Pre-Semester Course in Statistics

Tutor: Nan Hu

First version: 02.10.2018
This version: 27.09.2019
Slides author: Sebastian Schreiber

Overview

Introduction

Aims of the course

- Refresh basic statistics
- Introduce some more advanced topics (that are helpful in the first year GSEFM courses in my opinion)
- Build a solid foundation for the pre-semester course in Econometrics and for Mathematical Methods 1 (Probability Theory) $\longrightarrow$ this course will cover neither probability theory nor econometrics.

Disclaimer: This a is a theory course. Empirical topics such as hypothesis testing, confidence intervals, least squares estimation,.. are covered in the pre-semester course in Econometrics.

Structure of the course

Timing as usual: Start at 9:15 am, one-hour lunch break, end at 3 pm latest

- Sep 27: Lecture
- Sep 30: Morning: Lecture, Afternoon: Intro to the statistical software Stata
- October 1: Exercise session (most important session - have a look at the exercises in advance)

If you are familiar with the contents of this course (especially with the introductory stuff and the stata intro), feel free to use your time more efficiently, e.g. just briefly look at these slides at home, start refreshing your econometrics skills,... $\Rightarrow$ this course is not mandatory.

The slides and exercises can be find at:
https://sites.google.com/view/nanhu

Literature

Here are some chapters you might find helpful (Wooldridge and Stock/Watson also provide some exercises):

- Wooldridge - Introductory Econometrics: Appendices B,C
- Stock/Watson - Introduction to Econometrics: Chapters 2, 3.1
- Hamilton - Time Series Analysis: Appendix A.5

Lecture Material

These **slides** often contain long and detailed verbal explanations because I try to make them self-explanatory in case you prefer to study them on your own or want to look back at them after this week. Note that important concepts I introduce are **highlighted in bold** in the notes.

The **exercises** I assembled are meant to make you think about/beyond the concepts from the lecture and to strengthen your math skills. Try to work on them as diligent as possible. Feel free to work with the solution provided, try to understand all steps.

I expect the **exercise session** to be as follows: You tried all exercises at home and could hopefully solve them using the solutions. Yet, I will go through the exercises in detail again during the exercise session.

Random Variables

Random Variables - Overview

In this section, I will introduce basic ideas in Probability Theory with, roughly speaking, Random Variables as the starting point and the Law of Iterated Expectations as the 'ending point'.

Everything 'before' Random Variables, i.e. anything that builds the mathematical foundation of Random Variables (such as Sigma Algebras, Probability Spaces,..) will be covered thoroughly in the first year course *Mathematical Methods 1*.

Random Variables

Intuitive/nonmathematical approach: A **Random Variable (RV)** is a quantity/number, which is stochastic.

Semi-mathematical approach: A RV is a function that assigns a real number to each possible event of an experiment.

Mathematical approach: Will be covered in the lecture *Mathematical Methods 1* and is out of the scope of this precourse.

Random Variables - Example

Consider the example of tossing a coin two times in a row:

- The result (head/head , head/tail , tail/head , tail/tail) is not a RV since it is not a quantity
- The number of heads *or* tails is not a RV since it is not stochastic but equal to 2 in any case
- The number of heads is a RV (with possible **realisations** 0,1,2)
- When receiving a payoff of 5 per toss in case of heads and a payoff of 3 per toss in case of tail, this payoff is a RV (with possible realisations: 6,8,10)

Discrete vs. Continuous RVs

A RV X is said to be **discrete** if it can take on a finite or countably infinite number K of particular values, i.e. has the possible realizations $x_1, x_2, .., x_K$ or $x_1, x_2, ...$ Note that one uses capital letters to denote RVs and small letters to denote realizations.

A RV is **continuous** if it takes on an uncountable infinite number of realizations.

In the coin tossing example above, a discrete RV would be the number of heads tossed (even if we toss the coin infinitely many times!)
A continuous RV in this setting would be the distance of the tossed coin to another object in the room. (Can you imagine an easier example?)

Mixed RVs

Last but not least there are also mixed RVs[1], i.e. RVs that consist of both discrete and continuous elements. Roughly speaking, these are continuous RVs with jumps.

Example: Suppose you arbitrarily draw a real number R between 0 and 100. If $R \in [0, 50]$, you get a payoff P of $P = \frac{R}{2}$. If $R \in (50, 100]$, you get $P = 20$.
P in this case is a mixed RV.

Bearing in mind that mixed RVs exists, we will yet only consider discrete and continuous RVs in the remainder.

[1] not to be confused with mixture RVs/distributions and mixing RVs

Discrete RVs

Let's first look at discrete RVs. Suppose that we deal with a finite number of realisations.

The **Probability Distribution**[2] assigns a probability to each outcome $x_1, x_2, .., x_K$:

$$P(X = x_k) \in [0, 1] \ \forall k = 1, 2, .., K \tag{1}$$

whereby the probabilities must sum up to 1:

$$\sum_{k=1}^{K} P(X = x_k) = 1 \tag{2}$$

[2] also referred to as Probability Law

Discrete RVs

In order to pin down the stochastic behaviour of a RV, we make use of some functions.

The **Probability Mass Function (PMF)** is defined as follows:

$$f_X(x) = P(X = x) \tag{3}$$

That is, the PMF simply displays the probability of each value on the real axis.

Note in terms of notation that in $f_X(x)$ the capital X in the subscript indicates that f is the PMF of the RV X. This obviously is superfluous if we only deal with one RV, yet in case of multiple RVs one rather uses e.g. f_X, f_Y for distinction than giving each PMF another name such as f, g.

Discrete RVs

The **Cumulative Distribution Function (CDF)** is defined as follows:

$$F_X(a) = P(X \leq a) = \sum_{x_i \leq a} f(x_i) \tag{4}$$

The CDF therefore gives the probability of a RV X taking on a value smaller than or equal to the constant a.

In our discrete setting, one sees jumps at $x = x_i$.

Discrete RVs - Example

Consider again the introductory example of tossing a (fair) coin twice. Let the RV X be the number of heads tossed.

$$f_X(x) = \begin{cases} \frac{1}{4} & \text{if x=0,2} \\ \frac{1}{2} & \text{if x=1} \\ 0 & \text{otherwise} \end{cases}$$

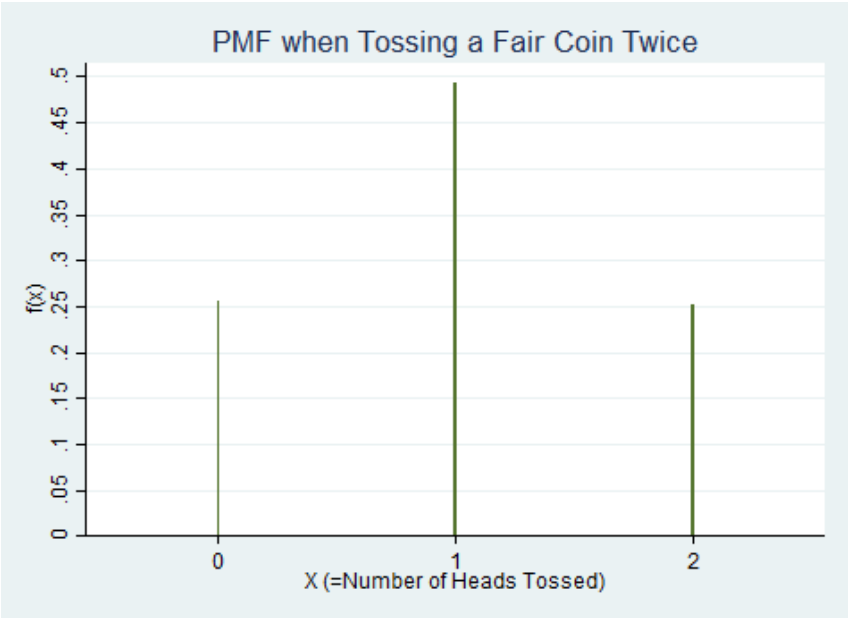

Figure 1: PMF when tossing a fair coin twice

Discrete RVs - Example

Now for the CDF, we obtain

$$F_X(x) = \begin{cases} 0 & \text{if } x<0 \\ \frac{1}{4} & \text{if } 0 \leq x <1 \\ \frac{3}{4} & \text{if } 1 \leq x <2 \\ 1 & \text{if } x \geq 2 \end{cases}$$

