlation equals 1):

$$U = aT + b \qquad a, b \text{ constants} \implies \text{ can depend on } \theta \text{ but not } \mathbf{x}$$

$$\text{i.e. } \frac{\partial \ell}{\partial \theta} = \mathcal{U}(\theta; \mathbf{x}) = a(\theta)T(\mathbf{x}) + b(\theta) \text{ for all } \mathbf{x}.$$

Integrating wrt θ, we obtain:

$$\log f(\mathbf{x}; \theta) = \ell(\theta; \mathbf{x}) = A(\theta)T(\mathbf{x}) + B(\theta) + h(\mathbf{x}),$$

where $A'(\theta) = a(\theta),\ B'(\theta) = b(\theta)$.

Hence,

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

Finally to ensure that $E(T) = \theta$, recall that $E\{\mathcal{U}(\theta; \mathbf{X})\} = 0$ and observe that in this case,

$$\begin{aligned} E\{\mathcal{U}(\theta; \mathbf{X})\} &= E[A'(\theta)T(\mathbf{X}) + B'(\theta)] \\ &= A'(\theta)E[T(\mathbf{X})] + B'(\theta) = 0 \\ \implies E[T(\mathbf{X})] &= \frac{-B'(\theta)}{A'(\theta)}. \end{aligned}$$

Hence, in order that $T(\mathbf{X})$ be unbiased for θ, we must have $\dfrac{-B'(\theta)}{A'(\theta)} = \theta$.

□

2.2.3 Exponential families of distributions

Definition. 2.2.3

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

$$f(\mathbf{x}; \theta) = \exp\{A(\theta)t(\mathbf{x}) + B(\theta) + h(\mathbf{x})\}$$

for all $x \in \mathcal{D} \in \mathbb{R}$, 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/prob functions becomes

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

In this case, a minimum variance unbiased estimator that achieves the CRLB can be seen to be the function:

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

which is the MVUE for $E(T) = \dfrac{-B'(\theta)}{A'(\theta)}$.

Example (Poisson example revisited)

Observe that the Poisson probability function,

$$\begin{aligned} p(x;\lambda) &= \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \exp\{(\log\lambda)x - \lambda - \log x!\} \\ &= \exp\{A(\lambda)t(x) + B(\lambda) + h(x)\}, \end{aligned}$$

where

$$\begin{aligned} A(\lambda) &= \log\lambda \\ B(\lambda) &= -\lambda \\ t(x) &= x \\ h(x) &= -\log x!. \end{aligned}$$

$\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} t(x_i)$ is the MVUE for

$$\begin{aligned} \frac{-B'(\lambda)}{A'(\lambda)} &= \frac{-(-1)}{1/\lambda} \\ &= \lambda. \end{aligned}$$

Example (2)

Consider the Exp(λ) distribution. The PDF is

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x}, \qquad x > 0 \\ &= \exp\{-\lambda x + \log \lambda\}; \end{aligned}$$

$\implies$ f is an exponential family with

$$\begin{aligned} A(\lambda) &= -\lambda \\ B(\lambda) &= \log \lambda \\ t(x) &= x \\ h(x) &= 0 \end{aligned}$$

We can also check that $E(X) = \dfrac{-B'(\lambda)}{A'(\lambda)}$. In particular, we have seen previously $E(X) = \dfrac{1}{\lambda}$ for $X \sim \text{Exp}(\lambda)$.

Now observe that $A'(\lambda) = -1$, $B'(\lambda) = \dfrac{1}{\lambda}$

$$\implies \frac{-B'(\lambda)}{A'(\lambda)} = \frac{1}{\lambda}, \text{ as required.}$$

It also follows that if $x_1, x_2, \ldots, x_n$ are *i.i.d.* Exp(λ) observations, then $\bar{X} = \dfrac{1}{n}\sum_{i=1}^{n} t(x_i)$ is the MVUE for $\dfrac{1}{\lambda} = E(X)$.

2.2.4 Sufficient statistics

Definition. 2.2.4

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

Remarks

(1) We will see that sufficient statistics capture all of the information in the data $\mathbf{x}$ that is relevant to θ.

(2) If we consider vector-valued statistics, then this definition admits trivial examples, such as $\mathbf{s} = \mathbf{x}$

$$\text{since } P(\mathbf{X} = \mathbf{x}|\mathbf{S} = \mathbf{s}) \quad = \quad \begin{cases} 1 & \mathbf{x} = \mathbf{s} \\ \\ 0 & \text{otherwise} \end{cases}$$

which does not depend on θ.

Example

Suppose $x_1, x_2, \ldots, x_n$ are *i.i.d.* Bernoulli-θ and let $s = \sum_{i=1}^{n} x_i$. Then S is sufficient for θ.

Proof.

$$\begin{aligned} P(\mathbf{x}) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i}(1-\theta)^{\sum(1-x_i)} \\ &= \theta^{s}(1-\theta)^{n-s}. \end{aligned}$$

Next observe that $S \sim \mathrm{B}(n, \theta)$

$$\begin{aligned} \implies P(s) &= \binom{n}{s}\theta^{s}(1-\theta)^{n-s} \\ \implies P(\mathbf{x}|s) &= \frac{P(\{\mathbf{X} = \mathbf{x}\} \cap \{S = s\})}{P(S = s)} \\ &= \frac{P(\mathbf{X} = \mathbf{x})}{P(S = s)} \\ &= \begin{cases} \dfrac{1}{\binom{n}{s}} & \text{if } \sum_{i=1}^{n} x_i = s \\ \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

□

Theorem. *2.2.4 The Factorization Theorem*

Suppose $x_1, \ldots, x_n$ have joint PDF/prob function $f(\mathbf{x}; \theta)$. Then S is a sufficient statistic for θ if and only if

$$f(\mathbf{x}; \theta) = g(\mathbf{s}; \theta) h(\mathbf{x})$$

for some functions g, h.

Proof.

