Lecture Notes for Statistics and Estimation Theory

Paul Ferrand

2019

Prerequisites

Linear algebra

We will consider matrices and vectors, noted as uppercase bold $\boldsymbol{A}$ and lowercase bold $\boldsymbol{v}$ respectively. A matrix is a collection of elements arranged in m rows and n columns. Vectors are matrices with a single column. We have

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{1m} & \dots & a_{mn} \end{pmatrix} \tag{1}$$

The transpose of a matrix is

$$\boldsymbol{A}^T = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{1m} & \dots & a_{mn} \end{pmatrix} \tag{2}$$

A matrix is symmetric if $\boldsymbol{A}^T = \boldsymbol{A}$, and it is square if $m = n$. You can add matrices and vectors if they have the same size. Matrices can be multiplied as follows. Assume that $\boldsymbol{A}$ is $m \times n$ and $\boldsymbol{B}$ is $n \times p$. Note that the number of columns of $\boldsymbol{A}$ needs to be equal to the number of columns of $\boldsymbol{B}$. Then

$$\boldsymbol{C} = \boldsymbol{A} \cdot \boldsymbol{B} \text{ and } c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \tag{3}$$

When $\boldsymbol{A} = \boldsymbol{a}^T$ is actually a *row* vector and $\boldsymbol{b}$ is a *column* vector, this multiplication outputs a scalar, and the operation is called a scalar product. It is sometimes denoted as

$$\boldsymbol{a}^T \boldsymbol{b} = \langle \boldsymbol{a} \,|\, \boldsymbol{b} \rangle \tag{4}$$

A matrix multiplication can thus be seen as a collection of scalar products between the rows of a matrix and the columns of another matrix—this can come in handy at times.

Scalar products Scalar products are used to check *orthogonality* in arbitrary spaces. In particular, we will say that 2 vectors are orthogonal if their scalar product is equal to 0. For vectors in 2 dimensions, this directly corresponds to the natural orthogonality that we learn since

childhood, and extends in arbitrary dimensions. We can also define the *norm* of a vector through scalar products. We will say that

$$\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}^T\boldsymbol{v}} \tag{5}$$

which in turns, allows to define an *angle* between arbitrary vectors through their cosine:

$$\cos\left[\angle(\boldsymbol{a},\boldsymbol{b})\right] = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\|\boldsymbol{a}\|\|\boldsymbol{b}\|} \tag{6}$$

Orthogonal vectors have their cosine equal to 0—once again, as we learned before. In contrast, *colinear vectors* have their cosines equal to 1, and thus the scalar product of colinear vectors is equal to the product of their norms.

Gram-Schmidt There is more to orthogonality; consider 2 vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ that are just arbitrary. We can split $\boldsymbol{b}$ into a part that is colinear with $\boldsymbol{a}$ and a part that is orthogonal to $\boldsymbol{a}$ as follows

$$\boldsymbol{b} = \frac{\boldsymbol{a}^T\boldsymbol{b}}{\|a\|^2}\boldsymbol{a} + \left(\boldsymbol{b} - \frac{\boldsymbol{a}^T\boldsymbol{b}}{\|a\|^2}\boldsymbol{a}\right) \tag{7}$$

The first term represent the colinear part and the second one between parentheses the orthogonal one. Through standard algebra you can verify that the above is obviously true. This can be used to create a set of orthogonal vectors through a set of arbitrary ones. Let such a set be denoted $\{\boldsymbol{u}_k\}$. We proceed iteratively, forming the set $\{\boldsymbol{v}_k\}$ as

$$\boldsymbol{v}_0 = \boldsymbol{u}_0 \tag{8}$$

$$\boldsymbol{v}_1 = \boldsymbol{u}_1 - \frac{(\boldsymbol{v}_0^T\boldsymbol{u}_1)}{\|v_0\|^2}\boldsymbol{v}_0 \tag{9}$$

$$\boldsymbol{v}_2 = \boldsymbol{u}_2 - \frac{(\boldsymbol{v}_0^T\boldsymbol{u}_2)}{\|v_0\|^2}\boldsymbol{v}_0 - \frac{(\boldsymbol{v}_1^T\boldsymbol{u}_2)}{\|v_1\|^2}\boldsymbol{v}_1 \tag{10}$$

$$\vdots \tag{11}$$

This operation is sometimes called an orthogonalization or a *Gram-Schmidt* procedure, and we say that $\{\boldsymbol{v}_k\}$ is a set of linearly independent vectors. One can indeed verify that $\boldsymbol{v}_i^T\boldsymbol{v}_j = 0$ by construction for any i, and j. If we normalize the vectors in the set, they will form a *basis*. The number of vectors in the basis is related to the dimension of the linear subspace described by the vectors, and is intuitively linked to the number of *degrees of freedom* of the linear space.

