Statistics TI Amsterdam 2020: lecture 1, online only

Peter Spreij

Universiteit van Amsterdam

Radboud Universiteit

August 2020

Outline

Outline

Webpage of the course

The course has a website,

https://staff.fnwi.uva.nl/p.j.c.spreij/onderwijs/TI/statistics.html

with *all* relevant information.

To find it, Google *Peter Spreij*, open his homepage, click there on Courses and proceed.

Some organizational details

- Lectures on location on Wednesdays (except the first lecture)
- Tutorials (TA sessions), with Aisha Schmidt and Saeed Badri on xxxdays
- Weekly homework, compulsory, starting from Lecture 2
- Literature. Main: book by Rice (2nd or 3rd edition), secondary: small set of additional notes, copies of a few slides and the slides of this presentation; see the webpage for links

Outline

Probability space

A probability space is a *triple* $(\Omega, \mathcal{F}, \mathbb{P})$.

Here is

- Ω (having elements denoted ω) a (non-empty) set, the sample space,
- $\mathcal{F}$ is a σ-algebra,
- $\mathbb{P}$ is a probability measure on $\mathcal{F}$.

What do these concepts mean?

Sample space

Ω is typically the set that lists all possible outcomes of an experiment.

Depending on the experiment, Ω could be

- the 2020 new TI students,
- all UvA students,
- a nonnegative integer,
- a real number,

and there is a lot more!

Events and σ-algebra

An event A is a subset of Ω, $A \subset \Omega$, but in principle not any subset. The collection of events is supposed to be a σ-algebra, $\mathcal{F}$:

- $\emptyset \in \mathcal{F}$,
- If $A \in \mathcal{F}$, then also its complement A^c is an element of $\mathcal{F}$,
- If $A_1, A_2, \ldots$ is a sequence of sets in $\mathcal{F}$, then also the union $\bigcup_{i=1}^{\infty} A_i$ belongs to $\mathcal{F}$.

Properties of events

- Finite unions like $A_1 \cup A_2$ belong to $\mathcal{F}$, whenever $A_1, A_2 \in \mathcal{F}$.
- Finite and countable intersections $A_1 \cap A_2$ and $\bigcap_{i=1}^{\infty} A_i$ belong to $\mathcal{F}$, if the A_i belong to it.
- In short all set theoretic operations applied to events yield events again, as long as they are performed at most countably often.

If the set Ω is finite or countable, one usually take the power set of Ω (all its subsets) as the collection of events $\mathcal{F}$.

Uncountable Ω

Is Ω is countably infinite, like $\Omega = \mathbb{R}$ or $\Omega = (0, 1)$, for technical reasons one takes a *smaller* collection than all subsets.

In the latter two examples, one usually takes the *Borel sets* (denoted $\mathcal{B}$), these are the sets that can be generated by at most countably often applied set theoretic operations to all open intervals.

For example, if $\Omega = \mathbb{R}$, then by definition an interval $(-\infty, a)$ is an element of $\mathcal{B}$, but then also $[a, \infty)$. Also every singleton belongs to $\mathcal{B}$, since $\{a\} = \cap_{n=1}^{\infty}(a - 1/n, a + 1/n)$. Other examples are $(-\infty, a]$, $(a, b]$, $[a, b)$, etc.

In fact any '*normal*' subset of $\mathbb{R}$ will be in $\mathcal{B}$, this is a tautology ...

Probability measure

Compare the notations $\mathbb{P}(A)$ and $f(x)$.

Indeed, a probability $\mathbb{P}$, also known as a *probability measure*, is a function too, defined on the collection of events $\mathcal{F}$, $\mathbb{P} : \mathcal{F} \to [0, 1]$. More precisely, we require

- $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$,
- for *disjoint* events $A_i \in \mathcal{F}$ it holds that $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Note that for disjoint A_1 and A_2, both in $\mathcal{F}$, we have the familiar rule $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$ (you check!). We also frequently use $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$ and $\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i)$ if $\cup_{i=1}^{\infty} B_i = \Omega$ (cutting A up in slices B_i).

Random variables

Elements of Ω can be 'anything' and you may not be able to perform computations with them. But, these you can do with

a *random variable* X, a function $X : \Omega \to \mathbb{R}$, that is *measurable*: $\{X \in B\} \in \mathcal{F}$ for every Borel set B.
Here $\{X \in B\}$ is shorthand notation for $\{\omega \in \Omega : X(\omega) \in B\}$.

Then every set $\{X \leq x\}$ is an element of $\mathcal{F}$ (here you take $B = (-\infty, x]$. In fact, it is possible to show that if all sets $\{X \leq x\}$ $(x \in \mathbb{R})$ are elements of $\mathcal{F}$, then X is measurable, a random variable.

