# 1 Probability Theory Basics

As elaborated on in subsequent chapters, in econometrics, we interpret data $x$ as realizations of a random variable (RV) $X$. This RV follows some probability distribution $p(X|\theta)$, which is indexed by a parameter $\theta$ that determines its shape and properties. In Sections 1.1 and 1.2, this chapter defines concepts such as RV, probability, expectation, variance, covariance, independence, etc., and it presents some handy rules for working with these objects. In Section 1.3, it then discusses useful concepts and results for the case when the size of our data sample grows to infinity.

The discussion starts by giving a definition of a RV that is sufficient for work in applied econometrics. The underlying theoretical foundations – an experiment with corresponding sample space and $\sigma$-algebra – are discussed in the Appendix. These are useful because they introduce some key concepts in an accessible, intuitive manner, they connect probability theory and statistics to set theory – a branch of mathematics –, and they are relevant for more theoretical work in econometrics. However, they are not necessary to understand the concepts discussed henceforth.

## 1.1 Univariate Random Variables

Intuitively, a random variable (RV) is a variable with stochastic outcomes, i.e. random realizations. For example, we might have a RV $X$ indicating whether it will rain tomorrow, with possible realizations 0 (no rain) and 1 (rain). The number of heads after two coin tosses is also a RV, with possible realizations 0, 1 and 2. If the number of possible outcomes is countable, as in these examples, we speak of a discrete RV.[1] In contrast, continuous RVs have uncountably many possible realizations. For example, the amount of precipitation tomorrow

[1] Countability means that we can list the outcomes. It does not imply that there is a finite set of outcomes. For example, if an RV can take on any value in $\mathbb{Z}$, it is a discrete RV, despite the fact that there are infinitely many integers.

can (in principle) take on any value in $\mathbb{R}_+$. In this section, we deal with univariate RVs, i.e. RVs that can take on a single, scalar value. We denote the RV with capital letters and a particular realization with lower-case letters, e.g. $X$ and $x$, respectively.

For discrete RVs, we can list the probabilities of all outcomes. For the weather-example, (under a bad forecast) we might have $\mathbb{P}[X = 1] = \mathbb{P}[\text{rain}] = 0.8$ and $\mathbb{P}[X = 0] = \mathbb{P}[\text{no rain}] = 0.2$. For the coinflip example – denoting this RV by $Y$ –, we have

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

The probability of each outcome must be in $[0, 1]$ and the probabilities of all outcomes must sum to 1, as one of the outcomes must happen for sure. Based on such a function that assigns a probability to each outcome, we can find the so-called cumulative distribution function (cdf), which shows the probability that the outcome of a RV is smaller than some value $z$. We obtain the cdf $F(z) = \mathbb{P}[X \leq z]$ by summing up the probabilities $\mathbb{P}[X = y]$ for all $x \leq z$. In the above two examples, this gives

$$F_X(z) = \mathbb{P}[X \leq z] = \begin{cases} 0 & \text{for } z < 0 \\ 0.8 & \text{for } z \in [0, 1) \\ 1 & \text{for } z \geq 1 \end{cases} , \quad F_Y(z) = \mathbb{P}[Y \leq z] = \begin{cases} 0 & \text{for } z < 0 \\ 0.25 & \text{for } z \in [0, 1) \\ 0.75 & \text{for } z \in [1, 2) \\ 1 & \text{for } z > 2 \end{cases} .$$

Note that these cdfs are defined for all $z \in \mathbb{R}$, despite the fact that both $X$ and $Y$ are discrete random variables. Based on the cdf, we can reconstruct the probabilities of the individual outcomes by looking at the jump-points of the cdf. See Fig. 1.1 for an illustration of the cdf and probability function of the RV $Y$.

For continuous RVs, we cannot list the probabilities of all outcomes. However, we can still describe the nature of their randomness by specifying their cdf. In other words, we define these RVs by specifying probabilities to sets of the form $(-\infty, z]$, $z \in \mathbb{R}$, i.e. specifying the probability that a RV takes on a value smaller than or equal to $z$, for all $z$. For a function to be a valid cdf, it has to satisfy certain properties.

**Definition 1.** *$F(x) = \mathbb{P}(-\infty, x]$ is a cumulative distribution function (cdf) if*[2]

[2]For more details, see Appendix. On the first condition: since $\lim_{x \to -\infty}(-\infty, x) = \emptyset$ and $\lim_{x \to \infty}(-\infty, x) = \mathbb{R}$, it is intuitive to give these two quantities the probabilities zero and one. The sec-

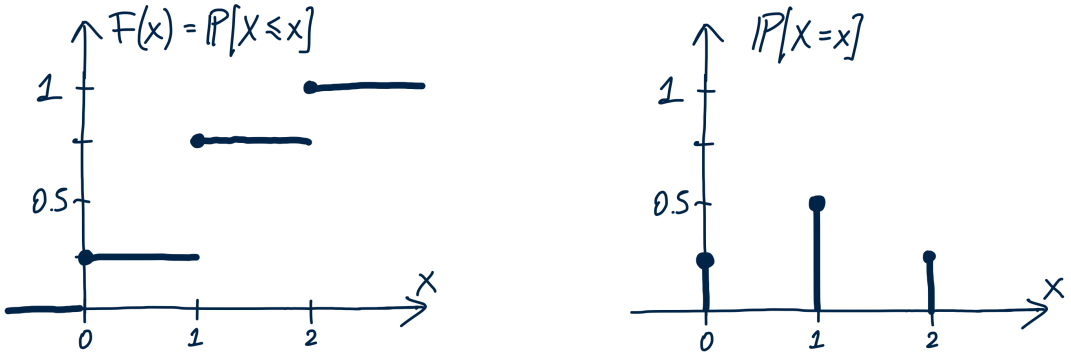


Figure 1.1: Representation of Probabilities for a Discrete Random Variable

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

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:* $\lim_{\epsilon\to 0} F(x_0 + \epsilon) = F(x_0) \quad \forall\, x_0 \in \mathbb{R}$.

Based on the cdf, we can define the probability density function (pdf), the counterpart of the probability function for continous RVs. Just as for discrete RVs summing up the probabilities of all realizations $x \leq x'$ gives $F(x')$, for continuous RVs integrating the pdf up to $x'$ gives $F(x')$.

**Definition 2.** *The probability density function (pdf) of a continuous RV $X$, $f_X$, is defined by*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \quad \forall\, x\ .$$

It holds that

$$f_X(x) \geq 0\ \forall\, x \quad \text{and} \quad \int_{-\infty}^{\infty} f_X(t)dt = \lim_{x\to\infty} F_X(x) = 1\ .^{3}$$

See Fig. 1.2 for an illustration of the cdf and pdf of a continuous RV $X$ (a standard Normal distribution; see Appendix).

**Definition 3.** *Two RVs $X$ and $Y$ are identically distributed (i.d.) if they have the same cdf,*

ond arises because $(-\infty, x)$ is a subset of $(-\infty, x']$, and the axioms of a probability function imply that $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$, as is intuitive. The third condition just ensures that, if $F(x)$ has a jump at $x$, then the value of $F$ exactly at $x$ is equal to the upper part of the curve, not the lower one. This has to do with the fact that the cdf shall denote probabilities of sets $(-\infty, x]$ and not $(-\infty, x)$, i.e. $x$ has to be included. See Fig. 1.1.

