Probability and Statistics

Yang Jiao (Joy)

The Graduate Institute of International and Development Studies, Geneva

Univariate Random Variables

Univariate random variables

Intuitively, a **random variable (RV)** is a variable with stochastic outcomes, i.e. random realizations.

If the number of possible outcomes is **countable** ($\neq$ finite), we speak of a **discrete RV**. For example, the number of heads after two coin tosses, the possible realizations are 0, 1 and 2.

In contrast, **continuous RV** has uncountably many possible realizations. For example, the amount of precipitation tomorrow can take on any value in $\mathbb{R}_+$.

We denote the RV with capital letters and a particular realization with lower-case letters, e.g. X and x, respectively.

In this section, we deal with **univariate RV** that can take on a single, scalar value.

Univariate random variables

For the coinflip example (discrete RV), we can list the probabilities of all outcomes

$$\mathbb{P}[Y = y] = \begin{cases} 0.25 & \text{for } y=0 \\ 0.5 & \text{for } y=1 \\ 0.25 & \text{for } y=2 \end{cases}$$

The probability of each outcome must be in [0, 1] and the probabilities of all outcomes must sum to 1.

Univariate random variables

$$F_X(z) = P[X \leq z]$$

We can then find the **cumulative distribution function (cdf)**, which shows the probability that the outcome of a RV is smaller than some value z.

We obtain the cdf $F_X(z) = \mathbb{P}[X \leq z]$ by summing up the $\mathbb{P}[X = x]$ for all all $x \leq z$. Note that the cdfs are defined for all $z \in \mathbb{R}$, despite the fact that the corresponding RV is discrete.

Univariate random variables

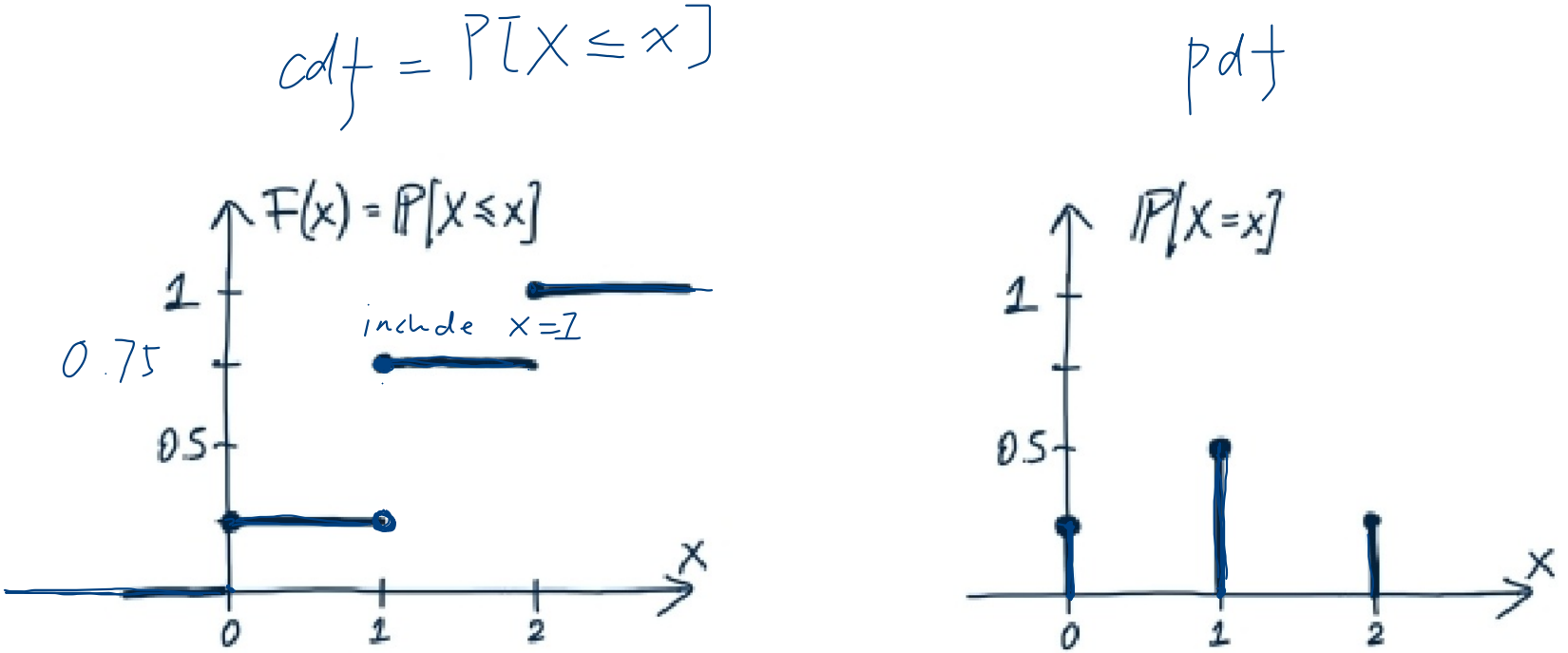


Figure 1.1: Representation of Probabilities for a Discrete Random Variable

Notes: Left plot shows the cdf, the right plot the probability function of the RV denoting the number of heads in two coinflips.

Univariate random variables

For continuous RVs, we cannot list the probabilities of all outcomes.

We define continuous RVs by specifying their cdf, i.e. the probability that a RV takes on a value in $(-\infty, z], z \in \mathbb{R}$.

Definition $F(x) = \mathbb{P}(-\infty, x]$ is a **cumulative distribution function (cdf)** if

(1) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$,

(2) $F(x)$ is non-decreasing : $F(x') \geq F(x) \quad \forall x' > x$,

(3) $F(x)$ is right continuous :
$\lim_{\epsilon \to 0} F(x_0 + \epsilon) = F(x_0) \quad \forall x_0 \in \mathbb{R}$.

Univariate random variables

cdf pdf

Just as for discrete RVs summing up the probabilities of all realizations $x \leq z$ gives $F_X(z)$, for continuous RVs integrating the pdf up to z gives $F_X(z)$.

Definition The **probability density function (pdf)** of a continuous *RV* X, f_X, is defined by

cdf pdf

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \quad \forall x \qquad \frac{d F_X(x)}{dx} = f_X(x)$$

It holds that

$$f_X(x) \geq 0 \ \forall x$$

and

$$\int_{-\infty}^{\infty} f_X(t)dt = \lim_{x \to \infty} F_X(x) = 1$$

Univariate random variables

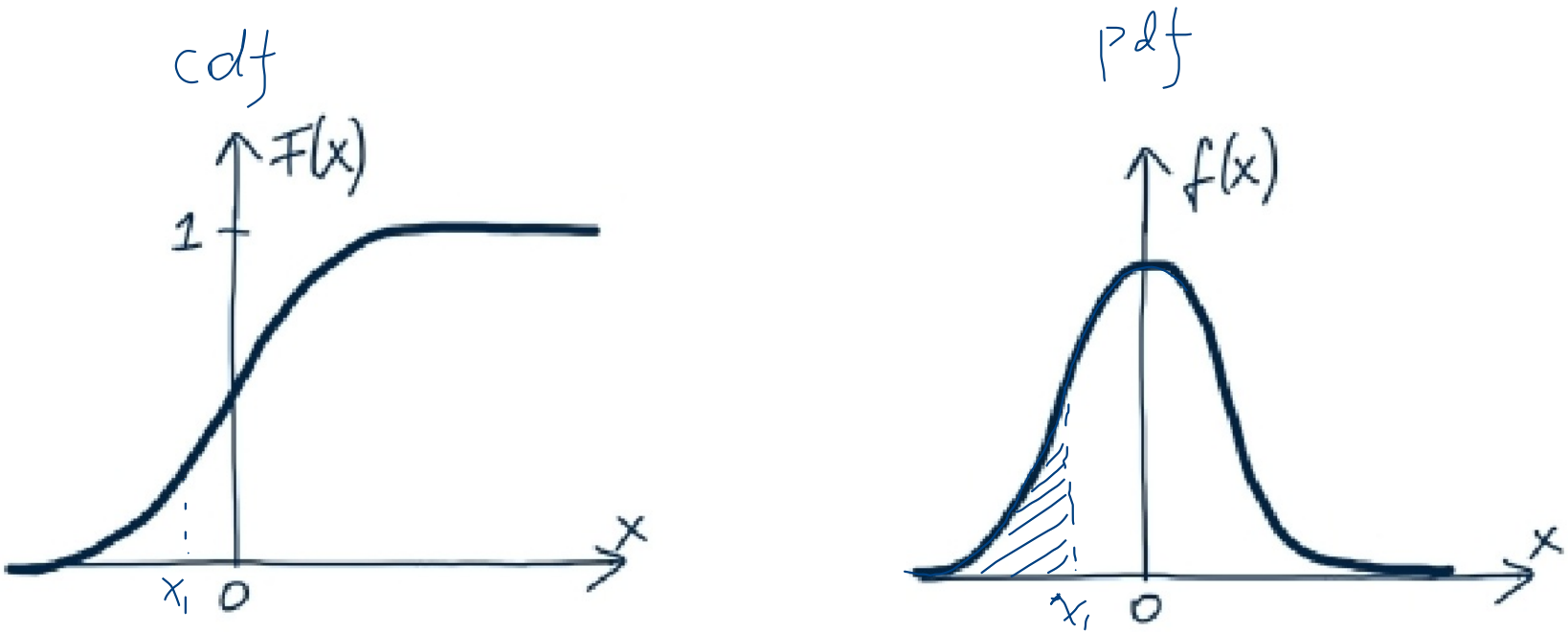


Figure 1.2: Representation of Probabilities for a Continuous Random Variable

Notes: Left plot shows the cdf, the right plot the pdf of a continuous RV (a standard Normal distribution.

Univariate random variables

Definition Two RVs X and Y are **identically distributed (i.d.)** if they have the same cdf, i.e. $F_X(z) = F_Y(z) \quad \forall z$.

Note that this does not imply that the two RVs X and Y are equal. For example, X could be the number of heads, Y the number of tails in the above example of two coin tosses.

The subsequent definitions and calculations focus on continuous RVs. They have intuitive counterparts for the discrete case, replacing e.g. integrals with sums.

X cdf pdf
Y = g(X)

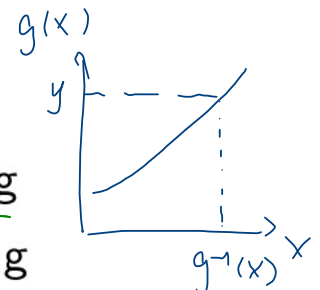
Transformations Sometimes we define a new RV Y based on a RV X : $Y = g(X)$. We can then find the cdf of Y from the cdf of X. For monotone g, we have

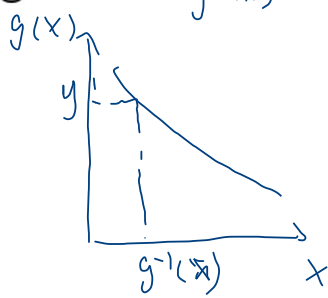
$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$$

$$= \begin{cases} \mathbb{P}\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right) & \text{if g increasing} \\ \mathbb{P}\left(X \geq g^{-1}(y)\right) = 1 - F_X\left(g^{-1}(y)\right) & \text{if g decreasing} \end{cases}$$

We can then find the pdf of Y by differentiating F_Y.

$P(X \geq g^{-1}(y)) + P(X \leq g^{-1}(y)) = 1$





Univariate random variables

Example : If $X \sim N(0,1)$ and we define $Y = g(X) = \mu + \sigma X$, $\sigma > 0$ then

$$X = g^{-1}(Y) = \frac{Y - \mu}{\sigma}$$

$$F_Y(y) = F_X\left(g^{-1}(y)\right) = F_X\left(\frac{y-\mu}{\sigma}\right),$$

and

$$\begin{aligned} f_Y(y) &= f_X\left(g^{-1}(y)\right) g^{-1'}(y) \\ &= f_X\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}. \end{aligned}$$

$X \sim N(0,1)$ $\quad Y = g(X) = \mu + \sigma X$

cdf, pdf $\quad X = g^{-1}(Y) = \dfrac{Y-\mu}{\sigma}$

cdf $\quad F_Y(y) = P[Y \le y] = P[\mu + \sigma X \le y] = P\left[X \le \dfrac{y-\mu}{\sigma}\right] = F_X\left(\dfrac{y-\mu}{\sigma}\right)$

pdf $= \dfrac{dF_Y(y)}{dy} = f_X\left(\dfrac{y-\mu}{\sigma}\right) \cdot \dfrac{1}{\sigma}$

$= \dfrac{1}{\sqrt{2\pi}} \exp\left\{-\dfrac{1}{2}\left(\dfrac{y-\mu}{\sigma}\right)^2\right\} \cdot \dfrac{1}{\sigma}$

Proposition **Probability Integral Transformation**

Let X have cdf F_X and define $Y = F_X(X)$. Then Y is uniformly distributed : $Y \sim \mathcal{U}[0,1]$, i.e. $F_Y(y) = y$ for $y \in [0,1]$

Exercise 1 : prove the proposition

Univariate random variables

discrete $\begin{cases} x=0 & 0.5 \\ x=1 & 0.5 \end{cases}$ joint $\begin{cases} y=1 & 0.5 \\ y=2 & 0.5 \end{cases}$

$E[X] = 0 \cdot 0.5 + 1 \cdot 0.5 = 0.5$

Moments The properties of a RV are commonly described by **moments**.

pdf $f_X(x)$

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

Definition The **expectation** of a RV X is given by $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$.
More generally, the expectation of a function h of a *RV* X is given by $\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx$.

