Handout 1
Probability Review[*]

Instructor: Vira Semenova

Note author: Danylo Tavrov, Vira Semenova

1 Introduction

Econometrics is the application of mathematics and statistical methods to social and economic data. This unification of economics, mathematics, and statistics adds *empirical* content to economic *theory*, allowing economists to estimate causal impacts of policies, test theories, and make forecasts.

1.1 Causal Questions

Forecasting (and nowcasting) is intensively used in the industry and the government. Economists working in investment banks, for example, produce quarterly forecasts for GDP, inflation, etc. The way they do this is by quantifying historical relationships between variables of interest (e.g., GDP, inflation) and some indicators (industrial production, etc.). But it's important to realize that these practitioners are not trying to assess a *causal* relationship between variables. They are just interested in patterns of temporal *correlation* between these variables and think that these relationships are stable enough over time to be informative.

More often, instead of simple correlations, econometrics is used in an attempt to answer causal questions, which usually bear important policy implications. For example,

- the effect of the minimum wage on employment;
- the return to education (how does an additional year of education change earnings);
- the price elasticity of cigarettes or gasoline (how effective will a tax on these goods be);
- the effect of increase in interest rates on output growth.

In addition, econometrics helps us make more sensible judgment over issues and claims that appear in popular media and readings, for example:

- lack of sleep may shrink your brain;
- dogs walked by men are more aggressive;
- Facebook users get worse grades in college.

All of the above issues deal with the so-called "causal effect." Ideally, we would like to have an *experiment* in a lab environment to produce a measurement. But almost always we only have *observational* data, which give us just a correlation (check out this website for illustrations of "correlation is not causation" principle). In order to extract a "causal effect," a list of problems inherent in observational data need to be addressed using econometric analysis. Most of the course deals with difficulties arising from using observational data to estimate causal effects.

[*] We thank Prof. Anna Mikusheva and numerous former ECON 140 and 141 GSIs for their course notes.

1.2 Review of Economic Data

1.2.1 Data Types

In economics, we use 3 types of data:

- *cross-sectional data*: they describe the activities of individuals, firms, or other units at a given point in time, such as given in the following table:

Country	Year	GDP	Interest rate	Europe
Brazil	2000	100	22	0
USA	2000	1000	4	0
Italy	2000	60	6	1

- *time series data*: they describe the movement of a variable over time. They can be daily, weekly, monthly, quarterly, or annual, such as given in the following table:

Country	Year	GDP	Interest rate	Europe
Brazil	2000	100	22	0
Brazil	2001	150	18	0
Brazil	2002	170	15	0
Brazil	2003	220	12	0

- *panel (longitudinal) data*: a combination of the above two data types, such as given in the following table:

Country	Year	GDP	Interest rate	Europe
Brazil	2000	100	22	0
Brazil	2001	150	18	0
Brazil	2002	170	15	0
Brazil	2003	220	12	0
USA	2000	1000	4	0
USA	2001	1150	3	0
USA	2002	1400	2	0
USA	2003	1900	2	0
Italy	2000	60	6	1
Italy	2001	130	5	1
Italy	2002	150	5	1
Italy	2003	190	3	1

1.2.2 Population and Sample

A *population* is the complete data collection to be studied, containing all the objects of interest.

A *sample* is a part of the population of interest, a sub-collection *selected* from a population. Empirical data we can obtain almost always constitute a sample rather than a population. That said, we want the sample to be *representative* of the population, so that the effect we find in the sample can be generalized to the population.

A *parameter* is a numerical measurement that describes a characteristic of a population, while a *statistic* is a numerical measurement that describes a characteristic of a sample. In general, we will use a statistic to *infer* something about a parameter.

A *data generating process* (*DGP*) is the joint probability distribution that is supposed to characterize the entire population, from which the sample has been drawn. While we do not know the probability distribution underlying the population of interest, we use the available data to get insights into its nature.

2 Random Variables

A *random variable* X is a *function* that assigns a real number to each possible outcome in some (unspecified) outcome space S. E.g., X can be a function that maps each person $\xi \in S$ to their height. We sometimes write $X(\xi)$ for $\xi \in S$, although usually we drop the argument and simply write X. We typically denote random variables with capital letters, and use small letters to denote their values, e.g. expression $X = x$ should read that a random variable X assumes a value $x \in \mathbb{R}$, which is a fixed constant.