[3] The first holds because $F_X$ is nondecreasing.

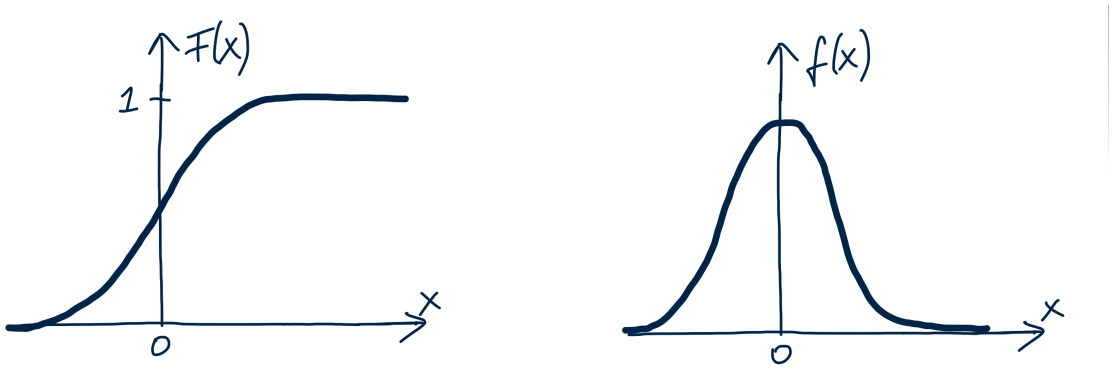


Figure 1.2: Representation of Probabilities for a Continuous Random Variable
*Notes:* Left plot shows the cdf, the right plot the pdf of a continuous RV (a standard Normal distribution.

*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, or $X$ could be the weather tomorrow and $Y$ the weather the day after tomorrow.

Because most distributions encountered in econometrics are continuous, the subsequent definitions and calculations focus on continuous RVs. They have intuitive counterparts for the discrete case, replacing e.g. integrals with sums.

**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

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

We can then find the pdf of $Y$ by differentiating $F_Y$.[4] See Appendix for an example. Also, Propositions 13 and 14 in the Appendix show how to derive $f_Y$ based on $f_X$ for general functions $g$ and for vector-valued $X$ and $Y$.

**Proposition 1** (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 $y \in [0,1]$.*[5]

[4] Thereby, $\frac{g^{-1}(y)}{\partial y} \leq 0$ if $g$ is decreasing s.t. $f_Y(y) \geq 0$ always.
[5] Proof: $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\left(F_X(X) \leq y\right) = \mathbb{P}\left(X \leq F_X^{-1}(y)\right) = F_X\left(F_X^{-1}(y)\right) = y$, as $F_X$ is increasing.

This result is useful for numerical sampling algorithms (see Section 7.2) and the evaluation of density forecasts in time series econometrics.

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

**Definition 4.** *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 $\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx$.*

In words, $\mathbb{E}[h(X)]$ is a pdf-weighted average of all possible realizations of $h(X)$. The expectation operator has the following properties:

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

The second property implies

$$\begin{aligned} h_1(x) \geq 0 \ \ \forall\, x \quad &\Rightarrow \quad \mathbb{E}\left[h_1(X)\right] \geq 0\,, \\ \text{and} \quad a \leq h_1(x) \leq b \ \ \forall\, x \quad &\Rightarrow \quad a \leq \mathbb{E}[h_1(X)] \leq b\,. \end{aligned}$$

**Definition 5.** *For $n \in \mathbb{Z}$, the nth moment of $X$ is $\mathbb{E}\left[X^n\right]$, while the nth central moment of $X$ is $\mathbb{E}\left[(X - \mathbb{E}[X])^n\right]$.*

Note that the first moment of $X$ is its expectation $\mathbb{E}[X]$, while the first central moment is zero.[6]

**Definition 6.** *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 $X$.*

The variance and standard deviation give us a sense of how much the realizations of $X$ can differ from the expectation of $X$. By the above properties of the expectation, it follows that $\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$.

**Definition 7.** *The moment-generating function (MGF) of $X$ is $M_X(t) = \mathbb{E}\left[exp\left\{tX\right\}\right]$.*

The MGF contains all the necessary information to construct moments of $X$ because it has

[6] The latter follows from the above properties of the expectation operator: $\mathbb{E}[(X - \mathbb{E}[X])] = \mathbb{E}[X] - \mathbb{E}[X] = 0$, because $\mathbb{E}[X]$ is simply a constant (like $c$ above).

the property

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

i.e. taking the $n$th derivative of $M_X(t)$ w.r.t. $t$ and evaluating it at $t = 0$, we get the $n$th (non-central) moment of $X$. Central moments can then be constructed as functions of non-central moments. For example, $\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Note that the MGF of $Y = aX + b$ is given by $M_{aX+b}(t) = \mathbb{E}\left[exp\left\{t(aX + b)\right\}\right] = \mathbb{E}\left[exp\left\{(ta)X\right\} exp\left\{tb\right\}\right] = M_X(at)exp\left\{bt\right\}$.

It is not guaranteed that a given moment of a RV $X$ exists in the sense that the corresponding integral could give infinity or be impossible to compute. However, a result shown in the Appendix states that if the $m$th central moment exists, then so do all lower-order moments.

**Proposition 2** (Jensen's Inequality)**.**
*For any convex function $g$, we have $\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$.*[7]

And hence, for any concave function $h$, $\mathbb{E}[h(X)] \leq h(\mathbb{E}[X])$. Remember the sign of the inequality by $\mathbb{E}\left[X^2\right] \geq \mathbb{E}[X]^2$ as $\mathbb{V}[X] \geq 0$.

## 1.2 Multivariate Random Variables

When dealing with multiple RVs, we can define a joint, marginal as well as conditional pdfs:

**Definition 8.** *Let $f_{X,Y}$ be the joint pdf of $X$ and $Y$. It satisfies*

$$f_{X,Y}(x, y) \geq 0 \quad \forall\, (x, y) \quad \textit{and} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1 \, .$$

*The marginal pdf of $X$ and the conditional pdf of $Y$ given $X$ are respectively given by*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \, , \quad \textit{and} \quad f_Y(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \, .$$

Note that you can think of the joint pdf of the two univariate/scalar-valued RVs $X$ and $Y$ – i.e. $X, Y \in \mathbb{R}$ – equivalently also as the pdf of the multivariate/vector-valued RV $Z = (X, Y)'$ – i.e. $Z \in \mathbb{R}^2$. The expectation of such a vector-valued RV $Z$ is simply the vector of the expectation of its individual elements, i.e. $\mathbb{E}[Z]$ consists of $\mathbb{E}[Z_i]$ for all elements $Z_i$ of $Z$.

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

Analogously to the univariate case, the variance of $Z$ is defined as

$$\mathbb{V}[Z] \equiv \mathbb{E}[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])'] = \mathbb{E}[ZZ'] - \mathbb{E}[Z]\mathbb{E}[Z]' \text{ .}^{8}$$

However, more interesting results emerge when we explicitly break up a multivariate RV into two components, $X$ and $Y$.