$h(X)$ pdf $f_X(x)$

$E[h(X)] = \int_{-\infty}^{\infty} h(x) \cdot f_X(x)\,dx$

$\mathbb{E}[h(X)]$ is a pdf-weighted average of all possible realizations of $h(X)$.

Univariate random variables

The expectation operator has the following properties :

- $\mathbb{E}\left[ah_1(X) + bh_2(X) + c\right] = a\mathbb{E}\left[h_1(X)\right] + b\mathbb{E}\left[h_2(X)\right] + c$
- $h_1(x) \geq h_2(x)\ \forall x \quad \Rightarrow \quad \mathbb{E}\left[h_1(X)\right] \geq \mathbb{E}\left[h_2(X)\right]$

The second property implies

- $h_1(x) \geq 0\ \forall x \quad \Rightarrow \quad \mathbb{E}\left[h_1(X)\right] \geq 0$
- $a \leq h_1(x) \leq b\ \forall x \quad \Rightarrow \quad a \leq \mathbb{E}\left[h_1(X)\right] \leq b$

Exercise 2 : prove the four properties (hint : using the definition of expectation)

$$E\left[ah_1(x)+bh_2(\alpha)+c\right]$$

$$=\int_{-\infty}^{+\infty}\left[ah_1(\alpha)+bh_2(x)+c\right]f_\alpha(\alpha)\,d\alpha$$

$$=\int_{-\infty}^{+\infty}ah_1(x)f_\alpha(x)\,dx+\int_{-\infty}^{+\infty}bh_2(x)f_\alpha(x)\,dx+\int_{-\infty}^{+\infty}cf_\alpha(x)\,dx$$

$$=a\int_{-\infty}^{+\infty}h_1(\alpha)f_x(\alpha)\,d\alpha+b\int_{-\infty}^{+\infty}f_\alpha(\alpha)h_2(\alpha)\,d\alpha+c\underbrace{\int_{-\infty}^{+\infty}f_\alpha(\alpha)\,d\alpha}_{=1}$$

$$=a\left[h_1(\alpha)\right]+b\left[h_2(\alpha)\right]+c$$

$$\int_{-\infty}^{+\infty} h_1(x) f_x(x)\,dx - \int_{-\infty}^{+\infty} h_2(x) f_x(x)\,dx$$

$$= \int_{-\infty}^{+\infty} \underbrace{[h_1(x) - h_2(x)]}_{\geq 0} f_x(x)\,dx$$

$\because h_1(x) \geq h_2(x)$, $f_x(x) \geq 0$

$\therefore \int_{-\infty}^{+\infty} [h_1(x) - h_2(x)] f_x(x)\,dx \geq 0$

$\therefore E[h_1(x)] \geq E[h_2(x)]$

$E[h_1(x)] \geq E[h_2(x)]$

$E[h_1(x)] - E[h_2(x)] \geq 0$

$$E[h_1(x)] = \int_{-\infty}^{+\infty} h_1(x) f_x(x) dx$$

$h_1(x) \geq 0$, $f_x(x) \geq 0$

$$\int_{-\infty}^{+\infty} h_1(x) f_x(x) dx \geq 0$$

$$E[h_1(x)] \geq 0$$

$h_1(x) \geq h_2(x)$

$\Rightarrow E[h_1(x)] \geq E[h_2(x)]$

let $h_2(x) = 0$

$\Rightarrow E[h_1(x)] \geq E[0] = 0$

let $h_1(x) = b$

$\Rightarrow E[b] = b \geq E[h_2(x)]$

$h_2(x) \geq a$

$\Rightarrow E[h_2(x)] \geq E[a] = a$

$a \le h_1(x) \le b$

$a f_x(x) \le h_1(x) f_x(x) \le b f_x(x)$

$$\rightarrow \int_{-\infty}^{+\infty} a f_x(x)\,dx \le \int_{-\infty}^{+\infty} h_1(x) f_x(x)\,dx \le \int_{-\infty}^{+\infty} b f_x(x)\,dx$$

$$\int_{-\infty}^{+\infty} f_x(x)\,dx = 1$$

$$\therefore a \le E[h_1(x)] \le b$$

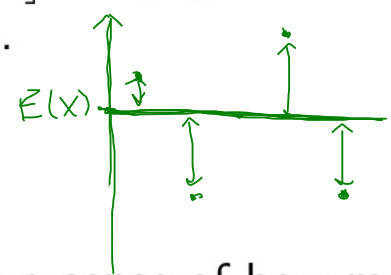
Univariate random variables

Definition For $n \in \mathbb{Z}$, the ***n*th moment** of X is $\mathbb{E}[X^n]$, while the ***n*th central moment** of X is $\mathbb{E}[(X - \mathbb{E}[X])^n]$.

Exercise 3 : find the first central moment of X

$$E[(X - E[X])] = E[X] - E[E[X]] = E[X] - E[X] = 0$$

Definition The second central moment of X is called the **variance** : $\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$. Its positive square root is the **standard deviation** of X.



Exercise 4 : prove the second equation

The variance and standard deviation give us a sense of how much the realizations of X can differ from the expectation of X.

The variance has the property

$$\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$$

$$E[aX + b] = aE[X] + b$$

Exercise 5 : prove the property

$$E[(X-E[X])^2] = E[X^2 - 2XE[X] + E[X]^2]$$

$$= E[X^2] - E[2XE[X]] + E[E[X]^2]$$

$$= E[X^2] - E[X]\cdot 2E[X] + E[X]^2$$

$$= E[X^2] - 2E[X]^2 + E[X]^2$$

$$= E[X^2] - E[X]^2$$

$V(Y) = E[X^2] - E[X]^2$

$V[aX+b] = E[((aX+b) - E[aX+b])^2]$

$= E[(aX+b)^2] - E[aX+b]^2$

$= E[a^2X^2 + 2abX + b^2] - (aE[X] + b)^2$

$= a^2E[X^2] + 2abE[X] + b^2 - (a^2E[X]^2 + 2abE[X] + b^2)$

$= a^2(E[X^2] - E[X]^2)$

$= a^2V[X]$

Univariate random variables

Definition The **moment-generating function (MGF)** of X is $M_X(t) = \mathbb{E}[\exp\{tX\}]$

We can obtain the nth non-central moment of X by

$$\mathbb{E}[X^n] = \frac{\partial^n}{\partial^n t} M_X(t)\Big|_{t=0}$$

Central moments can then be constructed as functions of non-central moments. For example, $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Exercise 6 : find the MGF of $Y = aX + b$

It is not guaranteed that a given moment of a RV X exists. However, if the m th central moment exists, then so do all lower-order moments.

Jensen's inequality

Proposition **Jensen's Inequality**

For any convex function g, we have $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$ [1]
For any concave function h, we have $\mathbb{E}[h(X)] \leq h(\mathbb{E}[X])$

Trick to remember : $\mathbb{E}\left[X^2\right] \geq \mathbb{E}[X]^2$ as $\mathbb{V}[X] \geq 0$

$V[X] = E[X^2] - E[X]^2 \geq 0 \Rightarrow E[X^2] \geq E[X]^2$ $y = X^2$ convex

$y = \ln X$ concave $E[\ln X] \leq \ln E[X]$

1. Proof : $g(x)$ convex implies $g(x) \geq g(x_0) + c(x - x_0)$, where $c = \partial g(x)/\partial x|_{x=x_0}$. Setting $x_0 = \mathbb{E}[X]$ and taking expectations gives the result.

Multivariate Random Variables

Multivariate random variables

Definition Let $f_{X,Y}$ be the **joint pdf** of X and Y. It satisfies

$$f_{X,Y}(x,y) \geq 0 \quad \forall (x,y)$$

$f_X(x)$

and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1.$$

The **marginal pdf** of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

and the **conditional pdf** of Y given X is

$$f_Y(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

$E[Y|X=1] = \sum y \cdot p(y|x=2)$

$= 1 \times 0.4 + 2 \cdot 0.1 = 0.4 + 0.2 = 0.6$

$E[Y|X=2] = 1 \times 0.2 + 2 \times 0.3 = 0.8$

$$E[Y|X] = \begin{cases} 0.6 & x=1 \\ 0.8 & x=2 \end{cases}$$

Example : A pizza restaurant caters to students. Each customer purchases either one or two slices of pizza and either one or two drinks during their meal. Let X be the number of pizza slices purchased, and Y be the number of drinks. The following joint pdf

$$\mathbb{P}[X=1, Y=1] = 0.4$$
$$\mathbb{P}[X=1, Y=2] = 0.1$$
$$\mathbb{P}[X=2, Y=1] = 0.2$$
$$\mathbb{P}[X=2, Y=2] = 0.3$$

$f_{Y|X}(1|1) = \frac{0.4}{0.5} = 0.8$

$= \frac{f_{X,Y}(1,1)}{f_X(1)}$

$f_X(x) = f_{X,Y}(x,1) + f_{X,Y}(x,2)$

$$= \begin{cases} 0.5 & x=1 \\ 0.5 & x=2 \end{cases}$$

is a valid probability function.

Conditional expectation

Definition The **conditional expectation** of Y given $X = x$ is

$$\mathbb{E}[Y \mid X = x] = \int y f_Y(y \mid x) dy$$

and anlogously

$$\mathbb{E}[h(Y) \mid X = x] = \int h(y) f_Y(y \mid x) dy$$

Note that $\mathbb{E}[Y \mid X = x]$ is a function of x.
We often write $\mathbb{E}[Y \mid X]$ without restricting ourselves to any particular realization of X. This $\mathbb{E}[Y \mid X]$ is a function of the RV X, and we can compute the expectation of it.

$$E[Y|X] = \begin{cases} E[Y|x_1] & x = x_1 \\ E[Y|x_2] & x = x_2 \end{cases}$$

Law of Iterated Expectations

Proposition **Law of Iterated Expectations (LIE)**

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]]$$

Proof :

$$\mathbb{E}[Y] = \int y f(y) dy = \int y \int f(x,y) dx dy = \iint y f(y \mid x) f(x) dx dy$$

$$= \int \left[\int y f(y \mid x) dy \right] f(x) dx = \mathbb{E}[\mathbb{E}[Y \mid X]]$$

The LIE states that taking the expectation of Y is the same as first taking the expectation conditional on X and then averaging over all X.

Exercise 7 : If $\mathbb{E}[U \mid X] = 0$, prove $\mathbb{E}[UX] = 0$.

$$E[UX] = E[E[UX|X]] = E[XE[U|X]] = E[X \cdot 0] = E[0] = 0$$

$E[Y|X] \Rightarrow$ function of X

X given, know value/realization of X

$Y = a + bX + cZ$

$E[a + bX + cZ|X] = a + bX + E[cZ|X]$

Independence

Definition Two *RV*s X and Y are **independent** if we can factorize the joint as the product of marginal pdfs : $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Properties : if X and Y are independent

- $f_{Y|X}(y \mid x) = f_Y(y)$ (knowing the realization of X conveys no information about Y)
- $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$ for all functions g and h
- $M_{X+Y}(t) = M_X(t)M_Y(t)$

X, Y independent

$E[XY] = E[X]E[Y]$

X, Y not necessarily independent

$E[X+Y] = E[X] + E[Y]$

$E[Y|X] = E[Y]$

$M_{X+Y}(t) = E[\exp(tX+tY)]$

$= E[\exp(tX)\cdot\exp(tY)]$

$= E[\exp(tX)]\cdot E[\exp(tY)]$

$= M_X(t)\cdot M_Y(t)$

Covariance

Definition The **covariance** between two scalar RVs X and Y is

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

The **correlation** (coefficient) is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{V[X]V[Y]}}$$

and satisfies

$$|\,\text{Corr}(X, Y)| \in [0, 1]$$

Covariance is the measure of the linear dependence of X and Y, the correlation coefficientt adjusts for the scale of the RVs X and Y.

A positive (negative) $\text{Corr}(X, Y)$ indicates that whenever X is above $\mathbb{E}[X]$, then Y tends to be above (below) $\mathbb{E}[Y]$, and the closer $\text{Corr}(X, Y)$ is to 1 (-1), the stronger this relationship.

Independent vs. linearly independent

$Cov(X,Y)=0 \Rightarrow X, Y$ linearly independent

weaker

independent

If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ and therefore $\text{Cov}(X, Y) = 0$, i.e. X and Y are also linearly independent.

However, the opposite does not hold necessarily : linear independence, $\text{Cov}(X, Y) = 0$, does not imply that two RVs are independent.