Omitted. □

Examples

(1) $x_1, x_2, \ldots, x_n$ *i.i.d.* $\mathrm{N}(\mu, \sigma^2)$, σ^2 known.

$$\begin{aligned}
\text{Then } f(\mathbf{x}; \mu) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{(-1/(2\sigma^2))(x_i - \mu)^2} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} x_i^2 - 2\mu \sum_{i=1}^{n} x_i + n\mu^2 \right) \right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{1}{2\sigma^2} \left(2\mu \sum_{i=1}^{n} x_i - n\mu^2 \right) \right\} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 \right) \\
&= \exp\left\{ \frac{1}{2\sigma^2} \left(2\mu \sum_{i=1}^{n} x_i - n\mu^2 \right) \right\} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 - \frac{n}{2} \log(2\pi\sigma^2) \right\} \\
\implies S &= \sum_{i=1}^{n} x_i \text{ is sufficient for } \mu.
\end{aligned}$$

(2) If $x_1, x_2, \ldots, x_n$ are *i.i.d.* with

$$\begin{aligned}
f(x) &= \exp\{A(\theta) t(x) + B(\theta) + h(x)\}, \\
\text{then } f(\mathbf{x}; \theta) &= \exp\left\{ A(\theta) \sum_{i=1}^{n} t(x_i) + nB(\theta) \right\} \exp\left\{ \sum_{i=1}^{n} h(x_i) \right\} \\
\implies S &= \sum_{i=1}^{n} t(x_i)
\end{aligned}$$

is sufficient for θ in the exponential family by the Factorization Theorem.

2.2.5 The Rao-Blackwell Theorem

Theorem. *2.2.5 Rao-Blackwell Theorem*

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 $\mathrm{Var}(T^*) \leq \mathrm{Var}(T)$. Moreover, $\mathrm{Var}(T^*) = \mathrm{Var}(T)$ iff $T^* = T$ with probability 1.

Proof. (I) Unbiasedness:

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

$$\implies E(T^*) = \theta.$$

(II) Variance inequality:

$$\begin{aligned} \mathrm{Var}(T) &= E\{\mathrm{Var}(T|S)\} + \mathrm{Var}\{E(T|S)\} \\ &= E\{\mathrm{Var}(T|S)\} + \mathrm{Var}(T^*) \\ &\geq \mathrm{Var}(T*), \end{aligned}$$

since $E\{\mathrm{Var}(T|S)\} \geq 0$.

Observe also that $\mathrm{Var}(T) = \mathrm{Var}(T^*)$

$$\begin{aligned} &\implies E\{\mathrm{Var}(T|S)\} = 0 \\ &\implies \mathrm{Var}(T|S) = 0 \text{ with prob. } 1 \\ &\implies T = E(T|S) \text{ with prob. } 1. \end{aligned}$$

(III) T^* is an estimator:

Since S is sufficient for θ,

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

which does not depend on θ.

□

Remarks

(1) The Rao-Blackwell Theorem can be used occasionally to construct estimators.

(2) The theoretical importance of this result is to observe that T^* will always depend on $\mathbf{x}$ only through S. If T is already a MVUE then $T^* = T \implies$ MVUE depends on $\mathbf{x}$ only through S.

Example
Suppose $x_1, \ldots, x_n \sim i.i.d.$ $\mathrm{N}(\mu, \sigma^2)$, σ^2 known. We want to estimate μ. Take $T = x_1$, as an unbiased estimator for μ.

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

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

If $\mathbf{x} = (x_1, \ldots, x_n)^T$ for $\mathbf{X} \sim N_n(\mu\mathbf{1}, \sigma^2 I)$, then $\begin{pmatrix} T \\ S \end{pmatrix} = A\mathbf{x}$, where A is a matrix

$$A = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ & & & \\ 1 & 1 & \ldots & 1 \end{pmatrix}$$

Hence,

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

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

$$= \; N\left(\frac{1}{n}s, \left(1 - \frac{1}{n}\right)\sigma^2\right)$$

$$\therefore\; E(T|S) = \frac{1}{n}\sum x_i = \bar{x} = T^*$$

is a better (or equal) estimator for μ.

Finally, observe $\mathrm{Var}(\bar{X}) = \dfrac{\sigma^2}{n} \leq \sigma^2 = \mathrm{Var}(X_1)$ with $<$ for $n > 1$, $\sigma^2 > 0$.

Remarks

(1) We have given only an incomplete outline of theory.

(2) There are also concepts of minimal sufficient & complete statistics to be considered.

2.3 Methods Of Estimation

2.3.1 Method Of Moments

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

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

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

Example

$X_1, \ldots, X_n \sim$ *i.i.d.* $\text{Exp}(\lambda) \implies E(X_i) = \dfrac{1}{\lambda}$; the method of moments estimator is defined as the solution to the equation

$$\bar{X} = \frac{1}{\tilde{\lambda}} \implies \tilde{\lambda} = \frac{1}{\bar{X}}.$$

Remark

The method of moments is appealing:

(1) it is simple;

(2) rationale is that $\bar{X}$ is BLUE for μ.

If $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$, the method of moments can be adapted as follows:

(1) let $\mu_k = \mu_k(\boldsymbol{\theta}) = E(X^k), \quad k = 1, \ldots, p$;

(2) let $m_k = \dfrac{1}{n}\sum_{i=1}^{n} x_i^k$.

The MoM estimator $\tilde{\boldsymbol{\theta}}$ is defined to be the solution to the system of equations

$$\begin{aligned} m_1 &= \mu_1(\tilde{\boldsymbol{\theta}}) \\ m_2 &= \mu_2(\tilde{\boldsymbol{\theta}}) \\ \vdots &\quad \vdots \\ m_p &= \mu_p(\tilde{\boldsymbol{\theta}}) \end{aligned}$$

Example

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

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

$$\begin{aligned} \mu_1(\boldsymbol{\theta}) &= E(X) = \mu \\ \mu_2(\boldsymbol{\theta}) &= E(X^2) = \sigma^2 + \mu^2 \end{aligned}$$

$$\begin{aligned} m_1 &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ m_2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{aligned}$$

$\therefore$ we need to solve for $\tilde{\mu}, \tilde{\sigma}^2$:

$$\tilde{\mu} = \bar{x}$$

$$\begin{aligned} \tilde{\sigma}^2 + \tilde{\mu}^2 = \frac{1}{n}\sum x_i^2 \implies \tilde{\sigma}^2 &= \frac{1}{n}\sum x_i^2 - \bar{x}^2 \\ &= \frac{1}{n}\sum (x_i - \bar{x})^2 \\ &= \frac{n-1}{n} s^2. \end{aligned}$$

Remark

The Method of Moments estimators can be seen to have good statistical properties. Under fairly mild regulatory conditions, the MoM estimator is

(1) Consistent, i.e., $\tilde{\theta} \to \theta$ in prob as $n \to \infty$, for all θ.

(2) Asymptotically Normal, i.e., $\dfrac{\tilde{\theta} - \theta}{\sqrt{\mathrm{Var}(\theta)}}. \xrightarrow[\mathcal{D}]{} \mathrm{N}(0,1)$ as $n \to \infty$.

However, the MoM estimator has one serious defect.