**Moments** Based on the conditional pdf $f_Y(y|x)$, we get the corresponding conditional expectation

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

and anlogously for $\mathbb{E}[h(Y)|X = x]$. Note that both $f_Y(y|x)$ and $\mathbb{E}[Y|X = x]$ are functions of $x$, the particular realization of $X$ that we condition on. We often write $\mathbb{E}[Y|X]$ for the expectation of $Y$ conditional on knowing (the realization of) the RV $X$ without restricting ourselves to any particular realization of $X$. This $\mathbb{E}[Y|X]$ is a function of the RV $X$, and we can compute the expectation of such functions (just as above we computed the expectation of some function $h$ of $X$, $\mathbb{E}[h(X)]$). This brings us to an important result.

**Proposition 3** (Law of Iterated Expectations (LIE))**.**
*We have* $\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right]$.

The proof of the LIE is instructive for its significance:

$$\begin{aligned}\mathbb{E}[Y] = \int y f(y) dy = \int y \int f(x,y) dx dy = \iint y f(y|x) f(x) dx dy \\ = \int \left[\int y f(y|x) dy\right] f(x) dx = \mathbb{E}\left[\mathbb{E}[Y|X]\right] \text{ .}\end{aligned}$$

The LIE states that taking the expectation of (i.e. averaging over all possible realizations of) $Y$ is the same as first taking the expectation conditional on $X$ and then averaging over all $X$. The LIE is a powerful tool. For example, if we know $\mathbb{E}[U|X] = 0$, it tells us that $\mathbb{E}[UX] = 0$.[9]

**Definition 9.** *Two RVs $X$ and $Y$ are <u>independent</u> if we can factorize the joint as the product of marginal pdfs: $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.*

[8] Let $\mathbb{E}[Z] = \mu$ and note that $\mathbb{E}[Z'] = \mathbb{E}[Z]' = \mu'$. The formula simplifies because

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

[9] This follows because $\mathbb{E}[XU] = \mathbb{E}[\mathbb{E}[XU|X]] = \mathbb{E}[X\mathbb{E}[U|X]] = \mathbb{E}[X \cdot 0] = 0$.

If $X$ and $Y$ are independent, then $f_Y(y|x) = f_Y(y)$ i.e. knowing the realization of $X$ conveys no information about $Y$. Under independence we also have

- $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$ for all functions $g$ and $h$, and therefore
- $M_{X+Y}(t) = M_X(t)M_Y(t)$.

**Definition 10.** *The covariance between two scalar RVs $X$ and $Y$ is*

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

*The correlation (coefficient) is then* $Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}}$.

Both measure the direction and strength of the (linear) relationship between $X$ and $Y$, i.e. they are measures of the linear dependence of $X$ and $Y$. Compared to the covariance, the correlation coefficient adjusts for the scale of the RVs $X$ and $Y$. A result in the Appendix shows that $|\text{Corr}(X,Y)| \in [0,1]$ for any two RVs. A positive $\text{Corr}(X,Y)$ indicates that whenever $X$ is above $\mathbb{E}[X]$, then $Y$ tends to be above $\mathbb{E}[Y]$, and the closer $\text{Corr}(X,Y)$ is to one, the stronger this relationship. Analogously, a negative $\text{Corr}(X,Y)$ indicates that then $Y$ tends to be below $\mathbb{E}[Y]$, and the closer $\text{Corr}(X,Y)$ is to $-1$, the stronger this relationship.

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.[10]

The covariance appears when we calculate the variance of the sum of two RVs: $\mathbb{V}[X + Y] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\text{Cov}(X,Y)$. More generally, $\mathbb{V}[a + bX + cY] = b^2\mathbb{V}[X] + c^2\mathbb{V}[Y] + 2bc\text{Cov}(X,Y)$.[11] As a result, for two uncorrelated RVs, the variance of the sum equals the sum of the variances.

The covariance also appears when we calculate the variance of a vector-valued RV. For

[10] The standard counterexample is $X \sim N(0,1)$ and $Y = X^2$. The two are clearly not independent as knowing one allows you to perfectly tell the other. Nevertheless, $\text{Cov}(X,Y) = 0$; since $\mathbb{E}[X] = 0$, we have $\text{Cov}(X,Y) = \mathbb{E}[XY] = \mathbb{E}[X^3]$ and the third moment is equal to zero for any Normal RV $X$.

[11] $\mathbb{V}[a+bX+cY] = \mathbb{E}\left[(a + bX + cY - \mathbb{E}[a + bX + cy])^2\right] = \mathbb{E}\left[(b(X - \mathbb{E}[X)) + c(Y - \mathbb{E}[Y]))^2\right] = b^2\mathbb{V}[X] + c^2\mathbb{V}[Y] + 2bc\text{Cov}(X,Y)$.

example, let $Z = (X, Y)' \in \mathbb{R}^2$, and let $\mathbb{E}[Z] = \mathbb{E}[(X, Y)'] = (\mu_x, \mu_y)' = \mu$. We have

$$\begin{aligned}\mathbb{V}[Z] = \mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])'\right] &= \mathbb{E}\left[\begin{bmatrix} X - \mathbb{E}[X] \\ Y - \mathbb{E}[Y] \end{bmatrix}\begin{bmatrix} X - \mathbb{E}[X] & Y - \mathbb{E}[Y] \end{bmatrix}\right] \\ &= \mathbb{E}\begin{bmatrix} (X - \mathbb{E}[X])(X - \mathbb{E}[X])' & (X - \mathbb{E}[X])(Y - \mathbb{E}[Y])' \\ (Y - \mathbb{E}[Y])(X - \mathbb{E}[X]) & (Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y]) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{V}[X] & \text{Cov}(X, Y) \\ \text{Cov}(X, Y) & \mathbb{V}[Y] \end{bmatrix}.\end{aligned}$$

Analogously to the case of two scalar RVs, 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 $Y$.

**Proportionality & Bayes' Formula**

**Definition 11.** *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 $f(x) \propto g(x)$.*

Proportionality is a useful concept for working with pdfs; due to the property of pdfs that they must integrate to one, we can completely specify a pdf by finding it up to proportionality. For example, suppose we know $f(x) \propto exp\{-\lambda x\}$ and $f(x)$ is a pdf defined on $x \in [0, \infty)$.[12] Then it must be that $f(x) = \lambda exp\{-\lambda x\}$ (an exponential distribution). This is because $\int_0^\infty exp\{-\lambda x\} = 1/\lambda$ and hence we must multiply the expression by $\lambda$ to get a valid pdf.

Note conditional pdfs are proportional to the joint pdf. For example,

$$f_Y(y|x) \propto f_{X,Y}(x, y),$$

and analogously for $f_X(x|y)$. The constant that makes sure that $f_Y(y|x)$ integrates to one is equal to $1/f_X(x) = \left(\int f(x, y)dy\right)^{-1}$ as $\int f_Y(y)dy = \int \frac{f(x,y)}{\int f(x,y)dy}dy = 1$. Of course, this is a constant only when we – as here – condition on $X = x$, i.e. it is a constant from the point of view of the RV $Y|(X = x)$.