The outcome space S need not contain numerical values. For example, the outcome of a coin toss—heads or tails—is not numerical, but a random variable $X \equiv$ "number of heads" is.

Random variable is fully characterized by a set of possible values and probabilities with which these values are taken. There are two types of random variables[1], discrete and continuous.

2.1 Discrete Random Variables

A *discrete* random variable takes on a finite (or countable) number of values. Examples include the number of children in a family, with $X \in \{0, 1, 2, \ldots\}$, or the number of times we get heads upon tossing a coin twice, with $X \in \{0, 1, 2\}$.

The *probability distribution* of a discrete random variable is the list of all possible values the variable can assume and probabilities associated with each of the outcomes from S. These probabilities sum to 1. Probabilities associated with each outcome are given by the *probability mass function* (PMF): $f(x) = \mathbb{P}(X = x)$.

A discrete random variable can be written as a list x_1, x_2, ..., with corresponding probabilities p_1, p_2, ..., where $p_k \equiv \mathbb{P}(X = x_k)$.

The *cumulative distribution function* (CDF) $F(x)$ of a random variable gives the probability of X taking on a value up to (and including) x: $F(x) = \mathbb{P}(X \leq x)$. CDF *cumulates* the total probability up to a certain value of the random variable.

Example. Let X be the number of heads after 2 coin tosses. Its PMF and CDF are given by the following table.

Outcome x	0	1	2
PMF $f(x) = \mathbb{P}(X = x)$	0.25	0.5	0.25
CDF $F(x) = \mathbb{P}(X \leq x)$	0.25	0.75	1

2.2 Continuous Random Variables

A *continuous* random variable can take on infinite values (values in any interval on the real line). E.g., the distribution of people's heights in the United States, measured in cm, is a continuous random variable.

The *probability density function* (PDF) $f(x)$ of a continuous random variable describes the probability that the random variable falls within a specific *interval* on the real line:

$$\int_a^b f(x)dx = \mathbb{P}(a \leq X \leq b) \ , \tag{2.1}$$

with $\int_{\mathbb{R}} f(x) = 1$. Note that the probability of any *specific* outcome (e.g., the height is *exactly* 187 cm) is zero.

The CDF of a continuous random variable is analogous to the discrete case:

$$F(x) = \int_{-\infty}^{x} f(z)dz \ . \tag{2.2}$$

Example. For people's heights, $F(180) = 0.9$ means that 90% of people in the U.S. have a height less than or equal to 180 cm.

From (2.1) and (2.2) it follows that

$$\mathbb{P}(a \leq X \leq b) = F(b) - F(a) \ .$$

[1] For the purposes of this class.

2.3 Joint Distributions

Several random variables realized together are characterized by their *joint distribution*. Here, we will focus on two continuous variables, although all the concepts can be easily extended to more variables, of different types.

The *joint probability density function* of random variables X and Y has the same interpretation as in the case of a single random variable:

$$\int_c^d \int_a^b f_{X,Y}(x,y)\,dxdy = \mathbb{P}\left(a \le X \le b\,, c \le Y \le d\right)\ . \tag{2.3}$$

The *joint cumulative distribution function* is defined as follows:

$$F_{X,Y}(x,y) = \mathbb{P}\left(X \le x\,, Y \le y\right) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u,v)\,dudv\ . \tag{2.4}$$

The *marginal distribution* of a random variable Y is the distribution of Y, averaged across all possible values of X:

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(X \le \infty\,, Y \le y\right) = \int_{-\infty}^{y} \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,dudv = \int_{-\infty}^{y} f_Y(v)\,dv\ , \\ f_Y(v) &= \int_{-\infty}^{\infty} f_{X,Y}(u,v)\,du\ . \end{aligned} \tag{2.5}$$

The *conditional distribution* of Y given $X = x$ is the distribution of Y holding X constant at x:

$$f_{Y|X=x}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}\ . \tag{2.6}$$

It is in itself a PDF, in the sense that $f_{Y|X=x}(y \mid x) \ge 0$ and $\int_{-\infty}^{\infty} f_{Y|X=x}(y \mid x)\,dy = 1$.