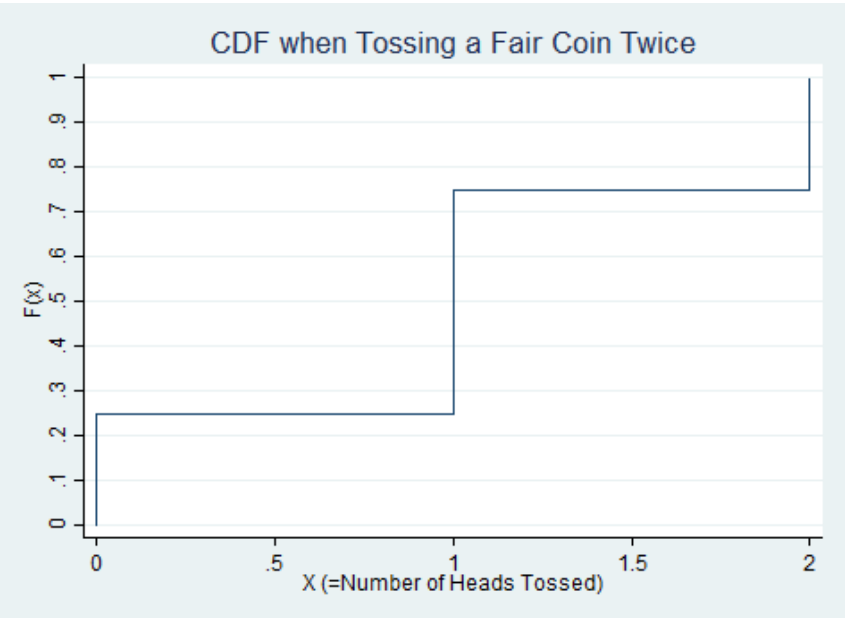

Figure 2: CDF when tossing a fair coin twice

Continuous RVs

Analogously to the discrete case, the CDF of a continuous RV X, $F_X(a)$, is defined as

$$F_X(a) \equiv P(X \leq a) = \int_{-\infty}^{a} f_X(x)dx \tag{5}$$

and therefore gives the probability that X takes on a value less than or equal to some constant a.

Continuous RVs

Due to $P(X = x_i) = 0 \ \forall i$, (2) does not hold in the continuous case and looking at the PMF makes no sense.
Instead, one can describe a continuous RV in terms of its **Probability Density Function (PDF)** $f_X(x)$ for which then holds

$$\int_{-\infty}^{\infty} f_X(x)dx = 1 \tag{6}$$

Note that this integral, and therefore the PDF, might not exist in some rare cases. We will not consider these and just assume a PDF to exist.

Using the PDF, we can calculate probabilities in the continuous case as follows:

$$P(a < X \leq b) = \int_a^b f_X(x)dx = F_X(b) - F_X(a) \tag{7}$$

$$P(X > a) = 1 - P(X \leq a) \tag{8}$$

Continuous RVs - Example

Below you find the PDF and the CDF of a RV that follows a continuous distribution (I chose a Chi square distribution with 10 degrees of freedom - you will get to know this distribution and the concept of degrees of freedom below, so don't worry too much about this now).

The formulae for the corresponding PDF/CDF are long and not worth mentioning here. Instead, I will just give a verbal example.

Continuous RVs - Example

X might be e.g. the lifetime in years (assuming that a year can be divided infinitely makes this distribution continuous) of a certain animal species: few die just after birth, most live around 10 years, some might live for over 30 years.

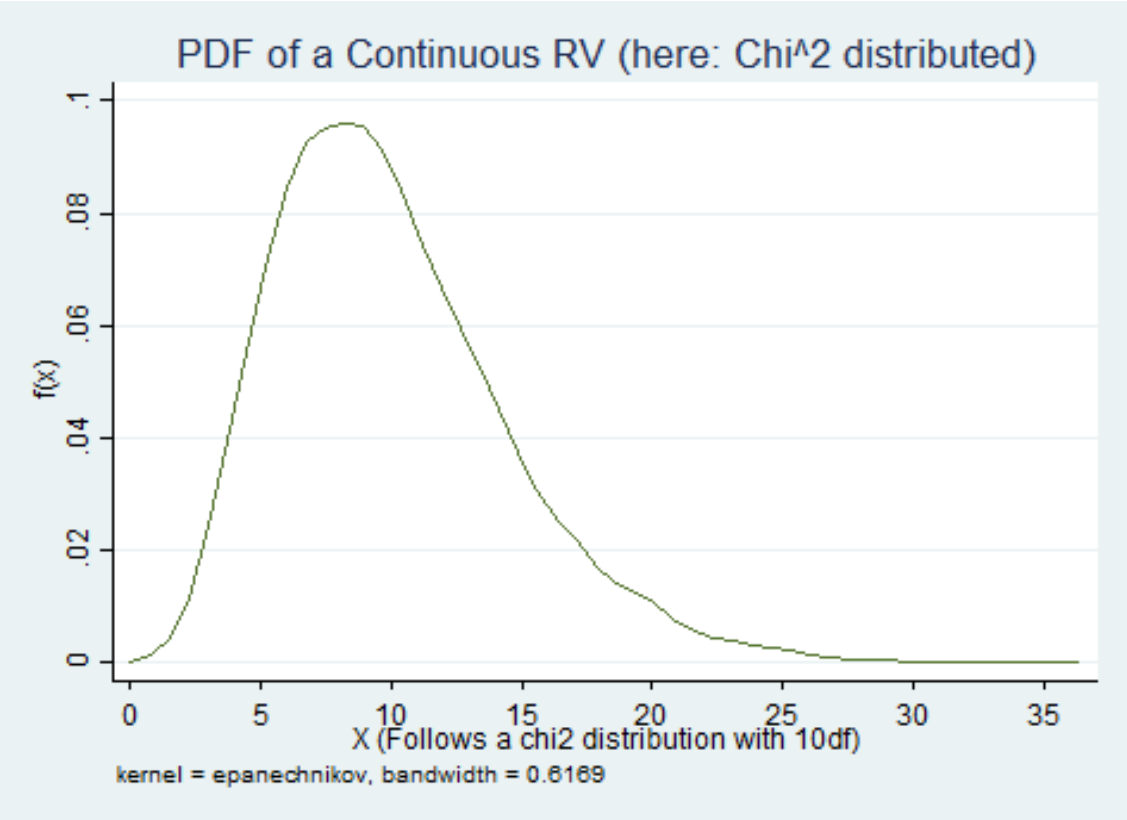

Figure 3: PDF of a continuous RV

Continuous RVs - Example

The corresponding CDF is

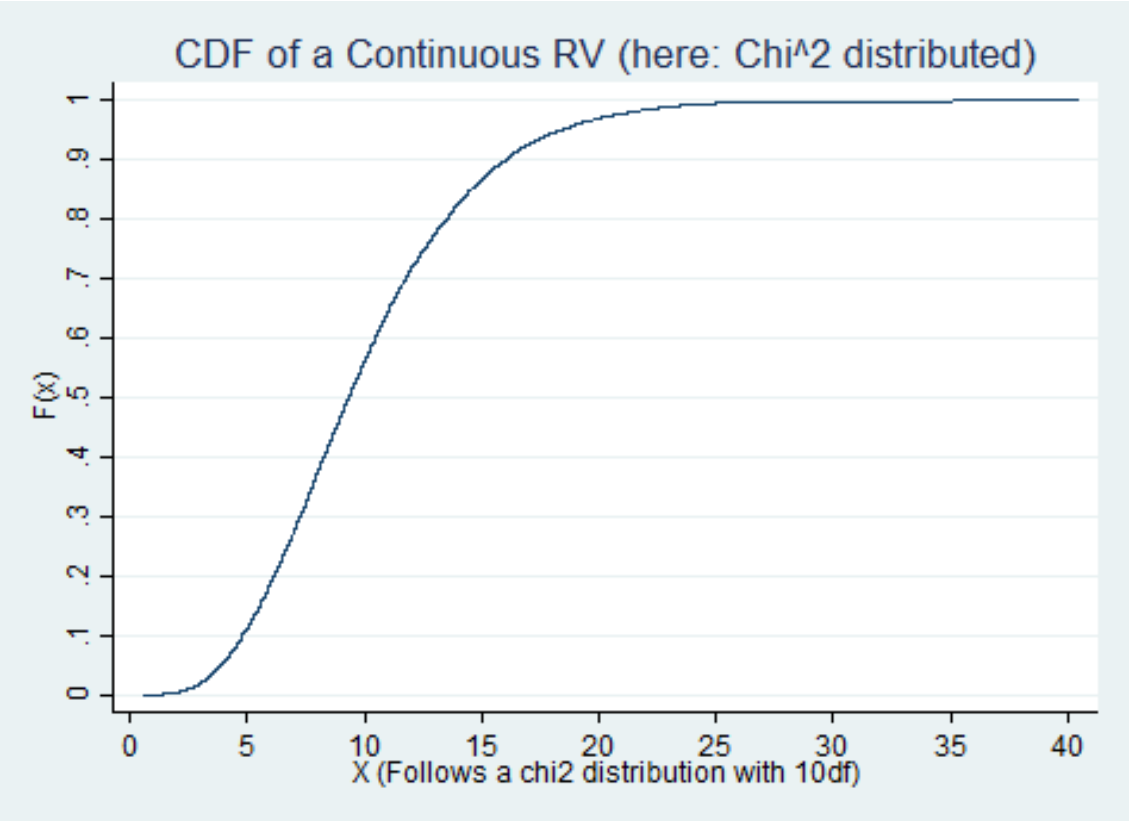

Figure 4: CDF of a continuous RV

Using (5) and (8) one can infer from the graph that e.g. the probability for an animal to live longer than 5 years is roughly $1 - 0.1 = 90\%$.