Suppose $X_1, \ldots, X_n$ is a random sample from F_{θ_X} & let $\tilde{\theta}_{\mathbf{X}}$ be MoM estimator.

Let $Y = h(X)$ for some invertible function $h(X)$, then $Y_1, \ldots, Y_n$ should contain the same information as $X_1, \ldots, X_n$.

$\therefore$ we would like $\tilde{\theta}_Y \equiv \tilde{\theta}_X$. Unfortunately this does <u>not</u> hold.

Example

$X_1, \ldots, X_n$ *i.i.d.* Exp(λ). We saw previously that $\tilde{\lambda}_X = \dfrac{1}{\bar{X}}$.

Suppose $Y_i = X_i^2$ (which is invertible for $X_i > 0$). To obtain $\tilde{\lambda}_Y$, observe $E(Y) = E(X^2) = \dfrac{2}{\lambda^2}$

$$\implies \tilde{\lambda}_Y = \sqrt{\frac{2n}{\sum X_i^2}} \neq \frac{1}{\bar{X}}.$$

2.3.2 Maximum Likelihood Estimation

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

Definition. The maximum likelihood estimate $\hat{\theta}$ is the solution to the problem $\max\limits_{\theta \in \Theta} \ell(\theta; \mathbf{x})$ i.e. $\hat{\theta} = \arg\max\limits_{\theta \in \Theta} \ell(\theta; \mathbf{x})$.

Remark

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

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

Example

If $X_1, X_2, \ldots, X_n$ are *i.i.d.* geometric-θ RV's, find $\hat{\theta}$.

Solution:

$$\begin{aligned}
\ell(\theta; \mathbf{x}) &= \log\left\{\prod_{i=1}^{n} \theta(1-\theta)^{x_i - 1}\right\} \\
&= \sum_{i=1}^{n} \log\theta + \sum_{i=1}^{n} (x_i - 1)\log(1-\theta) \\
&= n\{\log\theta + (\bar{x} - 1)\log(1-\theta)\} \\
\therefore \mathcal{U}(\theta; \mathbf{x}) &= \frac{\partial \ell}{\partial \theta} = n\left\{\frac{1}{\theta} - \frac{\bar{x} - 1}{1-\theta}\right\} \\
&= n\left(\frac{1 - \theta\bar{x}}{\theta(1-\theta)}\right).
\end{aligned}$$

To find $\hat{\theta}$, we solve for θ in $1 - \theta\bar{x} = 0$

$$\implies \hat{\theta} = \frac{1}{\bar{x}} \qquad \left[= \frac{n}{\sum_{i=1}^{n} x_i} = \frac{\#\text{ of successes}}{\#\text{ of trials}} \right]$$

2.3.3 Elementary properties of MLEs

(1) MLE's are invariant under invertible transformations of the data.

Proof. (Continuous *i.i.d.* RV's, strictly increasing/decreasing translation)
Suppose X has PDF $f(x;\theta)$ and $Y = h(X)$, where h is strictly monotonic;

$$\implies f_Y(y;\theta) = f_X(h^{-1}(y);\theta)|h^{-1}(y)'| \qquad \text{when } y = h(x).$$

Consider data $x_1, x_2, \ldots, x_n$ and the transformed version $y_1, y_2, \ldots, y_n$:

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

since $\log\left(\prod_{i=1}^{n}|h^{-1}(y_i)'|\right)$ does not depend on θ, it follows that $\hat{\theta}$ maximizes $\ell_X(\theta;\mathbf{x})$ iff it maximizes $\ell_Y(\theta;\mathbf{y})$.

□

(2) If $\phi = \phi(\theta)$ is a 1-1 transformation of θ, then the MLE's obey the transformation rule, $\hat{\phi} = \phi(\hat{\theta})$.

Proof. It can be checked that if $\ell_\phi(\phi;\mathbf{x})$ is the log-likelihood with respect to ϕ, then

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

It follows that $\hat{\theta}$ maximizes $\ell_\theta(\theta;\mathbf{x})$ iff $\hat{\phi} = \phi(\hat{\theta})$ maximizes $\ell_\phi(\phi;\mathbf{x})$.

□

(3) If $T(\mathbf{x})$ is a sufficient statistic for θ, then $\hat{\theta}$ depends on the data only as a function of $t(\mathbf{x})$.

Proof. By the Factorization Theorem, $T(\mathbf{x})$ is sufficient for θ iff

$$\begin{aligned} f(\mathbf{x};\theta) &= g(t(\mathbf{x});\theta)h(\mathbf{x}) \\ \implies \ell(\theta;\mathbf{x}) &= \log g(t(\mathbf{x});\theta) + \log h(\mathbf{x}) \end{aligned}$$

$\implies \hat{\theta}$ maximizes $\ell(\theta;\mathbf{x}) \Leftrightarrow \hat{\theta}$ maximizes $\log g(t(\mathbf{x});\theta)$

$\implies \hat{\theta}$ is a function of $T(\mathbf{x})$.

□

Example

Suppose $X_1, X_2, \ldots, X_n$ are *i.i.d.* Exp(λ); then

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

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

$$\begin{aligned} \mathcal{U}(\lambda,\mathbf{x}) = \frac{\partial \ell}{\partial \lambda} &= \frac{\partial}{\partial \lambda}(n\log\lambda - n\lambda\bar{x}) \\ &= \frac{n}{\lambda} - n\bar{x}; \end{aligned}$$

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

Note: as proved $\hat{\lambda}$ is a function of the sufficient statistic $\bar{x}$.

Let $Y_1, Y_2, \ldots, Y_n$ be defined by $Y_i = \log X_i$.

If $X \sim$Exp(λ) and $Y = \log X$, then we can find $f_Y(y)$ by taking $h(x) = \log x$ and using

$$\begin{aligned}
f_Y(y) &= f_X(h^{-1}(y))|h^{-1}(y)'| \qquad [h^{-1}(y) = e^y,\ h^{-1}(y)' = e^y] \\
&= \lambda e^{-\lambda e^y} e^y \\
\implies \ell_Y(\lambda; \mathbf{y}) &= \log\left\{\prod_{i=1}^{n} \lambda e^{-\lambda e^{y_i}} e^{y_i}\right\} \\
&= \log\left\{\lambda^n e^{-\lambda \sum_{i=1}^n e^{y_i}} e^{\sum_{i=1}^n y_i}\right\} \\
&= n\log\lambda - \lambda\sum_{i=1}^{n} e^{y_i} + \sum_{i=1}^{n} y_i \\
\implies \frac{\partial}{\partial\lambda}\ell_Y(\lambda; \mathbf{y}) &= \frac{n}{\lambda} - \sum_{i=1}^{n} e^{y_i} = 0 \\
\implies \hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} e^{y_i}} \\
&= \frac{n}{\sum_{i=1}^{n} e^{\log x_i}} \\
&= \frac{n}{\sum_{i=1}^{n} x_i} \\
&= \frac{1}{\bar{x}} \qquad \left(\frac{n}{n\bar{x}}\right).
\end{aligned}$$