More on random variables

- For random variables X the probabilities $\mathbb{P}(X \in B)$, short for $\mathbb{P}(\{X \in B\}) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$ are well defined.
- The rule $\mathbb{P}(X \in B_1 \cup B_2) = \mathbb{P}(X \in B_1) + \mathbb{P}(X \in B_2)$ for disjoint B_1, B_2 in $\mathcal{B}$.
- The probabilities $F(x) := \mathbb{P}(X \leq x)$ are the values of a function $F : \mathbb{R} \to [0, 1]$, called the *(cumulative) distribution function* of X. *Exercise:* show that F is non-decreasing and right-continuous, and $\lim_{x\to\infty} F(x) = 1$.
- Random *vectors* X will be considered as vectors of random variables X_i. A two-dimensional random vector is sometimes denoted as a row (X_1, X_2) or as a column $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$.

Outline

Discrete random variables

Let $x_1, x_2, \ldots$ be a finite or infinite sequence. A (measurable) function $X : \Omega \to \{x_1, x_2, \ldots\}$ is called a *discrete* random variable.

Indeed, the sets $\{X = x_i\}$ are in $\mathcal{F}$, and hence the probabilities $p_i := \mathbb{P}(X = x_i)$ are well defined. These form the *distribution* of X. The formula $p_i := \mathbb{P}(X = x_i)$ represents the *probability mass function*, masses p_i are put at the positions x_i.

Recall that the distribution function F of X is defined as $F(x) := \mathbb{P}(X \leq x)$, and that F is right-continuous.

By $F(x-)$ we denote $\lim_{y \uparrow x} F(y)$. Then $F(x-) = \mathbb{P}(X < x)$ and the *jump* of F at x is $\Delta F(x) := F(x) - F(x-) = \mathbb{P}(X = x) \geq 0$. In particular, we see that $\Delta F(x_i) = \mathbb{P}(X = x_i) = p_i$.

Note that $F(x) = \sum_{x_i \leq x} p_i$ and $F(b) - F(a) = \sum_{a < x_i \leq b} p_i$. [Shortly we will see integrals instead of sums.]

Example

Let $\Omega = \{hh, ht, th, tt\}$ and let $\mathbb{P}(\{\omega\}) = \frac{1}{4}$ for all ω, and $X(\omega)$ is the number of h's in ω. Then the distribution of X is represented by the following *table*.

x_i	0	1	2
p_i	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Note that $\sum_i p_i = 1$, and the graph of the distribution F is a 'staircase' that jumps at $0, 1, 2$, in particular F is not everywhere continuous. Make a picture of F!

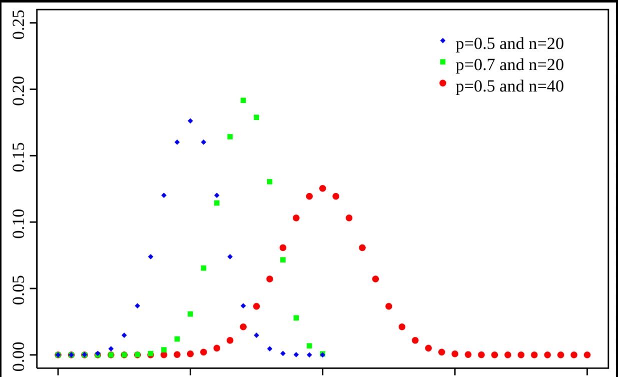
Examples of distributions

Here are some classical examples of distributions of random variables (more of them in Rice).

- Bernoulli distribution. $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$, $p \in [0, 1]$.
- Generalization: Binomial distribution $\text{Bin}(n, p)$. $\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$, $k \in \{0, \ldots, n\}$.
- Poisson(λ) distribution: $\mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!$, $k \in \{0, 1, \ldots\}$, $\lambda > 0$.