Relation between PMF/PDF and CDF

Let's clarify the relation between the functions we got to know so far. Any RV has a CDF. Now, one can derive the PMF (PDF) from the CDF, if the RV is discrete (continuous). Importantly, while the CDF always exists, the PMF/PDF might not exist. Yet, for simplicity, we will assume so in this course.

- Discrete case: One can derive the PMF (if it exists) from a discrete CDF by taking the difference between the right limit and the left limit of a point:

$$f_X(x) = \lim_{\varepsilon \to 0} F_X(x + \varepsilon) - \lim_{\varepsilon \to 0} F_X(x - \varepsilon) \tag{9}$$

- Continuous case: One can derive the PDF (if it exists) from a continuous CDF by taking the derivative of the CDF w.r.t. the variable of interest:

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{10}$$

Example - Discrete Case

Suppose

$$F_X(x;p) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

where $F_X(x;p)$ means that the distribution depends not only on x but also on the parameter p (we will look at this in greater detail below).

To figure out the PMF, use (9) for different values of x:

$$f_X(0.3) = \lim_{\varepsilon \to 0} F_X(0.3+\varepsilon) - \lim_{\varepsilon \to 0} F_X(0.3-\varepsilon) = (1-p) - (1-p) = 0$$

$$f_X(20) = \lim_{\varepsilon \to 0} F_X(20+\varepsilon) - \lim_{\varepsilon \to 0} F_X(20-\varepsilon) = 1 - 1 = 0$$

$$f_X(0) = \lim_{\varepsilon \to 0} F_X(0+\varepsilon) - \lim_{\varepsilon \to 0} F_X(0-\varepsilon) = (1-p) - 0 = 1-p$$

$$f_X(1) = \lim_{\varepsilon \to 0} F_X(1+\varepsilon) - \lim_{\varepsilon \to 0} F_X(1-\varepsilon) = 1 - (1-p) = p$$

Example - Discrete Case

One finds that the only 'critical' points, i.e. the x for which $f_X(x;p) \neq 0$, are 0 and 1, so we obtain

$$f_X(x;p) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$$

Example - Continuous Case

Suppose

$$F_X(x;\lambda) = 1 - e^{-\lambda x}$$

Using (10), i.e. taking the derivative w.r.t. x, we obtain

$$f_X(x;\lambda) = \lambda e^{-\lambda x}$$

CDF - properties

Any (discrete/continuous) CDF has the following properties:

- at least right-sided continuous

$$\lim_{\varepsilon \to 0} F_X(x + \varepsilon) = F_X(x) \qquad \forall\, \varepsilon > 0 \tag{11}$$

- monotonously increasing:

$$F_X(a) \leq F_X(b) \qquad \textit{for } a < b \tag{12}$$

- The limits are

$$\lim_{x \to -\infty} F_X(x) = 0 \qquad \text{and} \qquad \lim_{x \to \infty} F_X(x) = 1 \tag{13}$$

Example

Suppose the continuous RV X is characterized by the following CDF:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{27}(x-3)^3 + 1 & \text{if } 0 \leq x \leq 3 \\ 1 & \text{if } x > 3 \end{cases}$$

We want to find out $P(1 < X < 2)$. Using (7), we find

$$P(1 < X < 2) = F_X(2) - F_X(1) = \frac{1}{27}(2-3)^3 + 1 - \frac{1}{27}(1-3)^3 + 1 \approx 0.2593$$

Alternatively, using the PDF obtained using (10) yields the same result:

$$P(1 < X < 2) = \int_1^2 \frac{1}{9}(x-3)^2 dx \approx 0.2593$$

Note that since X is continuous, it does not matter whether we calculate $P(1 < X < 2)$, $P(1 \leq X \leq 2)$ or a mixture of both.

The Big Picture: Population vs. Sample

Before looking a what moments are, let's clarify what is meant when talking about population moments and sample moments.

A probability is a theoretical construct, since in the real world we do not observe probabilities, but relative frequencies. That is, in the real world we of course cannot observe the PDF and use it to calculate probabilities, but have to use the data sample we observe.

Therefore, population moments are mostly[3] unobserved and then estimated as good as possible by sample moments.
While we will look briefly at estimation issues below, we first establish the population moments.

[3] unless we have information about the whole population, which sometimes might be the case (e.g. ask every single citizen about his income) but often is impossible (e.g. in case of the exact temperature tomorrow)

Moments

Intuitively, **Moments** are expected values of powers of RVs. But let's start from scratch:

- For a continuous RV, the rth (population) moment is defined as

$$\int_{-\infty}^{\infty} x^r f_X(x) dx \tag{14}$$

- Note that I again just assume the integral in (14) to exist. I will keep this assumption for the remainder.
- Analogously, for a discrete RV, the rth (population) moment is defined as

$$\sum_{k=1}^{K} x_k^r P(X = x_k) \tag{15}$$

Expected Value - Definition

- The first population moment, known as the **Expected Value**, of a continuous RV X therefore is defined as follows

$$\mu_X = E[X] \equiv \int_{-\infty}^{\infty} x f_X(x) dx \tag{16}$$

  Alternative notation often encountered is E(X) or just EX. I generally use brackets for all statistical operators.

- In case of a discrete RV, we do not have to take the 'detour' via the PDF/PMF and can just sum up the realisations weighted by their respective probabilities:

$$\mu_X = E[X] \equiv \sum_{k=1}^{K} x_k P(X = x_k) \tag{17}$$

Expected Value - Properties

- Before looking at higher population moments, make sure you are familiar with the fact that $E[a + bX] = a + bE[X]$ (I will derive this for the continuous case, yet the discrete case works analogously):
- A generalization of (16) is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{18}$$

- now if $g(X) = a + bX$, we have

$$\begin{aligned} E[a + bX] &= \int_{-\infty}^{\infty} (a + bx) f_X(x) dx \\ &= \int_{-\infty}^{\infty} a f_X(x) dx + \int_{-\infty}^{\infty} bx f_X(x) dx \\ &= a \underbrace{\int_{-\infty}^{\infty} f_X(x) dx}_{=1 \text{ by } (6)} + b \underbrace{\int_{-\infty}^{\infty} x f_X(x) dx}_{=E[X] \text{ by } (16)} = a + bE[X] \end{aligned} \tag{19}$$

Variance - Definition

- The (population) **Variance** of a continuous RV X is defined as

$$\sigma_X^2 = Var[X] \equiv \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx = E[(X - \mu_X)^2] \tag{20}$$

  One can immediately see that this is not equal to the second population moment $\int_{-\infty}^{\infty} x^2 f_X(x) dx$, but to the *centralized* second population moment, where centralization means that the expected value is subtracted.

- If X is discrete, we have

$$\sigma_X^2 = Var[X] \equiv \sum_{k=1}^{K} (x_k - \mu_X)^2 P(X = x_k) \tag{21}$$

- The **Standard Deviation (SD)** σ_X is the square root of the variance:

$$\sigma_X = \sqrt{\sigma_X^2} \tag{22}$$

Variance - Properties

- A generalization of (20) is

$$Var[g(X)] = \int_{-\infty}^{\infty} (g(x) - E[g(X)])^2 f_X(x) dx \tag{23}$$

- Now if $g(X) = a + bX$, we have

$$\begin{aligned} Var[a + bX] &= \int_{-\infty}^{\infty} ((a + bx) - \underbrace{E[a + bX]}_{=a+b\mu_X \text{ by (19)}})^2 f_X(x) dx \\ &= b^2 Var[X] \end{aligned} \tag{24}$$

- An important result that is often used is

$$Var[X] = E[(X - \mu_X)^2] = E[X^2] - 2\mu_X E[X] + \mu_X^2 = E[X^2] - E[X]^2 \tag{25}$$

Higher Moments

Even though the expected value and the variance/SD are the most important moments, one might also want to consider the 3rd and 4th moment in some cases.