Finally, suppose we take $\theta = \log\lambda \implies \lambda = e^\theta$

$$\begin{aligned}
\implies f(\mathbf{x}; \theta) &= e^\theta e^{-e^\theta x} \qquad (\lambda e^{-\lambda x}) \\
\implies \ell_\theta(\theta; \mathbf{x}) &= \log\left(\prod_{i=1}^{n} e^\theta e^{-e^\theta x_i}\right) \\
&= \log\left\{e^{n\theta} e^{-e^\theta n\bar{x}}\right\} \\
&= n\theta - e^\theta n\bar{x}.
\end{aligned}$$

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= n - n\bar{x}e^{\theta} \\ \therefore \frac{\partial \ell}{\partial \theta} = 0 &\implies 1 = \bar{x}e^{\theta} \implies e^{\theta} = \frac{1}{\bar{x}} \\ &\implies \hat{\theta} = -\log \bar{x}. \end{aligned}$$

But $\hat{\lambda} = \dfrac{1}{\bar{x}} \implies \log \hat{\lambda} = \log\left(\dfrac{1}{\bar{x}}\right) = -\log \bar{x} = \hat{\theta}$, as required.

Remark: (Not examinable)

Maximum likelihood estimation and method of moments can both be generated by the use of estimating functions.

An estimating function is a function, $H(\mathbf{x}; \theta)$, with the property $E\{H(\mathbf{x}; \theta)\} = 0$.

H can be used to define an estimator $\tilde{\theta}$ which is a solution to the equation

$$H(\mathbf{x}; \theta) = 0.$$

For method of moments estimates, we can take

$$H(\mathbf{x}; \theta) = \bar{x} - E(X) \qquad [E(\bar{X}) \text{ if not } \textit{i.i.d.} \text{ case}].$$

To calculate the MLE:

(1) find the log-likelihood, then

(2) calculate the score & let it equal 0.

For maximum likelihood, we can use

$$H(\mathbf{x}; \theta) = \mathcal{U}(\theta; \mathbf{x}) = \frac{\partial \ell}{\partial \theta}.$$

Recall we showed previously that $E\{\mathcal{U}(\theta; \mathbf{x})\} = 0$.

2.3.4 Asymptotic Properties of MLEs

Suppose $X_1, X_2, X_3, \ldots$ is a sequence of *i.i.d.* RV's 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 $x \implies \theta_1 = \theta_2$.

We will show (in outline) that if $\hat{\theta}_n$ is the MLE (maximum likelihood estimator) based on $X_1, X_2, \ldots, X_n$, then:

(1) $\hat{\theta}_n \to \theta_0$ in probability, i.e., for each $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\hat{\theta_n} - \theta_0| > \epsilon) = 0.$$

(2)

$$\sqrt{ni(\theta_0)}(\hat{\theta}_n - \theta_0) \xrightarrow[\mathcal{D}]{} \mathrm{N}(0,1) \text{ as } n \to \infty$$

(Asymptotic Normality).

Remark

The practical use of asymptotic normality is that for large n,

$$\hat{\theta} \overset{\cdot}{\sim} \mathrm{N}\left(\theta_0, \frac{1}{ni(\theta_0)}\right).$$

Outline of consistency & asymptotic normality for MLEs:

Consider *i.i.d.* data $X_1, X_2, \ldots$ with common PDF/prob. function $f(x;\theta)$.

If $\hat{\theta}_n$ is the MLE based on $X_1, X_2, \ldots, X_n$, we will show (in outline) that $\hat{\theta}_n \to \theta_0$ in probability as $n \to \infty$, where θ_0 is the true value for θ.

Lemma. 2.3.1

Suppose f is such that $f(x;\theta_1) = f(x;\theta_2)$ for all $x \implies \theta_1 = \theta_2$. Then $\ell^*(\theta) = E\{\log f(X;\theta)\}$ is maximized uniquely by $\theta = \theta_0$.

Proof.

$$\begin{aligned}
\ell^*(\theta) - \ell^*(\theta_0) &= \int_{-\infty}^{\infty} (\log f(x;\theta)) f(x;\theta_0) dx - \int_{-\infty}^{\infty} (\log f(x;\theta_0)) f(x;\theta_0) dx \\
&= \int_{-\infty}^{\infty} \log\left(\frac{f(x;\theta)}{f(x;\theta_0)}\right) f(x;\theta_0) dx \\
&\leq \int_{-\infty}^{\infty} \left(\frac{f(x;\theta)}{f(x;\theta_0)} - 1\right) f(x;\theta_0) dx \\
&= \int_{-\infty}^{\infty} (f(x;\theta) - f(x;\theta_0)) dx \\
&= \int_{-\infty}^{\infty} f(x;\theta) dx - \int_{-\infty}^{\infty} f(x;\theta_0) dx = 0
\end{aligned}$$

Moreover, equality is achieved iff

$$\begin{aligned}
\frac{f(x;\theta)}{f(x;\theta_0)} &= 1 \qquad \text{for all } x \\
\implies f(x;\theta) &= f(x;\theta_0) \qquad \text{for all } x \\
\implies \theta &= \theta_0.
\end{aligned}$$

□

Lemma. 2.3.2

Let $\bar{\ell}_n(\theta;\mathbf{x}) = \frac{1}{n}\ell(\theta, x_1, \ldots, x_n)$, i.e., $\frac{1}{n}\times$ log likelihood based on $x_1, x_2, \ldots, x_n$.

Then for each θ, $\bar{\ell}_n(\theta;\mathbf{x}) \to \ell^*(\theta)$ in probability as $n \to \infty$.

Proof. Observe that

$$\begin{aligned}
\bar{\ell}_n(\theta;\mathbf{x}) &= \frac{1}{n}\log\prod_{i=1}^{n} f(x_i;\theta) \\
&= \frac{1}{n}\sum_{i=1}^{n} \log f(x_i;\theta) \\
&= \frac{1}{n}\sum_{i=1}^{n} L_i(\theta), \qquad \text{where } L_i(\theta) = \log f(x_i;\theta).
\end{aligned}$$

Since $L_i(\theta)$ are *i.i.d.* with $E(L_i(\theta)) = \ell^*(\theta)$, it follows by the weak law of large numbers that $\bar{\ell}_n(\theta;\mathbf{x}) \to \ell^*(\theta)$ in probability as $n \to \infty$.