From the definition of $f_Y(y|x)$, we can write the joint pdf as the product of conditional and

[12] In other words, $f(x) \propto exp\{-\lambda x\}\ \mathbf{1}\{x \geq 0\}$, i.e. $f(x)$ is proportional to $exp\{-\lambda x\}$ for $x \geq 0$ and it is simply zero for $x < 0$.

marginal:

$$f_{XY}(x,y) = f_Y(y|x)f_X(x) = f_X(x|y)f_Y(y) \ .$$

Together with the definition of $f_X(x)$, we then obtain Bayes' formula:

$$f_Y(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy} \propto f_X(x|y)f_Y(y) \ .$$

It has a wide variety of applications, including a whole field of Bayesian statistics (and econometrics) and microeconomic theory (learning). It tells us how, after having seen the outcome $X = x$, we can update the unconditional pdf of a RV $Y$, $f_Y(y)$ (the prior distribution or -belief), to the conditional pdf of $Y|(X = x)$, $f_Y(y|x)$ (the posterior distribution or -belief). Given that it is typically enough to know a pdf up to proportionality to conclude what the distribution is, to find the posterior $f_Y(y|x)$, we only have to multiply the prior $f_Y(y)$ with the so-called likelihood (of observing $X = x$ given a particular $y$), $f_X(x|y)$.

## 1.3 Asymptotic Theory

In econometrics, in the simplest case, we deal with a data sample $\{x_i\}_{i=1}^n$ of $n$ observations (realizations) $x_i$ of the same underlying RV $X_i$ which are drawn independently from some distribution $p(x|\theta)$ (i.e. the observations are i.i.d.). The discussion up to now is useful to analyze the moments and distribution of our data $\{x_i\}_{i=1}^n$ and functions thereof for a given sample size $n$. We speak of finite sample properties. However, oftentimes we are also interested in the properties of our data and functions of it as the sample size $n$ grows to infinity. In this case, we speak of asymptotic properties. To analyze them, a couple of concepts and results are important.

**Definition 12.** *A sequence of scalar RVs $\{X_n\}_{n=1}^{\infty}$ is said to converge in probability to a constant $c$ if*

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

*We write $X_n \xrightarrow{p} c$.*

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.[13] Sometimes it is useful to write $X_n \xrightarrow{p} c$ in functional form as $plim(X_n) = c$, where $plim$ stands for "probability limit". If $c = 0$, we say $X_n$ is $0_p(1)$, i.e. $X_n/1 \xrightarrow{p} 0$.

[13]Note that there is always (no matter how large $n$ is) a non-zero probability that $X_n \neq c$; it is the probability that is converging, not $X_n$ (which would be almost sure convergence, a stronger concept than convergence in probability).

For example, let $X_n$ be exponentially distributed with parameter $n$: $X_n \sim Exp(n)$, with cdf $F(x) = 1 - e^{-nx}$ and domain $\mathbb{R}_{++}$. 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 since $\mathbb{P}[X_n > \epsilon] = 1 - \mathbb{P}[X_n < \epsilon] = e^{-nx}$.

**Definition 13.** *A sequence of $k$-dimensional RVs $\{X_n\}_{n=1}^{\infty}$ converges in probability to a vector of constants $c$ if $Y_{j,n} \xrightarrow{p} c_j$ for $j = 1 : k$, where $X_{j,n}$ and $c_j$ are the $j$th element of $X_n$ and $c$, respectively.*[14]

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.

**Proposition 4** (Slutsky's Theorem I)**.**
*If $g : \mathbb{R}^k \to \mathbb{R}^m$ is continuous at $c$ and independent of $n$, then $X_n \xrightarrow{p} c \Rightarrow g(X_n) \xrightarrow{p} g(c)$.*

In other words, $plim(g(X_n)) = g(plim(X_n)) = g(c)$. For example, given that $X_n \xrightarrow{p} 0$ for $X_n \sim Exp(n)$ and that $g(x) = x^2$ is continuous at zero, we have $X_n^2 \xrightarrow{p} 0^2 = 0$. Importantly, based on Slutsky's theorem, we know that if $X_{1,n} \xrightarrow{p} c_1$ and $X_{2,n} \xrightarrow{p} c_2$, then

$$X_{1,n} + X_{2,n} \xrightarrow{p} c_1 + c_2 \ , \quad X_{1,n}' X_{2,n} \xrightarrow{p} c_1' c_2 \quad \text{and} \quad X_{1,n} X_{2,n}' \xrightarrow{p} c_1 c_2' \ .[15]$$

**Proposition 5** (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 $\bar{\mu}_n = \frac{1}{n}\sum_{i=1}^n \mu_i$. Then $\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$.*[16]

If $\mu_i = \mu \ \forall \ i$ s.t. $\bar{\mu}_n = \mu$, we can write simply $\bar{Y}_n \xrightarrow{p} \mu$. The WLLN says that a sample average of independent RVs converges in probability to the population mean (or the average of population means), i.e. as $n \to \infty$, its distribution gets more and more concentrated

[14] This is sometimes also defined by $||X_n - c|| \xrightarrow{p} 0$.

[15] This follows because Slutsky's theorem also holds for a vector-valued RV $X_n$, which we can split into two parts, $X_{1,n}$ and $X_{2,n}$, and because we can consider the functions which sum up or multiply these two parts, respectively, both of which are continuous. Of course, for the statement in the middle, the dimensions of $X_{1,n}$ and $X_{2,n}$ must fit.

[16] Formally, the following condition is added to the statement: $\mathbb{E}[|X_i|^{1+\delta}] < \Delta < \infty$ for some $\delta > 0$ and $\forall \ i$. If all $X_i$ have finite variances, it is satisfied for $\delta = 1$.

around it.

In practice, we often have a sample $\{x_i\}_{i=1}^n$, which we suppose to be i.i.d. with some mean $\mathbb{E}[x_i]$. Even without knowing $\mathbb{E}[X_i]$ we can conclude based on the WLLN that $\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{p} \mathbb{E}[X_i]$.

**Proposition 6** (Plug-In Property)**.**
*Suppose $X_n \xrightarrow{p} c$ and $g(X_n, Y_i)$ is a continuous function of $X_n$. Then $\frac{1}{n}\sum_{i=1}^n g(X_n, Y_i) \to \mathbb{E}[g(c, Y_i)]$.*

Convergence in probability refers to the case when a RV 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 RV has a distribution that in the limit equals some specific distribution.

**Definition 14.** *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 $F_X$ s.t. $\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \quad \forall\, x$ at which $F_X(x)$ is continuous. Then $X_n$ is said to converge in distribution to $X$. We write $X_n \xrightarrow{d} X$.*

As an example, let $Y_n$ be Pareto-distributed with cdf $F(y) = 1 - y^{-a}$ and support $y \geq 1$, and define $X_n = n(Y_n - 1)$. We have $X_n \xrightarrow{d} Exp(1)$ because

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

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.

**Proposition 7** (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 $X_n' Y_n \xrightarrow{d} c'Y$.*

**Proposition 8** (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 $g(X_n) \xrightarrow{d} g(X)$.*

Continuing the example from above, if we multiply $X_n = n(Y_n - 1)$, $Y_n$ Pareto distributed, by $k$, the CLT tells us that $kX_n \xrightarrow{d} kX \sim Exp(1/k)$, because $X \sim Exp(\lambda) \Rightarrow kX \sim Exp(\lambda/k)$. Similarly, we get $\mu - \sigma \log\left(\frac{exp\{-X_n\}}{1-exp\{-X_n\}}\right) \xrightarrow{d} \mu - \sigma \log\left(\frac{exp\{-X\}}{1-exp\{-X\}}\right)$, and it turns out that this limit-RV follows a logistic distribution with parameters $\mu$ and $\sigma$.