$E[X]=0 \quad V[X]=1$

Example : $X \sim N(0, 1)$ and $Y = X^2$
The two are clearly not independent as knowing one allows you to perfectly tell the other.
However, $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}\left[X^3\right] = 0$ (0)

Variance of the sum of two RVs

$E[X+Y] = E[X] + E[Y]$ always holds

$E[XY] = E[X]E[Y]$ X, Y independent

$$\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\,\mathrm{Cov}(X, Y)$$

More generally,

$$\mathbb{V}[a + bX + cY] = b^2\mathbb{V}[X] + c^2\mathbb{V}[Y] + 2bc\,\mathrm{Cov}(X, Y)$$

Exercise 8 : prove the above equation

If X and Y are uncorrelated, i.e. $\mathrm{Cov}(X, Y) = 0$, then

$$\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y]$$

$$(a+b)^2 = a^2 + 2ab + b^2$$

$$V[X+Y] = E[(X+Y-E[X+Y])^2]$$

$$= E[(X-E[X] + Y-E[Y])^2]$$

$$= E[(X-E[X])^2] + 2E[(X-E[X])(Y-E[Y])] + E[(Y-E[Y])^2]$$

$$= V(X) + 2cov(X,Y) + V(Y)$$

Variance of bivariate RV

Let $Z = (X, Y)' \in \mathbb{R}^2$ and let

$$\mathbb{E}[Z] = \mathbb{E}\left[(X, Y)'\right] = (\mu_x, \mu_y)' = \mu \qquad \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$

then

$$\mathbb{V}[Z] \equiv \mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])'\right] = \mathbb{E}\left[ZZ'\right] - \mathbb{E}[Z]\mathbb{E}[Z]'$$ [2]

also we have

$$\mathbb{V}[Z] = \mathbb{E}\left[\begin{array}{c} X - \mathbb{E}[X] \\ Y - \mathbb{E}[Y] \end{array}\right]\left[\begin{array}{cc} X - \mathbb{E}[X] & Y - \mathbb{E}[Y] \end{array}\right]$$

$$\begin{bmatrix} V(X) & Cov(X,Y) \\ Cov(X,Y) & V(Y) \end{bmatrix}$$

$$= \mathbb{E}\left[\begin{array}{cc} (X - \mathbb{E}[X])(X - \mathbb{E}[X])' & (X - \mathbb{E}[X])(Y - \mathbb{E}[Y])' \\ (Y - \mathbb{E}[Y])(X - \mathbb{E}[X]) & (Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y]) \end{array}\right]$$

$$= \left[\begin{array}{cc} \mathbb{V}[X] & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \mathbb{V}[Y] \end{array}\right]$$

2. $\mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])'\right] = \mathbb{E}\left[ZZ' - \mu Z' - Z\mu' - \mu\mu'\right] = \mathbb{E}\left[ZZ'\right] - \mu\mathbb{E}\left[Z'\right] - \mathbb{E}[Z]\mu' - \mu\mu' = \mathbb{E}\left[ZZ'\right] - \mu\mu'$

Covariance of bivariate RV

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

The covariance of two **multivariate** RVs X and Y is defined as

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])'\right] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]'$$

It is composed of the covariances between the individual elements of X and Y .

$$\text{Cov}[X, Y] = \begin{bmatrix} \text{cov}(X_1, Y_1) & \text{cov}(X_1, Y_2) \\ \text{cov}(X_2, Y_1) & \text{cov}(X_2, Y_2) \end{bmatrix}$$

Proportionality

Definition A function $f(x)$ is **proportional** to some other function $g(x)$ if we can write $f(x) = cg(x)$ for some constant independent of x. We write $f(x) \propto g(x)$.

$f(x) \propto g(x) =$

$f(x)$ is pdf

$\int_{-\infty}^{\infty} f(x)dx = 1$

$\int_{-\infty}^{\infty} cg(x)dx = 1 \Rightarrow c$

We can specify a pdf by finding it up to proportionality.

Example :

$f(x) \propto \exp\{-\lambda x\}$ and $f(x)$ is a pdf defined on $x \in [0, \infty)$

Since $\int_0^\infty \exp\{-\lambda x\} = 1/\lambda$ and pdfs must integrate to 1. We can get a valid pdf

$$f(x) = \lambda \exp\{-\lambda x\}$$

$\int_0^\infty c \exp(-\lambda x) = 1 \Rightarrow c\int_0^\infty \exp(-\lambda x) = -\frac{c}{\lambda}\exp(-\lambda x)\Big|_0^\infty = 0 - \left(-\frac{c}{\lambda}\cdot 1\right) = \frac{c}{\lambda} = 1$

$-\frac{1}{\lambda}\exp(-\lambda x)\Big|_0^\infty$ $\Rightarrow c = \lambda$

Bayes' formula

From the definition of conditional pdf :

$$f_{XY}(x,y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y)$$

Together with the definition of margianl pdf, we then obtain **Bayes' formula** :

$$f_Y(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x \mid y) f_Y(y)}{f_X(x)} \propto f_X(x \mid y) f_Y(y)$$

The Bayes formula tells us how to update the prior distribution $f_Y(y)$ to the posterior distribution (or -belief) $f_Y(y \mid x)$ by using the likelihood $f_X(x \mid y)$ and evidence $f_X(x)$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \propto \text{likelihood} \times \text{prior}$$

It is enough to find the pdf by knowing the function proportional to it, so we only have to know the product of prior and likelihood to get the posterior.

Bayes' formula

Example :

We observe some data x as draws of a RV X from probability distribution $p(X \mid \theta)$ indexed by a parameter θ. And we want to make inference about θ.

We treat θ as a RV. Starting from some prior belief about θ's distribution $p(\theta)$, we can update it to posterior belief $p(\theta \mid x)$ after observing the data sample $p(x \mid \theta)$ using Bayes' formula :

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)} \propto p(x \mid \theta)p(\theta)$$

This is the idea of Bayesian inference.

Common Probability Distributions

Uniform

Uniform $X \sim \mathcal{U}(a, b), \quad a, b \in \mathbb{R}$

- Domain : $[a, b]$
- pdf : $f(x) = \frac{1}{b-a}$
- cdf : $F(x) = \frac{x-a}{b-a}$
- $\mathbb{E}[X] = (a + b)/2$

Uniform

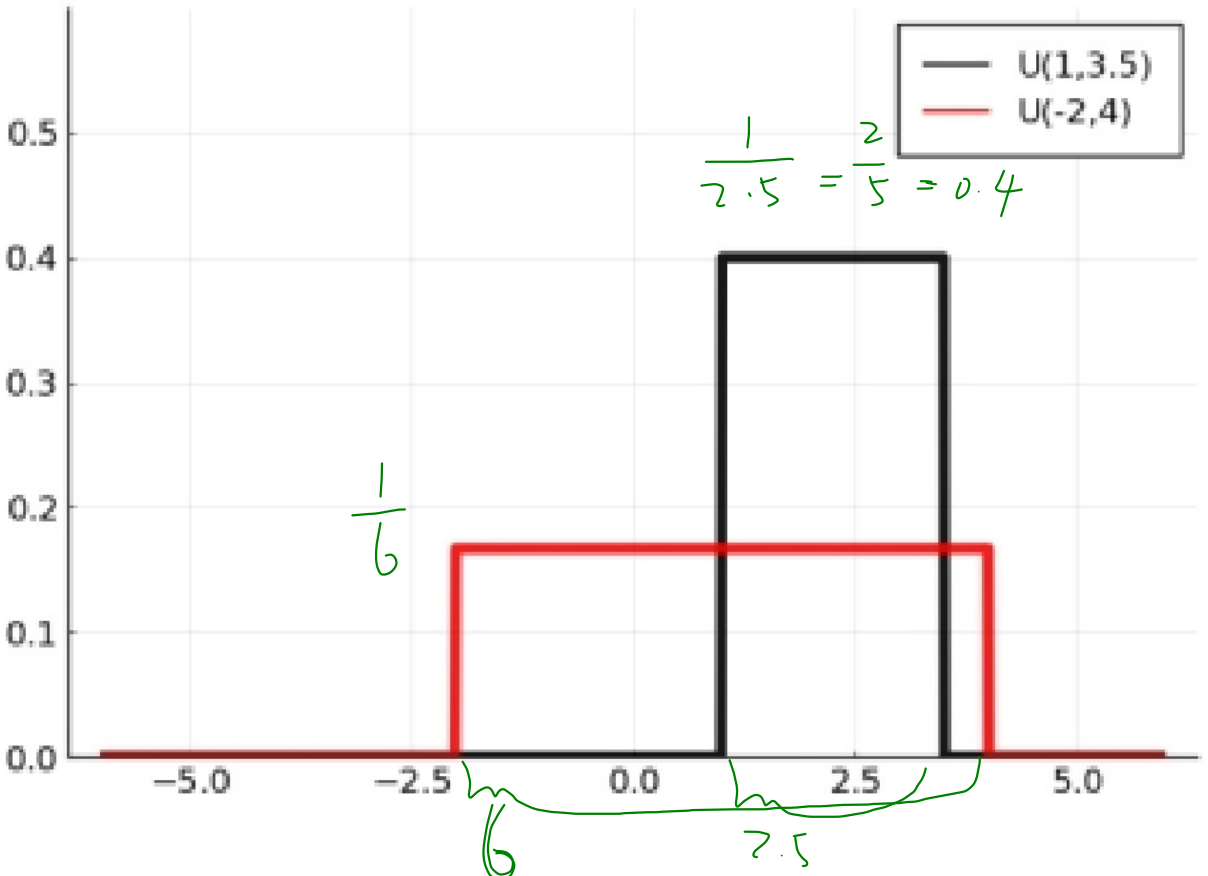


Figure D1: Uniform Distributions

Normal

Normal $X \sim N\left(\mu, \sigma^2\right), \quad \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}_{++}$

- Domain : $\mathbb{R}$
- pdf :

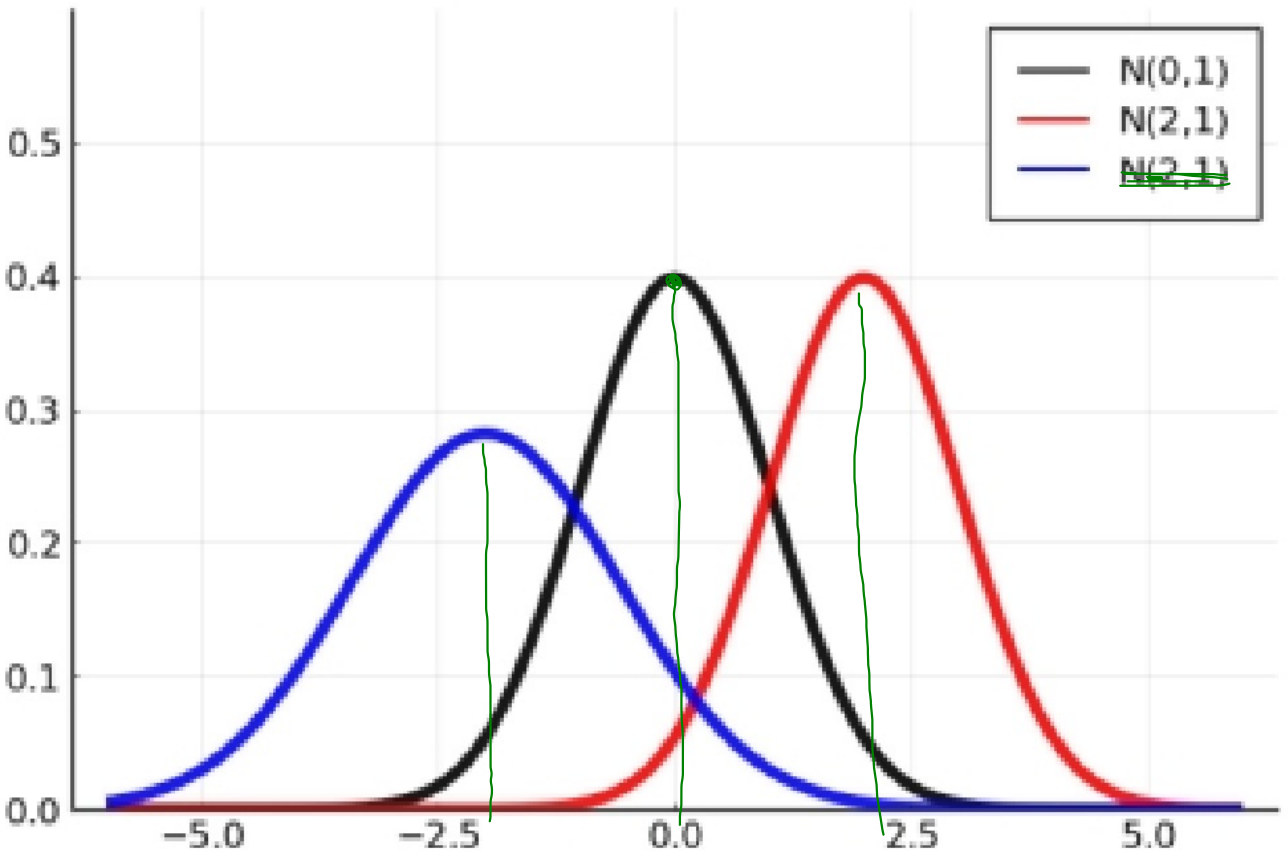
$$f(x) = \left(2\pi\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$