- The **Skewness** is the 3rd normalized (i.e. centralised around μ_X and divided by σ_X) moment:

$$\gamma_X \equiv \frac{E\left[(X-\mu_X)^3\right]}{\sigma_X^3} \tag{26}$$

- The **Kurtosis** is the 4th normalized (i.e. centralised around μ_X and divided by σ_X) moment:

$$\kappa_X \equiv \frac{E\left[(X-\mu_X)^4\right]}{\sigma_X^4} \tag{27}$$

Higher Moments - Illustrations

Consider again the PDF of a continuous variable from above:

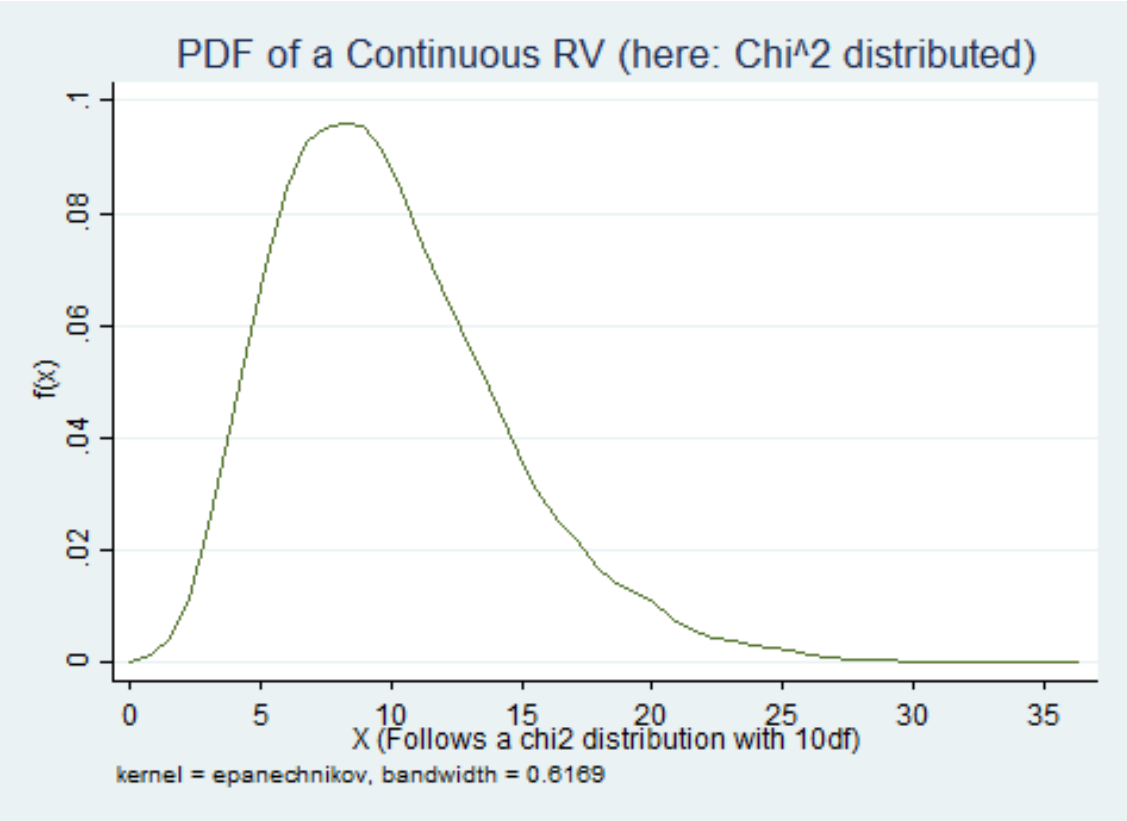

Figure 5: PDF of a continuous RV

The distribution *is skewed to the right* (or: *has a positive skew*).
Therefore, $\gamma_X > 0$.

Higher Moments - Illustrations

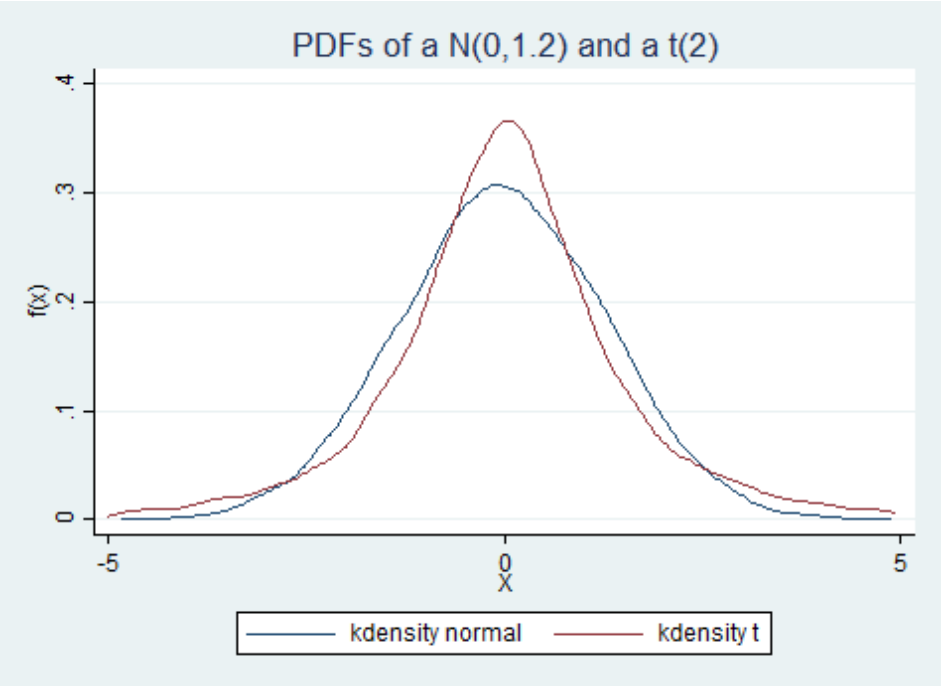

Figure 6: PDFs of two RVs

While the blue distribution has a kurtosis of 3, the red distribution has a kurtosis larger than 3. Roughly speaking, the more extreme deviations there are, i.e. the bigger the tails of the distribution, the larger the kurtosis.

Multidimensional Random Variables

Multidimensional Random Variables - Overview

Economists are usually interested in multidimensional RVs, or more precisely the relation between one variable and some other(s).

- While gathering data on inflation for a sample of countries might be of some interest, a dataset with both inflation and the rate of unemployment allows to test whether e.g. inflation is independent of unemployment.
- A stock price has few informative value if not compared to a benchmark such as a stock market index

In this section, the tools introduced above for one-dimensional RVs will be developed analogously for multidimensional RVs. After then defining conditional probabilities and moments, I will conclude this chapter with a proof of the Law of Iterated Expectations.

Joint Distributions

Consider two continuous RVs X and Y. Assuming that the **joint PDF** $f_{X,Y}(x, y)$ exists, one can define the **joint CDF** as

$$F_{X,Y}(a, b) = P(X \leq a, Y \leq b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x, y)\, dy\, dx \tag{28}$$

In case of discrete RVs, we have

$$F_{X,Y}(a, b) = \sum_{i|x_i \leq a} \sum_{j|y_j \leq b} f_{X,Y}(x_i, y_j) \tag{29}$$

Joint Distributions

- Now suppose we have two continuous RVs X and Y with their joint PDF $f_{X,Y}(x, y)$, but want to know the PDF of just X, i.e. $f_X(x)$. We find this **marginal density** by simply integrating the joint PDF with respect to y:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \tag{30}$$

- Discrete case:

$$P(X = x_i) = f_X(x_i) = \sum_j f_{X,Y}(x_i, y_j) \tag{31}$$

Conditional Distributions

- Conditional Probability

$$P(X = x_i | Y = y_i) = \frac{P(X = x_i \cap Y = y_i)}{P(Y = y_i)} \tag{32}$$

- Conditional PDF/PMF

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \tag{33}$$

- Conditional Expectation - Continuous

$$E[Y|X] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \tag{34}$$

- Conditional Expectation - Discrete

$$E[Y|X] = \sum_j y_j f_{Y|X}(y_j|x) \tag{35}$$

Independence

Together with the concept of conditional probabilities/expectations, the concept of **Independence** can be considered the most important idea in probability theory.

The RVs X and Y are independent iff

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) \tag{36}$$

or, in terms of conditional distributions, iff

$$f_X(x|y) = f_X(x), \qquad f_Y(y|x) = f_Y(y) \tag{37}$$

Example

Suppose $f_{X,Y}(x,y) = \begin{cases} x+y & \text{if } x,y \in [0,1] \\ 0 & \text{otherwise} \end{cases}$.

We want to calculate $f_X(x)$, $F_X(x)$, $E[X]$ and $Var[X]$ and check whether X and Y are independent or not.

To find $f_X(x)$, we use (30) and divide the integral into three parts:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \begin{cases} \int_{-\infty}^{0} 0\, dy = 0 \\ \int_{0}^{1} (x+y)\, dy = x + \frac{1}{2} \\ \int_{1}^{\infty} 0 dy = 0 \end{cases}$$

This leads us to

$$f_X(x) = \begin{cases} x + \frac{1}{2} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Example

To find $F_X(x)$ we use (5) and set the lower limit of the integral to 0 since $f_X(x)$ is 0 for smaller values (x now serves as upper limit so we choose z to be the variable of integration):

$$F_X(x) = \int_{-\infty}^{x} f_Z(z)dz = \int_{0}^{x} \left(z + \frac{1}{2}\right) dz = \frac{1}{2}x^2 + \frac{1}{2}x$$

This leads us to

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{2}x^2 + \frac{1}{2}x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}$$

Example

To find $E[X]$, use (16):

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \underbrace{\int_{-\infty}^{0} x\, 0\, dx}_{=0} + \int_{0}^{1} x\left(x + \frac{1}{2}\right) dx + \underbrace{\int_{1}^{\infty} x\, 0\, dx}_{=0} = \frac{7}{12}$$

Note that I also included the 0-terms for completeness, since it is a common mistake just to plug $(x + \frac{1}{2})$ into $\int_{-\infty}^{\infty} x f_X(x) dx$ (and then to conclude that the expected value does not exist since $\int_{-\infty}^{\infty} x(x + \frac{1}{2}) dx$ tends to infinity)$\Rightarrow$ remember to consider all cases of the PDF/CDF.