Verify that this indeed creates a set of orthogonal vectors.

Rank and identity matrix Back to matrices, we define the *rank* of a matrix as the number of linearly independent rows or columns, whichever is the smallest. There are alternative characterizations but

this one is practical. In particular, the number of linearly independent columns are linked to the descriptive power of the matrix since we can form a basis with the columns. Among specific matrices, we call the one that is neutral with respect to multiplication the identity matrix. It is a square matrix with ones on the diagonal, and it verifies that

$$\boldsymbol{A}\boldsymbol{I} = \boldsymbol{I}\boldsymbol{A} = \boldsymbol{A} \tag{12}$$

where the sizes of $\boldsymbol{I}$ are chosen so that the multiplication makes sense. This neutral element allows one to define the inverse of a matrix, which is *any* matrix such that

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I} \tag{13}$$

This definition only really makes sense when $\boldsymbol{A}$ is square, although there are *pseudo-inverses* in some other cases. When $\boldsymbol{A}$ is square, the inverse of the matrix will exist and be unique *if and only if (iif)* the rank of $\boldsymbol{A}$ is equal to n. If not, $\boldsymbol{A}$ is singular or *rank-deficient* and basically has no inverse. A matrix is said to be *orthogonal* if its columns and rows are orthogonal pairwise, and if

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^T \qquad \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I} \tag{14}$$

Determinant and trace Next we define the determinant of a square matrix as the scalar value constructed in a recursive manner as

$$\det \boldsymbol{A} = \sum_{j=1}^{n} a_{ij} C_{ij} \tag{15}$$

The values C_{ij} are the *cofactors at ij* of the matrix $\boldsymbol{A}$ obtained as

$$C_{ij} = (-1)^{i+j} \det \boldsymbol{M}_{ij} \tag{16}$$

The matrix $\boldsymbol{M}_{ij}$ is obtained by deleting the i-th row and j-th column of det A—hence the recursive definition. The cofactors are often considered as a matrix $\boldsymbol{C}$, which is important because the inverse of a matrix is defined through this matrix, as

$$\boldsymbol{A}^{-1} = \frac{\boldsymbol{C}^T}{\det \boldsymbol{A}} \tag{17}$$

It is the only general analytical formula to invert an arbitrary matrix; when $\boldsymbol{A}$ has some specific structure, there may be easier inverse formulas. We can also define the *trace* of a matrix as the sum of its diagonal elements

$$\mathrm{tr}\boldsymbol{A} = \sum a_{ii} \tag{18}$$

Both the determinant and the trace express profound things about a matrix; we will see shortly what that is.

Quadratic forms and definiteness A quadratic form is a function of a vector $\boldsymbol{x}$ defined through a matrix $\boldsymbol{A}$ as

$$Q(\boldsymbol{x}) = \sum_i \sum_j A_{ij} x_i x_j = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \tag{19}$$

The matrix $\boldsymbol{A}$ is supposed to be symmetric, which entails no loss in generality since any quadratic form may be expressed in this way. If for all $\boldsymbol{x} \neq \boldsymbol{0}$ we have

Exercise: show why.

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \geq 0 \tag{20}$$

the matrix $\boldsymbol{A}$ is said to be positive semi-definite. If the inequality is strict, $\boldsymbol{A}$ is said to be positive definite.

Properties of matrices, determinants and traces

- $(\boldsymbol{AB})^T = \boldsymbol{B}^T \boldsymbol{A}^T$
- $(\boldsymbol{A}^{-1})^T = (\boldsymbol{A}^T)^{-1} = \boldsymbol{A}^{-T}$
- $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1} \boldsymbol{A}^{-1}$
- $\det \boldsymbol{A}^T = \det \boldsymbol{A}$
- $\det(c\boldsymbol{A}) = c^n \det \boldsymbol{A}$
- $\det(\boldsymbol{AB}) = \det \boldsymbol{A} \det \boldsymbol{B}$
- $\det \boldsymbol{A}^{-1} = (\det \boldsymbol{A})^{-1}$
- $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$
- $\mathrm{tr}(\boldsymbol{A}^T \boldsymbol{B}) = \sum_i \sum_j A_{ij} B_{ij}$
- $\det \boldsymbol{I} = 1$
- If $\boldsymbol{D}$ is a diagonal matrix, then $\det \boldsymbol{D} = \prod_i d_{ii}$
- $\mathrm{tr}(c) = c$ for all scalars c

Some theorems

1. A square matrix $\boldsymbol{A}$ is invertible *if and only if* its determinant is non-zero.
2. A square matrix $\boldsymbol{A}$ is positive definite iif it can be written as $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}^T$ for some matrix $\boldsymbol{C}$ that is also full-rank and invertible.
3. A square matrix $\boldsymbol{A}$ is positive definite iif all its principal minors are positive[1]
4. If $\boldsymbol{C}$ is not full-rank or one of the principal minors is equal to 0, $\boldsymbol{A}$ is only positive semidefinite.