□

To summarise, we have proved that:

(1) $\bar{\ell}_n(\theta, \mathbf{x}) \to \ell^*(\theta)$;

(2) $\ell^*(\theta)$ is maximized when $\theta = \theta_0$.

Since $\hat{\theta}_n$ maximizes $\bar{\ell}_n(\theta, \mathbf{x})$, it follows that $\hat{\theta}_n \to \theta_0$ in probability.

Theorem. *2.3.1*

Under the above assumptions, let $\mathcal{U}_n(\theta; \mathbf{x})$ *denote the score based on* $X_1, \ldots, X_n$.

Then,

$$\frac{\mathcal{U}_n(\theta_0; \mathbf{x})}{\sqrt{ni(\theta_0)}} \xrightarrow[\mathcal{D}]{} N(0, 1).$$

Proof.

$$\begin{aligned}
\mathcal{U}_n(\theta_0; \mathbf{x}) &= \frac{\partial}{\partial \theta} \log \prod_{i=1}^{n} f(x_i; \theta)|_{\theta=\theta_0} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(x_i; \theta)|_{\theta=\theta_0} \\
&= \sum_{i=1}^{n} \mathcal{U}_i, \qquad \text{where } \mathcal{U}_i = \frac{\partial}{\partial \theta} \log f(x_i; \theta)|_{\theta=\theta_0}.
\end{aligned}$$

Since $\mathcal{U}_1, \mathcal{U}_2, \ldots$ are *i.i.d.* with $E(\mathcal{U}_i) = 0$ and $\mathrm{Var}(\mathcal{U}_i) = i(\theta)$, by the Central Limit Theorem,

$$\frac{\sum_{i=1}^{n} \mathcal{U}_i - nE(\mathcal{U})}{\sqrt{n \,\mathrm{Var}(\mathcal{U})}} = \frac{\mathcal{U}_n(\theta_0; \mathbf{x})}{\sqrt{ni(\theta_0)}} \xrightarrow[\mathcal{D}]{} \mathrm{N}(0, 1).$$

□

Theorem. *2.3.2*

Under the above assumptions,

$$\sqrt{ni(\theta_0)}(\hat{\theta}_n - \theta_0) \xrightarrow[\mathcal{D}]{} N(0, 1).$$

Proof. Consider the first order Taylor expansion of $\mathcal{U}_n(\hat{\theta}; \mathbf{x})$ about θ_0:

$$\mathcal{U}_n(\hat{\theta}_n; \mathbf{x}) \approx \mathcal{U}_n(\theta_0; \mathbf{x}) + \mathcal{U}_n'(\theta_0; \mathbf{x})(\hat{\theta}_n - \theta_0).$$

For large n

$$\implies \mathcal{U}_n(\theta_0; \mathbf{x}) \approx -\mathcal{U}_n'(\theta_0; \mathbf{x})(\hat{\theta}_n - \theta_0)$$

$$\frac{\mathcal{U}_n(\theta_0; \mathbf{x})}{\sqrt{ni(\theta_0)}} \xrightarrow[\mathcal{D}]{} \mathrm{N}(0, 1)$$

$$\implies \frac{-\mathcal{U}_n'(\theta_0; \mathbf{x})(\hat{\theta}_n - \theta_0)}{\sqrt{ni(\theta_0)}} \xrightarrow[\mathcal{D}]{} \mathrm{N}(0, 1).$$

Now observe that:

$$\frac{-\mathcal{U}_n'(\theta_0; \mathbf{x})}{n} \to i(\theta_0) \qquad \text{as } n \to \infty$$

by the weak law of large numbers, since:

$$\frac{-\mathcal{U}_n'(\theta_0; \mathbf{x})}{n} = \frac{1}{n}\sum_{i=1}^{n} -\frac{\partial^2}{\partial\theta^2} \log f(x_i; \theta|_{\theta=\theta_0})$$

$$\implies \frac{-\mathcal{U}_n'(\theta_0; \mathbf{x})}{ni(\theta_0)} \to 1 \qquad \text{in probability.}$$

$$\text{Hence,} \quad \frac{-\mathcal{U}_n'(\theta_0; \mathbf{x})}{\sqrt{ni(\theta_0)}}(\hat{\theta}_n - \theta_0) \approx \sqrt{ni(\theta_0)}(\hat{\theta}_n - \theta_0) \qquad \text{for large } n$$

$$\implies \sqrt{ni(\theta_0)}(\hat{\theta}_n - \theta_0) \xrightarrow[\mathcal{D}]{} \mathrm{N}(0, 1).$$

□

Remark"
The preceding theory can be generalized to include vector-valued coefficients. We will not discuss the details.

2.4 Hypothesis Tests and Confidence Intervals

Motivating example:

Suppose $X_1, X_2, \ldots, X_n$ are *i.i.d.* $\mathrm{N}(\mu, \sigma^2)$, 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.

2.4.1 Hypothesis testing

In general, consider a statistical problem with parameter

$$\theta \in \Theta_0 \cup \Theta_A, \qquad \text{where } \Theta_0 \cap \Theta_A = \phi.$$

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:

		Actual Status	
		H_0 true	H_A true
Test Result	Accept H_0	$\surd$	type II error (β)
	Reject H_0	type I error (α)	$\surd$

We would like both the type I and type II error rates to be as small as possible. However, these results conflict with each other. To reduce the type I error rate we need to "make it harder to reject H_0". To reduce the type II error rate we need to "make it easier to reject H_0".

The standard (Neyman-Pearson) approach to hypothesis testing is to control the type I error rate at a "small" value α and then use a test that makes the type II error as small as possible.

The equivalence between the confidence intervals and hypothesis tests can be formulated as follows: Recall that 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.

$\alpha = P(\text{reject } H_0 | H_0 \text{ true})$
$\beta = P(\text{retain } H_0 | H_A \text{ true})$
$1 - \beta = \text{power} = P(\text{reject } H_0 | H_A \text{ true})$, which is what we want.

Conversely, given a hypothesis test $H_0 : \theta = \theta_0$ with significance level α, it can be proved that the set $\{\theta_0 : H_0 : \theta = \theta_0 \text{ is accepted}\}$ is a $100(1-\alpha)\%$ confidence region for θ.

2.4.2 Large sample tests and confidence intervals

Consider a statistical problem with data $x_1, \ldots, x_n$, log-likelihood $\ell(\theta; \mathbf{x})$, score $\mathcal{U}(\theta; \mathbf{x})$ and information $i(\theta)$.