- $\mathbb{E}[X] = \mu$

  $\mathbb{V}[X] = \sigma^2$

  mode $= \mu$

Normal



N(-2,3)

Figure D2: Normal Distributions

Normal

3σ principles

X has the following probabilities of falling within 1, 2 or 3 standard deviations from the mean, respectively :

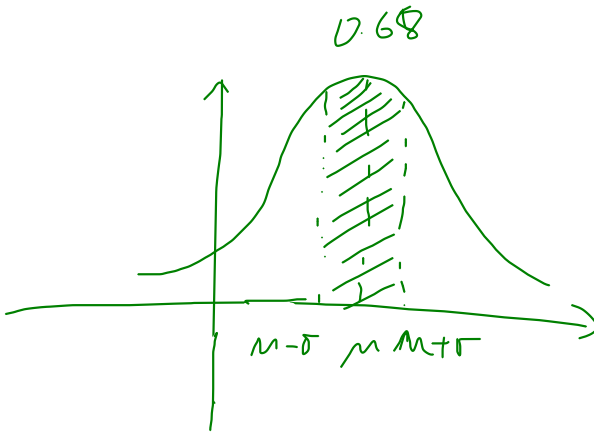
$Var(X) = \sigma^2$

$s.d(X) = \sqrt{Var(X)} = \sigma$

$$\mathbb{P}\left(\left|\frac{y-\mu}{\sigma}\right| < 1\right) \approx 0.68$$

$$\mathbb{P}\left(\left|\frac{y-\mu}{\sigma}\right| < 2\right) \approx 0.95$$

$$\mathbb{P}\left(\left|\frac{y-\mu}{\sigma}\right| < 3\right) \approx 0.997$$



Normal

Standard Normal $Z \sim N(0,1)$

For $Z \sim N(0,1)$, we have the MGF $M_Z(t) = \exp\left\{\frac{1}{2}t^2\right\}$

Proof : We have

$$\begin{aligned} M_Z(t) &= \mathbb{E}[\exp\{tz\}] \\ &= \int \exp\{tz\} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\} dx \\ &= \frac{1}{\sqrt{2\pi}} \int \exp\left\{-\frac{1}{2}\left(z^2 - 2t\right)\right\} dz \\ &= \exp\left\{\frac{1}{2}t^2\right\} \int \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(z-t)^2\right\} dz \\ &= \exp\left\{\frac{1}{2}t^2\right\}, \end{aligned}$$

the expression inside the integral is the pdf of a $N(t,1)$ and therefore has to integrate to one.

Normal

Note that

$$X \sim N\left(\mu, \sigma^2\right) \quad \Leftrightarrow \quad X = \mu + \sigma Z, \quad Z \sim N(0,1)$$

Using the property of MGF, we can get the MGF of X

$M_X(aX+b)$
$= e^{tb} M_{aX}(t)$

$$M_X(t) = \exp\left\{\frac{1}{2}\sigma^2 t^2\right\} \exp\{\mu t\}$$

Exercise 9 : prove the above expression

Normal

Note that

$$f(x) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(x^2 - 2x\mu\right)\right\}$$

$$\text{i.e. } f(x) = c\exp\left\{-\frac{1}{2\sigma^2}\left(x^2 - 2x\mu\right)\right\}$$

where c is a unique constant that makes sure that the expression integrates to 1 so that $f(x)$ is a valid pdf.

Thus, if for some RV Y

$$p(y) \propto \exp\left\{ay^2 + by\right\}$$

then we can deduce

$$Y \sim N\left(-\frac{b}{2a}, -\frac{1}{2a}\right)$$

Gamma

Gamma $X \sim G(\alpha, \beta), \quad \alpha, \beta \in \mathbb{R}_{++}$

- Domain : $\mathbb{R}_{++}$
- pdf :

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp\{-x/\beta\}$$

where $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} \exp\{-x\} dx$ is the Gamma-function

- $\mathbb{E}[X] = \alpha\beta$

$\mathbb{V}[X] = \alpha\beta^2$

Gamma

Exercise 10 : prove the Gamma-function has the following properties :

- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \quad$ for $\alpha > 0$
- $\Gamma(n) = (n-1)!$ for $n \in \mathbb{Z}$

Gamma

Gamma(α, β)

α peak

β variance

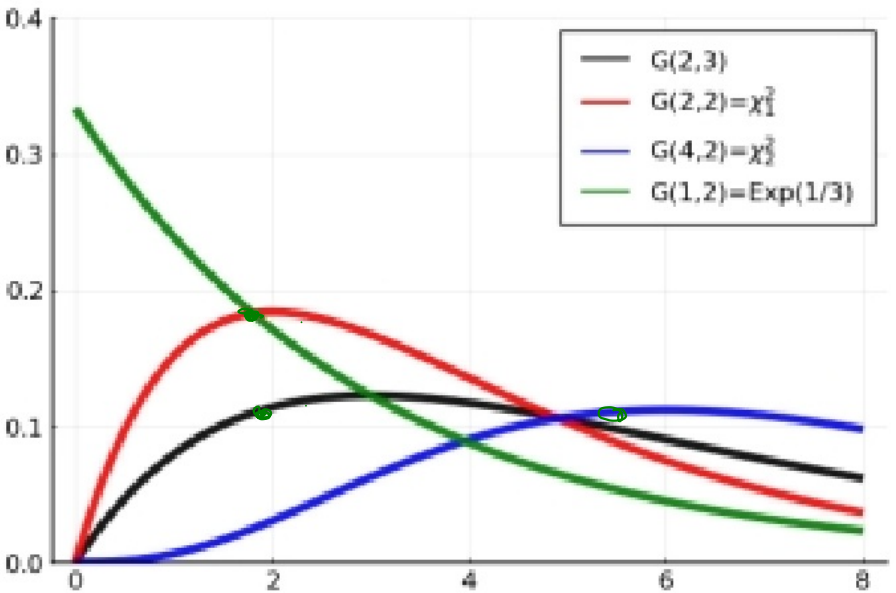


Figure D3: Gamma Distributions

The shape parameter α influences rather the peakedness, whereas the scale parameter β influences more the variance/spread.

Sometimes, a Gamma-distribution is parameterized also with its shape α and rate r, whereby $r = 1/\beta$.

Gamma : Exponential

The exponential and chi-squared distributions are special cases of the Gamma distribution :

Exponential $X \sim \text{Exp}(\lambda) \quad \Leftrightarrow \quad X \sim G(1, 1/\lambda)$

The pdf then simplifies to

$$f(x) = \lambda \exp\{-\lambda x\}$$

Gamma : Chi-squared

Chi-squared $X \sim \chi^2_\rho, \rho \in \mathbb{Z} \quad \Leftrightarrow \quad X \sim G\left(\frac{\rho}{2}, 2\right)$

X has the same distribution as the sum of ρ independent standard Normal RVs,

$$\{X_i\}_{i=1}^{\rho} \overset{\text{i.i.d.}}{\sim} N(0,1) \Rightarrow \sum_{i=1}^{\rho} X_i^2 \sim \chi^2_\rho$$

Inverse Gamma

$$X \sim G(\alpha, \beta) \quad \Leftrightarrow \quad \frac{1}{X} \sim \mathsf{IG}(\alpha, \beta)$$

Inverse Gamma $X \sim IG(\alpha, \beta), \quad \alpha, \beta \in \mathbb{R}_{++}$

- Domain : $\mathbb{R}_{++}$
- pdf :

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\{-\beta/x\}$$

- $\mathbb{E}[X] = \beta/(\alpha - 1)$ for $\alpha > 1$

  $\mathbb{V}[X] = \beta^2/(\alpha - 1)^2(\alpha - 2)$ for $\alpha > 2$

  $\text{mode} = \frac{\beta}{\alpha+1}$

Inverse Gamma

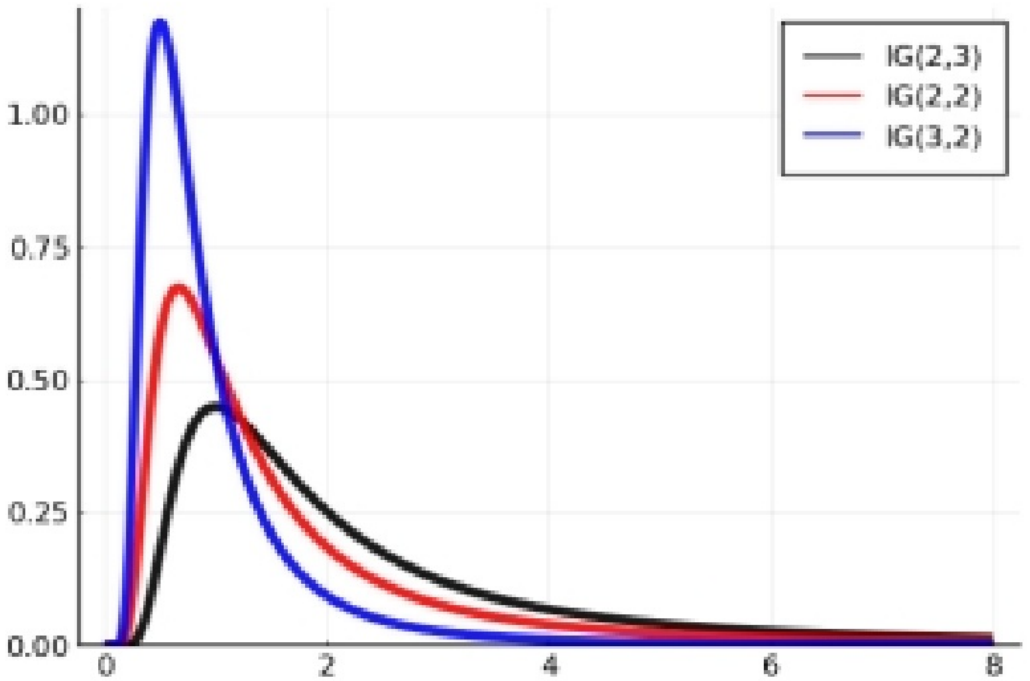


Figure D4: Inverse Gamma Distributions

Multivariate Normal

Multivariate Normal $X \sim N(\mu, \Sigma), \quad \mu \in \mathbb{R}^k, \Sigma_{k \times k}$ positive semi-definite

- Domain : $\mathbb{R}^k$
- pdf :

$$f(x) = (2\pi)^{-k/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}$$

- $\mathbb{E}[X] = \mu$

$\mathbb{V}[X] = \Sigma$

Multivariate Normal

Scalar $X \sim N(\mu, \sigma^2) \iff X = \mu + \sigma Z \quad Z \sim N(0,1)$

$\sigma \cdot \sigma = \sigma^2$

Analogously as for a univariate Normal distribution, we have

$$X \sim N(\mu, \Sigma) \quad \Leftrightarrow \quad X = \mu + \Sigma_{tr} Z, \quad Z \sim N(0, I)$$

$K\times 1 \quad K\times 1 \quad K\times K$

where Σ_{tr} is the Cholesky factor of Σ, i.e. it is a lower-triangular matrix s.t. $\Sigma_{tr}\Sigma'_{tr} = \Sigma$.

Hence, based on X, we can get

$$Z = \Sigma_{tr}^{-1}(X - \mu) \sim N(0, I)$$

Multivariate Normal

Let $X \sim N(\mu, \Sigma)$ and partition the vector X, μ and matrix Σ correspondingly,

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad X = (X_1', X_2')', \quad \mu = (\mu_1', \mu_2')', \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

The marginal pdfs of X_1 and X_2 are then also multivariate Normal with the corresponding elements of μ and Σ :

$$X_1 \sim N(\mu_1, \Sigma_{11}), \quad X_2 \sim N(\mu_2, \Sigma_{22}).$$

Note that this implies that every element of X follows a univariate Normal distribution.