[1] The principal minors are the determinants of the top-leftmost square matrices of increasing size.

5. If $\boldsymbol{A}$ is positive definite, then for some full rank $m \times n$ matrix $\boldsymbol{B}$ with $m \leq n$, $\boldsymbol{BAB}^T$ is also positive definite.

6. If $\boldsymbol{A}$ is positive definite, then its diagonal elements are positive and the determinant of $\boldsymbol{A}$ is positive. If $\boldsymbol{A}$ is only positive semi-definite, the previous properties are relaxed to non-negative—i.e. they can be equal to 0.

Matrix inversion lemma A very common and useful property is stated as follows

Proof of this?

$$(\boldsymbol{A} + \boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{DA}^{-1}\boldsymbol{B} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{DA}^{-1} \tag{21}$$

It relates the inverse of sum and product of matrices with the sum and products of the component inverse. Since the inverse is a complex operation, it is very useful numerically. As an example, it can be used to compute the so-called *rank-1* update of the inverse of the matrix $\boldsymbol{A}$

$$(\boldsymbol{A} + \boldsymbol{uu}^T)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{uu}^T\boldsymbol{A}^{-1}}{1 + \boldsymbol{u}^T\boldsymbol{A}^{-1}\boldsymbol{u}} \tag{22}$$

where $\boldsymbol{u}$ is a column vector and $\boldsymbol{uu}^T$ is the *outer product* of $\boldsymbol{u}$ with itself[2]. This matrix has rank 1 since its columns are all colinear with $\boldsymbol{u}$. This operation happens usually to integrate a newly measured value into a matrix $\boldsymbol{A}$ for which we already computed the inverse. Here we never invert a matrix, but we use the already known value of $\boldsymbol{A}^{-1}$ to compute the new value after adding a rank-1 matrix. If furthermore $\boldsymbol{A}$ is symmetric, then $\boldsymbol{A}^{-1} = \boldsymbol{A}^{-T}$ and denoting $\boldsymbol{v} = \boldsymbol{A}^{-1}\boldsymbol{u}$ we can rewrite (22) as

[2] $\boldsymbol{uu}^T$ is a matrix and is component at row i and column j is equal to $u_i u_j$

$$(\boldsymbol{A} + \boldsymbol{uu}^T)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{vv}^T}{1 + \boldsymbol{u}^T\boldsymbol{v}} \tag{23}$$

Eigendecompositions An eigenvector $\boldsymbol{v}$ of a matrix is the square satisfies

$$\boldsymbol{Av} = \lambda\boldsymbol{v} \tag{24}$$

for some scalar λ, which may be complex in general even if $\boldsymbol{A}$ only has real coefficients. We call λ the eigenvalue associated to the eigenvector $\boldsymbol{v}$. By convention, $\|\boldsymbol{v}\| = 1$. There are many properties on eigenvalues depending on the matrix./ In general, our interest will be on symmetric and possibly positive definite matrices, for which the eigenvalues are respectively real and positive, and whose eigenvectors are orthogonal if they correspond to different eigenvalues. In particular, this means that we can write

Exercise: what are the eigenvectors of the identity matrix?

$$\boldsymbol{A}\left[\boldsymbol{v}_1 \cdots \boldsymbol{v}_n\right] = \left[\lambda_1\boldsymbol{v}_1 \cdots \lambda_n\boldsymbol{v}_n\right] \tag{25}$$

or written differently

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Lambda} \tag{26}$$

with

$$\boldsymbol{V} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_n] \qquad \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

Since $\boldsymbol{V}$ is formed by orthogonal vectors you can rewrite (26) as

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T = \sum \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^T \tag{27}$$

EXERCISE *How does the eigen decomposition helps in computing the inverse of a matrix?*

We have

$$\boldsymbol{A}^{-1} = \boldsymbol{V}^{-T}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^{-1} = \boldsymbol{V}\boldsymbol{\Lambda}^{-1}\boldsymbol{V}^T = \sum \frac{1}{\lambda_i} \boldsymbol{v}_i \boldsymbol{v}_i^T$$

EXERCISE *What is the determinant of an orthogonal matrix? What can you say about the determinant of a square matrix with respect to its eigenvalues? What about the trace?*