Consider also a hypothesis $H_0 : \theta = \theta_0$. The following three tests are often considered:

(1) <u>The Wald Test:</u>

Test Statistic: $W = \sqrt{ni(\hat{\theta})}(\hat{\theta} - \theta_0)$

Critical Region: reject for $|W| \geq z(\alpha/2)$

(2) <u>The Score Test:</u>

Test Statistic: $V = \dfrac{\mathcal{U}(\theta_0; \mathbf{x})}{\sqrt{ni(\theta_0)}}$

Critical Region: reject for $|V| \geq z(\alpha/2)$

(3) <u>Likelihood-Ratio Test:</u>

Test Statistic: $G^2 = 2(\ell(\hat{\theta}) - \ell(\theta_0))$

Critical Region: reject for $G^2 \geq \chi^2_{1,\alpha}$

Example:

Suppose $X_1, X_2, \ldots, X_n$ *i.i.d.* $\text{Po}(\lambda)$, and consider $H_0 : \lambda = \lambda_0$.

$$\begin{aligned} \ell(\lambda; \mathbf{x}) &= \log \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} \\ &= n(\bar{x} \log \lambda - \lambda) - \log \prod_{i=1}^{n} x_i! \end{aligned}$$

$$\mathcal{U}(\lambda;\mathbf{x}) = \frac{\partial \ell}{\partial \lambda} = n\left(\frac{\bar{x}}{\lambda} - 1\right) = \frac{n(\bar{x}-\lambda)}{\lambda} \implies \hat{\lambda} = \bar{x}$$

$$\mathcal{I}(\lambda) = ni(\lambda) = E\left(-\frac{\partial^2 \ell}{\partial \lambda^2}\right) = E\left(\frac{n\bar{x}}{\lambda^2}\right) = \frac{n}{\lambda}$$

$$\begin{aligned}
\implies W &= \sqrt{ni(\hat{\lambda})}(\hat{\lambda} - \lambda_0) \\
&= \frac{\hat{\lambda} - \lambda_0}{\sqrt{\hat{\lambda}/n}} \\
V &= \frac{\mathcal{U}(\lambda_0;\mathbf{x})}{\sqrt{ni(\lambda_0)}} \\
&= \frac{n(\bar{x}-\lambda_0)}{\lambda_0(\sqrt{n/\lambda_0})} \\
&= \frac{\bar{x}-\lambda_0}{\sqrt{\lambda_0/n}} \\
G^2 &= 2(\ell(\hat{\lambda}) - \ell(\lambda_0)) \\
&= 2n(\bar{x}\log\hat{\lambda} - \hat{\lambda}) - 2n(\bar{x}\log\lambda_0 - \lambda_0) \\
&= 2n\left(\bar{x}\log\frac{\hat{\lambda}}{\lambda_0} - (\hat{\lambda} - \lambda_0)\right).
\end{aligned}$$

Remarks

(1) It can be proved that the tests based on W, V, G^2 are asymptotically equivalent for H_0 true.

(2) As a by-product, it follows that the null distribution of G^2 is χ^2_1. (Recall from Theorems 2.3.1, 2.3.1 that the null distribution for W, V are both $\mathrm{N}(0,1)$).

(3) To understand the motivation for the three tests, it is useful to consider their relation to the log-likelihood function. See Figure 20.

We have introduced the Wald test, score test and the likelihood test for $H_0 : \theta = \theta_0$ vs. $H_A : \theta = \theta_a$. These are large-sample tests in that asymptotic distributions for the test statistic under H_0 are available. It can also be proved that the LR statistic and the

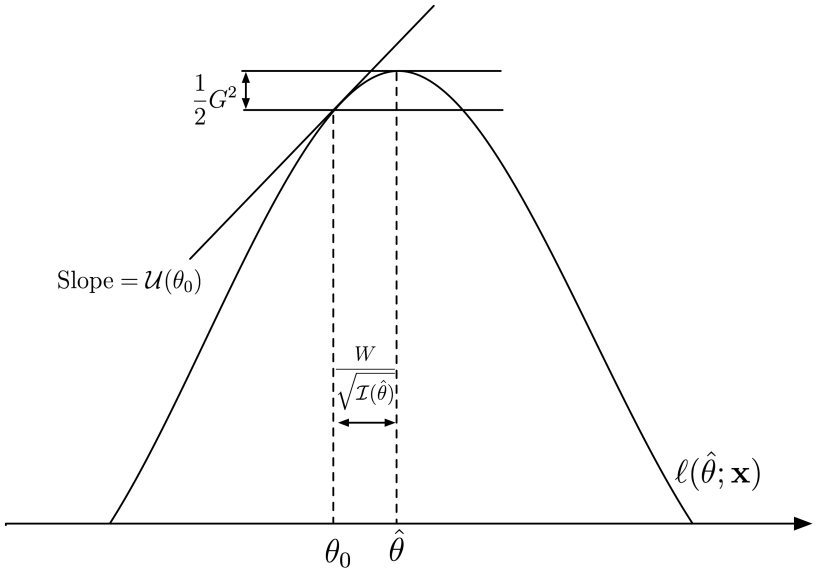


Figure 20: Relationships of the tests to the log-likelihood function

score test statistic are invariant under transformation of the parameter, but the Wald test is not.

Each of these three tests can be inverted to give a confidence interval (region) for θ:

Wald Test
Solve for θ_0 in $W^2 \leq z(\alpha/2)^2$.

$$\begin{aligned} \text{Recall,} \qquad W &= \sqrt{ni(\hat{\theta})}(\hat{\theta} - \theta_0) \\ \implies \hat{\theta} - \frac{z(\alpha/2)}{\sqrt{ni(\hat{\theta})}} &\leq \theta_0 \leq \hat{\theta} + \frac{z(\alpha/2)}{\sqrt{ni(\hat{\theta})}} \\ \text{i.e.,} \qquad \hat{\theta} &\pm \frac{z(\alpha/2)}{\sqrt{ni(\hat{\theta})}}. \end{aligned}$$

Score Test

Need to solve for θ_0 in $V^2 = \left(\frac{\mathcal{U}(\theta_0; \mathbf{x})}{\sqrt{ni(\theta_0)}} \right)^2 \leq z(\alpha/2)^2$.

LR Test
Solve for θ_0 in $2(\ell(\hat{\theta}; \mathbf{x}) - \ell(\theta_0; \mathbf{x})) \leq \chi^2_{1,\alpha} = z(\alpha/2)^2$.

Example:

$X_1, \ldots, X_n$ *i.i.d.* $\text{Po}(\lambda)$.