Multivariate Normal

$X_1 \sim N(\mu_1, \Sigma_1)$

$X_2 \sim N(\mu_2, \Sigma_2)$

For the conditional, we get $X_1 \mid X_2 \sim N\left(\mu_{1|2}, \Sigma_{1|2}\right)$ with

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(X_2 - \mu_2\right), \quad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

Two Normal RVs are independent iff they are uncorrelated, i.e. iff $\mathrm{Cov}\left(X_1, X_2\right) = \Sigma_{12} = \Sigma_{21}' = 0$

Independent

↓ ↑

uncorrelated

(linearly independent)

(or = 0)

↑ ↓

Multivariate Normal

$$\underset{k\times 1}{Z} \sim N(0, I) \quad \Rightarrow \quad \underset{1\times k\ k\times 1}{Z'Z} = \sum_{j=1}^{k} Z_i^2 = \sim \chi_k^2$$

$$X \sim N(\mu, \Sigma) \quad \Rightarrow \quad (X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi_k^2$$

$Z = \Sigma_{tr}^{-1}(X-M) \sim N(0, I)$ $\quad (X-M)'(\Sigma_{tr}')^{-1}\Sigma_{tr}^{-1}(X-M)$

$= (X-M)'\Sigma^{-1}(X-M)$

Multivariate Normal

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma^2 \cdot \text{corr}(\,) & \\ & \sigma^2 & \\ & & \ddots \\ & & & \sigma^2 \end{bmatrix} \qquad \text{corr} = \frac{\text{cov}}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{\text{cov}}{\sigma \cdot \sigma} = \frac{\text{cov}}{\sigma^2}$$

Note that under $\Sigma = \sigma^2 V$ the pdf of $X \sim N(\mu, \Sigma)$ can be written in two ways :

$= \sigma^2 V \qquad V \begin{bmatrix} 1 & \text{corr} & \\ \text{corr} & 1 & \\ & & \ddots \\ & & & 1 \end{bmatrix}$

$$f(x) = (2\pi)^{-k/2} \left|\sigma^2 V\right|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)' \left[\sigma^2 V\right]^{-1} (x-\mu)\right\}$$

$$= \left(2\pi\sigma^2\right)^{-k/2} |V|^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)' V^{-1} (x-\mu)\right\}$$

Multivariate Normal

Analogously as for the univariate Normal distribution, note that

$$f(x) \propto \exp\left\{-\frac{1}{2}x'\Sigma^{-1}x - 2\mu'\Sigma^{-1}x\right\}$$

Therefore, if for some RV Y we have

$$p(y) \propto \exp\left\{-\frac{1}{2}x'Ax - 2b'Ax\right\}$$

then we can deduce

$$Y \sim N\left(b, A^{-1}\right)$$

Matrix-Variate Normal

Matrix-Variate Normal

$X \sim MN(\mu, U, V), \mu_{n\times k} \in \mathbb{R}^{nk}, U_{n\times n}, V_{k\times k}$ p.s.d.

- Domain : $X \in \mathbb{R}^{nk}$ is an $n \times k$ matrix
- pdf :

$$f(x) = (2\pi)^{-nk/2}|V|^{-n/2}|U|^{-p/2}\exp\left\{-\frac{1}{2}\text{tr}\left[V^{-1}(X-\mu)'U^{-1}(X-\mu)\right]\right\}$$

- $\mathbb{E}[X] = \mu$

U/V determines the variance among rows/columns of X (which gets scaled by the trace (sum of diagonal elements) of V/U) :

$$\mathbb{E}\left[(X-\mu)(X-\mu)'\right] = \text{Utr}[V], \quad \mathbb{E}\left[(X-\mu)'(X-\mu)\right] = \text{Vtr}[U].$$

Asymptotic Theory

Asymptotic theory

In econometrics, in the simplest case, we deal with a data sample $\{x_i\}_{i=1}^n$ of n observations (realizations) x_i.

We interpret $\{x_i\}_{i=1}^n$ as independent draws from some distribution $p(X_i \mid \theta)$ (i.e. the observations are i.i.d.).

The discussion up to now analyzes the moments and distribution of $\{x_i\}_{i=1}^n$ for a given sampel size n - finite sample properties.

We are also interested in the **asymptotic properties** as the sample size n grows to infinity.

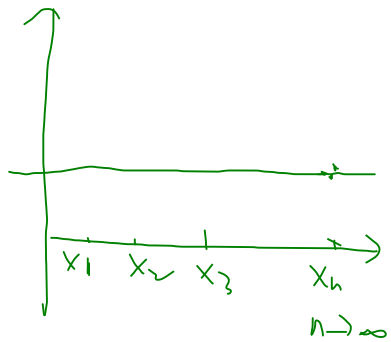
Converge in probability

Definition A sequence of scalar RVs $\{X_n\}_{n=1}^{\infty}$ is said to **converge in probability to a constant c** if

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} \mathbb{P}\left[|X_n - c| > \epsilon\right] = 0.$$

We write $X_n \xrightarrow{p} c$ or $\text{plim}\,(X_n) = c$.

In other words, the probability of being outside an interval of $\pm\epsilon$ around c converges to zero, for any $\epsilon > 0$, i.e. no matter how small we make this interval.



If $c = 0$, we say X_n is $o_p(1)$, i.e. $X_n \xrightarrow{p} 0$

Converge in probability

$\text{plim } X_n = 0 \Rightarrow \forall \varepsilon \quad \lim_{n\to\infty} P[|X_n - 0| > \varepsilon] = 0$

$\lim_{n\to\infty} P[X_n > \varepsilon] = \lim_{n\to\infty} e^{-n\varepsilon} = 0$

Example :

let X_n be exponentially distributed with parameter $n : X_n \sim \text{Exp}(n)$, with cdf $F(x) = 1 - e^{-nx}$ and domain $\mathbb{R}_{++}$.

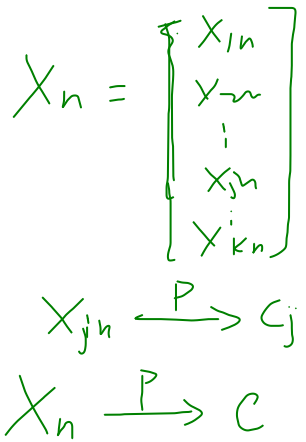
$\text{plim } X_n = 0 \quad P[X \le x] \quad P[X > x] = 1 - F(x)$

We get that $X_n \xrightarrow{p} 0$, i.e. X_n is $0_p(1)$:

$$\lim_{n\to\infty} \mathbb{P}[|X_n - 0| > \epsilon] = \lim_{n\to\infty} \mathbb{P}[X_n > \epsilon] = \lim_{n\to\infty} e^{-n\epsilon} = 0,$$

because $X_n > 0$ and $\mathbb{P}[X_n > \epsilon] = 1 - \mathbb{P}[X_n < \epsilon] = e^{-nx}$.

Converge in probability



Definition A sequence of k-dimensional RVs $\{X_n\}_{n=1}^{\infty}$ **converges in probability to a vector of constants** c if

$$X_{j,n} \xrightarrow{p} c_j \quad \text{for} \quad j = 1:k$$

where $X_{j,n}$ and c_j are the jth element of X_n and c, respectively.

In other words, a vector-valued RV converges in probability to a vector c if each of its elements converges in probability to the corresponding element of c.

Unless otherwise specified, the following results apply for scalar- as well as vector-valued RVs.

Slutsky's Theorem I

Proposition **Slutsky's Theorem I**

If $g : \mathbb{R}^k \to \mathbb{R}^m$ is continuous at c and independent of n, then

$$X_n \xrightarrow{p} c \Rightarrow g(X_n) \xrightarrow{p} g(c)$$

or in other words

$$\text{plim}\,(g(X_n)) = g(\text{plim}\,(X_n)) = g(c)$$

Example : Given that $X_n \xrightarrow{p} 0$ for $X_n \sim \text{Exp}(n)$ and that $g(x) = x^2$ is continuous at zero, we have $X_n^2 \xrightarrow{p} 0^2 = 0$

Slutsky's Theorem I

If $X_{1,n} \xrightarrow{p} c_1$ and $X_{2,n} \xrightarrow{p} c_2$, then

- $X_{1,n} + X_{2,n} \xrightarrow{p} c_1 + c_2$
- $X'_{1,n} X_{2,n} \xrightarrow{p} c'_1 c_2$
- $X_{1,n} X'_{2,n} \xrightarrow{p} c_1 c'_2$

Weak Law of Large Numbers (WLLN)

Proposition **Weak Law of Large Numbers (WLLN)**

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent RVs with finite means $\mathbb{E}[X_i] = \mu_i$.

Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i$

If $\mathbb{E}\left[|X_i|^{1+\delta}\right] < \Delta < \infty$ for some $\delta > 0$ and $\forall i$

Then

$$\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$$

If $\mu_i = \mu \ \forall i$ s.t. $\bar{\mu}_n = \mu$, we can write simply

$$\bar{Y}_n \xrightarrow{p} \mu$$

In other words, a sample average of independent RVs converges in probability to the average of population means, i.e. as $n \to \infty$, its distribution gets more and more concentrated

Plug-In Property

Proposition **Plug-In Property**

Suppose $X_n \xrightarrow{p} c$ and $g(X_n, Y_i)$ is a continuous function of X_n. Then

$$\frac{1}{n}\sum_{i=1}^{n} g(X_n, Y_i) \xrightarrow{p} \mathbb{E}[g(c, Y_i)]$$

Converge in distribution

Convergence in probability refers to the case when a sequence of RVs has a distribution that in the limit is ever more tightly concentrated around a point.

In contrast, convergence in distribution refers to the case when a sequence of RVs has a distribution that in the limit equals some specific distribution.

Converge in distribution

Definition Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of RVs and let $F_{X_n}(x)$ denote the cdf of X_n. Suppose $\exists a$ cdf F_X such that

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \forall x \text{ at which } F_X(x) \text{ is continuous}$$

Then X_n is said to **converge in distribution to** X.
We write $X_n \xrightarrow{d} X$.

RV distribution

Probability C

Constant

Converge in distribution

Example :

$$F_{X_n}(x) = P[X_n \le x] = P[n(Y_n - 1) \le x] = P\left[Y_n \le \frac{x}{n} + 1\right] = F_{Y_n}\left(\frac{x}{n} + 1\right)$$

let Y_n be Pareto-distributed with cdf $F(y) = 1 - y^{-n}$ and support $y \ge 1$, and define $X_n = n(Y_n - 1)$.

$$= 1 - \left(\frac{x}{n} + 1\right)^{-n}$$

We have $X_n \xrightarrow{d} \text{Exp}(1)$ because

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

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

$$F_{X_n}(x) = \mathbb{P}[X_n \le x] = \mathbb{P}[Y_n \le 1 + x/n] = 1 - \frac{1}{(1 + x/n)^n},$$

and

$$\lim_{n\to\infty} 1 - \frac{1}{(1 + x/n)^n} = 1 - 1/e^x = 1 - e^{-x}$$

which is the cdf of Exp(1), the standard exponential distribution.

Slutsky's Theorem II

Proposition **Slutsky's Theorem II**

Suppose $X_n \xrightarrow{p} c$ and $Y_n \xrightarrow{d} Y$. Then

$$X_n + Y_n \xrightarrow{d} Y + c$$

and

$$X_n' Y_n \xrightarrow{d} c' Y$$

Continuous Mapping Theorem

Proposition **Continuous Mapping Theorem**

Suppose $X_n \xrightarrow{d} X$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ is a continuous function and independent of n. Then

$$g(X_n) \xrightarrow{d} g(X)$$

Continuous Mapping Theorem

Example : Continuing the example from above, $X_n \xrightarrow{d} \text{Exp}(1)$

$$\mu - \sigma \log\left(\frac{\exp\{-X_n\}}{1-\exp\{-X_n\}}\right) \xrightarrow{d} \mu - \sigma \log\left(\frac{\exp\{-X\}}{1-\exp\{-X\}}\right) \qquad X \sim \text{Exp}(1)$$

this limit-RV follows a logistic distribution with parameters μ and σ.

Liapunov's Central Limit Theorem (CLT)

Proposition **Liapunov's Central Limit Theorem (CLT)**

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent scalar RVs with $\mathbb{E}[X_i] = \mu_i$ and $\mathbb{V}[X_i] = \sigma_i^2 > 0$.

Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mu_i$ and $\bar{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2$.

If the following conditions are met :

1. $\mathbb{E}\left[|X_i - \mu_i|^{2+\delta}\right] < \Delta < \infty$ for some $\delta > 0$ and $\forall i$
2. $\bar{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 > \eta > 0 \quad \forall n$ large

Then

$$\frac{\sqrt{n}\left(\bar{X}_n - \bar{\mu}_n\right)}{\bar{\sigma}_n} \xrightarrow{d} N(0,1)$$

Liapunov's Central Limit Theorem (CLT)

This result can be thought of as $\bar{X}_n \xrightarrow{d} N\left(\bar{\mu}_n, \frac{1}{n}\bar{\sigma}_n^2\right)$, but this statement is ill-defined because it would say that the limit distribution as $n \to \infty$, is a function of n.

If $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2 \forall$, then we can write

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N\left(0, \sigma^2\right)$$

If in addition $\mu = 0$, we can write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N\left(0, \sigma^2\right)$$

Lindeberg-Lévy CLT

The following CLT holds for vector-valued RVs, but it requires the RVs to be identically distributed :

Proposition **Lindeberg-Lévy CLT**

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of k-dimensional i.i.d. RVs with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \Sigma$. Then

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \Sigma).$$

If $\mu = 0$, we can write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N(0, \Sigma)$$

Delta Method

Proposition **Delta Method**

If $\sqrt{n}(X_n - c) \xrightarrow{d} X$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ is continuously differentiable, then

$$\sqrt{n}(g(X_n) - g(c)) \xrightarrow{d} G \cdot X$$

where $G = \left.\frac{\partial g(x)}{\partial x'}\right|_{x=c}$ is a $m \times k$ matrix.