Relation between Binomial and Poisson: if $n \to \infty$, $np \to \lambda$ then $\binom{n}{k} p^k (1 - p)^{n-k} \to e^{-\lambda}\lambda^k/k!$

Binomial pmfs



Probability mass function for the binomial distribution

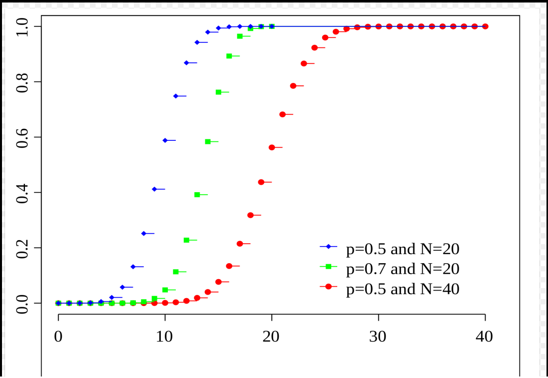
More details

Tayste - Own work

Binomial distribution probability mass function

File: Binomial distribution pmf.svg
Created: 2 March 2008

Binomial Cdfs



Cumulative distribution function for the binomial distribution

More details

Tayste - Own work

Binomial distribution cumulative distribution function

File: Binomial distribution cdf.svg
Created: 2 March 2008

Definitions

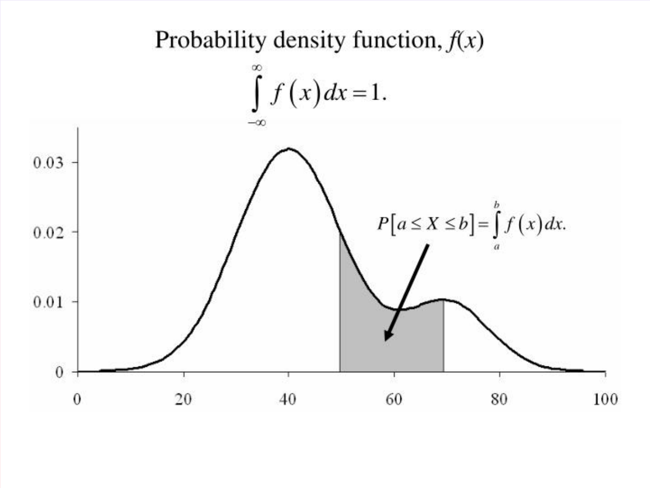
A random variable X is called *continuous* if its distribution function F is (everywhere) continuous. Note that in such a case one has $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$ and hence $\mathbb{P}(X \leq x) = \mathbb{P}(X < x)$.

If there exists a nonnegative function f on $\mathbb{R}$ such that $F(x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u$ for all $x \in \mathbb{R}$, then f is called a *(probability) density* of X. Note that $\int_{-\infty}^{+\infty} f(u)\,\mathrm{d}u = 1$.

Such an f cannot be unique, if you change f at one point u (with $u < x$), then $F(x)$ stays the same. Usually we take a 'nice' version of f: if F is differentiable at x, we take $f(x) = F'(x)$.

The distribution of X is the collection of all probabilities $\mathbb{P}(X \in B)$, for $B \in \mathcal{B}$. Each of these is an integral, $\mathbb{P}(X \in B) = \int_B f(u)\,\mathrm{d}u$. [In fact, $\mathbb{P}^X$ defined by $\mathbb{P}^X(B) := \mathbb{P}(X \in B)$, $B \in \mathcal{B}$ is a probability measure on $\mathcal{B}$.]

Probabilities as an area under the pdf



Gamma function

Gamma integral. $\Gamma(\alpha) := \int_0^\infty u^{\alpha-1} e^{-u} \, \mathrm{d}u$, for $\alpha > 0$.
Properties: $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$. If α is an integer, $\Gamma(\alpha) = (\alpha-1)!$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ (see later).

Make a change of variable in the integral, $u = \lambda x$. Then $\Gamma(\alpha) = \lambda^\alpha \int_0^\infty x^{\alpha-1} e^{-\lambda x} \, \mathrm{d}x$.

It follows that the function f with $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ for $x \geq 0$ and $f(x) = 0$ for $x < 0$ is a density. The corresponding distribution is the $\Gamma(\alpha, \lambda)$ distribution, also denoted Gamma(α, λ) distribution.

Special case 1: $\alpha = 1$, *exponential* distribution, $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0$.
Special case 2: $\alpha = \lambda = \frac{1}{2}$, also called χ^2_1 distribution (see later).

Normal distribution

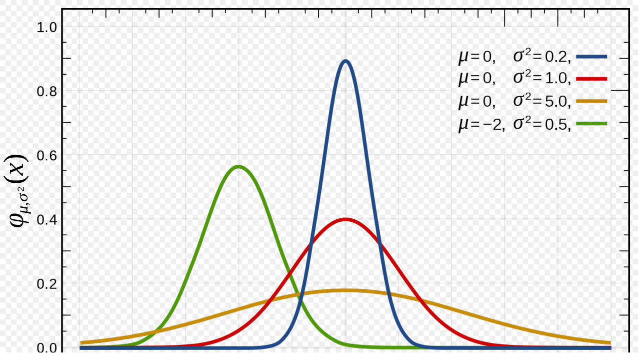
A random variable is said to have the $N(\mu, \sigma^2)$ distribution if it has density $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Special case. $\mu = 0$, $\sigma^2 = 1$: standard normal distribution,
$f_{0,1}(x) =: \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.
Special notation for the distribution function:

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)\,\mathrm{d}u.$$

There exists no simple formule for Φ in terms of 'well known functions'.
Property (check!): $\Phi(-x) = 1 - \Phi(x)$, only need table for $x > 0$.

