Review of some probability concepts: random vectors, large-sample results

(A quick tour)

N. Torelli

Fall 2020

University of Trieste

Random vectors

Random vectors

In statistics multiple variables are usually observed, and vectors of random variables (**random vectors**) are required. The two-dimensional case is useful to illustrate the main concepts, and will be used here.

but I can define this also for discrete random variables

For continuous r.v., the **joint (probability) density function** extends the one-dimensional case: it is the $f(x, y)$ function such that, for any $A \subset \mathbb{R}^2$

A is the support of these variables

$$\Pr\{(X, Y) \in A\} = \int\int_A f(x, y) dx\, dy \,.$$

Note that $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx\, dy = 1$.

The probability density function defines the **joint distribution** of the random vector (X, Y).

Marginal distribution

The joint distribution embodies information about each components, so that the distribution of X, ignoring Y, can be obtained from $f(x, y)$.

The *marginal* density function of X is given by

if we have the joined density function of multiple vars, I can obtain the marginal distr of one var by process called MARGINALIZATION

$\sum_{y \in \text{Support}} f(x,y) = f(x)$

can be also reduced to the discrete case by substituting the integral over |R with a sum over all possible values of y

$$f(x) = \int_{-\infty}^{\infty} f(x, y) dy \, ,$$

and similarly for the other variable.

(Note: here and elsewhere we always use the symbol f for any p.d.f., identifying the specific case by the argument of the function).

Conditional distribution

The *conditional density function* of Y given $X = x_0$ updates the distribution of Y by incorporating the information that $X = x_0$.

It is given by the important formula

It's the pdf with one of the variables held fixed

$$f(y|X = x_0) = \frac{f(x_0, y)}{f(x_0)}\,, \qquad \text{provide } f(x_0) > 0\,.$$

The simplified notation $f(y|x_0)$ is often employed.

The conditional p.d.f. is properly defined, since $f(y|X = x_0) \geq 0$ and $\int_{-\infty}^{\infty} f(y|x_0)dy = 1$.

A symmetric definition applies to X given $Y = y_0$.

Conditional distribution: useful properties

In the two dimensional case, it is readily possible to write

$$f(x, y) = f(x)\, f(y|x)\,.$$

Extensions to higher dimensions require some care:

$$\begin{aligned}
f(x, y, z) &= f(x, y|z)\, f(z) \\
f(x, y|z) &= f(x|z)\, f(y|x, z) \\
f(x, y, z) &= f(x|y, z)\, f(y, z) \\
f(x, y, z) &= f(x|y, z)\, f(y|z)\, f(z) \\
f(x_1, x_2, \ldots, x_n) &= f(x_1)\, f(x_2|x_1)\, f(x_3|x_2, x_1)\, \ldots f(x_n|x_{n-1}, \ldots, x_2, x_1)
\end{aligned}$$

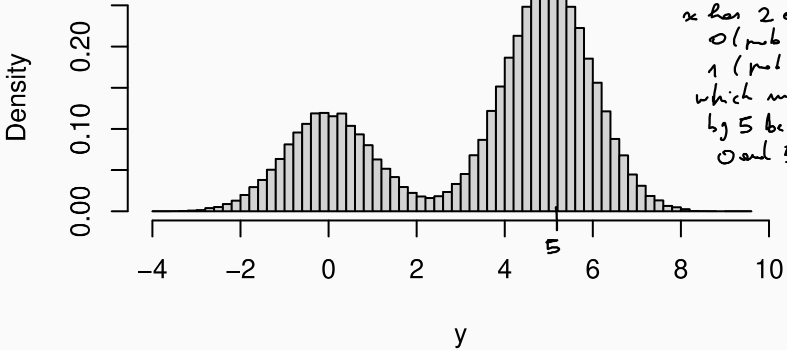
R lab: simulation from joint distributions (a mixture)

```r
x <- rbinom(10^5, size = 1, prob = 0.7)
y <- rnorm(10^5, m = x * 5, s = 1) ### Y| X = x ~ N(x * 5, 1)
hist.scott(y, main = "", xlim = c(-4, 10))
```