Example : If $\sqrt{n}\sum_{i=1}^{n} X_i \xrightarrow{d} N(0, \sigma^2)$, then the Delta method tells us that

$g(x) = \exp(2x)$ $\quad G = \left.\frac{\partial g(x)}{\partial x}\right|_{x=c=0} = 2\exp(2x)\Big|_{x=0} = 2$

$$\sqrt{n}\sum_{i=1}^{n} \exp\{2X_i\} \xrightarrow{d} 2\exp\{2 \cdot 0\} N(0, \sigma^2) = N(0, 4\sigma^2)$$

Asymptotic Theory $\{X_i\}_{i=1}^n \quad n \to \infty$

converge in probability $X_n \xrightarrow{P} c$ (constant)

converge in distribution $X_n \xrightarrow{d} X$ (distribution)

X_i scalar case

Slutsky's Theorem I

$$X_n \xrightarrow{P} c \Rightarrow g(X_n) \xrightarrow{P} g(c)$$

$$X_{1n} \xrightarrow{P} c_1 \quad X_{2m} \xrightarrow{P} c_2 \Rightarrow X_{1n} \pm X_{2m} \xrightarrow{P} c_1 + c_2$$

$$X_{1n} X_{2m} \xrightarrow{P} c_1 c_2$$

Slutsky's Theorem II

$$X_n \xrightarrow{P} c \quad Y_n \xrightarrow{d} Y \Rightarrow X_n + Y_n \xrightarrow{d} Y + c$$

$$X_n Y_n \xrightarrow{d} cY$$

Continuous mapping theorem

$$X_n \xrightarrow{d} X \Rightarrow g(X_n) \xrightarrow{d} g(X)$$

WLLN (converge in probability)

$\{X_i\}_{i=1}^n \quad E[X_i] = \mu_i$

$\Rightarrow \bar{X}_n - \bar{\mu}_n \xrightarrow{P} 0$

if $E[X_i] = \mu$

$\Rightarrow \bar{X}_n - \mu \xrightarrow{P} 0$

$\bar{X}_n \xrightarrow{P} \mu$

Lindeberg-Lévy
Lapunov CLT (converge in distribution)

$\{X_i\}_{i=1}^n$ scalar Vector but i.i.d

$E[X_i] = \mu_i \quad V[X_i] = \sigma_i^2$

$E[X_i] = \mu \quad V[X_i] = \Sigma$

$\Rightarrow \dfrac{\sqrt{n}(\bar{X}_n - \bar{\mu}_n)}{\bar{\sigma}_n} \xrightarrow{d} N(0,1)$

if $E[X_i] = \mu \quad V[X_i] = \sigma^2$

$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$

if $\mu = 0 \quad \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma)$

$\dfrac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{d} N(0, \sigma^2)$

$\dfrac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{d} N(0, \Sigma)$

plug-in property

$$X_n \xrightarrow{P} c \Rightarrow \frac{1}{n}\sum_{i=1}^{n} g(X_n, Y_i) \xrightarrow{P} E[g(c, Y_i)]$$

Delta method

$$\sqrt{n}(X_n - c) \xrightarrow{d} X \Rightarrow \sqrt{n}(g(X_n) - g(c)) \xrightarrow{d} \left.\frac{\partial g(x)}{\partial x}\right|_{x=c} \cdot X$$

Statistical Inference

Statistical Inference

In econometrics, we interpret data x as draws of a random variable X from probability distribution $p(X \mid \theta)$ indexed by a parameter θ.

Key problem : Suppose we observe a particular realization x. What can we say about θ ?

Statistical Inference

Notations in this chapter :

Data sample $\{x_i\}_{i=1}^n$: n observations (realizations) x_i of the same underlying RV X_i which are drawn independently from some distribution $p(x \mid \theta)$ (i.e. the observations are i.i.d.).

We take our observations together into the $n \times 1$ vector $x = (x_1, \ldots, x_n)'$.

For simplicity, we assume that θ and X_i (and hence x_i) are scalars.

We will predominantly deal with continuous random variables and write $p(x)$ for the pdf, $F(x)$ or $\mathbb{P}[X \leq x]$ for the cdf.

Point Estimation

Point Estimation

RV parameter unkown, we wanna estimate

$X \sim p(x|\theta)$

RV

$\hat{\theta} = \delta(X)$ estimator

rule of estimation

Definition A **point estimator** $\delta(X)$ is a mapping from the sample space of X to the parameter space $\Theta : \delta : \mathcal{X} \to \Theta$.

We denote the point estimator as $\hat{\theta}$, $\hat{\theta} = \delta(X)$ is a function of the data X.

~~x~~ realization (data)

$\hat{\theta} = \delta(x)$ estimate

If we evaluate the point estimator at the particular realization x we observe, we call the result an **point estimate**.

Loosely speaking, the **point estimator** is a rule how we form our best guess based on a RV X, whereas the **point estimate** is the best guess we obtain given a particular observation x of X.

Point Estimation

There are different estimators that exploit different aspects of $p(x \mid \theta)$.

- **Frequentist/Classical Inference**
  Treat θ as a fixed but unknown parameter.
  $\Rightarrow$ *Least Squares (LS) estimator, Method of Moments (MM) estimator, Maximum Likelihood (ML) estimator*

$X \sim N(\theta, 1)$

$p(x|\theta)$

not RV

- **Bayesian Inference**
  Treat θ as a RV. Update prior belief to posterior belief after observing the sample x.
  $\Rightarrow$ *Bayesian estimator*

$\theta \sim$ distribution

$p(\theta)$

Least Squares

Assume the conditional mean $\mathbb{E}[X_i \mid \theta]$ is equal to θ, i.e. we do not need to know the full distribution of $X_i \mid \theta$.

The **Least Squares (LS) estimator** minimizes the sum of squared deviations of x_i from their conditional mean $\mathbb{E}[X_i \mid \theta] = \theta$:

$$\hat{\theta}_{LS} = \arg\min_{\theta \in \Theta} \sum_{i=1}^{n} (x_i - \theta)^2$$

this leads to

$$\hat{\theta}_{LS} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Method of Moments

sample moments = population moments $\begin{cases} \text{1st} \\ \text{2nd} \\ \vdots \\ \text{nth} \end{cases}$

① $X_i | \theta \sim Dist(\theta, \square)$

$\{x_i\}_{i=1}^n$ data

$$\frac{1}{n}\sum x_i = \hat{\theta}_{MM} \Rightarrow \hat{\theta}_{MM} = \frac{1}{n}\sum x_i$$

② $X_i | \theta \sim Dist(\theta, 1)$

$\{x_i\}_{i=1}^n$

population second moment $E[X_i^2|\theta] = V[X_i|\theta] + E[X_i|\theta]^2 = 1 + \theta^2$

||

sample second moment $\frac{1}{n}\sum X_i^2 = 1 + \hat{\theta}_{MM}^2 \Rightarrow \hat{\theta}_{MM} = \sqrt{\frac{1}{n}\sum X_i^2 - 1}$

$\theta = (\mu, \sigma)$

$X_i \mid \theta \sim \text{Dist}(\mu, \sigma^2)$

$\{x_i\}_{i=1}^n$

First moment

Sample = population

$\frac{1}{n}\sum X_i = \hat{\mu}_{MM}$ (1)

Second moment

Sample = Population

$\frac{1}{n}\sum X_i^2 = E[X_i^2 \mid \theta] = V[X_i \mid \theta] + E[X_i \mid \theta]^2 = \hat{\mu}^2_{MM} + \hat{\sigma}^2_{MM}$ (2)

$\Rightarrow \hat{\mu}_{MM} = \frac{1}{n}\sum X_i$

$\hat{\sigma}_{MM} = \sqrt{\frac{1}{n}\sum X_i^2 - \left(\frac{1}{n}\sum X_i\right)^2}$

Method of Moments

The **Method of Moments (MM) estimator** solves for $\hat{\theta}_{MM}$ by setting empirical moments of $\{x_i\}_{i=1:n}$ equal to the corresponding population moments in $p(x \mid \theta)$.

For example, let our RV X_i and parameter θ be related as follows : $X_i \mid \theta \sim Dist(\theta, 1)$

Using the mean, we get the same result as under LS, i.e.

$$\mathbb{E}\left[X_i \mid \hat{\theta}_{MM}\right] = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \Rightarrow \quad \hat{\theta}_{MM} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Method of Moments

Under $\mathbb{V}[X \mid \theta] = 1$, we also have

$$\mathbb{E}\left[X_i^2 \mid \hat{\theta}\right] = \mathbb{V}\left[X_i \mid \theta\right] + \mathbb{E}\left[X_i \mid \hat{\theta}\right]^2 = 1 + \theta^2$$

and we can define $\hat{\theta}_{MM}$ also as the quantity that equalizes the second moment :

$$\mathbb{E}\left[X_i^2 \mid \hat{\theta}_{MM}\right] = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \Rightarrow \hat{\theta}_{MM} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - 1}$$

Method of Moments

Similarly to LS, the MM estimator does not require the researcher to specify the whole distribution of $X_i \mid \theta$, but only enough moments so that we can solve for θ.

If θ were two dimensional (and X_i still a scalar), then we would need to specify two (scalar) moments, e.g. $\mathbb{E}[X_i \mid \theta]$ and $\mathbb{V}[X_i \mid \theta]$.

Exercise Solve for the MM estimator $\hat{\theta} = (\hat{\mu}_{MM}, \hat{\sigma}^2_{MM})$ if we assume $X_i \mid \theta \sim Dist(\mu, \sigma^2)$

Maximum Likelihood

The **Maximum Likelihood (ML) estimator** maximizes the likelihood of observing the sample $x = (x_1, \ldots, x_n)'$ out of all the possible draws of $X = (X_1, \ldots, X_n)'$:

$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta \mid x) = \arg\max_{\theta \in \Theta} \ell(\theta \mid x),$$

where $\mathcal{L}(\theta \mid X) = p(X \mid \theta)$ is the pdf of the RV $X \mid \theta$ and $\ell(\theta \mid X) = \log \mathcal{L}(\theta \mid X)$ is its log.

$X_i \mid \theta \sim N(\theta, 1)$ i.i.d

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

$\{x_i\}_{i=1}^n$

likelihood we observe $L(\theta \mid x) = \prod_{i=1}^n f(x_i \mid \theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{1}{2}(x_i - \theta)^2\right\}$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{\frac{1}{2}\sum (x_i - \theta)^2\right\}$$

take log $\ell(\theta \mid x) = \log L(\theta \mid x) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2$

$\hat{\theta}_{ML} = \arg\max_\theta \ell(\theta \mid x) = \frac{1}{n}\sum x_i$

Maximum Likelihood

For example, under $X_i \mid \theta \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, we get

$$\mathcal{L}(\theta \mid x) = p(x \mid \theta) = \prod_{i=1}^{n} p(x_i \mid \theta) = \prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} \exp\left\{ \frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2 \right\}$$

this leads to

$$\ell(\theta \mid x) = -\frac{n}{2} \ln(2\pi) + \frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2$$

which, once again, gives $\hat{\theta}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Maximum Likelihood

ML estimator requires assuming the distribution (the pdf) of the underlying $X_i \mid \theta$.

Invariance property : if $\hat{\theta}$ is the ML estimator of θ, then $f(\hat{\theta})$ is the ML estimator of $f(\theta)$.

Finite Sample Properties

Definition A point estimator $\delta(X)$ of θ is **unbiased** if $\mathbb{E}[\delta(X) \mid \theta] = \theta$.

For example, under $X_i \mid \theta \sim N(\theta, 1)$, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is unbiased :

$$\mathbb{E}[\hat{\theta} \mid \theta] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i \mid \theta\right] = \frac{1}{n}\sum_{i=1}^{n} \theta = \theta$$

In contrast, the estimator $\hat{\theta}_* = \frac{1}{n-1}\sum_{i=1}^{n} X_i$ is not unbiased :

$$\mathbb{E}\left[\hat{\theta}_* \mid \theta\right] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n} X_i\right] = \mathbb{E}\left[\frac{n}{n-1}\hat{\theta}\right] = \frac{n}{n-1}\theta \neq \theta$$

Finite Sample Properties

We can also compute

$$\mathbb{V}[\hat{\theta} \mid \theta] = \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i \mid \theta] = \frac{1}{n^2}\sum_{i=1}^{n} 1 = \frac{1}{n}$$

and

$$\mathbb{V}\left[\hat{\theta}_* \mid \theta\right] = \mathbb{V}\left[\frac{n}{n-1}\hat{\theta}\right] = \frac{n^2}{(n-1)^2}\mathbb{V}[\hat{\theta}] = \frac{n}{(n-1)^2}$$

The quantity $\mathbb{E}[\hat{\theta} \mid \theta]$ tells us the average value we expect to get if we were to randomly draw different samples $x = (x_1, \ldots, x_n)'$ of size n and compute the point estimator $\hat{\theta} = \delta(x)$ for these samples.

The quantity $\mathbb{V}[\hat{\theta} \mid \theta]$ tells us the dispersion we would get in all of these point estimators.

Finite Sample Properties

Provided that $X_i \mid \theta$ is Normally distributed, we know that $\hat{\theta}$ is itself Normally distributed, because it is the average of Normally distributed RVs. We get

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i \sim N\left(\theta, n^{-1}\right).$$

normal holds if $X_i|\theta$ is normal

Asymptotic Properties

We can also analyze the properties of a point estimator $\hat{\theta}$ as the sample size $n \to \infty$ (large samples).