To find $Var[X]$, we first find calculate $E[X^2]$ and then use (25):

$$E\left[X^2\right] = \int_{0}^{1} x^2\left(x + \frac{1}{2}\right) dx = \frac{5}{12}$$

$$Var[X] = E[X^2] - E[X]^2 = \frac{11}{144}$$

Example

Now check for independence. We want to check whether (36) holds. Since we already know $f_{X,Y}(x, y)$ and $f_X(x)$, we only need to calculate $f_Y(y)$. One could do this as above, but we just make use of the symmetry between X and Y, so since $f_X(x) = x + \frac{1}{2}$,

$$f_Y(y) = y + \frac{1}{2}$$

Plugging into (36) yields

$$x + y \neq \left(x + \frac{1}{2}\right)\left(y + \frac{1}{2}\right)$$

Therefore, X and Y are not independent.

IID

We often deal with sequences of RVs. For example, X_t is a sequence of RVs $X_1, X_2, ...$ Think of e.g. X_1 being the temparature today, X_2 in one year and so forth. When now claiming that X_t is **iid**, i.e. that $X_t \overset{iid}{\sim} somedistribution$, we suppose that the temperature in one year is independent of the temperature today while still following the same distribution.

Another example: $X_t \overset{iid}{\sim} N(0,1)$ means that $X_1, X_2, ..$ is an independently and identically distributed sequence of RVs with mean 0 and variance 1.

Example

Let's get acquainted with some of the concepts introduced on the previous slides.

Suppose $X_1, \ldots, X_n$ and $\varepsilon_1, \ldots, \varepsilon_n$ are sequences of iid continuous RVs (we do not assume a particular distribution, but only that they are iid). We further do not claim anything about the relation between the two RVs.

Make sure you understand the setting: Due to iid, knowing e.g. x_6 does not deliver any information about x_9, yet x_3 and ε_3 might interact in some way, i.e. are not claimed to be independent.

We formally want to show the intuitive result that

$$E[\varepsilon_i | x_1, \ldots, x_i, \ldots, x_n] = E[\varepsilon_i | x_i]$$

If you wonder why I don't use capital Xs here: I use $E[\varepsilon_i | x_i]$ shorthand for $E[\varepsilon_i | X_i = x_i]$ and so on.

Example

Note that for the sake of readability I will omit the subscripts in the PDFs, i.e. $f_{\varepsilon_i|x_1,..,x_i,..,x_n}(\varepsilon_i|x_1,..,x_i,..,x_n)$ becomes $f(\varepsilon_i|x_1,..,x_i,..,x_n)$.

$$
\begin{aligned}
& E[\varepsilon_i|x_1,..,x_i,..,x_n] && \text{rewrite using (34)} \\
&= \int_{-\infty}^{\infty} \varepsilon_i f(\varepsilon_i|x_1,..,x_i,..,x_n) d\varepsilon_i && \text{use (33)} \\
&= \int_{-\infty}^{\infty} \varepsilon_i \frac{f(\varepsilon_i, x_1,..,x_i,..,x_n)}{f(x_1)\cdot..\cdot f(x_i)\cdot..\cdot f(x_n)} d\varepsilon_i && \text{use (36) since iid} \\
&= \int_{-\infty}^{\infty} \varepsilon_i \frac{f(\varepsilon_i, x_i)\cdot f(x_1)\cdot..\cdot f(x_{i-1})\cdot f(x_{i+1})\cdot..f(x_n)}{f(x_1)\cdot..\cdot f(x_i)\cdot..\cdot f(x_n)} d\varepsilon_i && \text{cross out} \\
&= \int_{-\infty}^{\infty} \varepsilon_i \frac{f(\varepsilon_i, x_i)}{f(x_i)} d\varepsilon_i && \text{use (33)} \\
&= \int_{-\infty}^{\infty} \varepsilon_i f(\varepsilon_i|x_i) d\varepsilon_i && \text{rewrite using (34)} \\
&= E[\varepsilon_i|x_i]
\end{aligned}
$$

Covariance - Definition

- A two-dimensional generalization of (16) is

$$E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)dy\,dx \tag{38}$$

- If we now specify $g(X,Y) = (X-\mu_X)(Y-\mu_Y)$, we obtain the **Covariance** between X and Y:

$$Cov[X,Y] \equiv E[(X-\mu_X)(Y-\mu_Y)] \tag{39}$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{X,Y}(x,y)dy\,dx \tag{40}$$

Covariance - Definition

- A simplified formula to calculate the covariance is

$$\begin{aligned} Cov[X, Y] &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E[XY - X\mu_Y - \mu_X Y + \mu_X \mu_Y] \\ &= E[XY] - E[X]E[Y] \end{aligned} \tag{41}$$

- Reformulating (41) yields a formula we lack so far, namely the expected value of a product of two RVs:

$$E[XY] = E[X]E[Y] + Cov[X, Y] \tag{42}$$

  which simplifies to $E[XY] = E[X]E[Y]$ if X and Y are independent.

- Furthermore, the **(Coefficient of) Correlation** between X and Y is defined as

$$\rho_{XY} \equiv \frac{Cov[X, Y]}{\sigma_X \sigma_Y} \tag{43}$$

Independence vs. Correlation

Bear in mind that $X \perp\!\!\!\perp Y$ (X and Y are independent) implies $Cov[X, Y] = 0$, but the opposite does not hold. Consider the following counterexample:

Suppose

$$Z \perp\!\!\!\perp Y \qquad \text{and} \qquad \mu_Z = \mu_Y = 0$$

Let

$$X \equiv Z \cdot Y$$

Note that the value of X obviously depends on Y, so X and Y are not independent:

$$X \not\perp\!\!\!\perp Y$$

Yet, they are uncorrelated:

$$Cov[X, Y] = E[XY] - E[X]\underbrace{E[Y]}_{=0} \overset{X=ZY}{=} E[ZY^2] \overset{Z \perp\!\!\!\perp Y}{=} \underbrace{E[Z]}_{=0} E[Y^2] = 0$$

Some important Properties

You definitely should become familiar with the following essential properties of RVs (as usual capital letters denote RVs and small letters constants)

$$E[aX + bY] = aE[X] + bE[Y] \tag{44}$$

$$Var[aX + bY] = a^2 Var[X] + b^2 Var[Y] + 2abCov[X, Y] \tag{45}$$

$$Cov[aX + bY, cU + dV] = acCov[X, U] + adCov[X, V] + bcCov[Y, U] + bdCov[Y, V] \tag{46}$$

While we already saw the proofs for the expected value and the variance (see (19) and (24)), we omit the proof for the covariance, which works just the same way.

Law of Iterated Expectations

Last but not least we derive the important **Law of Iterated Expectations** using the results introduced so far.
In the proof, we add subscripts to the expectation operators that should clarify w.r.t. which variable an expected value is calculated.[4]

LIE in its most general form:

$$E_X[E_{Y|X,Z}[Y|X]|Z] = E_Y[Y|Z] \tag{47}$$

A special case that often suffices is

$$E_X[E_{Y|X}[Y|X]] = E_Y[Y] \tag{48}$$

or, in short

$$E[E[Y|X]] = E[Y] \tag{49}$$

[4] Take care that such subscripts usually have a different meaning, namely to condition on these variables, e.g. $E[X|Y] = E_Y[X]$)

LIE - Proof of the Special Case

$$
\begin{aligned}
& E_X[E_{Y|X}[Y|X]] && \text{|use (16) and (34)} \\
&= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \right] f_X(x) dx && \text{|rearrange terms} \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \underbrace{f_{Y|X}(y|x) f_X(x)}_{= f_{X,Y}(x,y) \text{ by (33)}} dy\, dx && \text{|rearrange terms} \\
&= \int_{-\infty}^{\infty} y \underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,y) dx}_{= f_Y(y) \text{ by (30)}} dy \\
&= \underbrace{\int_{-\infty}^{\infty} y f_Y(y) dy}_{= E_Y[Y] \text{ by (16)}} \\
&= E_Y[Y] && (50)
\end{aligned}
$$

LIE - Application

In the simple (=one regressor) linear regression model $y = \alpha + \beta x + u$, one usually makes the zero conditional mean assumption $E[u|x] = 0$. Using the LIE, one can derive some important results:

- $E[u] = 0$

$$\begin{aligned} E[u|x] &= 0 \qquad |E[(\cdot)] \\ \underbrace{E[E[u|x]]}_{=E[u] \text{ by LIE}} &= \underbrace{E[0]}_{=0} \end{aligned}$$

- $E[g(x)u] = 0$

$$\begin{aligned} E[u|x] &= 0 \qquad |\cdot g(x) = E[g(x)|x] \\ E[g(x)u|x] &= 0 \\ \underbrace{E[E[g(x)u|x]]}_{=E[g(x)u] \text{ by LIE}} &= 0 \end{aligned} \tag{51}$$

LIE - Application

To give a more intuitive example, suppose the RV G is your grade in an exam and the RV K is your knowledge on the subject the exam covers. This can either be high (then, $K = K_H$) or low (then, $K = K_L$).