If $\boldsymbol{V}$ is orthogonal, then $\boldsymbol{V}^T = \boldsymbol{V}^{-1}$ which is equivalent to

$$\det \boldsymbol{V} = (\det \boldsymbol{V})^{-1} \Leftrightarrow \det \boldsymbol{V} = \pm 1$$

In turn, this means that

$$\det \boldsymbol{A} = \det \boldsymbol{V} \det \boldsymbol{\Lambda} (\det \boldsymbol{V}^T) = \det \boldsymbol{\Lambda} = \prod \lambda_i$$

For the trace, remark that

$$\operatorname{tr}(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T) = \operatorname{tr}(\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Lambda}) = \operatorname{tr}(\boldsymbol{\Lambda}) = \sum \lambda_i$$

Probability

We will consider a simplistic view of probability, one that serves our interest without getting in the way or introducing hard concepts unless they provide a very strong intuition about a problem. A *random variable* can be considered as a mathematical object that can take any value on a specific domain. This domain may be purely discrete, purely continuous, or something in-between. This last possibility creates a host of practical issues that require something called *measure theory* to solve adequately. As such, we will try to avoid these random variables and only consider the pure ones, remembering that there is a unifying view behind these.

Random variables are usually noted as upper-case letters as X and Y, and the values they take on their domains as lowercase letters. A *realization* of a random variable takes a value, which in the discrete case is naturally noted as the *event* $X = x$. In a continuous domain, there is an uncountable infinity of possible values for the random

variable to take, and therefore there is not much interest in a specific values, but rather in ranges of values. In this case, we are more interested in events $a \leq X \leq b$ for some $a < b$. If the domain of the r.v. is $(-\infty, \infty)$ we will mostly consider the event $X \leq x$.

Cumulative and probability functions To each event we associate a probability, meaning that we are able to measure the relative weight of the event among all possible realizations of the random variable, assuming that the sum of event weights are equal to 1. For discrete variables, this measure of weight or probability *law* is given as a function of the $x \in \mathcal{D}$ where $\mathcal{D}$ is the definition domain, as

$$\mathbb{P}(X = x) \tag{28}$$

This function can sometimes be parametrized, something that will be of great interest in the rest of the course. A common discrete probability law is the Poisson law, where for an integer $k > 0$ we have

$$\mathbb{P}(X = k) = \frac{\theta^k}{k!} e^{-\theta} \tag{29}$$

In the continuous case, the weight of the event is related to the range and defined through a continuous *probability density function* or *p.d.f.* $f(x)$, so that

$$\begin{aligned} F(b) - F(a) = \mathbb{P}(a \leq x \leq b) &= \int_a^b f(x)dx \\ &= \int_{-\infty}^b f(x)dx - \int_{-\infty}^a f(x)dx \end{aligned} \tag{30}$$

The cumulative distribution function $F(x)$ is thus defined as

$$F(x) = \int_{-\infty}^x f(y)dy \tag{31}$$

Once again, care must be taken but when this integral is proper, and $f(x)$ is well behaved, it usually poses no issue. Note that you should definitely not think about $f(x)$ as the probability of the event $X = x$; as said before, the probability of this event is virtually zero and uninteresting. In particular, although in the discrete case we have by construction that $P(X = x) \leq 1$ since the sum of the probability weights over all the possible events sums to 1, in the continuous case, assuming the domain is $(-\infty, \infty)$, we want

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{32}$$

which does not preclude $f(x)$ to be stricly larger than 1 for some x.

Show that this is true for the Poisson law; solution uses the fact that

$$e^\theta = \sum_{k=0}^{\infty} \frac{\theta^k}{k!}$$

Usual laws Among the laws used the most for continuous r.v., the most common one is definitely the Gaussian or so called *normal* law, denoted as $\mathcal{N}(\mu, \sigma^2)$ and whose p.d.f. is

$$f(x; \sigma, \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{33}$$

We include the parameters in the function using the notation $f(x; \mu, \sigma)$. As we gravitate towards estimation, this notation will prove useful. Another common law is the continuous distribution between endpoints a and b, usually written as $\mathcal{U}(a, b)$

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

The exponential law $\mathcal{E}(\theta)$ has density

$$f(x; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \tag{35}$$

and is defined on $[0, \infty]$. The parameter θ is sometimes defined as $\mu = \theta^{-1}$. In this case, the density is

$$f(x; \mu) = \mu e^{-\mu x}. \tag{36}$$

Finally, the χ^2 distribution regularly appears throughout probability and statistics. In particular, if you have r.vs. X_i distributed as standard Gaussian $\mathcal{N}(0, 1)$ and are independent[3] then the sum of their squares

[3] We'll see later what that means exactly

$$Y = \sum_{i=0}^{n} X_i^2 \tag{37}$$

follows a $\chi^2(n)$ law. The parameter K is the number of independent Gaussians in the sum and is dubbed *degrees of freedom* in most literature. The p.d.f. of a χ^2 random variable is

$$f(y; n) = \begin{cases} \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} \exp\left(-\frac{1}{2}y\right) & y \geq 0 \\ 0 & y < 0 \end{cases} \tag{38}$$

The function $\Gamma(\cdot)$ is the Gamma integral; it is a common special function that behaves like the continuous equivalent of the factorial function.