Example. Consider a discrete case. Let Y be the random variable taking value 1 if the person is employed and 0 otherwise. Likewise, the X be the random variable taking value 1 if the person has college degree and 0 otherwise. Suppose that the distribution for the U.S. labor force over 25 years of age in 2012 is given by the following table:

	$Y=0$	$Y=1$	Total
$X=0$	0.053	0.586	0.639
$X=1$	0.015	0.346	0.361
Total	0.068	0.932	1.000

Here, column and row titled “Total” correspond to marginal distributions of X and Y, respectively.

If a person has a college diploma, the probability of being employed is

$$\mathbb{P}\left(Y = 1 \mid X = 1\right) = \frac{\mathbb{P}\left(X = 1\,, Y = 1\right)}{\mathbb{P}\left(X = 1\right)} = \frac{0.346}{0.361} = 0.958\ .$$

If a person does not work, the probability of being employed is

$$\mathbb{P}\left(Y = 1 \mid X = 0\right) = \frac{\mathbb{P}\left(X = 0\,, Y = 1\right)}{\mathbb{P}\left(X = 0\right)} = \frac{0.586}{0.639} = 0.917\ .$$

Thus, a person with college education is more likely to be employed than the one without.

3 Characteristics of Random Variables

Probability distributions are often summarized in terms of their *expectations* (*means*) and *variances*.

3.1 Expectation

The *expectation* (mean, expected value) of a random variable X is the first moment of X[2], denoted by $\mathbb{E}[X]$ or μ_X[3], and equals the *weighted average* of the possible values, where their probabilities play the role of weights:

- for discrete case:

$$\mathbb{E}[X] = \sum_{i=1}^{N} x_i \mathbb{P}(X = x_i) \ ; \tag{3.1}$$

- for continuous case:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x)\, dx \ . \tag{3.2}$$

Example. Given a constant a, $\mathbb{E}[a] = a$, because a constant value occurs with probability 1, so we can use (3.1).

It can be shown[4] that, for some function $g(\cdot)$,

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

3.2 Variance

The *variance* of a random variable X is the second central moment of X[5], denoted by $\text{Var}(X)$ or σ_X^2[6], is a measure of *dispersion* of the distribution of X (how stretched or squeezed it is):

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \ . \tag{3.3}$$

In particular,

- for discrete case:

$$\text{Var}(X) = \sum_{i=1}^{N} (x_i - \mu_X)^2 \, \mathbb{P}(X = x_i) \ ; \tag{3.4}$$

- for continuous case:

$$\text{Var}(X) = \int_{\mathbb{R}} (x - \mu_X)^2 f(x) dx \ . \tag{3.5}$$

In other words, the variance is the mean deviation of X from $\mathbb{E}[X]$.

Variance can also be defined as "the expectation of the square minus the square of the expectation":

$$\begin{aligned} \text{Var}(X) = \mathbb{E}[X - \mathbb{E}[X]]^2 &= \mathbb{E}\left[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]\,\mathbb{E}[X] + (\mathbb{E}[X])^2 \\ &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \ . \end{aligned} \tag{3.6}$$

The positive square root of $\text{Var}(X)$ is called the *standard deviation*, $\sigma_X = \sqrt{\sigma_X^2}$. It is another measure of dispersion, which is more convenient to work with because its units of measurement are the same as those of X.

[2] The n^{th} *moment* of X equals to $\mathbb{E}[X^n]$.

[3] Or simply μ if X is obvious from the context.

[4] Using an application of the change of variables, which we omit here because it is discussed in probability theory courses.

[5] The n^{th} *central moment* of X is defined as $\mathbb{E}[(X - \mathbb{E}[X])^n]$.

[6] Or simply σ^2 if X is obvious from the context.