Normal pdfs



A selection of Normal Distribution Probability Density Functions (PDFs). Both the mean, μ, and variance, σ^2, are varied. The key is given on the graph.

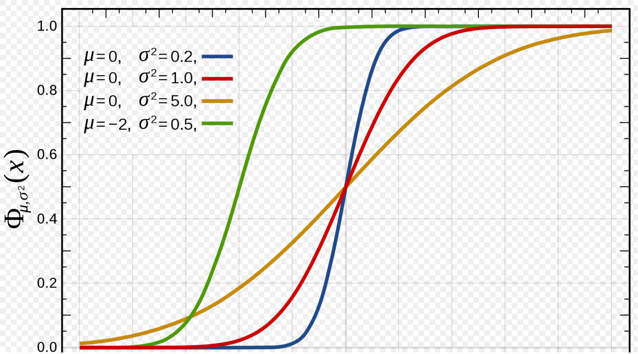
More details

Inductiveload - self-made, Mathematica, Inkscape

File: Normal Distribution PDF.svg

Created: 2 April 2008

Normal Cdfs



A selection of Normal Distribution Cumulative Density Functions (CDFs). Both the mean, μ, and variance, σ^2, are varied. The key is given on the graph.

More details

Inductiveload - self-made, Mathematica, Inkscape

File: Normal Distribution CDF.svg
Created: 3 February 2008

Linear transformation

Let X have a continuous distribution with differentiable distribution function F_X and density $f_X = F_X'$ and put $Y = aX + b$ with $a \neq 0$.

Then Y also has a density, f_Y say, and $f_Y(y) = f_X(\frac{y-b}{a})\frac{1}{|a|}$.

Fundamental approach *via the distribution function F_Y of Y*, for the case $a < 0$ (the case $a > 0$ is similar):

$$\mathbb{P}(Y \leq y) = \mathbb{P}(aX+b \leq y) = \mathbb{P}(X \geq \frac{y-b}{a}) = 1-F_X(\frac{y-b}{a}).$$

Differentiation (chain rule!) gives

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

Monotone transformations

Let X have a density f_X and $Y = g(X)$ where g is a strictly monotone function. Let h be the inverse function of g. Compute for *decreasing* g (then also h is decreasing)

$$F_Y(y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \geq h(y)) = 1 - F_X(h(y)).$$

Differentiate to get

$$f_Y(y) = -f_X(h(y))h'(y) = f_X(h(y))|h'(y)|,$$

a formula which is also valid for increasing g (with similar proof).

Sometimes the calculus rule $h'(y) = \frac{1}{g'(h(y))}$ may come in handy.

Linear transformation in the normal case

Let $Y = aX + b$ and X have the $N(\mu, \sigma^2)$ distribution, so with density $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Then

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\frac{y-b}{a}-\mu)^2}{2\sigma^2}} \frac{1}{|a|} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-b-\mu a)^2}{2a^2\sigma^2}} \frac{1}{\sqrt{a^2}} \\ &= \frac{1}{\sqrt{2\pi a^2\sigma^2}} e^{-\frac{(y-(b+\mu a))^2}{2a^2\sigma^2}}. \end{aligned}$$

It follows that *also* Y has a normal distribution, $N(a\mu + b, a^2\sigma^2)$.

Standardization

Let $\sigma > 0$ and make the special choice $a = \frac{1}{\sigma}$, $b = -\frac{\mu}{\sigma}$. Then

$$Y = \frac{X - \mu}{\sigma} \text{ and } Y \text{ is } N(0, 1).$$

Use (recall $\sigma > 0$):

$$\mathbb{P}(X \leq x) = \mathbb{P}(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}) = \mathbb{P}(Y \leq \frac{x - \mu}{\sigma}) = \Phi(\frac{x - \mu}{\sigma}).$$

The distribution function of X that is $N(\mu, \sigma^2)$ can be expressed in terms of the single function Φ; only on 'table' (for the standard normal distribution) is needed for all normal distributions.

A nonlinear nonmonotone transformation

Let X have a continuous distribution with density f_X and $Y = X^2$. We want the density f_Y of Y, compute this (again) via the distribution function in $y > 0$.