The LIE now states that

$$E[G] = E[E[G|K]] = P(K = K_H)E[G|K = K_H] + P(K = K_L)E[G|K = K_L]$$

That is, your expected grade will be a weighted average of your grade when your knowledge of the exam content is high and of your grade when it is low.

Particular Distributions

Motivation - Distributions

So far, we have talked in abstract terms about RVs and their properties. Especially, we have never specified a particular PMF/PDF/CDF. We will now do so by considering particular distributions. Note that most of the distributions I introduce are not crucially important for the courses/your research, but you should have heard of them.

First some general words about distributions: In the general expression

$$a \sim b(c, d, e)$$

- a is the RV of interest
- $\sim$ means 'distributed as'
- b is an abbreviation for a distribution
- (c, d, e) means that this distribution has three parameters. If we know these, we know the exact distribution, i.e. all moments,...

Motivation - Distributions

In this section I will often omit proofs or results used for them. Yet, I will provide links to look them up in case you are interested, mostly on `https://proofwiki.org`.

Furthermore, I will not always state PMF/PDF, CDF and moments, but sometimes omit some of these, if I consider it to be out of scope.

Let's now start with some particular discrete distributions.

Uniform - Discrete

Ex. Roll a die

- PMF

$$f_X(x; n) = P(X = x) = \begin{cases} \frac{1}{n} & \text{if } x = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases} \tag{52}$$

- Moments

$$E[X] = \sum_{x=1,2,..} xP(X = x) = \frac{1}{n} \underbrace{\sum_{k=1}^{n} k}_{= \frac{n(n+1)}{2}} = \frac{n+1}{2} \tag{53}$$

Note: See `here` if you are unfamiliar with the result $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$

$$Var[X] = \frac{n^2 - 1}{12} \qquad \text{(Proof } \texttt{here}\text{)} \tag{54}$$

Bernoulli

If $X \sim Bn(p)$, X is restricted to the values 0 and 1, e.g. you are either employed or unemployed. The distribution is fully characterized by the parameter p, which gives the probability of $x = 1$, e.g. of employment in the above example.

- PMF:

$$f_X(x; p) = \begin{cases} 1 - p & \text{if } x = 0 \\ p & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{55}$$

or, alternatively

$$f_X(x; p) = \begin{cases} p^x (1 - p)^{1-x} & \text{if } x = 0, 1 \\ 0 & \text{otherwise} \end{cases} \tag{56}$$

Bernoulli

- CDF:

$$F_X(x;p) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases} \tag{57}$$

- Moments:

$$E[X] = \sum_{i=1}^{K} x_i P(X = x_i) = 1 \cdot p + 0 \cdot (1-p) = p \tag{58}$$

$$Var[X] = E[X^2] - E[X]^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = p \cdot (1-p) \tag{59}$$

Note that (9) holds, as already shown in the example on p.26.

Binomial

When $X \sim Bi(n, p)$, X is the number of successes when n independent Bernoulli experiments with success probability p are carried out. Therefore, the binomial distribution is the generalization of the bernoulli distribution, which is obtained when $n = 1$.

- PMF:

$$f_X(x; p, n) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{if } x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases} \tag{60}$$

where

$$\binom{n}{x} \equiv \frac{n!}{x!(n-x)!} \tag{61}$$

- Moments (see `here` and `here` for proofs)

$$E[X] = np \tag{62}$$

$$Var[X] = np(1-p) \tag{63}$$

Poisson Distribution

If $X \sim Po(\lambda)$, X can be interpreted as the number of successes of a large number ($n \to \infty$) of Bernoulli Experiments with a small success probability ($p \to 0$) (see `here` for the proof).

- PMF:

$$f_X(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \text{for } x = 0, 1, 2, \ldots \tag{64}$$

- Moments (Proofs are exercises):

$$E[X] = \lambda \tag{65}$$

$$Var[X] = \lambda \tag{66}$$

Continuous Distributions

Now let's consider some particular continuous distributions.

Continuous Uniform Distribution

Suppose $X \sim U(a, b)$, i.e. X uniformly taking on values in $[a, b]$.

- PDF:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{67}$$

- CDF:

$$F_X(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b \end{cases} \tag{68}$$

Note that (10) holds.

Continuous Uniform Distribution

- Expected Value

$$E[X] = \int_a^b x f_X(x) dx = \frac{b-a}{2} \tag{69}$$

- Variance

$$Var[X] = E[X^2] - E[X]^2 = \int_a^b \frac{x^2}{b-a} dx - \left(\frac{b-a}{2}\right)^2 = \frac{(b-a)^2}{12} \tag{70}$$

Normal

Suppose X is **normally distributed**, i.e. $X \sim N(\mu_X, \sigma_X^2)$.

- PDF:

$$f_X(x; \mu_X, \sigma_X) = \frac{1}{\sigma_X \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2} \tag{71}$$

- Moments (deriving this is tedious):

$$E[X] = \mu_X \tag{72}$$

$$Var[X] = \sigma_X^2 \tag{73}$$

Normal

You should know the following properties of the normal distribution by heart[5]:

$$X \sim N(\mu_X, \sigma_X^2) \quad \Rightarrow \quad aX + b \sim N(a\mu_X + b, a^2\sigma_X^2) \tag{74}$$

$$X \sim N(a\mu_X + b, a^2\sigma_X^2) \quad \Rightarrow \quad \frac{X - b}{a} \sim N(\mu_X, \sigma_X^2) \tag{75}$$

Using these properties, one can carry out the following manipulations referred to as **standardization**:

$$\begin{aligned} X &\sim N(\mu_X, \sigma_X^2) && | -\mu_X \\ X - \mu_X &\sim N(0, \sigma_X^2) && | : \sigma_X \\ \frac{X - \mu_X}{\sigma_X} &\sim N(0, 1) \end{aligned} \tag{76}$$

[5]The 'trick' is just to remember that if constants are factored out of a variance, they are squared, cfr. (45)

Normal

If we consider the RV $Z \equiv \frac{X-\mu_X}{\sigma_X}$, we have that $Z \sim N(0,1)$ and Z is said to be **standard normally distributed**.

Note that one then often labels the PDF of a standard normal distribution not $f_Z(z;0,1)$ but $\phi(z)$ and the CDF not $F_Z(z;0,1)$ but $\Phi(z)$.
Additional to (7) and (8), the following holds due to symmetry around 0:

$$P(Z < -z) = P(Z > z) \tag{77}$$

Distributions based on normally distributed RVs - Overview

The following distributions based on normally distributed RVs are very important in hypothesis testing.

- Let $(X_1, X_2, .., X_n)$ be iid standard normally distributed RVs, i.e. $X_i \overset{iid}{\sim} N(0,1)\ \forall i = 1, \ldots, n$. Let $Y \equiv \sum_{i=1}^{n} X_i^2$. Then, Y follows a **Chi-Square Distribution** with n **degrees of freedom (df)**:

$$Y \sim \chi^2(n) \tag{78}$$

- Let $X \overset{iid}{\sim} N(0,1)$, $Y \sim \chi^2(n)$ and X and Y independent. Let $Z \equiv \frac{X}{\sqrt{\frac{Y}{n}}}$. Then, Z follows a **t Distribution** with n df:

$$Z \sim t(n) \tag{79}$$

- Let $Y_1 \sim \chi^2(n_1)$ and $Y_2 \sim \chi^2(n_2)$ with Y_1 and Y_2 independent. Let $Z \equiv \frac{Y_1/n_1}{Y_2/n_2}$. Then, Z follows an **F Distribution** with n_1 numerator df and n_2 denominator df:

$$Z \sim F(n_1, n_2) \tag{80}$$

Example

Suppose $X \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$. Let's show that the famous one sample t test statistic follows a $t(n-1)$ distribution:[6]

$$t = \frac{\bar{x}_n - \mu_X}{\frac{s}{\sqrt{n}}} = \frac{\bar{x}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \frac{\sigma_X}{s}$$

Now one can show that $\bar{x}_n \sim N(\mu_X, \frac{\sigma_X^2}{n})$:

$$\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$E[\bar{x}_n] = \frac{1}{n} \sum_{i=1}^{n} \underbrace{E[x_i]}_{=\mu_X} = \frac{1}{n} n\mu_X = \mu_X$$

$$Var[\bar{x}_n] = \frac{1}{n^2} \sum_{i=1}^{n} \underbrace{Var[x_i]}_{=\sigma_X^2} = \frac{1}{n^2} n\sigma_X^2 = \frac{\sigma_X^2}{n}$$