3.3 Covariance

The relationship between two random variables X and Y is often summarized in terms of their *covariance*, denoted by $\mathrm{Cov}\left(X, Y\right)$ or σ_{XY}. It is the mean product of their deviations from their individual means:

$$\mathrm{Cov}\left(X, Y\right) = \mathbb{E}\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right] \ . \tag{3.7}$$

In particular,

- for discrete case:
$$\mathrm{Cov}\left(X, Y\right) = \sum_{i=1}^{N}\sum_{j=1}^{K}\left(x_j - \mu_X\right)\left(y_i - \mu_Y\right)\mathbb{P}\left(X = x_j \ , Y = y_i\right) \ ; \tag{3.8}$$
- for continuous case:
$$\mathrm{Cov}\left(X, Y\right) = \int_{\mathbb{R}}\int_{\mathbb{R}}\left(X - \mu_X\right)\left(Y - \mu_Y\right)f_{X,Y}(x, y)\,dxdy \ . \tag{3.9}$$

It shows *linear* relationship between two random variables:

- if higher values of X predict higher values of Y, the covariance between X and Y is positive;
- if higher values of X predict lower values of Y, the covariance between X and Y is negative;
- if there is no linear association, covariance is 0.

One explanation, perhaps slightly informal, of this relationship is as follows. Suppose a random variable Y can be approximated as a linear function of X: $Y \approx a + bX$. Then,

$$\mathrm{Cov}\left(X, Y\right) = \mathrm{Cov}\left(X, a + bX\right) = b\mathrm{Cov}\left(X, X\right) = b\mathrm{Var}\left(X\right) \ .$$

Since variance is always non-negative, the sign of covariance tells us the sign of the slope of the linear function that best approximates Y. Therefore, we claim that when covariance is positive, bigger values of X correspond to bigger values of Y, and so forth.

Covariance can also be defined as "the expectation of the product minus the product of expectations":

$$\begin{aligned}\mathrm{Cov}\left(X, Y\right) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])(Y - \mathbb{E}\left[Y\right])\right] &= \mathbb{E}\left[XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y\right] \\ &= \mathbb{E}\left[XY\right] - \mu_Y\mathbb{E}\left[X\right] - \mu_X\mathbb{E}\left[Y\right] + \mu_X\mu_Y \\ &= \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] \ . \end{aligned} \tag{3.10}$$

Correlation maps covariance to an interval $[-1, 1]$, making it easier to interpret. It is defined as

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}\left(X, Y\right)}{\sqrt{\mathrm{Var}\left(X\right)\mathrm{Var}\left(Y\right)}} \ . \tag{3.11}$$

Correlation is sometimes denoted by ρ_{XY}.

3.4 Useful Properties

Some useful properties of expectations, variances, and covariances are as follows. For constants $a, b, c, d \in \mathbb{R}$ and random variables X, Y:

- expectation is a *linear operator*:
$$\mathbb{E}\left[aX + bY + c\right] = a\mathbb{E}\left[X\right] + b\mathbb{E}\left[Y\right] + c \ ; \tag{3.12}$$

- variance is *not* a linear operator:

$$\mathrm{Var}\left(aX+bY+c\right)=a^2\mathrm{Var}\left(X\right)+b^2\mathrm{Var}\left(Y\right)+2ab\mathrm{Cov}\left(X,Y\right)\ ; \tag{3.13}$$

- special case of the above:

$$\mathrm{Var}\left(aX+b\right)=a^2\mathrm{Var}\left(X\right)\ ;$$

- $\mathrm{Cov}\left(a+bX,c+dY\right)=bd\mathrm{Cov}\left(X,Y\right)$;
- $\mathrm{Cov}\left(X,X\right)=\mathrm{Var}\left(X\right)$;
- $\mathrm{Corr}(X,X)=1$.