**Proposition 9** (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* $\bar{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. *Then*

$$\frac{\sqrt{n}(\bar{X}_n - \bar{\mu}_n)}{\bar{\sigma}_n} \xrightarrow{d} N(0,1) \text{ .}^{17}$$

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 $\lim_{n\to\infty} F_{X_n} = N\left(\bar{\mu}_n, \frac{1}{n}\bar{\sigma}_n^2\right)$, i.e. the limit distribution, obtained as $n \to \infty$, is a function of $n$! However, if $\mu_i = \mu$ and $\sigma_i^2 = \sigma^2 \ \forall \ \ i$, then $\bar{\mu}_n = \mu$ and $\bar{\sigma}_n^2 = \sigma^2$, and we can simply write $\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} N(0, \sigma^2)$. If in addition $\mu = 0$, we can write $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \xrightarrow{d} N(0, \sigma^2)$.

While the WLLN tells us that the average of independent RVs converges in probability to their mean (or average of means, if they are different), the CLT tells us that, when properly standardized, this average converges to a Normal distribution. In other words, as $n$ increases, the distribution of a sample mean $\bar{X}_n$ not only gets more and more concentrated around $\bar{\mu}_n$ (as it does under the WLLN), bet it also gets "more Normal".

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

**Proposition 10** (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*

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

Analogously to before, this can be thought of as $\bar{X}_n \xrightarrow{d} N(\mu, \frac{1}{n}\Sigma)$, while if $\mu = 0$, we can indeed write $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \xrightarrow{d} N(0, \Sigma)$.

In practice, we often have a sample $\{x_i\}_{i=1}^n$, which we suppose to be i.i.d. with some mean $\mathbb{E}[X_i]$ and some variance $\mathbb{V}[X_i] = \mathbb{E}[X_i X_i'] - \mathbb{E}[X_i]\mathbb{E}[X_i]'$. Even without knowing these

[17] Formally, the following conditions are required:

1. $\mathbb{E}[|X_i - \mu_i|^{2+\delta}] < \Delta < \infty$ 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.

The first is satisfied if the third moment of each $X_i$ exists, the latter essentially just requires our sample average of variances, $\bar{\sigma}_n^2$, to be positive.

moments, we can conclude based on the CLT that $\sqrt{n}\left(\bar{X}_n - \mathbb{E}[X_i]\right) \xrightarrow{d} N(0, \mathbb{V}[X_i])$. If $\mathbb{E}[X_i] = 0$, we get that $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \xrightarrow{d} N(0, \mathbb{E}[X_i X_i'])$.

**Proposition 11** (Delta Method)**.**
*If* $\sqrt{n}\left(X_n - c\right) \xrightarrow{d} X$ *and* $g : \mathbb{R}^k \to \mathbb{R}^m$ *is continuously differentiable, then* $\sqrt{n}\left(g(X_n) - g(c)\right) \xrightarrow{d} G \cdot X$*, where* $G = \left.\frac{\partial g(x)}{\partial x'}\right|_{x=c}$ *is a* $m \times k$ *matrix.*

For example, if $\sqrt{n}\frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{d} N(0, \sigma^2)$, then the Delta method tells us that $\sqrt{n}\frac{1}{n}\sum_{i=1}^n exp\{2X_i\} \xrightarrow{d} 2exp\{2 \cdot 0\}N(0, \sigma^2) = N(0, 4\sigma^2)$. The Delta method, however, applies more generally for a $k$-dimensional RV $X_n$ and a vector-valued function $g$ with $m$ outputs (essentially we look at $m$ scalar-valued functions of $X_n$).

# Appendix

## Theoretical Underpinnings of Random Variables

This section provides some background on probability theory, just enough to understand where RVs are coming from. Some claims made in this section are not straightforward. They are marked with an asterisk and the corresponding proofs are presented at the end of this section.

**Events & Probabilities** Imagine an experiment, like a coinflip or the weather tomorrow. Let $\Omega$ be the sample space, i.e. the set of all possible elementary outcomes $\omega$ of this experiment. In the coinflip example, we would have $\Omega = \{\text{heads, tails}\}$. For tomorrow's weather, we might have $\Omega = \{\text{rain, snow, no precipitation}\}$. Let $A, B \subseteq \Omega$ be events, i.e. sets/collections of possible outcomes. For example, $A = \{\text{rain, snow}\} \subset \Omega$ is the event that no picnic is possible tomorrow. This section defines a probability function on the sample space $\Omega$ and establishes some useful rules when working with events and probabilities.

Based on two events $A$ and $B$ we can define a bunch of other events, as illustrated in Fig. 1.3:

- $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$ is the union of $A$ and $B$, i.e. the set of all outcomes $\omega$ which are either in $A$ or in $B$ (or both).
- $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$ is the intersection of $A$ and $B$, i.e. the set of all outcomes $\omega$ with are both in $A$ and in $B$.
- $A^c = \{\omega \in \Omega : \omega \notin A\}$ is the complement of $A$, i.e. the set of all outcomes $\omega$ which are not in $A$)

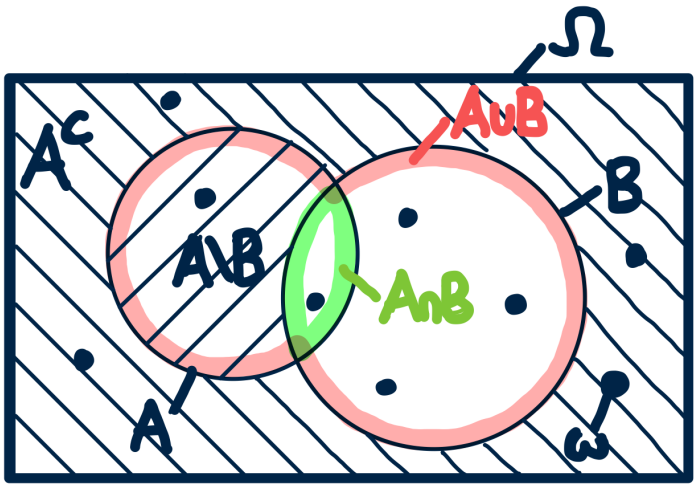


Figure 1.3: Sample Space, Events and Outcomes in a Venn-Diagram
*Notes:* Illustration of the sets $\Omega$, $A$, $B$, $A \cup B$, $A \cap B$, $A^c$ and $A \backslash B$ in a so-called Venn-diagram.

- $A \backslash B = A \cap B^c$ is "$A$ without $B$", i.e. the set of all outcomes $\omega$ which are in $A$ but not in $B$.

For any 3 events $A, B, C$, we have the following rules (which are easy to verify using the above definitions or a Venn-diagram):

- Commutativity: $A \cup B = B \cup A, \quad A \cap B = B \cap A$
- Associativity: $A \cup (B \cup C) = (A \cup B) \cup C, \quad A \cap (B \cap C) = (A \cap B) \cap C$
- Distributivity: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C), \quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- De Morgan's laws: $(A \cup B)^c = A^c \cap B^c, \quad (A \cap B)^c = A^c \cup B^c$

**Definition 15.** *Two events $A$ and $B$ are called disjoint if they have no elementary outcomes in common: $A \cap B = \emptyset$, i.e. the intersection of $A$ and $B$ is the empty set.*
*A collection of events $A_1, A_2, \ldots$ is called pairwise disjoint if $A_i \cap A_j = \emptyset \quad \forall\, i \neq j$.*
*If $A_1, A_2, \ldots$ are pairwise disjoint and their union constitutes the whole sample space, i.e. $\bigcup_i A_i = \Omega$, then the collection $A_1, A_2, \ldots$ is said to form a partition of $\Omega$.*

For example, the sets $A_i = (i-1, i]$ for $i \in \mathbb{Z}$ form a partition of $\mathbb{R}_+ = (0, \infty)$.