Expectation, moments and quantiles The expectation of a random variable is defined as

$$\mathbb{E}[X] = \sum_i x_i \mathbb{P}(X = x_i) \tag{39}$$

if X is a discrete r.v. and takes on values in a set $\{x_i\}$ and

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{40}$$

if X is a continuous r.v. The integration is limited to the domain if it is less than the entire real line. This basic definition is usually called the mean, although expectation is actually defined for any function $g(X)$, for example in the continuous case[4]

Note that the expectation may not exist, in which case usually the integral definition will diverge or be improper. Sometimes you can salvage this with measure theory, but there exists even common cases where you can't—e.g. Cauchy r.v.s.

[4] Unless specified, the discrete case mirrors all definitions.

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx. \tag{41}$$

Since it is defined as an integral, the expectation is linear, in the sense that if $g(X) = aX + b$ for some scalars a and b, we have

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b. \tag{42}$$

Among possible functions $g(\cdot)$, monomials x^k are called k-th order moments, and

$$\mathbb{E}\left[(X - \mathbb{E}[X])^k\right] \tag{43}$$

are called *centered moments*. The centered moment of order 2 is called *variance* and holds a very special place, so it sometimes has its own notation

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \tag{44}$$

$$\begin{aligned}
&\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\
=&\mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\right] \\
=&\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2
\end{aligned}$$

There is an interesting interplay between the mean and variance. Consider a r.v. X such that $\mathbb{E}[X^2] < \infty$. Then for all c in $\mathbb{R}$ we have that

$$\begin{aligned}
\mathbb{E}\left[(X - c)^2\right] &= \mathbb{E}[X^2] - 2\mathbb{E}[X]c + c^2 \\
&= \mathbb{E}[X]^2 - 2\mathbb{E}[X]c + c^2 + \mathrm{Var}(X) \\
&= (\mathbb{E}[X] - c)^2 + \mathrm{Var}(X)
\end{aligned}$$

This also says that there is an extremal for the mean with respect to the variance, in the sense that $\mu = \mathbb{E}[X]$ iif

$$\mathbb{E}\left[(X - \mu)^2\right] = \min_{c \in \mathbb{R}} \mathbb{E}\left[(X - c)^2\right] \tag{45}$$

Any other choice of c beyond the mean leads to a higher value for $\mathbb{E}\left[(X - c)^2\right]$.

Among other characterizations of an r.v. an interesting one is the *quantile*, defined as the solution[5] to

[5] This definition only really makes sense for well-behaved continuous r.v.

$$F(q_p) = p \tag{46}$$

for some $0 \leq p \leq 1$. Within quantiles the value $M = q_{1/2}$ is called the *median* and has special properties. As per the definition, we have that

$$\mathbb{P}(X \geq M) \geq 1/2 \text{ and } \mathbb{P}(X \leq M) \geq 1/2 \tag{47}$$

Like the mean, the median characterizes the *position* or *location* of the r.v., whereas values like the variance characterize its *scale*. Unlike the mean however the median is always defined. It also has an extremal property, where

TODO: understand the derivation

$$\mathbb{E}\left[|X - M|\right] = \min_{c \in \mathbb{R}} \mathbb{E}\left[|X - c|\right] \tag{48}$$

However this point may not be unique, since in the very general case there may be many possible values of a verifying (47).

Inequalities There are interesting inequalities in probability theory that should be known. If anything, remember that they exist so that you can check whether they can help you derive bounds on quantities of interest involving random variables. They are also used in proofs a lot, for obvious reason: in mathematics, bounding a value is a very powerful statement.