3.5 Conditional Moments

The *conditional expectation* of Y given $X=x$ is defined as follows:

- for discrete case:

$$\mathbb{E}\left[Y\mid X=x\right]=\sum_{i=1}^{N}y_i\mathbb{P}\left(Y=y_i\mid X=x\right)\ ; \tag{3.14}$$

- for continuous case:

$$\mathbb{E}\left[Y\mid X=x\right]=\int_{\mathbb{R}}yf_{Y|X=x}(y\mid x)\,dy\ . \tag{3.15}$$

The *conditional variance* of Y given $X=x$ is defined as follows:

- for discrete case:

$$\mathrm{Var}\left(Y\mid X=x\right)=\sum_{i=1}^{N}\left(y_i-\mathbb{E}\left[Y\mid X=x\right]\right)^2\mathbb{P}\left(Y=y_i\mid X=x\right)\ ; \tag{3.16}$$

- for continuous case:

$$\mathrm{Var}\left(Y\mid X=x\right)=\int_{\mathbb{R}}\left(y-\mathbb{E}\left[Y\mid X=x\right]\right)^2f_{Y|X=x}(y\mid x)\,dy\ . \tag{3.17}$$

3.6 Independence

Random variables X and Y are said to be *independent* if the realization of one does not affect the realization of the other.

Formally, $X\perp Y$ if $f_{X,Y}(x,y)=f_X(x)f_Y(y)$ for all possible values of x and y.

An equivalent definition involves the concept of conditional distribution: $X\perp Y$ if $\mathbb{P}\left(Y=y\mid X=x\right)=\mathbb{P}\left(Y=y\right)$ and $\mathbb{P}\left(X=x\mid Y=y\right)=\mathbb{P}\left(X=x\right)$ for all possible values of x and y. As a consequence, we have that, for independent random variables, $\mathbb{P}\left(X=x\ ,Y=y\right)=\mathbb{P}\left(X=x\right)\mathbb{P}\left(Y=y\right)$.

If X and Y are independent, they are also *mean independent*: $\mathbb{E}\left[XY\right]=\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$. It follows that:

- $\mathrm{Cov}\left(X,Y\right)=\mathbb{E}\left[XY\right]-\mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]=0$;
- $\mathrm{Corr}(X,Y)=0$.

In other words, if $X\perp Y$, X and Y are uncorrelated, but in general *the converse is not true* (correlation is only a measure of linear association).

3.7 Law of Iterated Expectations

The *law of iterated expectations* (law of total expectations, tower rule) states that

$$\mathbb{E}[Y] = \mathbb{E}_X[\mathbb{E}_Y[Y \mid X]] \ , \tag{3.18}$$

if all involved expectations exist. In other words, the mean of Y is the weighted average of the conditional distribution of Y given X, weighted by the probability distribution of X.

Intuitively, if you want to compute the average height of all humans, you could compute it the standard way (LHS of the equation), or you could compute the average heights for each country, and take the weighted mean of *those* averages, with the weights being probabilities that a person belongs to a given country.

One leading example of application of the law of iterated applications that will be used repeatedly throughout the course is as follows. Suppose that $\mathbb{E}[U \mid X] = 0$ for random variables U and X. Then, it follows that

$$\begin{aligned}\mathbb{E}[U] &= \mathbb{E}[\mathbb{E}[U \mid X]] = \mathbb{E}[0] = 0 \ ; \\ \mathbb{E}[UX] &= \mathbb{E}[\mathbb{E}[UX \mid X]] = \mathbb{E}[\mathbb{E}[U \mid X] \cdot X] = \mathbb{E}[0 \cdot X] = 0 \ .\end{aligned}$$

Note, however, that it *does not* follow, in general, that if $\mathbb{E}[U] = 0$ then $\mathbb{E}[U \mid X] = 0$.

3.8 Law of Iterated Variance

The *law of iterated variance* (law of total variance) states that

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y \mid X)] + \text{Var}(\mathbb{E}[Y \mid X]) \ . \tag{3.19}$$

The first part is sometimes referred to as the “explained” part of the variance (since it is “explained” by the variance of X), and the second part is referred to as the unexplained variance.

4 Some Useful Distributions

4.1 Bernoulli Distribution

A *Bernoulli random variable* assumes only two values: 1, with probability p, and 0, with probability $q \equiv 1 - p$. It is often used to describe binary variables (e.g. rain or no rain, heads or not heads).