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) \\ &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

By differentiation (chain rule!),

$$f_Y(y) = f_X(\sqrt{y})\frac{1}{2\sqrt{y}} - f_X(-\sqrt{y})\frac{-1}{2\sqrt{y}} = \frac{1}{2\sqrt{y}}(f_X(\sqrt{y}) + f_X(-\sqrt{y})).$$

Square of $N(0,1)$ is χ_1^2

Let X have the $N(0,1)$ and $Y = X^2$. We want the density f_Y of Y. Previous result becomes

$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}}(\phi(\sqrt{y}) + \phi(-\sqrt{y})) \\ &= \frac{1}{2\sqrt{y}}\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{\sqrt{y}^2}{2}} + e^{-\frac{(-\sqrt{y})^2}{2}}\right) \\ &= \frac{1}{\sqrt{2\pi}} y^{-\frac{1}{2}} e^{-\frac{y}{2}} = \frac{(\frac{1}{2})^{\frac{1}{2}}}{\sqrt{\pi}} y^{\frac{1}{2}-1} e^{-\frac{y}{2}}. \end{aligned}$$

This is the $\Gamma(\frac{1}{2}, \frac{1}{2})$ density (the χ_1^2 density) and we also see that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Outline

Introduction and notation

A random vector is a vector X of random variables X_i. A two-dimensional random vector is sometimes denoted as a row $X = (X_1, X_2)$ or as a column $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$. You can guess how this would look in higher dimensions.

We also often write (X, Y) or $\begin{pmatrix} X \\ Y \end{pmatrix}$ in the two-dimensional case for random variables X and Y (and note the ambiguous use of the notation X ...)

Example

Let $\Omega = \{hhh, hht, hth, thh, htt, tht, tth, ttt\}$ and $\mathbb{P}(\{\omega\}) = \frac{1}{8}$ for all ω. Let $X(\omega)$ denote the number of h's in the first position of ω and $Y(\omega)$ the total number of h's in ω. The values of X and Y can jointly be represented with corresponding ω's.

$x \backslash y$	0	1	2	3
0	ttt	tht, tth	thh	
1		htt	hht, hth	hhh

Example continued, with probabilities

Assigning the probabilities in the previous table gives the *joint distribution* of (X, Y):

$x \backslash y$	0	1	2	3	
0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	0	$\frac{1}{2}$
1	0	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	$\frac{1}{2}$
	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

In the right and lower *margins*, containing the row and column subtotals, one recognizes the *marginal distributions* of X and Y respectively.

General notation for discrete (X, Y)

We assume that a *discrete* vector (X, Y) assume values (x_i, y_j) (sometimes also shortly written as (x, y)), where the x_i and y_j may come from a finite or a countably infinite set.

- The $p(x_i, y_j) := \mathbb{P}(X = x_i, Y = y_j)$ (which is short for $\mathbb{P}(\{X = x_i\} \cap \{Y = y_j\})$) represents the *joint probability mass function* and the *joint distribution* of the vector (X, Y).
- The *marginal distribution* of X is given by $\mathbb{P}(X = x_i) = \sum_j p(x_i, y_j)$, similar expression for the marginal of Y.
- In general one has $\mathbb{P}((X, Y) \in A) = \sum_{(x_i, y_j) \in A} p(x_i, y_j)$.
- The *joint (cumulative) distribution function* F (sometimes written as $F_{X,Y}$) of the vector (X, Y) is given by $F(x, y) := \mathbb{P}(X \leq x, Y \leq y) = \sum_{x_i \leq x, y_j \leq y} p(x_i, y_j)$.

Continuous random vectors

A random vector (X, Y) is called continuous if the joint distribution function F is continuous. We will always assume that there exists a *joint density* f, a nonnegative function $f : \mathbb{R}^2 \to \mathbb{R}$ such that for all (Borel) sets $A \in \mathbb{R}^2$ one computes probabilities as double integrals

$$\mathbb{P}((X, Y) \in A) = \int\int_A f(u, v)\,\mathrm{d}u\mathrm{d}v.$$

Note that $\int\int_{\mathbb{R}^2} f(u, v)\,\mathrm{d}u\mathrm{d}v = 1$.