x is a vector cointaining ~70% ones and ~30% zeros. it is a bernoulli variable

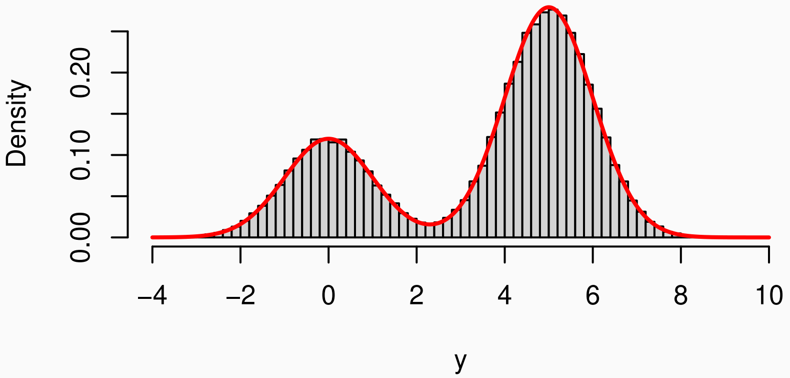
y is a var with distro $\mathcal{N}(5x, 1)$

x has 2 exp value 0 (prob 30%) and 1 (prob 70%), which multiplied by 5 becomes 0 and 5



R lab: simulation from joint distributions (cont'd.)

```
xx <- seq(-4, 10, l = 1000)
ff <- 0.3 * dnorm(xx, 0) + 0.7 * dnorm(xx, 5)
### This is a mixture of normal distributions
hist.scott(y, main = "",  xlim = c(-4, 10))
lines(xx, ff, col = "red", lwd = 2)
```



Bayes theorem

From the factorization of the joint distribution it readily follows that

$$f(x,y) = f(x)\,f(y|x) = f(y)\,f(x|y)$$

from which we obtain the **Bayes theorem**

$$f(x|y) = \frac{f(x)\,f(y|x)}{f(y)}\,.$$

$$=\big(f(x)f(y|x)\big)/\Big(\underbrace{\int_{-\infty}^{+\infty}dx\,\underbrace{f(x)f(y|x)}_{f(x,y)}}_{f(y)}\Big)$$

This is a cornerstone of statistics, leading to an entire school of statistical modelling.

Independence and conditional independence

No dependence on x

=> f(y|x)=f(y). using f(y|x)=f(x,y)/f(x) => f(x,y)=f(x)f(y)

When $f(y|x)$ does not depend on the value of x, the r.v. X and Y are *independent*, and

$$f(x, y) = f(y)\, f(x)$$

More in general, n r.v. are independent if and only if

$$f(x_1, x_2, \ldots, x_n) = f(x_1)\, f(x_2) \ldots f(x_n)\,.$$

this variables are independent but also EQUALLY DISTRIBUTED (f is the same for each one of them)

Conditional independence arises when two r.v. are independent given a third one:

$$f(y, x|z) = f(x|z)\, f(y|z)$$

An important part of statistical modelling exploits some sort of conditional independence.

Example of conditional independence: the Markov property

The general factorization defined above

$$f(x_1, x_2, \dots, x_n) = f(x_1)\, f(x_2|x_1)\, f(x_3|x_2, x_1)\, \dots f(x_n|x_{n-1}, \dots, x_2, x_1)$$

will simplify considerably when the *first order Markov property* holds:

$$f(x_i|x_1, \dots, x_{i-1}) = f(x_i|x_{i-1})$$

the dependence
is only on the last variable before
the one that I'm studying

which means that X_i is independent of $X_1, \dots, X_{i-2}$ given X_{i-1}. We get

$$f(x_1, x_2, \dots, x_n) = f(x_1) \prod_{i=2}^{n} f(x_i|x_{i-1})\,.$$

When the variables are observed over time, this means that the process has *short memory*, a property quite useful in the statistical modelling of **time series**.

Mean and variance of linear transformations

For two r.v. X and Y and two constants a, b we get suppose at first X and Y are not independent

$$E(a\,X + b\,Y) = a\,E(X) + b\,E(Y)\,.$$

The result follows from the more general one

$$E\{g(X,Y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\,f(x,y)\,dx\,dy\,.$$

For variances we need first to introduce the **covariance** between X and Y

$$\mathrm{cov}(X,Y) = E\{(X-\mu_x)\,(Y-\mu_y)\} = E(X\,Y) - \mu_x\,\mu_y\,,$$ it can be >,< or = to 0

where $\mu_x = E(X)$ and $\mu_y = E(Y)$. Then

$$\mathrm{var}(a\,X + b\,Y) = a^2\,\mathrm{var}(X) + b^2\,\mathrm{var}(Y) + 2\,a\,b\,\mathrm{cov}(X,Y)\,.$$

Note: for X, Y independent it follows that $\mathrm{cov}(X,Y) = 0$. The reverse is not true, unless the joint distribution of X and Y is multivariate normal.