Definition point estimator $\delta(X)$ of θ is **consistent** if $\delta(X) \xrightarrow{p} \theta$.

We know that $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is consistent, because by the WLLN

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X_i \mid \theta] = \theta$$

The estimator $\hat{\theta}_* = \frac{1}{n-1}\sum_{i=1}^{n} X_i$ is also consistent by the WLLN and Slutsky theorem

$$\hat{\theta}_* = \frac{1}{n-1}\sum_{i=1}^{n} X_i = \frac{n}{n-1}\hat{\theta} \xrightarrow{p} \theta$$

Asymptotic Properties

$X_i | \theta \sim Dist(\theta, 1)$

Based on the CLT, we also know the asymptotic distribution of $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$:

$$\sqrt{n}(\hat{\theta} - \theta) \mid \theta \xrightarrow{d} N(0, 1),$$

While for the finite sample distribution we had to know the distribution of the data $X_i \mid \theta$, we found this asymptotic distribution under minimal assumptions : i.i.d. of our observations $\{x_i\}_{i=1}^{n}$.

If we do not know the distribution of the data, we can approximate the finite sample distribution based on the asymptotic one, by saying that

$$\sqrt{n}(\hat{\theta}-\theta) \mid \theta \overset{\text{approx.}}{\sim} N(0, 1) \quad \text{or, equivalently, } \hat{\theta} \mid \theta \overset{\text{approx.}}{\sim} N\left(\theta, n^{-1}\right)$$

Hypothesis Testing

$y = \alpha + \beta x$

$\hat{\beta}$ $H_0 : \beta = 0$ Reject H_0, β significantly non-zero

$H_1 : \beta \neq 0$

Hypothesis Testing

Often, we are interested in testing whether θ lies in some set (interval) $\Theta_0 \subset \Theta$.

We test the zero hypothesis

$$\mathscr{H}_0 = \theta \in \Theta_0 = \{\theta_1, \theta_2, \theta_3\}$$

against the alternative hypothesis

$$\mathscr{H}_1 : \theta \in \Theta_1$$

. Often, Θ_1 is simply the complement of Θ_0.

If Θ_0 contains a single value θ_0, then $\mathscr{H}_0 : \theta = \theta_0$ is called a **point hypothesis**, otherwise it is called a **composite hypothesis** (and analogously for $\mathscr{H}_1$).

Hypothesis Testing

Definition A **hypothesis test** $\varphi \in \{0, 1\}$ is a rule that specifies when we reject and when we accept (do not reject) $\mathcal{H}_0$, with $\varphi = 0$ indicating rejection.

A hypothesis test can be viewed as a point estimator of the object $\mathbf{1}\{\theta \in \Theta_0\}$.

Testing procedures φ can be **evaluated** using the following loss function :

$$L(\theta, \varphi) = \begin{cases} 1 & \text{if } \varphi \neq \mathbf{1}\{\theta \in \Theta_0\} \quad \text{(i.e. we are wrong)} \\ 0 & \text{otherwise} \qquad\qquad\quad \text{(i.e. we are right)} \end{cases}.$$

$H_0: \beta = 0$
$H_1: \beta \neq 0$

$H_0: \theta = \theta_0$
$H_1: \theta \neq \theta_0$

	H_0 is true	H_0 is false
accept H_0	✓ $1-\alpha$	× $1-\beta$ → type II error
reject H_0	× α	✓ β

α → size → type I error

β → power

Hypothesis Testing

This loss function leads to the following frequentist risk $R(\theta, \varphi) = \mathbb{E}[L(\theta, \varphi(X)) \mid \theta]$ which depends on θ :

- if $\theta \in \Theta_0$, then $R(\theta, \varphi) = \mathbb{P}[$ reject $\mathscr{H}_0 \mid \mathscr{H}_0$ is true $] \equiv \alpha(\theta)$ (**type I error**)
- if $\theta \in \Theta_1$, then $R(\theta, \varphi) = \mathbb{P}[$ accept $\mathscr{H}_0 \mid \mathscr{H}_0$ is false $] \equiv 1 - \beta(\theta)$ (**type II error**)

We call $\beta(\theta) = \mathbb{P}[$ reject $\mathscr{H}_0 \mid \mathscr{H}_0$ is false $]$ as the **power** of the test φ.

Note that $\alpha(\theta)$ and $\beta(\theta)$ are in fact the same function defined on the whole sample space Θ :

$$\beta(\theta) = \mathbb{P}[\text{ reject } \mathscr{H}_0 \mid \theta] = \mathbb{P}[\varphi = 0 \mid \theta]$$

People just write it as $\alpha(\theta)$ if $\theta \in \Theta_0$, while they write it as $\beta(\theta)$ if $\theta \in \Theta_1$.

Hypothesis Testing

A good test has a power function near 1 for $\theta \in \Theta_1$ and near 0 for $\theta \in \Theta_0$.

Usually, we put a constraint on the type I error by specifying a size requirement and then search for the test with the highest power.

Definition For $\alpha \in (0,1)$, a test is a **size α test** if $\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$ ($\beta(\theta_0) = \alpha$ for point hypothesis), i.e. if its type I error is equal to α.

Hypothesis Testing

There are different tests that use different **test-statistics**. The generic form is

$$\varphi(x;\alpha) = \mathbf{1}\{T(x) < c_\alpha\}$$

i.e. we reject if our realized test-statistic $T(x)$ is larger than some critical value c_α, whereby different choices of the size α translate into different values for c_α.

Under this form, the power function becomes
$\beta(\theta) = \mathbb{P}[T(x) \geq c_\alpha \mid \theta]$.

t-test

Definition Suppose $\hat{\theta} \mid \theta \sim N\left(\theta, v^2\right)$ and we are testing a point hypothesis $\mathcal{H}_0 : \theta = \theta_0$. Under the alternative $\mathscr{H}_1 : \theta \neq \theta_0$, the **two-sided t-test** is

$$\varphi_t(x) = \mathbf{1}\left\{\left|\frac{\hat{\theta} - \theta_0}{v}\right| < c\right\}$$

Under the alternative $\mathscr{H}_1 : \theta > \theta_0$, the one-sided t-test is

$$\varphi_t(x) = \mathbf{1}\left\{\frac{\hat{\theta} - \theta_0}{v} > c\right\}$$

while under the alternative $\mathscr{H}_1 : \theta < \theta_0$, the (one-sided) t-test is

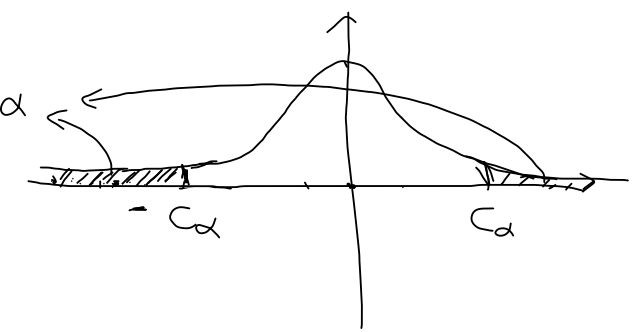
$$\varphi_t(x) = \mathbf{1}\left\{\frac{\hat{\theta} - \theta_0}{v} < c\right\}$$

$H_0 : \theta = \theta_0$

$H_1 : \theta \neq \theta_0$

$\hat{\theta} \mid \theta \sim N(\theta, V^2)$

under $H_0 : \theta = \theta_0 \Rightarrow \hat{\theta} \mid \theta \sim N(\theta_0, V^2)$



type I error $\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true})$

$$t\text{-statistic} = \frac{\hat{\theta} - \theta_0}{V} \overset{H_0}{\sim} N(0, 1)$$

t-test

For example, let $X \mid \theta \sim N(\theta, 1)$, and assume for simplicity that we have a single observation x

Then $\hat{\theta} = x$ has the distribution $\hat{\theta}|\theta = X|\theta \sim N(\theta, 1)$

If $\mathscr{H}_0$ is true, then this distribution is $N(\theta_0, 1)$ and therefore $\frac{\hat{\theta}-\theta_0}{1} \sim N(0, 1)$.

The two-sided t-test rejects $\mathscr{H}_0$ if this statistic is (on either side) too far from its supposed mean of zero (i.e. if x is too far from its supposed mean θ_0) and accepts otherwise.

t-test

To find c, we set the type I error equal to te desired type I error α :

$$\begin{aligned}
\alpha &= \beta(\theta_0) \\
&= \mathbb{P}[\varphi_t = 0 \mid \theta = \theta_0] \\
&= 1 - \mathbb{P}\left[-c \leq \hat{\theta} - \theta_0 \leq c \mid \theta = \theta_0\right] \\
&= 1 - \mathbb{P}[-c \leq Z \leq c \mid \theta = \theta_0] \\
&= 1 - [\Phi(c) - \Phi(-c)] \\
&= 2(1 - \Phi(c))
\end{aligned}$$

To get a test of size 10%($\alpha = 0.1$), we take $c = 1.64$. Different values of c yield t-tests of different sizes α.

t-test

If $\mathscr{H}_0$ is false. If the true value of θ is some $\tilde{\theta}$, then $\hat{\theta} \sim N(\tilde{\theta}, 1)$ and so $\frac{\hat{\theta}-\theta_0}{1} \sim N\left(\tilde{\theta}-\theta_0, 1\right)$.

The power of this test is

$$\begin{aligned}
\beta(\tilde{\theta}) &= \mathbb{P}\left[\varphi_t = 0 \mid \theta = \tilde{\theta}\right] \\
&= 1 - \mathbb{P}\left[-c \leq \hat{\theta} - \theta_0 \leq c \mid \theta = \tilde{\theta}\right] \\
&= 1 - \mathbb{P}\left[-c - \left(\tilde{\theta} - \theta_0\right) \leq Z \leq c - \left(\tilde{\theta} - \theta_0\right) \mid \theta = \tilde{\theta}\right] \\
&= 1 - \left[\Phi\left(c - \left(\tilde{\theta} - \theta_0\right)\right) - \Phi\left(-c - \left(\tilde{\theta} - \theta_0\right)\right)\right]
\end{aligned}$$

t-test

trade-off α & β

small α small type I error $P[\text{reject } H_0 \mid H_0 \text{ true}]$ small

$\Downarrow$

small β small power $P[\text{reject } H_0 \mid H_0 \text{ false}]$ small

Under H_0

$t\text{-stat} \sim N(0,1)$

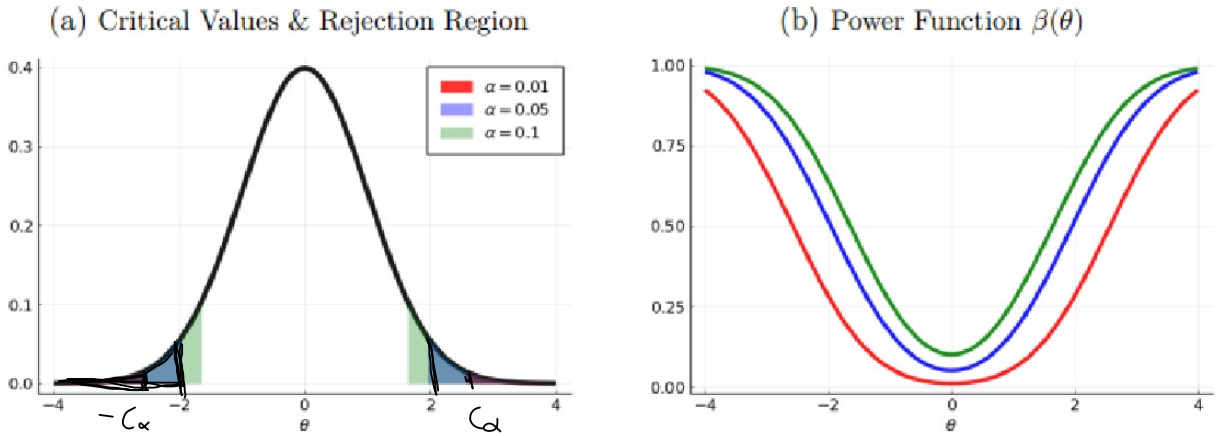


Figure 2.1: Two-Sided t-Tests

Notes: Illustration of the critical values, rejection region and power function for a two-sided t-test with sizes $\alpha \in \{0.01, 0.05, 0.1\}$.

$H_0: \beta = 0$

$H_1: \beta \neq 0$

reject H_0

Likelihood Ratio Test

Definition The **likelihood ratio (LR) test** is

$$\varphi_{LR}(x) = \mathbf{1}\left\{\frac{\sup_{\theta\in\Theta_1} p(x \mid \theta)}{\sup_{\theta\in\Theta_0} p(x \mid \theta)} < c\right\}$$

For a point zero-hypothesis $\mathscr{H}_0 : \theta = \theta_0$ with the alternative $\mathscr{H}_1 : \theta \neq \theta_0$, the LR test simplifies to

$$\varphi_{LR}(x) = \mathbf{1}\left\{\frac{p\left(x \mid \hat{\theta}_{ML}\right)}{p\left(x \mid \theta_0\right)} < c\right\}$$

(Under continuous parameter spaces like $\mathbb{R}$, the probability that $\hat{\theta}_{ML}$ is exactly equal to θ_0 is zero.)