Joint distribution function

Special case, $A = (-\infty, x] \times (-\infty, y]$, leads to the double and *iterated* integrals

$$\begin{aligned}
\mathbb{P}(X \le x, Y \le y) &= F(x, y) \\
&= \iint_{(-\infty,x]\times(-\infty,y]} f(u, v)\,\mathrm{d}u\mathrm{d}v \\
&= \int_{(-\infty,y]} \Big(\int_{(-\infty,x]} f(u, v)\,\mathrm{d}u \Big)\mathrm{d}v \\
&= \int_{(-\infty,x]} \Big(\int_{(-\infty,y]} f(u, v)\,\mathrm{d}v \Big)\mathrm{d}u
\end{aligned}$$

Differentiation

If F is twice continuously differentiable, then we obtain from

$$F(x,y) = \int_{(-\infty,x]} \Big(\int_{(-\infty,y]} f(u,v)\,\mathrm{d}v \Big) \mathrm{d}u$$

and $\frac{\partial F}{\partial x} = \int_{(-\infty,y]} f(x,v)\,\mathrm{d}v$ the relation

$$\frac{\partial^2 F}{\partial y \partial x} = \frac{\partial}{\partial y} \int_{(-\infty,y]} f(x,v)\,\mathrm{d}v = f(x,y).$$

Of course, also

$$\frac{\partial^2 F}{\partial x \partial y} = f(x,y).$$

Marginal distribution functions

Special case, $A = (-\infty, x] \times \mathbb{R}$, leads to the iterated integral for the marginal distribution of X,

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}(X \leq x, Y < \infty) \\ &= \int_{(-\infty,x]} \Big(\int_{\mathbb{R}} f(u,v) \, \mathrm{d}v \Big) \mathrm{d}u.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(X < \infty, Y \leq y) \\ &= \int_{(-\infty,y]} \Big(\int_{\mathbb{R}} f(u,v) \, \mathrm{d}u \Big) \mathrm{d}v.\end{aligned}$$

Marginal densities

Put (like computing marginal sums in the discrete case)

$$f_X(u) = \int_{\mathbb{R}} f(u,v)\,\mathrm{d}v = \int_{-\infty}^{+\infty} f(u,v)\,\mathrm{d}v.$$

Then

$$\mathbb{P}(X \le x) = \int_{(-\infty,x]} f_X(u)\,\mathrm{d}u = \int_{-\infty}^{x} f_X(u)\,\mathrm{d}u.$$

It follows that f_X is the marginal density of X. Similarly,

$$f_Y(v) = \int_{\mathbb{R}} f(u,v)\,\mathrm{d}u = \int_{-\infty}^{+\infty} f(u,v)\,\mathrm{d}u$$

gives the marginal density of Y.

Bivariate normal distribution

A random variable (X, Y) has a bivariate normal distribution if it has density $f(x, y)$ given by $(\sigma_X, \sigma_Y > 0)$

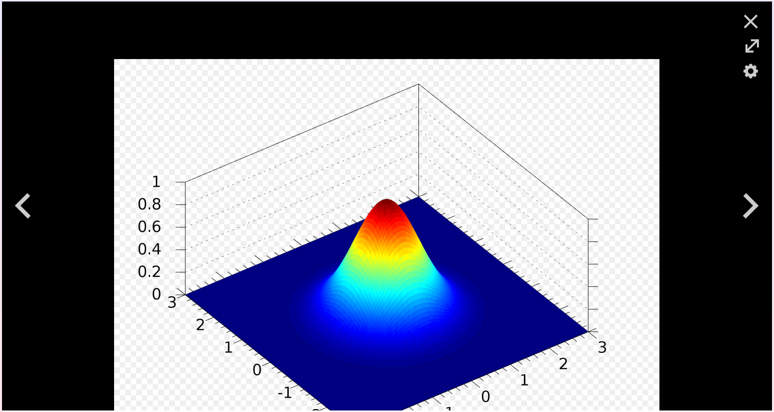
$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right)\right).$$

Observe the parameters $\mu_X, \mu_y, \sigma_x^2, \sigma_Y^2$ and ρ. The parameters will get a meaning later.

Useful notation:

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}.$$

A bivariate normal pdf



Gaussian curve with a two-dimensional domain

More details

Krishnavedala - Own work

File: Gaussian 2d.svg

Created: 27 September 2014

Isometric plot of a two dimensional Gaussian function. GNU Octave source code graphics_toolkit ("gnuplot"); % force use of gnuplot backend instead of FLTK for plot. Generates smaller SVG file [X, Y] = meshgrid(-3:.05:3, -3:.05:3); % smaller step size increases resolution and smoothness but increases file size Z = exp(-X.^2 - Y.^2); surf(X, Y, Z); view(-36, 56); shading flat; % remove edge lines on plot but keep color patches print('Gaussian_2d.svg')

Normal marginals

The bivariate normal has (by tedious computations!) an attractive consequence.