Mean vector

For a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$, the **mean vector** is just

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}.$$

The mean vector has the same properties of the scalar case, so that for example $E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y})$, and for $\mathbf{A}$ and $\mathbf{b}$ a $n \times n$ matrix and a $n \times 1$ vector, respectively, it follows that

$$E(\mathbf{A}\,\mathbf{X} + \mathbf{b}) = \mathbf{A}\,E(\mathbf{X}) + \mathbf{b}\,.$$

Variance-covariance matrix

The variance-covariance matrix of the random vector $\mathbf{X}$ collects all the variances (on the main) diagonal and all the pairwise covariances (off the main diagonal), being the $n \times n$ symmetric semi-definite matrix ← This follow from def of covariance, which is symmetric

$$\boldsymbol{\Sigma} = E\{(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)^\top\} = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \cdots & \mathrm{cov}(X_1,X_n) \\ \mathrm{cov}(X_1,X_2) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2,X_n) \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{cov}(X_1,X_n) & \mathrm{cov}(X_2,X_n) & \cdots & \mathrm{var}(X_n) \end{pmatrix}.$$

Useful properties:

linear tsf applied to X

$$\boldsymbol{\Sigma}_{\mathbf{A}\,\mathbf{X}+\mathbf{b}} = \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}^\top$$

$$\boldsymbol{\Sigma}_{\mathbf{X}^\top\mathbf{A}\,\mathbf{X}} = \boldsymbol{\mu}_x^\top\mathbf{A}\,\boldsymbol{\mu}_x + \mathrm{tr}(\mathbf{A}\,\boldsymbol{\Sigma})$$

quadratic tsf applied to X → this is a quadratic form

Transformation of random variables and random vectors

Given a continuous r.v. X and a transformation $Y = g(X)$, with g an invertible function, it readily follows that

distribution of y $\rightarrow$ $$f_y(y) = f_x\{g^{-1}(y)\} \left| \frac{dx}{dy} \right| .$$

The result is extended to two continuous random vectors with the same dimension

$$f_{\mathbf{Y}}(\mathbf{Y}) = f_{\mathbf{X}}\{g^{-1}(\mathbf{Y})\} \, |\mathbf{J}| \, ,$$

with $J_{ij} = \partial x_i / \partial y_j$.

For discrete r.v., the results are simpler, with no need of including the Jacobian of the transformation.

The multivariate normal distribution

The multivariate normal distribution

independent
and identically
distributed
variables

Start from a set of n i.i.d. $Z_i \sim \mathcal{N}(0, 1)$, so that $E(\mathbf{z}) = \mathbf{0}$ and covariance matrix $\mathbf{I}_n$. If $\mathbf{B}$ is $m \times n$ matrix of coefficients and $\boldsymbol{\mu}$ a m-vector of coefficients, then the m-dimensional random vector $\mathbf{X}$

$$\mathbf{X} = \mathbf{B}\,\mathbf{z} + \boldsymbol{\mu}$$

has a **multivariate normal distribution** with covariance matrix $\boldsymbol{\Sigma} = \mathbf{B}\,\mathbf{B}^\top$.

The notation is

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

Joint p.d.f.

that is that $\vec{Z} \sim \mathcal{N}(\vec{0}, \vec{1})$
$\Rightarrow \vec{X} \sim \mathcal{N}(\vec{\mu}, \Sigma = DB^T)$

Using basic results on transformation of random vectors, starting from the joint p.d.f of $Z_1, Z_2, \ldots, Z_n$ we obtain

$$f_{\mathbf{X}}(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^m |\mathbf{\Sigma}|}} \, exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\}, \qquad \text{for } \mathbf{X} \in \mathbb{R}^m,$$