Likelihood Ratio Test

Consider the example before, $X \sim N(\theta, 1)$, with single realization x, testing $\mathscr{H}_0 : \theta = \theta_0$ against $\mathscr{H}_1 : \theta \neq \theta_0$. Then $\hat{\theta} = x$. We get

$$T(x) = \frac{p\left(x \mid \hat{\theta}_{ML}\right)}{p\left(x \mid \theta_0\right)} = \frac{(2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x-\hat{\theta})^2\right\}}{(2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x-\theta)^2\right\}} = \exp\left\{\frac{1}{2}\left(\hat{\theta}-\theta_0\right)^2\right.$$

Define $\tilde{T}(x) = 2 \cdot \ln\left(\frac{p\left(x|\hat{\theta}_{ML}\right)}{p(x|\theta_0)}\right) = (x-\theta_0)^2$ and $\tilde{c} = 2\ln(c)$. If $\mathscr{H}_0$ is true, $X - \theta_0 \sim N(0,1)$ and so $(X-\theta_0)^2 \sim \chi^2_1$.

We then find $\tilde{c}$ from Chi-squared distribution with one degree of freedom so as to set the type I error to α :

$$\mathbb{P}\left[T(X) \geq c \mid \mathscr{H}_0\right] = \mathbb{P}\left[\tilde{T}(X) \geq \tilde{c} \mid \mathscr{H}_0\right] = \alpha$$

Likelihood Ratio Test

Definition A test φ_α with size α and power function $\beta(\theta)$ is a **uniformly most powerful** size α test if it maximizes the power uniformly on Θ_1 among all tests with size α, i.e. if $\beta(\theta) \geqslant \beta'(\theta)$ for all $\theta \in \Theta_1$ and for all power functions $\beta'(\theta)$ of size α tests φ'.

The Neyman-Pearson Lemma (beyond scope) shows that if both $\mathscr{H}_0$ and $\mathscr{H}_1$ are point hypotheses, the LR test is uniformly most powerful.

Hypothesis Testing

Conduct a hypothesis test :

- Choose the test-type.
- Choose the desired size of the test α.
- Find the right critical value
  - equal $\mathbb{P}[T(X) > c \mid \mathcal{H}_0]$ to α.
  - If we cannot find the distribution of $T(x)$, we can apply a monotonic transformation of both $T(x)$ and c.
  - If we cannot proceed analytically, then we do numerically : we repeatedly draw $X \mid \theta$ from its supposed distribution under $\mathscr{H}_0$ and simulate the distribution of $T(X)$.

Hypothesis Testing

Definition Take a test $\varphi(x;\alpha)$ based on a test statistic $T(x)$ and let α be its size. Then ***p*-value** $= \sup \alpha$ s.t. $\varphi(x;\alpha) = 1$, i.e. we accept $\mathscr{H}_0$.

(a) Two-sided test with $T(X) \in \mathbb{R}$ (b) Test with $T(X) \in \mathbb{R}_{++}$

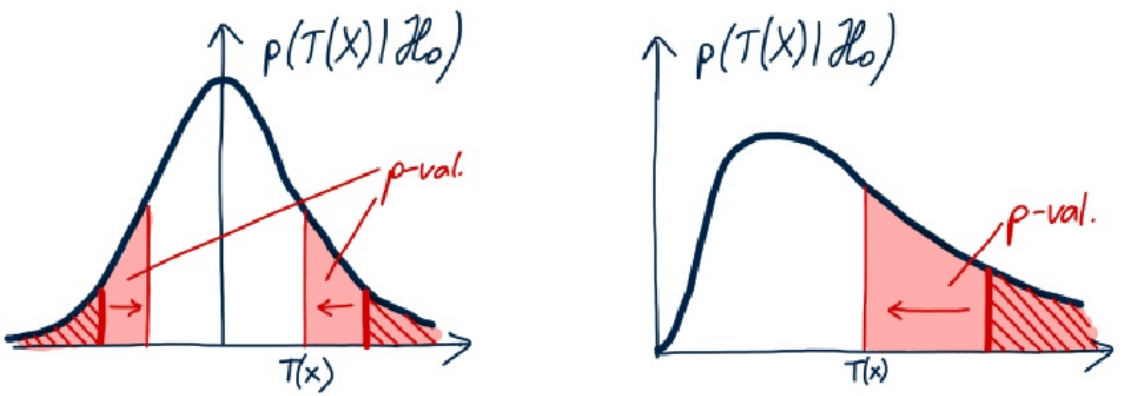


Figure 2.2: p-Values

Notes: Illustration of the critical values with corresponding rejection region (thick red line) and p-values (red area) for two different tests; left the two-sided t.test and right a test with positive support of the test-statistic (e.g. the LR test in the example in the text, where $T(X) \sim \chi^2$ under $\mathscr{H}_0$).

p-value is often used as a decision rule ; we reject if the *p*-value is smaller than a desired size α, and accept otherwise.

Hypothesis Testing

Example $X \sim N(\theta, 1)$, and we test $\mathscr{H}_0 : \theta = \theta_0$ vs. $\mathscr{H}_1 : \theta \neq \theta_0$ using the test $\varphi(X; \alpha) = \mathbf{1}\{T(X) < c_\alpha\}$ based on the statistic $T(X) = \left|\hat{\theta} - \theta_0\right|$, with $\hat{\theta} = X$. Then

$$\begin{aligned} p &= \sup_\alpha \mathbb{P}\left[\varphi(X; \alpha) = 0 \mid \mathcal{H}_0\right] \\ &= \sup_\alpha \mathbb{P}\left[|X - \theta_0| > c_\alpha \mid \mathscr{H}_0\right] \\ &= \sup_\alpha 2 \cdot \mathbb{P}\left[X - \theta_0 > c_\alpha \mid \mathscr{H}_0\right] \\ &= \sup_\alpha 2\left(1 - \Phi\left(c_\alpha\right)\right). \end{aligned}$$

This expression decreases with c_α. Therefore, the smallest c_α we can take s.t. we still accept $\mathcal{H}_0$ is $c_\alpha = T(x) = |x - \theta_0|$. We obtain $p = 2\left(1 - \Phi\left(|x - \theta_0|\right)\right)$.

Confidence Sets

$X|\theta \sim N(\theta, \sigma^2)$ θ is unkown σ^2 is known

$\{x_i\}_{i=1}^n$ i.i.d

LS/MM/ML $\hat{\theta} = \frac{1}{n}\sum x_i \sim N(\theta, \frac{1}{n}\sigma^2)$

$H_0: \theta = \theta_0 \Rightarrow \hat{\theta} \sim N(\theta_0, \frac{1}{n}\sigma^2)$

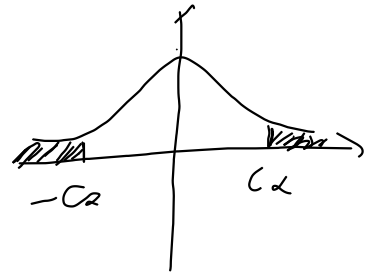
$H_1: \theta \neq \theta_0$

① construct test - statistics $T(\hat{\theta})$

✓ { t-test $\frac{\hat{\theta}-\theta_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$

LR-test $\left(\frac{\hat{\theta}-\theta_0}{\sigma/\sqrt{n}}\right)^2 \overset{H_0}{\sim} \chi_1^2$

② choose α e.g. 10% 5% 1% $\Rightarrow C_\alpha$



③ Decision rule

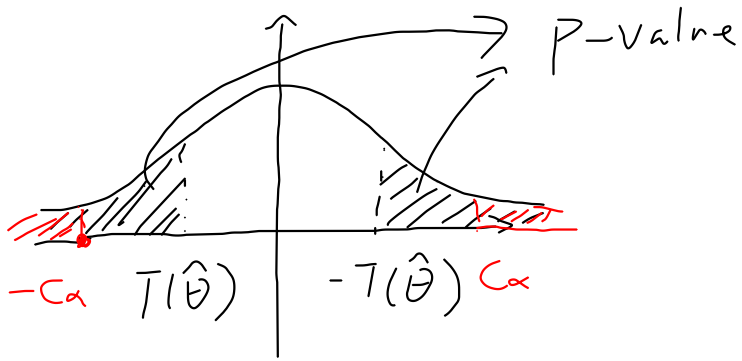
A $\left|T^t(\hat{\theta})\right| = \left|\frac{\hat{\theta}-\theta_0}{\sigma/\sqrt{n}}\right| < C_\alpha^t$ accept H_0

$> C_\alpha^t$ reject H_0

$$\left|T^{LR}(\hat{\theta})\right| = \left(\frac{\hat{\theta}-\theta_0}{\sigma/\sqrt{n}}\right)^2 < C_\alpha^t \quad \text{accept } H_0$$

$$> C_\alpha^t \quad \text{reject } H_0$$

B Calculate P-value



P-value $> \alpha$ accept H_0

$< \alpha$ reject H_0

[L] Calculate confidence intervals

t-test $-C_\alpha^t < \frac{\hat{\theta}-\theta_0}{\sigma/\sqrt{n}} < C_\alpha^t \Rightarrow \hat{\theta} - \frac{\sigma}{\sqrt{n}} C_\alpha^t < \theta_0 < \hat{\theta} + \frac{\sigma}{\sqrt{n}} C_\alpha^t$ accept H_0

$\left[\hat{\theta} - \frac{\sigma}{\sqrt{n}} C_\alpha^t,\ \hat{\theta} + \frac{\sigma}{\sqrt{n}} C_\alpha^t\right]$

LR-test $\left(\frac{\hat{\theta}-\theta_0}{\sigma/\sqrt{n}}\right)^2 < C_\alpha^{LR} \Rightarrow \hat{\theta} - \frac{\sigma}{\sqrt{n}}\sqrt{C_\alpha^{LR}} < \theta_0 < \hat{\theta} + \frac{\sigma}{\sqrt{n}}\sqrt{C_\alpha^{LR}}$

accept H_0

Confidence Sets

A **confidence set** $C(X) \subseteq \Theta$ is a (random) set that should cover the true θ with a prespecified probability :

$$\inf_{\theta \in \Theta} \mathbb{P}[\theta \in C(X) \mid \theta] = 1 - \alpha$$

For a scalar θ, we speak of confidence intervals.

The probability is taken with respect to the RV X, with θ fixed. Hence, if we look at many different, random realizations of X, the set $C(X)$ should contain the true θ in $100(1-\alpha)\%$ of cases (even if the true θ is such that this probability is lowest).

Confidence Sets

We can construct $C(X)$ based on a test φ, testing the point-hypothesis $\mathscr{H}_0 : \theta = \theta_0$.

Let $C(x)$ contain all the values for θ_0 that we would accept given our realization x :

$$C(x) = \{\theta_0 \in \Theta : \varphi(x; \theta_0) = 1\}$$

Then we can get the desired coverage probability

$$\begin{aligned}
\inf_{\theta \in \Theta} \mathbb{P}[\theta \in C(X) \mid \theta] &= \inf_{\theta \in \Theta} \mathbb{P}[\varphi(x; \theta) = 1 \mid \theta] \\
&= 1 - \underbrace{\sup_{\theta \in \Theta} (\mathbb{P}[\varphi(x; \theta) = 0 \mid \theta])}_{\text{type I error ; reject } \theta \text{ given } \theta \text{ is true}} \\
&= 1 - \alpha
\end{aligned}$$

Confidence Sets

For the same test statistic, there are (infinitely) many confidence sets one could construct. Typically, we want the confidence set to be as small (short) as possible (conditioning on a given coverage probability).

Essentially, constructing C(X) from the acceptance region of a point-hypothesis test leads to the smallest C(X).

Confidence Sets

Example $X \sim N(\theta, 1)$, and we test $\mathscr{H}_0 : \theta = \theta_0$ vs. $\mathscr{H}_1 : \theta \neq \theta_0$.

Using the two-sided t-test $\varphi_t(x; \theta_0) = \mathbf{1}\{|x - \theta_0| < c_\alpha^t\}$, we accept $\mathscr{H}_0$ if $-c_\alpha^t \leqslant x - \theta_0 \leqslant c_\alpha^t$, which shows that the set of all θ_0 we would accept is

$$C_\alpha^t(x) = \left[x - c_\alpha^t, x + c_\alpha^t\right]$$

In the case of a size $\alpha = 0.05$ test, we get the 95% confidence interval

$$C_{0.05,t}(x) = [x - 1.96, x + 1.96]$$

Under the LR test, $\varphi_{LR}(x; \theta_0) = \mathbf{1}\left\{(x - \theta_0)^2 \leq c_\alpha^{LR}\right\}$, which means we accept $\mathcal{H}_0$ if $-\sqrt{c_\alpha^{LR}} \leq (x - \theta_0) \leq \sqrt{c_\alpha^{LR}}$ and yields

$$C_\alpha^{LR}(x) = \left[x - \sqrt{c_\alpha^{LR}}, x + \sqrt{c_\alpha^{LR}}\right]$$