Proposition 1 (Markov inequality)**.** *Let $h(\cdot)$ be an increasing and positive function and X a non-negative r.v. such that $\mathbb{E}[h(X)] < \infty$. Then for any a such that $h(a) > 0$ we have*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}\left[h(X)\right]}{h(a)} \tag{49}$$

Proof. $h(\cdot)$ is an increasing positive function, so

$$\mathbb{P}(X \geq a) \leq \mathbb{P}(h(X) \geq h(a)) = \int \mathbb{1}_{h(X) \geq h(a)} f(x) dx \tag{50}$$

This notation uses indicators rather than bound; notice that

$$P(h(X) \geq h(a)) = \mathbb{E}\left[\mathbb{1}_{h(X) \geq h(a)}\right]$$

which is the basis of a whole way to define probability theory through expectations.

Then

$$\int \mathbb{1}_{h(X) \geq h(a)} f(x) dx = \mathbb{E}\left[\mathbb{1}_{h(X) \geq h(a)}\right] \leq \mathbb{E}\left[\frac{h(X)}{h(a)} \mathbb{1}_{h(X) \geq h(a)}\right] \tag{51}$$

Then

$$\frac{\mathbb{E}\left[h(X)\mathbb{1}_{h(X) \geq h(a)}\right]}{h(a)} = \frac{\mathbb{E}\left[h(X)\right]}{h(a)} - \frac{\mathbb{E}\left[h(X)\mathbb{1}_{h(X) < h(a)}\right]}{h(a)} \tag{52}$$

The last term is positive since $h(\cdot)$ is positive; removing it thus leads to the expected result. □

Proposition 2 (Chebyshev inequality)**.** *Let X be an r.v. with $\mathbb{E}[X^2] < \infty$. Then assuming $a > 0$*

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}\left[X^2\right]}{a^2} \qquad \mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var} X}{a^2} \tag{53}$$

Proof. Use Markov's inequality. □

Proposition 3 (Hölder inequality). *Let $1 < r < \infty$ and $1/r + 1/s = 1$. Let X and Y be 2 r.v. such that $\mathbb{E}\left[|X|^r\right] < \infty$ and $\mathbb{E}\left[|Y|^s\right] < \infty$. Then $\mathbb{E}\left[|XY|\right] < \infty$ and*

$$\mathbb{E}\left[|XY|\right] \leq \mathbb{E}\left[|X|^r\right]^{\frac{1}{r}} \mathbb{E}\left[|Y|^s\right]^{\frac{1}{s}} \tag{54}$$

Proof. Remark that

$$\frac{1}{r}\log a + \frac{1}{s}\log b \leq \log\left(\frac{a}{r} + \frac{b}{s}\right) \tag{55}$$

for $a > 0$ and $b > 0$, since log is a concave function. This is equivalent to

$$a^{\frac{1}{r}} b^{\frac{1}{s}} \leq \frac{a}{r} + \frac{b}{s} \tag{56}$$

Then, substitute

$$a = \frac{|X|^r}{\mathbb{E}\left[|X|^r\right]} \qquad b = \frac{|Y|^s}{\mathbb{E}\left[|Y|^s\right]} \tag{57}$$

which leads to

$$|XY| \leq \mathbb{E}\left[|X|^r\right]^{\frac{1}{r}} \mathbb{E}\left[|Y|^s\right]^{\frac{1}{s}} \left(\frac{|X|^r}{r\mathbb{E}\left[|X|^r\right]} + \frac{|Y|^s}{s\mathbb{E}\left[|Y|^s\right]}\right) \tag{58}$$

Taking the expectation on both sides gives the proof. □

Proposition 4 (Lyapunov inequality). *Let $0 < v < t$ and $1/r + 1/s = 1$. Let X $\mathbb{E}\left[|X|^t\right] < \infty$. Then $\mathbb{E}\left[|X|^v\right] < \infty$ and*

$$\mathbb{E}\left[|X|^v\right]^{\frac{1}{v}} \leq \mathbb{E}\left[|X|^t\right]^{\frac{1}{t}} \tag{59}$$

Proof. Use Hölder's with $X = X^v$, $Y = 1$, $r = t/v$. □

This inequality implies that

$$\mathbb{E}[|X|] \leq \mathbb{E}[|X|^2]^{\frac{1}{2}} \leq \cdots \leq \mathbb{E}[|X|^k]^{\frac{1}{k}}$$

For real variables this means that $\mathbb{E}[|X|] < \infty$ is always true if $\mathbb{E}[X^2] < \infty$.