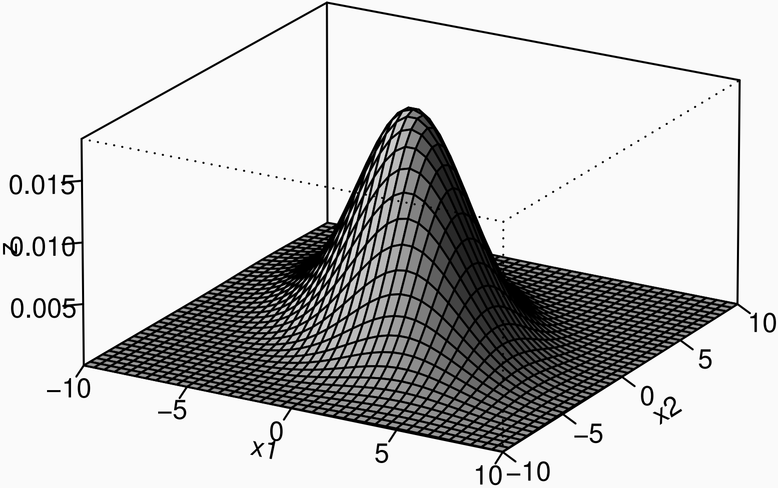
provide that $\mathbf{\Sigma}$ has full rank m. The result can be extended to *singular* $\mathbf{\Sigma}$ by recourse to the *pseudo-inverse of* $\mathbf{\Sigma}$: this is used, for example, in the analysis of *compositional data*.

A useful property which holds only for this distribution: *two r.v. with multivariate normal distribution and* **zero covariance** *are* **independent**.

This won't true in general, where INDEP VAR $\Rightarrow$ COV $= 0$
COV $= 0 \not\Rightarrow$ INDEP VAR
For multivariate normal distribution INDEP VAR $\Leftrightarrow$ COV $= 0$

Example: bivariate case

We take $\mu_1 = \mu_2 = 0$, $\sigma_1^2 = 10$, $\sigma_2^2 = 10$, $\sigma_{12} = 15$



Linear transformations

It is simple to verify that if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{A}$ is a $k \times m$ matrix of constants then

$$\mathbf{A}\,\mathbf{X} \sim \mathcal{N}(\mathbf{A}\,\boldsymbol{\mu}, \mathbf{A}\,\boldsymbol{\Sigma}\,\mathbf{A}^\top)\,.$$

A special case is obtained when $k = 1$, in that for a m-dimensional vector $\mathbf{a}$

$$\mathbf{a}^\top\,\mathbf{X} \sim \mathcal{N}(\mathbf{a}^\top\,\boldsymbol{\mu}, \mathbf{a}^\top\,\boldsymbol{\Sigma}\,\mathbf{a})\,.$$

Note that for suitable choices of $\mathbf{a}$ (when all the elements 0s or 1s) it follows that **the marginal distribution of any subvector of X is multivariate normal**.

when putting to = 0 some of the components of X, by suitable choices of a

Normality of the marginal distributions, instead, does not imply multivariate normality.

Conditional distributions

Consider two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with multivariate normal joint distribution, and partition their joint covariance matrix as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix},$$

and similarly for the mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_x, \boldsymbol{\mu}_y)^\top$.

Using results on *partitioned matrices*, it follows that the **conditional distributions are multivariate normal**.

For instance

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\,\boldsymbol{\Sigma}_{xx}^{-1}\,(\mathbf{X} - \boldsymbol{\mu}_x), \boldsymbol{\Sigma}_{yy} - \boldsymbol{\Sigma}_{yx}\,\boldsymbol{\Sigma}_{xx}^{-1}\,\boldsymbol{\Sigma}_{xy})\,.$$

Statistics

Random sample

The collection of r.v. $X_1, X_2, \ldots, X_n$ is said to be a **random sample** of size n if they are *independent and identically distributed*, that is

- $X_1, X_2, \ldots, X_n$ are independent r.v.
- They have the same distribution, namely the same c.d.f.

The concept is central in statistics, and it is the suitable mathematical model for the outcome of sampling units from a very large population. The definition is, however, more general.

(For more details: https://www.probabilitycourse.com/chapter8/8_1_1_random_sampling.php)

$x_1, \ldots, x_n$ i.i.d all distributed as one r.v. X with certain distr function

should be known when dealing with large population t is a r.v. ⟸ iid

A **statistic** is a r.v. defined as a function of a set of r.v. $t = g(Y_1, .., Y_n)$

Obvious examples are the sample mean and variance of data $\underbrace{y_1, y_2, \ldots, y_n}_{\text{Sample}}$

$E(\bar{y}) = E(y)$

y in my population

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i\,, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\,.$$

$v(\bar{y}) = \frac{\sigma^2}{n}$, $\sigma^2 = var(y)$ ⟹ var of $\bar{y}$ gets smaller with the size of the population

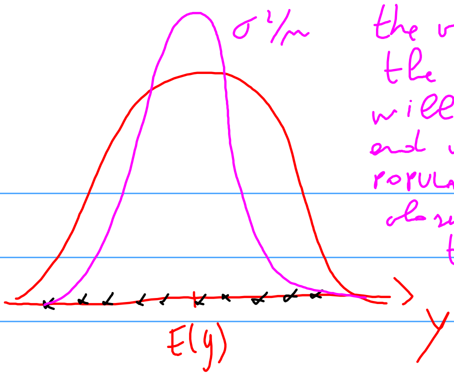
Consider a random vector $\mathbf{Y}$ with p.d.f. $f_{\boldsymbol\theta}(\mathbf{Y})$ depending on a vector $\boldsymbol\theta$ (which is the *parameter*, as we will see).