With all these rules and definitions, we can now define a $\sigma$-algebra ("sigma-algebra"), denoted by $\mathcal{F}$. You can think of it as the set of all events and collections of events to which a probability can be assigned. Formally, probabilities are defined on the $\sigma$-algebra $\mathcal{F}$ rather

than the sample space $\Omega$. Intuitively, this is because we would like to assign probabilities to events $A$ and collections of events $\{A_1, A_2, ...\}$ rather than (just) elementary outcomes $\omega$ (and in some cases the latter is not even possible, e.g. for uncountable sample spaces $\Omega$[18]). However, for most practical purposes and to develop intuition, you can think of probabilities being defined on $\Omega$ (i.e. ignore the definition of $\mathcal{F}$, and in the following, wherever it appears, replace it in your mind with $\Omega$).

**Definition 16.** *Let $\mathcal{F}$ be a collection of subsets of $\Omega$. $\mathcal{F}$ is called a $\sigma$-algebra if*

1. *it contains the empty set: $\emptyset \in \mathcal{F}$,*
2. *for any set $A$ it contains, it also contains its complement: $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$,*
3. *for any collection of sets it contains, it also contains their union: $A_1, A_2 \ldots \in \mathcal{F} \Rightarrow \bigcup_i A_i \in \mathcal{F}$.*

Note that these three conditions imply

- $\mathcal{F}$ contains the sample space itself: $\Omega \in \mathcal{F}$ (by 1 and 2),
- for any collection of sets it contains, it also contains their intersection: $\bigcap_i A_i \in \mathcal{F}$ (by 2 and 3).*

For example, in the example of tossing a coin twice, we have $\Omega = \{HH, HT, TH, TT\}$, and one $\sigma$-algebra on it is

$$\mathcal{F} = \Big\{\Omega, \emptyset, \\ HH, \{HT, TH, TT\}, HT, \{HH, TH, TT\}, TH, \{HH, HT, TT\}, TT, \{HH, HT, TH\}, \\ \{HH, HT\}, \{TH, TT\}, \{HH, TH\}, \{HT, TT\}, \{HH, TT\}, \{HT, TH\}\Big\}.$$

Defining a probability function on $\mathcal{F}$, we can assign probabilities to elementary outcomes, triple-sets, double-sets such as the event "no tail" or "first toss is head", etc. Note that other $\sigma$-algebras can be defined on the same $\Omega$, for example

$$\mathcal{F}' = \Big\{\Omega, \emptyset, \{HH, HT\}, \{TH, TT\}, \{HH, TH\}, \{HT, TT\}, \{HH, TT\}, \{HT, TH\}\Big\}.$$[19]

Which $\sigma$-algebra is constructed depends on what events we would like to assign probabilities

[18] For example, if $\Omega = \mathbb{R}$, there are uncountably many outcomes $\omega \in \mathbb{R}$.

[19] The smallest $\sigma$-algebra, $\mathcal{F}'' = \{\Omega, \emptyset\}$, is uninteresting.

to. Formally, a probability function maps each element of a $\sigma$-algebra $\mathcal{F}$ into a number between zero and one.

**Definition 17.** *Given $\Omega$ and an associated $\mathcal{F}$, a probability function $\mathbb{P} : \mathcal{F} \to [0,1]$ is a function with domain $\mathcal{F}$ that satisfies the Kolmogorov axioms:*

1. *probabilities are non-negative: $\mathbb{P}(A) \geq 0 \quad \forall\, A \in \mathcal{F}$,*
2. *the sample space has probability one: $\mathbb{P}(\Omega) = 1$,*
3. *we can simply add up probabilities of disjoint sets to get the probability of their union: if $A_1, A_2, \ldots \in \mathcal{F}$ are pairwise disjoint, then $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$.*

*The triplet $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space.*

Note that these three conditions imply that:*

- the probabilities of an event and its complement add up to one: $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$,
- probabilities are no larger than one: $\mathbb{P}(A) \leq 1$,
- the empty set has probability zero: $\mathbb{P}(\emptyset) = 0$.

Using these elementary rules for probabilities, together with the rules for working with sets defined above, we obtain the following three equations:*

- $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$,
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$,
- $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.

Often, we can update the sample space and therefore the probability measure based on new information we received about the experiment. For example, if we know event $B$ occurred, we would like to condition all probability statements on that.

**Definition 18.** *Let $A, B \subseteq \Omega$ and $\mathbb{P}(B) > 0$. The conditional probability of $A$ given $B$ is $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.*

Intuitively, $B$ becomes the sample space: we have $\mathbb{P}(B|B) = 1$. From this definition, it follows that

- $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$ and $\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$, and therefore

- $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$.

**Definition 19.** *Two events $A, B$ are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.*

Note that if $A$ and $B$ are independent, then so are $A$ and $B^c$, $A^c$ and $B$, and $A^c$ and $B^c$.[20]

**Definition 20.** *The events $A_1, \ldots, A_n$ are mutually independent if for any subcollection $A_{i_1}, \ldots, A_{ik}$, $k \leq n$, we have $\mathbb{P}\left(\bigcap_{j=1}^{k} A_{ij}\right) = \prod_{j=1}^{k} \mathbb{P}(A_{ij})$.*

**Proposition 12** (Bayes' Rule)**.**
*Let $A_1, ..., A_n$ be a partition of the sample space and let $B$ be any set. Then*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}.$$[21]

**Random Variables** Intuitively, a random variable (RV) $X$ is a variable that maps each outcome $\omega$ in the sample space $\Omega$ to a number (i.e. an element in $\mathbb{R}$). In other words, it is a function that puts our experiment into numbers. In the weather example with $\Omega =$ {rain, snow, no precipitation}, we might have a RV $X$ indicating whether it will be dry, with possible realizations 0 and 1. If $X$ instead denotes the number of heads in an experiment where we toss a coint twice – $\Omega = \{HH, HT, TH, TT\}$ –, it has the possible realizations 0, 1 and 2. Because the realized outcome $\omega$ of the experiment $\Omega$ is stochastic, so is the realization $x$ of the RV $X$.

To make a more formal definition of an RV that applies both for continuous and discrete RVs, we need to define a probability measure on $\mathbb{R}$, which in turn requires constructing a $\sigma$-algebra on it. Let $\mathcal{G}$ contain all open intervals in $\mathbb{R}$: $\mathcal{G} = \{(a, b) : a, b \in \mathbb{R}\}$. The Borel $\sigma$-algebra $\mathcal{B}(\mathbb{R})$ is the smallest $\sigma$-algebra that contains all sets $A \in \mathcal{G}$ (and by the above definition of a $\sigma$-algebra also their complements, unions, intersections, etc.). For discrete sample spaces $\Omega$, we can list the probabilities of some or all events in the corresponding $\sigma$-algebra $\mathcal{F}$. For continuous sample spaces, we can not possibly list the probabilities of all elements in $\mathcal{B}(\mathbb{R})$ because there are uncountably many! Instead, we use cumulative distribution functions (cdf), which assign probabilities to sets of the form $(-\infty, x]$, $x \in \mathbb{R}$, as defined in the main

[20] For example, for the first pair we have $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A)[1 - \mathbb{P}(B)] = \mathbb{P}(A)P(B^c)$. Similar calculations apply for the other two pairs.