Recall Wald statistic is $W = \dfrac{\hat{\lambda} - \lambda_0}{\sqrt{\hat{\lambda}/n}}$

$$\begin{aligned} &\implies \hat{\lambda} \pm z(\alpha/2)\sqrt{\hat{\lambda}/n} \\ &\iff \bar{x} \pm z(\alpha/2)\sqrt{\bar{x}/n}. \end{aligned}$$

Score test:
Recall that the test statistic is $V = \dfrac{\bar{x} - \lambda_0}{\sqrt{\lambda_0/n}}$. Hence to find a confidence interval, we need to solve for λ_0 in the equation

$$\begin{aligned} V^2 &\leq z(\alpha/2)^2 \\ \implies \frac{(\bar{x} - \lambda_0)^2}{\lambda_0/n} &\leq z(\alpha/2)^2 \\ \implies (\bar{x} - \lambda_0)^2 &\leq \frac{\lambda_0 z(\alpha/2)^2}{n} \\ \implies \lambda_0^2 - \left(2\bar{x} + \frac{z(\alpha/2)^2}{n}\right)\lambda + \bar{x}^2 &\leq 0, \end{aligned}$$

which is a quadratic in λ_0 and hence can be solved:

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

For the LR test, we have

$$G^2 = 2n\left\{\bar{x}\log\frac{\bar{x}}{\lambda_0} - (\bar{x} - \lambda_0)\right\}$$

$\rightarrow$ can solve numerically for λ_0 in $G^2 \leq z(\alpha/2)^2$. See Figure 21.

2.4.3 Optimal tests

Consider simple null and alternative hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{and} \quad H_a : \theta = \theta_a$$

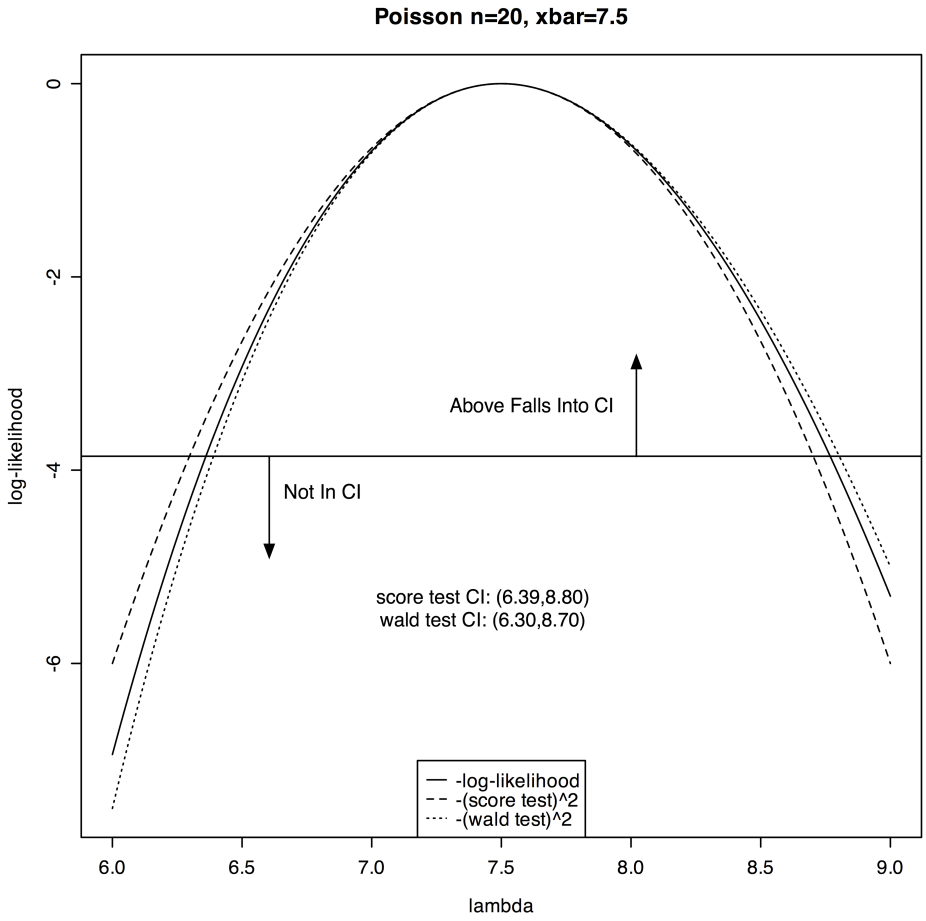


Figure 21: The 3 tests

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 $1 - \beta = P(\text{reject } H_0 | H_0 \text{ false}) \rightarrow$ so we want high power.

Theorem. *2.4.1 (Neyman-Pearson Lemma)*

Consider the test $H_0 : \theta = \theta_0$ *vs.* $H_a : \theta = \theta_a$ *defined by the rule*

$$\textit{reject } H_0 \textit{ for } \frac{f(\mathbf{x}; \theta_0)}{f(\mathbf{x}; \theta_a)} \leq k,$$

for some constant k.

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 $\mathcal{C}$ be the critical region for the LR test and let $\mathcal{D}$ be the critical region (RR) for any other test.

Let $\mathcal{C}_1 = \mathcal{C} \cap \mathcal{D}$ and $\mathcal{C}_2, \mathcal{D}_2$ be such that

$$\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \quad , \quad \mathcal{C}_1 \cap \mathcal{C}_2 = \phi$$

$$\mathcal{D} = \mathcal{C}_1 \cup \mathcal{D}_2 \quad , \quad \mathcal{C}_1 \cap \mathcal{D}_2 = \phi$$