If a statistic $t(\mathbf{Y})$ is such that $f_{\boldsymbol\theta}(\mathbf{Y})$ can be written as

$$f_{\boldsymbol\theta}(\mathbf{Y}) = h(\mathbf{Y})\, g_{\boldsymbol\theta}\{t(\mathbf{Y})\}\,,$$

where h does not depend on $\boldsymbol\theta$, and g depends on $\mathbf{Y}$ only through $t(\mathbf{Y})$, then t is a **sufficient statistic** for $\boldsymbol\theta$: all the *information* available on $\boldsymbol\theta$ contained in $\mathbf{Y}$ is supplied by $t(\mathbf{Y})$.

The concepts of information and sufficiency are central in statistical inference.



the variance of the mean $\bar{y}$ will be smaller and with INCREASING POPULATION I'll get closer and closer to the true mean of y

EX. $Y \sim N(\mu, \sigma^2)$

μ unknown

OBS. $y_1, \dots, y_n$ iid

$\rightarrow$ $\underbrace{\bar{y} \text{ sample mean}}$

contains all the info relevant about the distribution of the variable

MINIMAL SUFFICIENT STATISTICS: save the more infos possible in the less space possible

Example: sufficient statistic for the normal distribution

Given a vector of independent normal r.v. $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, it follows that $\boldsymbol{\theta} = (\mu, \sigma^2)$ and

$$
\begin{aligned}
f_{\boldsymbol{\theta}}(\mathbf{Y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\,\pi}\,\sigma} \exp\left\{-\frac{1}{2\,\sigma^2}(y_i - \mu)^2\right\} \\
&= \frac{1}{\left(\sqrt{2\,\pi}\right)^n \sigma^n} \exp\left\{-\frac{1}{2\,\sigma^2} \sum_i (y_i - \mu)^2\right\}.
\end{aligned}
$$

By some simple algebra, it is possible to show that the two-dimensional statistic $t(\mathbf{Y}) = (\bar{y}, s^2)$ is sufficient for (μ, σ^2).

Complements & large-sample results

Moment generating function

The **moment generating function** (m.g.f.) characterises the distribution of a r.v. X, and it is defined as

$$M_X(t) = E(e^{tX}), \qquad \text{for } t \text{ real}.$$

$\boxed{M_X(t) = \int_{\mathbb{R}} e^{tx} f(x)\,dx}$

The name derives from the fact the k^{th} derivative of the m.g.f. at $t = 0$ gives the k^{th} uncentered moment:

$$\frac{d^k M_X(t)}{d\,t^k}\Big|_{t=0} = E(X^k).$$

k-th moment

$m_k = E(X^k)$

$= \int_{\mathbb{R}} x^k f(x)\,dx$

Two useful properties:

- If $M_X(t) = M_Y(t)$ for some small interval around $t = 0$, then X and Y have the same distribution.
- If X and Y are independent, $M_{X+Y}(t) = M_X(t)\,M_Y(t)$.

take ft $E(e^{tx})$

Taylor exp: $e^{tx} = 1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \dots$

$\Rightarrow E\left(1 + tx + \frac{t^2x^2}{2!} + \frac{t^3x^3}{3!} + \dots\right) =$

$= 1 + t\,E(x) + t^2\frac{E(x^2)}{2!} + t^3\frac{E(x^3)}{3!} + \dots$

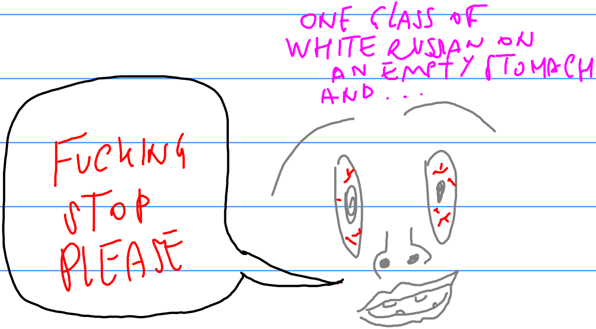
$\Rightarrow \left.\frac{d M_x(t)}{dt}\right|_{t=0} = E(x)$