[21] This follows since $\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j) = \sum_j \mathbb{P}(A_j \cap B) = \mathbb{P}(\bigcup_j (A_j \cap B)) = \mathbb{P}(\Omega \cap B) = \mathbb{P}(B))$.

text.

**Definition 21.** *A random variable (RV) $X : \Omega \to \mathbb{R}$ is a mapping (i.e. a function) from the sample space $\Omega$ to the real line $\mathbb{R}$ s.t. $\forall\ B \in \mathcal{B}(\mathbb{R})$, $X^{-1}(B) \in \mathcal{F}$, i.e. for all sets $B$ on the real line to which we can assign probability, the inverse-image of $B$ (i.e. the set of events that make $X$ take on values in $B$) is a subset of $\Omega$ to which we can assign probability. We say $X$ is $\mathcal{F}\backslash\mathcal{B}(\mathbb{R})$-measurable.*

The condition in this definition appears because we would like to make probability statements about $X$ of the form $F_X(x) \equiv \mathbb{P}(X(\omega) \in B)$ for some $B = (-\infty, x]$ and we can do so only if the inverse-image of $B$ – i.e. the subset of $\Omega$ that gives $X(\omega) \in B$ – is in the $\sigma$-algebra $\mathcal{F}$:

$$\mathbb{P}(X(\omega) \in B) = \mathbb{P}\left(\omega \in X^{-1}(B)\right) = \mathbb{P}(X^{-1}(B))$$

is only well-defined if $X^{-1}(B) \in \mathcal{F}$, since $\mathbb{P}$ is defined on $\mathcal{F}$.

For example, take the experiment of tossing a coin twice – $\Omega = \{HH, HT, TH, TT\}$ – and let $X$ denote the number of heads. For $B = (-\infty, 1.5]$, we have $F_X(1.5) \equiv \mathbb{P}(X(\omega) \in B) = \mathbb{P}\left(\omega \in X^{-1}(B)\right) = \mathbb{P}(\omega \in \{HT, TH, TT\}) = 3/4$. Doing so for all intervals $(-\infty, x], x \in \mathbb{R}$, we would get the whole cdf of $X$, $F_X(x)$.

**Proofs of Claims Marked by Asterisks**

**Claim.** *For any $\sigma$-algebra $\mathcal{F}$, we have $A_1, A_2, ... \in \mathcal{F} \Rightarrow \bigcap_i A_i \in \mathcal{F}$*

**Proof:**

$$\begin{aligned}
& & A_1, A_2, ... \in \mathcal{F} & \\
& \Rightarrow & A_1^c, A_2^c, ... \in \mathcal{F} & \text{by condition 2} \\
\Rightarrow & \bigcup_i A_i^c = \left(\bigcap_i A_i\right)^c \in \mathcal{F} & & \text{by condition 3 and De Morgan's laws} \\
& \Rightarrow & \bigcap_i A_i \in \mathcal{F} & \text{by condition 2 .} \quad \blacksquare
\end{aligned}$$

**Claim.** *The Kolmogorov axioms from the definition of a probability function imply the following statements:*

- *the probabilities of an event and its complement add up to one: $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$,*
- *probabilities are no larger than one: $\mathbb{P}(A) \leq 1$,*

- *the empty set has probability zero:* $\mathbb{P}(\emptyset) = 0$.

**Proof:**

- This follows from $\mathbb{P}(A \cup A^c) = \mathbb{P}(\Omega) = 1$ (by axiom 2) and $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$ (by axiom 3).
- This follows from $\mathbb{P}(A) = 1 - \mathbb{P}(A^4) \leq 1$ (by first statement) as $\mathbb{P}(A^c) \geq 0$ (by axiom 1).
- We have $\mathbb{P}(\Omega \cup \emptyset) = \mathbb{P}(\Omega) = 1$ (by axiom 2) and $\mathbb{P}(\Omega \cup \emptyset) = \mathbb{P}(\Omega) + \mathbb{P}(\emptyset) \leq 1$ (by axiom 3 and second statement). Together with axiom 2, they imply $\mathbb{P}(\emptyset) \leq 0$, which together with $\mathbb{P}(\emptyset) \geq 0$ (by axiom 1) implies $p(\emptyset) = 0$. ■

**Claim.** *1.* $\mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$*,*

*2.* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$*,*

*3.* $A \subseteq B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.

**Proof:**

1. For any 2 sets $A, B$, $B = B \cap \Omega = B \cap (A \cup A^c) = (B \cap A) \cup (B \cap A^c)$, whereby $(B \cap A) \cap (B \cap A^c) = \emptyset$. Therefore, $\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$.
2. $A \cup B = (A \cup A^c) \cap (A \cup B) = A \cup (A^c \cap B)$. As a result, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
3. $A \subseteq B \Rightarrow \mathbb{P}(A) = \mathbb{P}(A \cap B) \Rightarrow \mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A) \geq 0$ by axiom 1. ■

Based on the second statement in the claim, we can derive two famous inequalities that provide bounds for $\mathbb{P}(A \cup B)$:

- Bonferroni's inequality: $\mathbb{P}(A \cup B) \geq \max\{\mathbb{P}(A) + \mathbb{P}(B) - 1, 0\}$,[22]
- Boole's inequality: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

## Density Transformations

In the main text, we have discussed how to construct the cdf of $Y = g(X)$ based on the cdf of $X$ when $g$ is monotone (i.e. the mapping between the sample spaces of $X$ and $Y$ is

[22] It follows from $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$.

bijective/one-to-one). For example, if $X \sim N(0,1)$ and we define $Y = g(X) = \mu + \sigma X$ with $X = g^{-1}(Y) = \frac{Y-\mu}{\sigma}$, then

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

and

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

The following two propositions show how to do that i) for more general functions $g$ in the case of univariate RVs and ii) for bijective $g$ in the case of multivariate (vector-valued) RVs.

**Proposition 13** (Density Transformation: Univariate Random Variables)**.**
*Let $Y = g(X)$ and define the sample space of $X$ as $\mathcal{X} = \{x : f_X(x) \geq 0\}$. Suppose $\exists$ a partition $A_1, \ldots, A_k$ of $\mathcal{X}$ s.t. $g(X)$ is continuous on each $A_i$. Also, suppose $\exists$ functions $g_1, \ldots, g_k$ defined an $A_1, \ldots, A_k$, respectively, s.t.*

1. *$g(x) = g_i(x)$, $\forall x \in A_i$;*
2. *$g_i(x)$ is monotone on $A_i$;*
3. *$\mathcal{Y} = \{y : g_i(x) \ , \quad x \in A_i\}$ is the same $\forall\ i$*
4. *$g_i^{-1}(y)$ has a continuous derivative $\forall\ i$*

*Then $f_Y(y) = \sum_{i=0}^{k} f_X(g_i^{-1})|\frac{g^{-1}(y)}{\partial y}|$ for $y \in \mathcal{Y}$ (and zero otherwise).*

For example, let $X \sim N(0,1)$ and take $Y = g(x) = X^2$. Then $A_1 = (-\infty, 0), A_2 = (0, \infty)$ and $g_1(x) = g_2(x) = x^2$ with derivatives $g_1^{-1}(y) = -\sqrt{y}, g_2^{-1}(y) = \sqrt{y}$ for $y \geq 0$ (i.e. $\mathcal{Y} = \mathbb{R}_+$). We then get

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

Note that $Y$ is chi-squared distributed with one degree of freedom: $Y \sim \chi_1^2$.