The marginal distributions of X and Y are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$. This statement has no converse, it may happen that X and Y are $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, but (X, Y) is **not** bivariate normal. Later more about ρ.

Outline

Definitions and characterization

- Two events E and F are called *independent* if the product rule $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$ holds.
- Two random variables (vectors) X and Y are independent if all events $\{X \in A\}$ and $\{Y \in B\}$ are independent, $\mathbb{P}(\{X \in A\} \cap \{Y \in B\}) = \mathbb{P}(\{X \in A\})\mathbb{P}(\{Y \in B\})$ for all (Borel) sets A, B.
- Characterization: Random variables X and Y are independent iff $\mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}) = \mathbb{P}(\{X \leq x\})\mathbb{P}(\{Y \leq y\})$ for all (Borel) sets x, y. [Proof omitted]
  Stated otherwise, the product rule for the distribution functions holds, $F(x, y) = F_X(x)F_Y(y)$ for all x, y.

Independence and densities

Suppose $F(x, y) = F_X(x)F_Y(y)$ for all x, y. Differentiate the product w.r.t. x to get $\frac{\partial F}{\partial x}(x, y) = f_X(x)F_Y(y)$. Differentiate once more, $\frac{\partial^2 F}{\partial y \partial x}(x, y) = f_X(x)f_Y(y)$.
If the latter product rule holds for all x, y, then by integration $F(x, y) = F_X(x)F_Y(y)$.

Assume the joint density f of (X, Y) exists. Then X and Y are independent iff $f(x, y) = f_X(x)f_Y(y)$ for all x, y.

Discrete X and Y are independent iff
$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$ for all x, y.

Independence and densities

Assume the joint density f of (X, Y) exists. Suppose $F(x, y) = F_X(x)F_Y(y)$ for all x, y. Differentiate the product w.r.t. x to get $\frac{\partial F}{\partial x}(x, y) = f_X(x)F_Y(y)$. Differentiate once more, $\frac{\partial^2 F}{\partial y \partial x}(x, y) = f_X(x)f_Y(y)$, so $f(x, y) = f_X(x)f_Y(y)$.
If the latter product rule holds for all x, y, then by integration $F(x, y) = F_X(x)F_Y(y)$.

It follows that X and Y are independent iff $f(x, y) = f_X(x)f_Y(y)$ for all x, y.

Discrete X and Y are independent iff $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$ for all x, y.

Transformations and independence

Proposition

If X and Y are independent and $U = g(X)$, $V = h(Y)$ for (measurable) functions g and h, then also U and V are independent.

Proof.

Let $u, v \in \mathbb{R}$, $A = \{g \leq u\} = \{x : g(x) \leq u\}$ and $B = \{h \leq v\}$. Then $\{U \leq u\} \cap \{V \leq v\} = \{g(X) \leq u\} \cap \{h(Y) \leq v\} = \{X \in A\} \cap \{Y \in B\}$.

Hence $\mathbb{P}(\{U \leq u\} \cap \{V \leq v\}) = \mathbb{P}(\{X \in A\})\mathbb{P}(\{Y \in B\}) = \mathbb{P}(\{U \leq u\}\mathbb{P}(\{V \leq v\})$. The result follows. $\square$

Independence and bivariate normality

The bivariate normal distribution has a simple characterization for independence of its marginals. Here the parameter ρ comes in.

Proposition

Let (X, Y) be bivariate normal. Then X and Y are independent random variables iff $\rho = 0$.

Proof.

We consider (in self-evident notation) $\frac{f(x,y)}{f_X(x)f_Y(y)}$ which has to be identically equal to one in case of independence. W.l.o.g. we assume $\mu_X = \mu_Y = 0$ and $\sigma_X = \sigma_Y = 1$. The mentioned identity then takes place iff $\rho^2 x^2 + \rho^2 y^2 - 2\rho xy$ is identically equal to zero which happens iff $\rho = 0$, $\square$

Sums of discrete independent random variables

Let X, Y be independent and $Z = X + Y$. Then ('cut up in slices')

$$\begin{aligned}\mathbb{P}(Z = z) = \mathbb{P}(X + Y = z) &= \sum_x \mathbb{P}(X + Y = z, X = x) \\ &= \sum_x \mathbb{P}(Y = z - x, X = x) \\ &= \sum_x \mathbb{P}(Y = z - x)\mathbb{P}(X = x) \\ &= \sum_x p_Y(z - x)p_X(x).\end{aligned}$$

Formula known as convolution formula. The summation is taken over those x for which the probabilities make sense (and do not equal zero).