Proposition 5 (Jensen inequality). *If $g(\cdot)$ is a convex function and $\mathbb{E}[|X|] < \infty$, then*

$$g(\mathbb{E}\left[X\right]) \leq \mathbb{E}[g(X)] \tag{60}$$

Proof. Since $g(\cdot)$ is convex, for any point a in $\mathbb{R}$ there is an affine function $\ell_a(x)$ which is dominated by $g(x)$ and for which $\ell_a(a) = g(a)$. If $g(\cdot)$ is differentiable, this is the tangent of $g(\cdot)$ at a. Then for the r.v. X we have that

$$\ell_a(X) \leq g(X) \Longrightarrow \mathbb{E}\left[\ell_a(X)\right] \leq \mathbb{E}\left[g(X)\right]$$

Since $\ell_a(\cdot)$ is linear,

$$\mathbb{E}\left[\ell_a(X)\right] = \ell_a\left(\mathbb{E}[X]\right)$$

Then setting $a = \mathbb{E}[X]$ finishes the proof. □

Proposition 6 (Cauchy-Schwarz inequality). *Let X and Y be 2 r.v. with finite variance. Then*

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[|XY|]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2] \tag{61}$$

Proof. Note that the second inequality can be obtained through Hölder's. By Jensen's inequality, we have that

$$|\mathbb{E}[X]| \leq \mathbb{E}[|X|] \tag{62}$$

which can be used to prove the first inequality. □

Independence Let us start by defining the joint repartition of a couple of continuous r.v. X and Y as

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y) \tag{63}$$

Notice once again that we are weighing a set of values defined by bounds on said values. The distribution function provides the weight to assign to the set, in a way. If everyone is well behave, we can define the joint p.d.f. as

$$f_{X,Y}(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y} \tag{64}$$

The density should verify

$$\iint_{\mathbb{R}^2} f_{X,Y}(x, y) dx dy = 1 \tag{65}$$

as in the single r.v. case—we are still normalizing our probabilities to 1. The *marginal* p.d.f. of X and Y are obtained by integrating over a single variable

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx \tag{66}$$

Knowing the marginals does not give you the joint distribution, but the opposite is true. In the discrete case the joint p.d.f. is naturally defined through $\mathbb{P}(X = x, Y = y)$.

We can now define independence as follows: we say that 2 r.v. are independent iif their *joint* p.d.f. factors in separate parts depending only on their respective variables, i.e.

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \tag{67}$$

The separate terms are in fact the marginals; you can check this by definition. We can revisit our previous statement: knowing the marginals gives you the joint distribution iif the variables are independent. If we also look at the joint expectation of X and Y through the straightforward definition, we see that

$$\mathbb{E}[XY] = \iint xy f(x, y) dy dx \tag{68}$$

Whenever X and Y are independent, this can be turned into

$$\begin{aligned}\mathbb{E}[XY] &= \iint xyf(x)f(y)dydx \\ &= \int xf(x)\left(\int yf(y)dy\right)dx \\ &= \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

The separation of the integrals is known as Fubini's theorem or Tonelli's variant. It requires conditions on $f(x)$ and $f(y)$ which are a bit beyond the scope of this class. We will just say that for most non-pathological cases at hand they are indeed satisfied. Note here that while independence implies that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, having $\mathbb{E}[X]\mathbb{E}[Y]$ is not a sufficient condition for independence of X and Y.

Conditioning If the variable are not independent, we can develop on how they depend on one another. Here, the joint p.d.f. will always tell you everything you would need, but in practice it is not always the most practical to manipulate or observe. In this case, you can try to study the conditioned r.v. $Y|X$, with its p.d.f. defined as

Check that $Y|X$ is indeed an r.v., and check what happens when X and Y are independent.

$$f(y|x) = \begin{cases} \dfrac{f(x,y)}{f(x)} & \text{if } f(x) > 0 \\ f(y) & \text{if } f(x) = 0 \end{cases} \tag{69}$$

We also define the conditional expectation as a function of x through

$$\mathbb{E}(Y|X = x) = \int yf(y|x)dy \tag{70}$$

This last value is a scalar for any x. However, X is itself a r.v. so in full generality we may also see $\mathbb{E}[Y|X]$ as a r.v. As an expectation, it also has many properties that we expect, among which linearity, as well as

TODO: can the last one be shown without measure getting in the way? It's more or less OK in the discrete case.

1. $\mathbb{E}[Y|X] = \mathbb{E}[Y]$ if X and Y are independent
2. $\mathbb{E}[h(X)|X] = h(X)$ for well-behaved functions $h(\cdot)$
3. $\mathbb{E}[h(X,Y)|X = x] = \mathbb{E}[h(x,Y)|X = x]$

Covariance Dependence and conditioning can be a bit annoying to manipulate in practice, and in many cases they are very hard to characterize. Let us define the covariance between 2 r.v. with finite variance as

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{71}$$

There are some specific and useful properties for the covariance

Blitz exercises

1. $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$
2. $\mathrm{Cov}(aX, bY) = ab\mathrm{Cov}(X, Y)$
3. $\mathrm{Cov}(X + a, Y) = \mathrm{Cov}(X, Y)$
4. $\mathrm{Cov}(Y, X) = \mathrm{Cov}(X, Y)$
5. $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$
6. $\mathrm{Cov}(X, Y) = 0$ if X and Y are independent

Random vectors Beyond couples of random variables, we will encounter a lot of random vectors within this course. A random vector is just a collection of r.v. indexed and stacked into a vector $\boldsymbol{x} = (X_1, \ldots, X_n)^T$. We can define its mean as the vectors of means of the individual components, i.e.