**Proposition 14** (Density Transformation: Multivariate Random Variables)**.**
*Let $\{X_i\}_{i=1:n}$ be RVs with pdf $f_X(x_1, ..., x_n)$ and support $\mathcal{X}$. Let $Y_i = t_i(X_1, ..., X_n)$ for $i = 1 : n$ and suppose the transformation from $\{X_i\}_{i=1:n}$ to $\{Y_i\}_{i=1:n}$ is one-to-one. Then it can be solved for $X_i = t_i^{-1}(Y_1, ..., Y_n)$ for $i = 1 : n$, and the pdf of $\{Y_i\}_{i=1:n}$ is*

$$f_Y(y_1, ..., y_n) = f_X(t_1^{-1}(y), ..., t_n^{-1}(y))|J| \,, \quad \text{where } J = \begin{vmatrix} \begin{bmatrix} \frac{\partial t_1^{-1}}{\partial y_1} & \cdots & \frac{\partial t_1^{-1}}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial t_n^{-1}}{\partial y_1} & \cdots & \frac{\partial t_n^{-1}}{\partial y_n} \end{bmatrix} \end{vmatrix} .$$

## Further Probability-Theoretic Results

The following two theorems establish under what conditions we can exchange the expectation with the limit or differentiation operators.

**Proposition 15** (Dominated Convergence)**.**
*Suppose $h(x, \theta)$ is continuous at $\theta_0 \quad \forall\, x$ and $\exists$ a function $g(x)$ satisfying*

1. $|h(x, \theta)| \leq g(x) \quad \forall\, x, \theta$;
2. $\int_{-\infty}^{\infty} g(x)dx < \infty$

*Then* $\lim_{\theta\to\theta_0} \int_{-\infty}^{\infty} h(x, \theta_0)\, dx = \int_{-\infty}^{\infty} \lim_{\theta\to\theta_0} h(x, \theta)dx$.
As a result, $\lim_{\theta\to\theta_0} \mathbb{E}[k(x, \theta)] = \mathbb{E}[\lim_{\theta\to\theta_0} k(x, \theta)]$ (take $h(x, \theta) = f(x)k(x, \theta)$).

**Proposition 16.**
*Suppose that:*

1. *$h(x, \theta)$ is differentiable w.r.t. $\theta$ at $\theta = \theta_0$, i.e.* $\lim_{\delta\to 0} \frac{h(x_1\theta_0 + \delta) - h(x_1\theta_0)}{\delta} = \frac{\partial h(x_0\theta)}{\partial\theta}|_{\theta=\theta_0}$
2. *$\exists\, g(x, \theta)$ and $\delta_0 > 0$ s.t.* $\left|\frac{h(x, \theta_0 + t) - h(x, \theta_0)}{\delta}\right| \leq g(x, \theta_0) \quad \forall\, x$ *and* $|\delta| \leq \delta_0$
3. $\int_{-\infty}^{\infty} g(x, \theta_0)\, dx < \infty$

*Then* $\frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} h(x, \theta)dx\Big|_{\theta=\theta_0} = \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} h(x, \theta)\Big|_{\theta=\theta_0} dx$

Taking again $h(x, \theta) = f(x)k(x, \theta)$ yields $\frac{\partial}{\partial\theta}\mathbb{E}[k(x, \theta)]\big|_{\theta=\theta_0} = \mathbb{E}\left[\frac{\partial}{\partial\theta} k(x, \theta)\big|_{\theta=\theta_0}\right]$.

The following claim helps us prove Höldner's inequality, which in turn establishes several important results.

**Claim.** $\forall\ a, b \geq 0$ *and* $p, q$ *s.t.* $\frac{1}{p} + \frac{1}{q} = 1$*, we have* $\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$*.*

**Proof:** Define $g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab \Rightarrow \frac{\partial g(a)}{\partial a} = a^{p-1} - b = 0 \Rightarrow b = a^{pp-1} \Rightarrow g\left(a^p\right) = \frac{1}{p}a^p + \frac{1}{q}a^{(p-1)q} - a^p = \left(\frac{1}{p} + \frac{1}{q}\right)a^p - a^p = 0$, as $\frac{1}{p} + \frac{1}{q} = \frac{p+q}{pq} = 1 \Rightarrow p = q(p-1)$.
This is a minimum since $\frac{\partial^2 q(a)}{\partial^2 a} < 0$ as $p < 1$. ■

**Proposition 17** (Höldner's inequality)**.**
$|\mathbb{E}[XY]| \leq \mathbb{E}[|XY|] \leq \mathbb{E}\left[|X|^p\right]^{\frac{1}{p}} \mathbb{E}\left[|Y|^q\right]^{\frac{1}{q}}$ *for* $\frac{1}{p} + \frac{1}{q} = 1\ ,\quad p, q > 1.$

**Proof:** First inequality follows from $-|XY| \leq XY \leq |XY|$. For the second one, set $a = \frac{|X|}{\mathbb{E}[|X|^p]^{\frac{1}{p}}}$, $b = \frac{|Y|}{\mathbb{E}[|Y|^q]^{\frac{1}{q}}}$, apply previous claim and take expectations. ■

Setting $p = q = 2$ gives the Cauchy-Schwarz inequality:

$$|\mathbb{E}[XY]| \leq \mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]}\ .$$

In turn, applying the latter to the RVs $X - \mathbb{E}[X]$ and $Y - \mathbb{E}[Y]$ (instead of $X$ and $Y$) establishes $|\text{Corr}(X, Y)| \leq 1$. Note that we get $\text{Corr}(X, Y) = 1$ iff $(X - \mathbb{E}[X]) = c(Y - \mathbb{E}[Y]), c \in \mathbb{R}$.

Taking $Y = 1$, we get $|\mathbb{E}[X]| \leq \mathbb{E}[|X|] \leq \mathbb{E}\left[|X|^p\right]^{1/p}$ for $1 < p < \infty$. Then, setting $p = 2$, we can see that if $\mathbb{E}\left[X^2\right]$ is finite, so is $\mathbb{E}[X]$. Taking the RV $Z$ s.t. $|Z|^r = |X|$, we get $\mathbb{E}[|Z|^r] \leq \mathbb{E}[|Z|^{rp}]^{\frac{1}{p}}$, which leads to Liapunov's inequality:

$$\mathbb{E}[|Z|^r]^{\frac{1}{r}} \leq \mathbb{E}[|Z|^s]^{\frac{1}{s}}\ ,\quad 1 < r < s < \infty\ .$$

It establishes that if some higher order moment exists (is finite), then so do all moments of lower order.