Sums of continuous independent random variables

Let X, Y be independent with densities f_X and f_Y and $Z = X + Y$. Then (again a convolution formula)

$$\begin{aligned} f_Z(z) &= \int_{\mathbb{R}} f_Y(z-x) f_X(x) \,\mathrm{d}x \\ &= \int_{\mathbb{R}} f_Y(y) f_X(z-y) \,\mathrm{d}y. \end{aligned}$$

As in the discrete case, this is an abstract formula, to be used for computations in discrete situations.

Binomial example

Let X, Y be independent, X is $\mathrm{Bin}(n, p)$ and Y is $\mathrm{Bin}(m, p)$, and $Z = X + Y$. Then Z is $\mathrm{Bin}(n + m, p)$. *In particular we can also view X as the sum of n independent Bernoulli random variables.*

Let z be an integer between 0 and $n + m$. Then (with $\binom{n}{k} = 0$ for $k < 0$ or $k > n$)

$$
\begin{aligned}
\mathbb{P}(Z = z) &= \sum_{x=0}^{z} \mathbb{P}(Y = z - x)\mathbb{P}(X = x) \\
&= \sum_{x=0}^{z} \binom{m}{z-x} p^{z-x}(1-p)^{m-z+x} \binom{n}{x} p^{x}(1-p)^{n-x} \\
&= p^{z}(1-p)^{n+m-z} \sum_{x=0}^{z} \binom{m}{z-x}\binom{n}{x} \\
&\overset{!}{=} p^{z}(1-p)^{n+m-z} \binom{n+m}{z}.
\end{aligned}
$$

Gamma example

Let X, Y be independent, X is Gamma(α, λ) and Y is Gamma(β, λ), and $Z = X + Y$. Then Z is Gamma$(\alpha + \beta, \lambda)$.

Let $z > 0$. Then

$$\begin{aligned}
f_Z(z) &= \int_{\mathbb{R}} f_Y(z-x) f_X(x) \, \mathrm{d}x \\
&= \int_0^z \frac{\lambda^\beta}{\Gamma(\beta)} (z-x)^{\beta-1} e^{-\lambda(z-x)} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \, \mathrm{d}x \\
&= \frac{\lambda^{\beta+\alpha}}{\Gamma(\beta)\Gamma(\alpha)} e^{-\lambda z} \int_0^z (z-x)^{\beta-1} x^{\alpha-1} \, \mathrm{d}x \\
&\overset{x=zu}{=} \frac{\lambda^{\beta+\alpha}}{\Gamma(\beta)\Gamma(\alpha)} e^{-\lambda z} z^{\alpha+\beta-1} \int_0^1 (1-u)^{\beta-1} u^{\alpha-1} \, \mathrm{d}u \\
&= \text{`the desired expression'}.
\end{aligned}$$

More examples

Let X, Y be independent, X is Poisson(λ) and Y is Poisson(μ), and $Z = X + Y$. Then Z is is Poisson($\lambda + \mu$). [Summation as before.]

Let X, Y be independent, X is $N(\mu, \sigma^2)$ and Y is $N(\nu, \tau^2)$, and $Z = X + Y$. Then Z is $N(\mu + \nu, \sigma^2 + \tau^2)$. [Tedious computations, see Rice.]

Transformation rule (hardly used)

If X is an n-dimensional random vector and

$Y = g(X) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_n(x) \end{pmatrix}$, where $g : \mathbb{R}^n \to \mathbb{R}^n$ is invertible (with inverse h) and differentiable, then

$$f_Y(y) = \frac{f_X(h(y))}{|J(h(y))|},$$

where

$$J(x) = \det \begin{pmatrix} \frac{\partial}{\partial x_1} g_1(x) & \cdots & \frac{\partial}{\partial x_n} g_1(x) \\ \vdots & & \vdots \\ \frac{\partial}{\partial x_1} g_n(x) & \cdots & \frac{\partial}{\partial x_n} g_n(x) \end{pmatrix}.$$

See Rice for examples (especially for bivariate normal distributions).

Transformation rule, linear case

Let X be a n-dimensional random vector and $Y = AX + b$, where A is an invertible matrix and b a n-dimensional vector. Then $g(x) = Ax + b$ and $J(x) = \det(A)$. Hence

$$f_Y(y) = \frac{f_X(A^{-1}(y-b))}{|\det(A)|}.$$

If X is bivariate normal ($n = 2$), a (tedious) computation shows that also Y is bivariate normal, with corresponding μ-vector and Σ-matrix

$$A\mu + b \text{ and } A\Sigma A^\top$$

respectively. More (without tedious computations) on this later in the course.

Final remark

Be sure to have studied these slides and the corresponding parts in Rice before the start of lecture 2.