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{x}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix} \tag{72}$$

Similarly, we can look at the covariance between the components of the vector, e.g. $\mathbb{E}[X_i X_j]$ for $1 \leq i \leq n$ and $1 \leq j \leq n$. This also has a nice matrix through what is sometimes called the *dyadic* or *outer* product

$$\boldsymbol{C} = \mathbb{E}\left[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T\right] \tag{73}$$

and we take the expectation on each component of the matrix.

Exercise: show that the covariance matrix is positive semidefinite. Consider a r.v. $\boldsymbol{c}$ that is centered w.l.o.g. and has covariance matrix $\boldsymbol{C}$. For any $\boldsymbol{x}$ we have that $\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x} = \boldsymbol{x}^T\mathbb{E}[\boldsymbol{c}\boldsymbol{c}^T]\boldsymbol{x} = \mathbb{E}[(\boldsymbol{x}^T\boldsymbol{c})(\boldsymbol{c}^T\boldsymbol{x})]$. This is the variance of some random r.v. $z = \boldsymbol{x}^T\boldsymbol{c}$, and since the variance is always positive we deduce that $\boldsymbol{C}$ is positive definite.

Among the possible random vectors an astounding majority of practical—or should we say solvable—problems are related to the multivariate Gaussian distribution, which is the extension of the Gaussian distribution to larger vectors using the aforementioned mean and covariance. The p.d.f. of this distribution is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{N}{2}}\sqrt{|\boldsymbol{C}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{C}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right] \tag{74}$$

Exercise: what happens where the components of $\boldsymbol{x}$ are independent?

You can see that this indeed looks like a *p.d.f.*; the matrix operations within the exponential are a quadratic form between the *centered* vector $\boldsymbol{x}$ and the *inverse* covariance matrix $\boldsymbol{C}$. This interpretation will come in handy for us later.

A very important property of the multivariate Gaussian is that all linear transformations of a multivariate Gaussian random vector are still multivariate Gaussian random vectors. In other words, if we define

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \tag{75}$$

and $\boldsymbol{x}$ follows a multivariate Gaussian distribution, then so does $\boldsymbol{y}$. Obviously, the parameters of the distribution are different, but we can compute them easily using the linearity of the expectation:

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] + \boldsymbol{b} \tag{76}$$

and

$$\mathbb{E}\left[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^T\right] = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T \tag{77}$$

These expressions are actually not conditioned on the fact that $\boldsymbol{y}$ and $\boldsymbol{x}$ are multivariate Gaussian; they are properties of the expectation for any random vectors, and analogous to the scalar cases $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ and $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

EXERCISE *Assuming $X_1, \ldots, X_n$ are independent variables, let*

$$X_{\min} = \min(X_1, \ldots, X_n) \qquad X_{\max} = \max(X_1, \ldots, X_n)$$

Show that

$$\mathbb{P}(X_{\min} > x) = \prod_i \mathbb{P}(X_i > x) \qquad \mathbb{P}(X_{\max} < x) = \prod_i \mathbb{P}(X_i < x)$$

If $X_1, \ldots, X_n$ are uniform random variables, compute $\mathbb{E}[X_{\max}]$ and $\mathrm{Var}(X_{\max})$.

EXERCISE *Let X be a positive r.v. with finite expectation. Show that*

Hint: the expectation of an indicator is the probability.

$$\mathbb{E}[X] = \int_0^\infty (1 - F(x))dx = \int_0^\infty P(X > x)dx \tag{78}$$

EXERCISE *Let X, Y_1 and Y_2 be independent r.v. such that Y_1 and Y_2 are $\mathcal{N}(0,1)$. Let*

$$Z = \frac{Y_1 + XY_2}{\sqrt{1 + X^2}}$$

Show that Z is $\mathcal{N}(0,1)$ using $\mathbb{P}(Z \leq z | X = x)$.

EXERCISE *Let X and N be 2 r.v. with finite absolute expectation and N takes values in the strictly positive integers. Let $X_1, X_2, \ldots$ be the sequence of r.v. with the same law as X. Using conditioning, show Wald's inequality*

$$\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[N]\mathbb{E}[X]$$