The PDF of this random variable is as follows:

$$f(x) = \begin{cases} p^x(1-p)^{1-x} \ , & x \in \{0, 1\} \\ 0 \ , & \text{otherwise} \end{cases} \ . \tag{4.1}$$

Note that the single parameter p is enough to characterize the entire distribution:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^{2} x_i \mathbb{P}(X = x_i) = 0 \cdot p^0(1-p)^{1-0} + 1 \cdot p^1(1-p)^{1-1} = p \ , \\ \text{Var}(X) &= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = p - p^2 = p(1-p) \ ,\end{aligned}$$

because $\mathbb{E}\left[X^2\right] = \mathbb{E}[X]$ since both 0 and 1, raised to the second power, remain the same.

4.2 Uniform Distribution

The probability density function of a *uniformly distributed random variable* assumes a constant non-zero value in the interval $[a, b]$, and is zero everywhere else. Since the PDF has to integrate to one, this value equals to $\frac{1}{b-a}$:

$$f(x) = \begin{cases} \dfrac{1}{b-a} \ , & x \in [a, b] \\ 0 \ , & \text{otherwise} \end{cases} \ . \tag{4.2}$$

The CDF is then given by (2.2):

$$F(x) = \int_{-\infty}^{x} f(z)\,dz = \begin{cases} \displaystyle\int_{-\infty}^{x} 0\,dz\ , & x < a \\ \displaystyle\int_{-\infty}^{a} 0\,dz + \int_{a}^{x} \frac{dz}{b-a}\ , & a \le x \le b \\ \displaystyle\int_{-\infty}^{a} 0\,dz + \int_{a}^{b} \frac{dz}{b-a} + \int_{b}^{x} 0\,dz\ , & b < x \end{cases}$$

$$= \begin{cases} 0\ , & x < a \\ \dfrac{x-a}{b-a}\ , & a \le x \le b \\ 1\ , & b < x \end{cases}\ .$$

Note that

$$\mathbb{E}\left[x\right] = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{a}^{b} \frac{x}{b-a}\,dx = \frac{1}{b-a}\cdot\frac{1}{2}(b^2-a^2) = \frac{a+b}{2}\ ,$$

$$\mathrm{Var}\left(x\right) = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 = \int_{a}^{b} \frac{x^2}{b-a}\,dx - \frac{(a+b)^2}{4} = \frac{1}{b-a}\cdot\frac{1}{3}(b^3-a^3) - \frac{(a+b)^2}{4} = \frac{(a-b)^2}{12}\ .$$

4.3 Normal Distribution

Normal distribution is the most important and most frequently used distribution in econometrics. The estimators we will cover in this course all have a normal distribution as the number of observations $n \to \infty$.

The PDF of the normally distributed random variable is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}\ , \quad x \in \mathbb{R}\ . \tag{4.3}$$

As such, normal distribution $N(\mu, \sigma^2)$ is completely defined by its mean, μ, and variance, σ^2. Its PDF is bell shaped: single-peaked and symmetric about the mean.

The CDF is given by

$$F(x) = \int_{-\infty}^{x} f(z)dz\ . \tag{4.4}$$

This integral can't be evaluated in closed form. Instead, its values are computed numerically.

For any normal distribution, the probability of assuming values within a given number of standard deviations from the mean is constant. Therefore:

- approximately 68% of the observations lie within one standard deviation from the mean:

$$\mathbb{P}\left(\mu - \sigma \le X \le \mu + \sigma\right) \cong 0.68\ ;$$

- approximately 95% of the observations lie within two (1.96, to be precise) standard deviations from the mean:

$$\mathbb{P}\left(\mu - 2\sigma \le X \le \mu + 2\sigma\right) \cong 0.95\ .$$

An important property of normal random variables is as follows: if $X \sim N(\mu, \sigma^2)$, then $Y = a + bX \sim N(a + b\mu, b^2\sigma^2)$. In other words, a *linear transformation* of a normal random variable is also a normal random variable.

The special case of a normal distribution with $\mu = 0$ and $\sigma = 1$ is referred to as the *standard normal distribution*. Random variables with this distribution are typically denoted by Z. Their PDF is

$$\phi\left(z\right) = \frac{1}{\sqrt{2\pi}} e^{-\dfrac{z^2}{2}}\ , \tag{4.5}$$

and their CDF is

$$\Phi(z) = \int_{-\infty}^{z} \phi(t)\, dt \; . \tag{4.6}$$

Let $Y \sim N(\mu, \sigma^2)$. Then, $Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$. Indeed, as noted above, normal distribution is fully characterized by its mean and variance, and linear transformation of a normal variable is a normal variable. Therefore, we immediately conclude that Z must be a normal random variable, and we only need to compute its mean to variance.