$\vdots$

$\left.\frac{d^k M_x}{dt^k}\right|_{t=0} = E(x^k)$

$\Rightarrow$ M IS THE GENERATING FT OF MOMENTS

⚠ The MOM GEN FT could not exist for some distributions



Given

Prob distr ft $\longrightarrow$ Moment gen ft

Moment gen ft $\xrightarrow{?}$ Pdf

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

$$\Rightarrow \underbrace{\int_{\mathbb{R}} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}_{N(\mu,\sigma^2)} = e^{t\mu + \frac{\sigma^2 t^2}{2}} = M(t)$$

Suppose I have $M(t) = \exp\left(t\mu + \frac{\sigma^2 t^2}{2}\right)$

$\sim \exp\left(t^2/2\right)$

$\Rightarrow$ I can $\underbrace{\text{RECOGNIZE}}_{\text{how?}}$ $X \sim N(0,1)$

By using MOMENTS

The central limit theorem

For i.i.d. r.v. $X_1, X_2, \ldots, X_n$ with mean μ and finite variance σ^2, the **central limit theorem** states that for large n the distribution of the r.v. $\overline{X}_n = \sum_{i=1}^n X_i/n$ is approximately

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n).$$

More formally, the theorem says that for any $x \in \mathbb{R}$ the c.d.f. of $Z_n = (\overline{X}_n - \mu)/\sqrt{\sigma^2/n}$ satisfies

$$\lim_{n\to\infty} F_{Z_n}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz\,. = \Phi(z)$$

cdf of a standard normal

The proof is simple, and it uses the m.g.f.

The theorem generalizes to multivariate and non-identical settings.

It has a central importance in statistics, since it supports the normal approximation to the distribution of a r.v. that can be viewed as the sum of other r.v.

dot means an ASYMPTOTIC APPROX

$Z_n \dot{\sim} N(0,1)$ for big n and

n iid r.v.

The law of large numbers

Consider i.i.d. (independent and identically distributed) r.v. $X_1, \ldots, X_n$, with mean μ and $(E|X_i|) < \infty$.

The **strong law of large numbers** states that, for any positive ϵ

I look at CONVERGENCE OF THE $|\bar{X}_n - \mu|$

$$\Pr\left(\lim_{n\to\infty} |\overline{X}_n - \mu| < \epsilon\right) = 1\,,$$

namely $\overline{X}_n$ *converges almost surely to* μ.

With the further assumption $\mathrm{var}(X_i) = \sigma^2$, the **weak law of large numbers** follows

$$\lim_{n\to\infty} \Pr\left(|\overline{X}_n - \mu| \geq \epsilon\right) = 0\,.$$

I look at the CONVERGENCE OF PROBABILITY

Proof of the weak law of large numbers

First we recall the *Chebyshev's inequality*: given a r.v. X such that $E(X^2) < \infty$ and a constant $a > 0$, then

$$\Pr(|X| \geq a) \leq \frac{E(X^2)}{a^2}\,.$$

We apply the inequality to the case of interest, so that

$$\Pr\left(|\overline{X}_n - \mu| \geq \epsilon\right) \leq \frac{E\{(\overline{X}_n - \mu)^2\}}{\epsilon^2} = \frac{\mathrm{var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\,\epsilon^2}\,,$$

which tends to zero when $n \to \infty$.

The result may hold also for non-i.i.d. cases, provided $\mathrm{var}(\overline{X}_n) \to 0$ for large n.

Jensen's inequality

This is another useful result, that states that for a r.v. X and a concave function g

$$g\{E(X)\} \geq E\{g(X)\}\,.$$

NOT TRUE for gene-ric functions

(Note: a concave function is such that

$$g\{\alpha\, x_1 + (1-\alpha)\, x_2\} \geq \alpha\, g(x_1) + (1-\alpha)\, g(x_2)\,,$$

for any x_1, x_2, and $0 \leq \alpha \leq 1$).

An example is $g(x) = -x^2$, so that

$$-E(X)^2 \geq -E(X^2) \quad\Rightarrow\quad E(X)^2 \leq E(X^2)\,,$$

which is obviously true since $E(X^2) = \mathrm{var}(X) + E(X)^2$.

And this is ≥ 0