[6] I abbreviate the sample variance $s_{X,n}$ with s. See (89) below

Example

Therefore, $\frac{\bar{x}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$ is $N(0,1)$ distributed:

$$\bar{x}_n \sim N(\mu_X, \frac{\sigma_X^2}{n}) \qquad | - \mu_X$$

$$\bar{x}_n - \mu_X \sim N(0, \frac{\sigma_X^2}{n}) \qquad | : \frac{\sigma_X}{\sqrt{n}}$$

$$\frac{\bar{x}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \sim N(0,1)$$

One can show less trivially (see e.g. `here`) that $\frac{s_{X,n}}{\sigma_X} \sim \sqrt{\frac{\chi^2_{n-1}}{n-1}}$. So all in all,

$$t = \frac{\bar{x}_n - \mu_X}{\frac{s}{\sqrt{n}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \overset{(79)}{\sim} t(n-1)$$

Lognormal

I will not go into detail here, but only present the first two moments. Let $X \sim N(\mu_X, \sigma_X^2)$. Let $Y = e^X$. Then,

$$Y \sim logN\left(e^{\mu_X + \frac{1}{2}\sigma_X^2}, e^{2\mu_X + \sigma_X^2}\left(e^{\sigma_X^2} - 1\right)\right) \tag{81}$$

Lognormal Distribution - Example

Suppose $\tau = e^{\theta}$ where $\theta \sim N(3, 8)$.

One can calculate $E[\tau]$ as

$$E[\tau] = E[e^{\theta}] \overset{(*)}{=} e^{E[\theta]+\frac{1}{2}Var[\theta]} = e^{3+\frac{1}{2}8} = e^{7}$$

Note especially that $E[e^{\theta}] \neq e^{E[\theta]}$.

Now for the Variance,

$$Var[\tau] = Var[e^{\theta}] \overset{(**)}{=} e^{2E[\theta]+Var[\theta]}(e^{Var[\theta]} - 1) = e^{2\cdot 3+8}(e^{8} - 1) = e^{22} - e^{14}$$

That is, $\tau \sim logN(e^{7}, e^{22} - e^{14})$

You should know $(*)$ and $(**)$ by heart.

Exponential

- PDF:

$$f_X(x, \lambda) = \lambda e^{-\lambda x} \qquad \text{for } x \geq 0,\ \lambda > 0 \tag{82}$$

- CDF:

$$F_X(x, \lambda) = \int_0^x \lambda e^{-\lambda k} dk = \lambda \left[-\frac{1}{\lambda} e^{-\lambda k} \right]_0^x = 1 - e^{-\lambda x} \tag{83}$$

- Moments:

$$E[X] = \frac{1}{\lambda} \tag{84}$$

$$Var[X] = \frac{1}{\lambda^2} \tag{85}$$

$$\tag{86}$$

Large Sample Analysis and Intro to Estimation

Sample Moments

Recapitulate what we did so far: We have only considered population moments like the expected value / population variance /..
Now a sample moment is an estimate (unless our sample consists of the entire population) of a population moment based on realisations/observations/data $x_1, x_2, .., x_K$.

Sample Moments

The first sample moment is the **sample mean** (or sample average). Let's introduce the common notation:

- In the **cross-sectional** case, where one observes individuals $i = 1, 2, .., n$, the sample mean is

$$\bar{x}_n \equiv \frac{1}{n} \sum_{i=1}^{n} x_i \tag{87}$$

- In the **time-series/longitudinal** case, where one has observations at time points $t = 1, 2, .., T$, the sample mean is

$$\bar{x}_T \equiv \frac{1}{T} \sum_{t=1}^{T} x_t \tag{88}$$

In the remainder, I will use the cross-sectional notation.

Sample Moments

- The **sample variance** in the cross-sectional case is

$$s^2_{X,n} \equiv \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \tag{89}$$

- The **sample covariance** between X and Y reads as

$$s_{XY,n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) \tag{90}$$

Note that if $s_{XY,n} = 0$, we call X and Y **orthogonal**, which is the sample equivalent to 'uncorrelated' in the population case.

Motivation for the Theorems

We now want to link sample moments and population moments. Yet, due to the randomness involved, statements such as $\lim_{n\to\infty} \bar{x}_n = E[X]$ do not hold. For example, its is theoretically possible to obtain only heads when tossing a coin infinitely many times.

Therefore, we need a weaker statement than convergence, namely convergence in Probability. This concept admits that it is possible to e.g. toss heads an infinite number of times in a row, but it takes into account that the *Probability* that this occurs goes to 0 as the number of trials goes to infinity.

Disclaimer: I will try to provide intuition for the theorems, not to proof the rigorously. I therefore will provide few information on assumptions made and so on. Note also that there are numerous versions of the theorems, but I only introduce the most prominent ones (those assuming X is iid)

Weak Law of Large Numbers

In the following slides, I will present two of the most important theorems in statistics.

The **Weak Law of Large Numbers (WLLN)** states that if X is iid and its expected value and variance exist and are finite,

$$\bar{x}_n \xrightarrow{p} E[X] \tag{91}$$

or in alternative notations

$$\operatorname*{plim}_{n\to\infty} \bar{x}_n = \mu_X \tag{92}$$

$$P(|\bar{x}_n - \mu_X| \geq \varepsilon) \to 0 \qquad \text{for } n \to \infty \text{ and } \varepsilon > 0 \text{ arbitrarily small} \tag{93}$$

For the covariance (and therefore for the variance), the result is similar:

$$\operatorname*{plim}_{n\to\infty} s_{XY,n} = \sigma_{XY} \qquad (\Rightarrow \operatorname*{plim}_{n\to\infty} s^2_{X,n} = \sigma^2_X) \tag{94}$$

Strong Law of Large Numbers

Less-used and only included here for the sake of completeness is the Strong Law of Large Numbers (SLLN):

$$P(\lim_{n\to\infty} \bar{x}_n = \mu_X) = 1 \tag{95}$$

or

$$\bar{x}_n \xrightarrow{a.s.} E[X] \tag{96}$$

Note that the SLLN makes a slightly stronger statement (sample mean will be equal to expected value with probability one) using a slightly stronger mode of convergence (almost sure convergence)

Central Limit Theorem

To tackle the **Central Limit Theorem (CLT)**, we introduce another mode of convergence, namely convergence in distribution, denoted $\xrightarrow{d}$ (also referred to as convergence in Law, denoted $\xrightarrow{L}$). Since it is implied by convergence in Probability, one also talks of weak convergence.

The CLT states that if X is iid, $E[X] = \mu_X$ and $Var[X] = \sigma_X^2 < \infty$,

$$\sqrt{n}(\bar{x}_n - \mu_X) \xrightarrow{d} N(0, \sigma_X^2) \tag{97}$$

or

$$\sqrt{n}(\bar{x}_n - \mu_X) \overset{a}{\sim} N(0, \sigma_X^2) \tag{98}$$

where $\overset{a}{\sim}$ stands for 'asymptotically distributed'.
Yet,

$$\sqrt{n}(\bar{x}_n - \mu_X) \sim N(0, \sigma_X^2)$$

would be wrong since the distribution is not exactly normal but only approximately.

Central Limit Theorem - Appraisal

Note that this is quite a powerful statement: X can follow any (discrete/continuous) distribution, it only must be iid[7] and have a finite second moment. Then, as n gets bigger, the sample mean tends to a normal distribution.
If you ever asked yourself why the normal distribution is so important: This is one of main reasons why.

[7] There even exist other versions of the CLT that allow for non-iid observations

Central Limit Theorem

Recall the manipulations in (76). Analogously, we can represent (98) as

$$\begin{aligned} \sqrt{n}(\bar{x}_n - \mu_X) &\xrightarrow{d} N(0, \sigma_X^2) && | : \sqrt{n} \\ (\bar{x}_n - \mu_X) &\xrightarrow{d} N(0, \frac{\sigma_X^2}{n}) && | + \mu_X \\ \bar{x}_n &\xrightarrow{d} N(\mu_X, \frac{\sigma_X^2}{n}) \end{aligned} \tag{99}$$

meaning that the sample mean approximately follows a normal distribution, or as

$$\begin{aligned} \sqrt{n}(\bar{x}_n - \mu_X) &\xrightarrow{d} N(0, \sigma_X^2) && | : \sigma_x \\ \frac{\bar{x}_n - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} &\xrightarrow{d} N(0, 1) \end{aligned} \tag{100}$$

meaning that the standardized sample mean approximately follows a standard normal distribution.