The mean is computed as follows:

$$\mathbb{E}[Z] = \mathbb{E}\left[\frac{Y-\mu_Y}{\sigma_Y}\right] = \frac{1}{\sigma_Y}\mathbb{E}[Y-\mu_Y] = \frac{1}{\sigma_Y}(\mathbb{E}[Y]-\mu_Y) = 0\;.$$

The variance is computed as follows:

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Var}\left(\frac{Y-\mu_Y}{\sigma_Y}\right) = \frac{1}{\sigma_Y^2}\mathrm{Var}(Y-\mu_Y)\\ &= \frac{1}{\sigma_Y^2}\left(\mathrm{Var}(Y) + \mathrm{Var}(\mu_Y) - 2\mathrm{Cov}(Y,\mu_Y)\right)\\ &= \frac{1}{\sigma_Y^2}\left(\sigma_Y^2 + 0 - 0\right)\\ &= 1\;.\end{aligned}$$

Therefore, using a table for standard normal distribution, one can find probabilities for any normal random variable:

$$\mathbb{P}(Y \le a \mid \mu, \sigma) = \mathbb{P}\left(Z \le \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)\;.$$

4.4 Multivariate Normal Distribution

Let $\mathbf{X} = (X_1, X_2, \ldots, X_k)^\top$ be the vector of random variables with means $\boldsymbol{\mu} = (\mu_{X_1}, \ldots, \mu_{X_k})^\top$ and a covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{X_1}^2 & \mathrm{Cov}(X_1,X_2) & \ldots & \mathrm{Cov}(X_1,X_k)\\ \mathrm{Cov}(X_1,X_2) & \sigma_{X_2}^2 & \ldots & \mathrm{Cov}(X_2,X_k)\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{Cov}(X_1,X_k) & \mathrm{Cov}(X_2,X_k) & \ldots & \sigma_{X_k}^2 \end{pmatrix}.$$

Then, $\mathbf{X}$ is said to have a *multivariate normal distribution*, $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$, if it has the following PDF:

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}\;, \tag{4.7}$$

where $|\Sigma|$ is the determinant of Σ. Note that Σ needs to be positive definite.

The multivariate normal distribution has the following properties:

- if k variables have a multivariate normal distribution, any linear combination is also normally distributed. In fact, this is an alternative definition of a multivariate normal distribution: two random variables X and Y are jointly normal if $aX + bY$ has a normal distribution $\forall a, b \in \mathbb{R}$;
- the marginals of a multivariate normal are also normal. In particular, for $\mathbf{X} = (X_1, ..., X_k)^\top$, the marginal distribution of X_1 is given by $X_1 \sim N(\mu_{X_1}, \sigma_{X_1}^2)$;
- the conditional distribution of Y given $X = x$ is also normally distributed, with mean $\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ and variance $(1-\rho^2)\sigma_Y^2$, where ρ is correlation between X and Y;
- zero covariance *does imply* independence. One can see this by applying $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ to bivariate normal random variables with $\Sigma = \begin{pmatrix}\sigma_X^2 & 0\\ 0 & \sigma_Y^2\end{pmatrix}$.

4.5 Chi-Squared Distribution

A lot of useful distributions can be constructed from random variables that are normally distributed. One example is χ^2-distribution.

By definition, a random variable distributed as χ_m^2 is the sum of m independent standard normal random variables raised to the second power. Parameter m is called the number of *degrees of freedom*.

The pdf of $X \sim \chi_m^2$ is as follows:

$$f(x \mid m) = \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} x^{\frac{m}{2}-1} e^{-\frac{x}{2}} \ ,$$

where $x \geq 0$, $m > 0$, and $\Gamma(\alpha)$ is the gamma function.

Its mean and variance are $\mathbb{E}\left[X\right] = m$ and $\text{Var}\left(X\right) = 2m$, respectively.