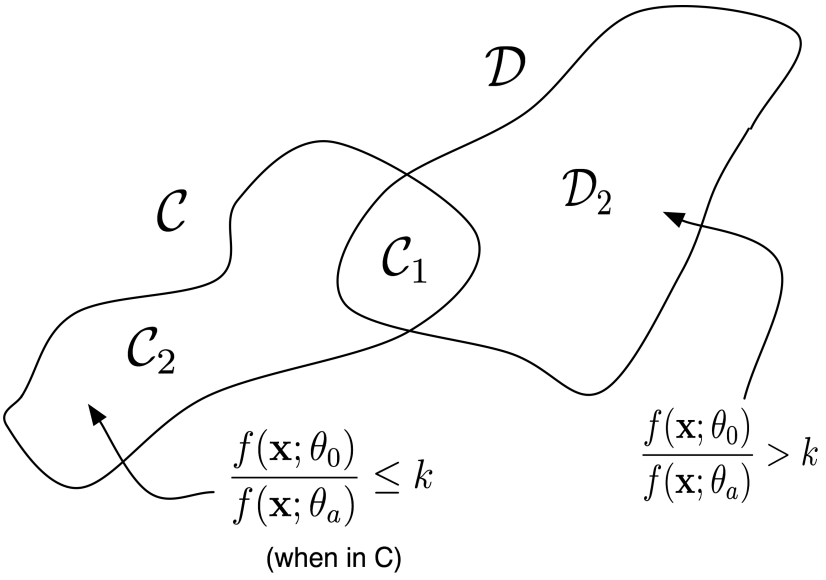


Figure 22: Neyman-Pearson Lemma

See Figure 22 and observe $\alpha \leq \alpha^*$

$$\begin{aligned}
\implies \int_{\mathcal{D}} f(\mathbf{x};\theta_0)dx_1 \ldots dx_n &\leq \int_{\mathcal{C}} f(\mathbf{x};\theta_0)dx_1 \ldots dx_n \\
\implies \int_{\mathcal{D}_2} f(\mathbf{x};\theta_0)dx_1 \ldots dx_n &\leq \int_{\mathcal{C}_2} f(\mathbf{x};\theta_0)dx_1 \ldots dx_n. \\
\text{Now} \quad (1-\beta^*) - (1-\beta) &= \int_{\mathcal{C}} f(\mathbf{x};\theta_a)dx_1 \ldots dx_n - \int_{\mathcal{D}} f(\mathbf{x};\theta_a)dx_1 \ldots dx_n \\
\implies (1-\beta^*) - (1-\beta) &= \int_{\mathcal{C}_1} f(\mathbf{x};\theta_a)dx_1 \ldots dx_n + \int_{\mathcal{C}_2} f(\mathbf{x};\theta_a)dx_1 \ldots dx_n \\
&\quad - \int_{\mathcal{C}_1} f(\mathbf{x};\theta_a)dx_1 \ldots dx_n - \int_{\mathcal{D}_2} f(\mathbf{x};\theta_a)dx_1 \ldots dx_n, \\
&\quad \text{since} \quad \mathcal{D} = (\mathcal{C}_1 \cup \mathcal{D}_2), \\
&= \int_{\mathcal{C}_2} f(\mathbf{x};\theta_a)dx_1 \ldots dx_n - \int_{\mathcal{D}_2} f(\mathbf{x};\theta_a)dx_1 \ldots dx_n \\
&\geq \frac{1}{k}\int_{\mathcal{C}_2} f(\mathbf{x};\theta_0)dx_1 \ldots dx_n - \frac{1}{k}\int_{\mathcal{D}_2} f(\mathbf{x};\theta_0)dx_1 \ldots dx_n \\
&\geq 0, \qquad \text{as required.}
\end{aligned}$$

See Figure 23.

Moreover, equality is achieved only if

$$\mathcal{D}_2 \text{ is empty } (= \phi).$$

□

Example:

Suppose $X_1, X_2, \ldots, X_n$ are *i.i.d.* $\mathrm{N}(\mu, \sigma^2)$, σ^2 given, and consider

$$H_0 : \mu = \mu_0 \text{ vs. } H_a : \mu = \mu_a, \qquad \mu_a > \mu_0$$

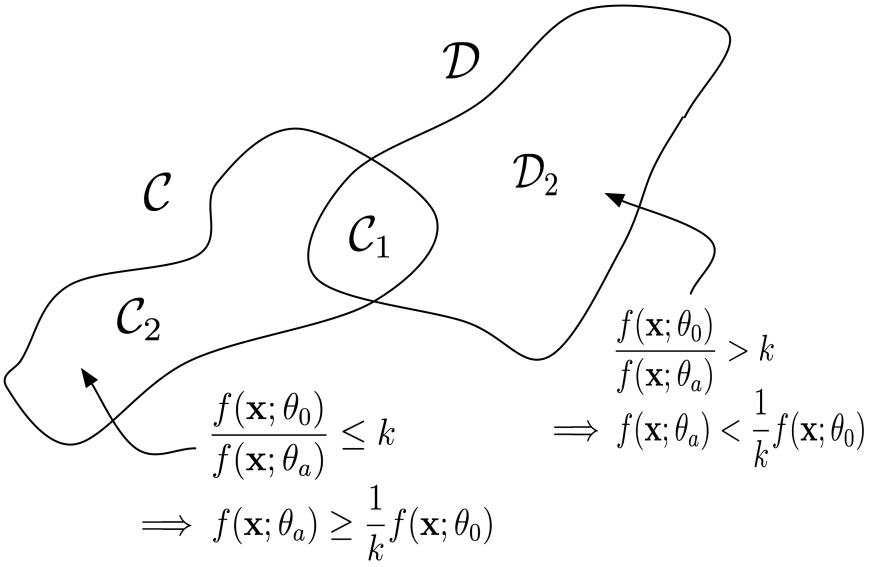


Figure 23: Neyman-Pearson Lemma

$$\begin{aligned}
\text{Then } \frac{f(\mathbf{x};\mu_0)}{f(\mathbf{x};\mu_a)} &= \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu_0)^2\right\}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu_a)^2\right\}} \\
&= \frac{\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\mu_0 + n\mu_0^2\right)\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}\mu_a + n\mu_a^2\right)\right\}} \\
&= \exp\left\{\frac{1}{2\sigma^2}(2n\bar{x}(\mu_0-\mu_a) - n\mu_0^2 + n\mu_a^2)\right\}.
\end{aligned}$$

For a constant k,

$$\begin{aligned}
\frac{f(\mathbf{x};\mu_0)}{f(\mathbf{x};\mu_a)} &\leq k \\
\Leftrightarrow (\mu_0-\mu_a)\bar{x} &\leq k^* \\
\Leftrightarrow \bar{x} &\geq c,
\end{aligned}$$

for a suitably chosen c (rejected when $\bar{x}$ is too big).

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,

$$\text{reject } H_0 \text{ if } \hat{z} \geq z(\alpha)$$

is the Neyman-Pearson LR test in this case.

Remarks

(1) This result 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) This can be extended to the case of

$$H_0 : \mu \leq \mu_0 \text{ vs. } H_A : \mu > \mu_0$$

In this case we take

$$\alpha = \max_{H_0} P(\text{rejecting } |\mu = \mu_0).$$

(3) This construction fails when we consider two-sided alternatives

$$\text{i.e., } \quad H_0 : \mu = \mu_0 \text{ vs. } H_A : \mu \neq \mu_0$$

$\implies$ no uniformly most powerful test exists for that case.