Central Limit Theorem - Illustration

In the Figure below, I calculated the sample mean of 1,000 RVs that follow a Poisson distribution with $\lambda = 6$, i.e. $\bar{x}_{1000}$.
Then I repeated this step 10,000 times to obtain a distribution of the sample mean, which is the blue line in the figure. Comparing it with the red line, namely a normal distribution, shows that CLT 'worked'. Using (99), one can calculate that $\bar{x}_{1000} \xrightarrow{d} N(E[X], \frac{Var[X]}{n}) \overset{by(65)}{\sim} N(6, \frac{6}{1000})$

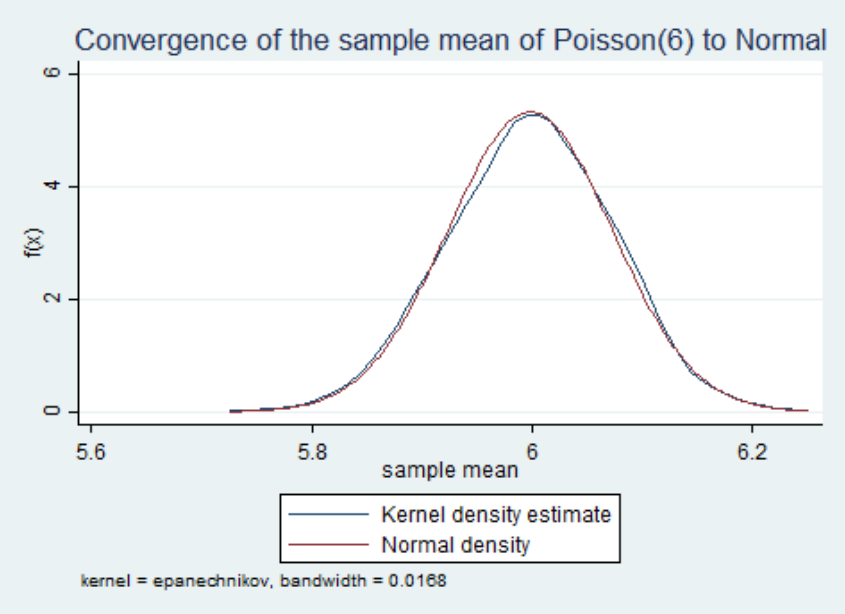

Figure 7: CLT: Convergence of $\bar{x}_{1000}$, where $X \sim Po(6)$, to $N(6, \frac{6}{1000})$

Central Limit Theorem - Illustration

Even though the CLT applies roughly speaking in 99.9% of the cases, there are situation where it doesn't.
Consider e.g. a Cauchy distribution: This distribution has no moments and thus violates the CLT assumptions of finite variance and existent first moment. We therefore see no convergence:

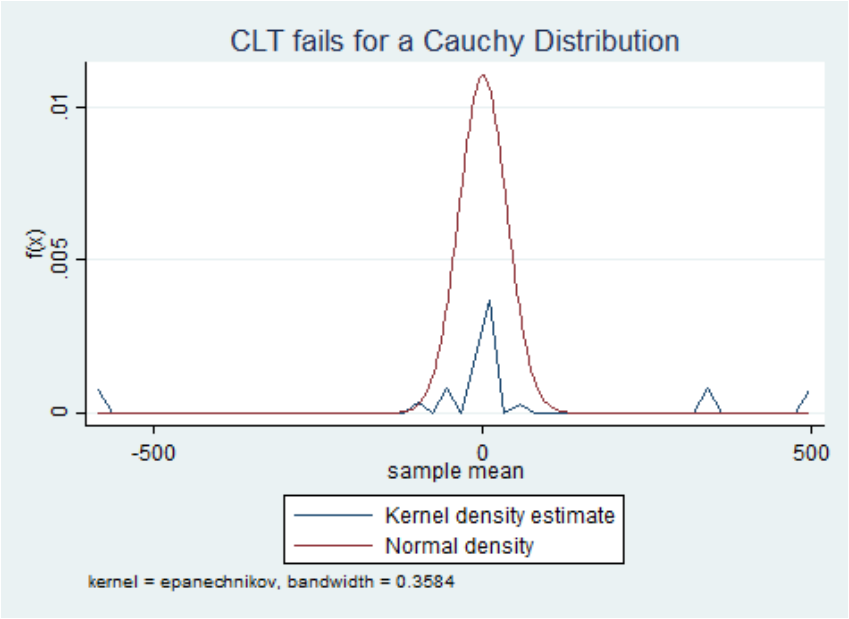

Figure 8: CLT fails for a Cauchy since this distribution has no moments

Modes of Convergence - Overview

As the previous limit theorems illustrate, when dealing with RVs we need more modes of convergence than $\rightarrow$, which is too strong when stochastics come into play.

You should bear in mind that their order is as follows:

$$\rightarrow \text{ implies } \xrightarrow{a.s.} \text{ implies } \xrightarrow{p} \text{ implies } \xrightarrow{d} \tag{101}$$

Properties of Estimators

We need some criteria to distinguish good estimators from bad ones. To establish the properties, I rely on θ to denote a parameter of interest to be estimated (e.g. p if $X \overset{iid}{\sim} Bn(p)$) and on $\hat{\theta}$ do denote an estimator for it.

- **Unbiasedness**

$$E[\hat{\theta}] = \theta \tag{102}$$

- Asymptotical Unbiasedness (implied by unbiasedness)

$$\lim_{n\to\infty} E[\hat{\theta}] = \theta \tag{103}$$

- **Consistency**: $\lim_{n\to\infty} E[\hat{\theta}] = \theta$ and $\lim_{n\to\infty} Var[\hat{\theta}] = 0$, or, in short

$$\operatorname{plim}_{n\to\infty} \hat{\theta} = \theta \tag{104}$$

- **Efficiency**: θ^* is the estimator among all unbiased estimators for which

$$Var[\theta^*] < Var[\theta] \tag{105}$$

Example

To illustrate that consistency and unbiasedness are distinct concepts, consider the following four estimators for the expected value μ when $X \overset{iid}{\sim} N(\mu, \sigma^2)$:

- $\bar{x}_n = \frac{1}{n}\sum_{j=1}^{n} x_j$ is both unbiased and consistent:

$$E[\bar{x}_n] = \frac{1}{n}\sum_{j=1}^{n}\underbrace{E[x_j]}_{=\mu} = \frac{1}{n}\mu\sum_{j=1}^{n} = \mu$$

$$\operatorname*{plim}_{n\to\infty} \bar{x}_n = \mu \qquad \text{by WLLN}$$

- x_1, i.e. using just the first observation, is unbiased but inconsistent:

$$E[x_1] = \mu$$

$$\operatorname*{plim}_{n\to\infty} x_1 = x_1 \neq \mu \qquad \text{by WLLN}$$

Example

- $\bar{x}_n + \frac{1}{n}$ is biased but consistent:

$$E[\bar{x}_n + \frac{1}{n}] = \underbrace{E[\bar{x}_n]}_{=\mu} + \underbrace{E[\frac{1}{n}]}_{=\frac{1}{n}} = \mu + \frac{1}{n} \neq \mu$$

(Yet, $\bar{x}_n + \frac{1}{n}$ is asymptotically unbiased since $\lim_{n\to\infty} E[\bar{x}_n + \frac{1}{n}] = \mu$)

$$\operatorname{plim}_{n\to\infty} \bar{x}_n + \frac{1}{n} = \mu \qquad \text{by WLLN}$$

- $\frac{1}{2}\bar{x}_n$ obviously is both a biased and inconsistent estimator for μ

Note also that $\bar{x}$ is efficient (you will not find an unbiased estimator of μ with a smaller variance.

Degrees of freedom

You need to know the idea of degrees of freedom (df), yet I will not consider a mathematical approach here because this would be out of scope. Roughly speaking, the number of df is equal to the number of observations n minus the number of parameters that had to be estimated in advance.

Ex.: Suppose we want to estimate the population variance σ_X^2, but don't know the population mean μ_X and have to use the sample mean $\bar{x}_n$ as an estimator for it. One can show that the following estimator for σ_X^2 is biased:

$$s_{X,n}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$$

$$E[s_{X,n}^2] = \frac{n-1}{n}\sigma_X^2 \neq \sigma_X^2$$

Degrees of freedom

Yet, if we adjust for the one df that is lost (this is known as Bessel's Correction) since one first has to estimate $\bar{x}_n$, we obtain an unbiased estimator:[8]

$$\hat{\sigma}^2_{X,n} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \tag{106}$$

$$E[\hat{\sigma}^2_X] = \frac{n}{n-1}E[s^2_{X,n}] = \frac{n}{n-1}\frac{n-1}{n}\sigma^2_X = \sigma^2_X$$

Now if we were to know the true population mean μ_X we would not need to estimate it and thus obtain an unbiased estimator without the df-adjustment:

$$\sigma^2_X = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2$$

$$E[\sigma^2_X] = \sigma^2_X$$

[8] Pay attention here with the notation: E.g. Wooldridge calls the corrected estimator for the variance not $\hat{\sigma}^2_{X,n}$ but $s^2_{X,n}$ and does not look at what we called $s^2_{X,n}$ at all.