Econometrics A

Handout 0: Quick Review of Probability and Statistics

Laura Mayoral

IAE and BGSE

Insead, Winter 2021

This session's goal

- Our goal today is to make sure that we all speak the same language (the language of probability and random variables)

- To this effect, we will (quickly) review some basic concepts that should be well-known for everybody by the end of this week.

- If you find difficulties understanding this handout, please come to talk to me and I will give you additional materials that will help you to review this section.

Roadmap

1. Basic notions of probability: probability, random variables, distributions, moments.

2. Asymptotic Theory: a quick introduction.

3. Estimators and properties.

4. Hypotheses testing

1. Basic notions of probability: probability, random variables, distributions, moments.

Probability and statistics in econometrics

- Remember our goal in this course:

- to determine whether a change in one variable causes a change in another variable.

- These variables are interpreted as random $\longrightarrow$

- we should consider probabilistic notions to formalise the sense in which a change in one variable causes the other variable.

- In the following we review some basic notions of probability and statistics that will be very important to understand the main principles of econometrics

Some probability background

- See: Greene (Appendix);

Definition of Probability

- Consider an experiment that has various possible outcomes.

Each possible outcome is represented as a point in a set. Each of these points are elementary events.

- Other events can be formed by combining elementary events.

- Sample space, Ω: the set that contains all elementary events.

Definition of probability

- Probabilities will be assigned to the elementary events according to certain axioms.

- Let Ω be the sample space, A be an event and P$(.)$ is a probability assignment. The three axioms that define a probability are:

- $0 \leq P(A) \leq 1$

- $P(\Omega) = 1$

- If A_1, $A_2, \ldots$ are disjoint events, then $P(\cup_j A_j) = \sum_j P(A_j)$

Example

- Consider the random experiment of throwing a dice.

- $\Omega = \{1, 2, 3, 4, 5, 6\}$.

- Each of these elements are the elementary events.

- Other events can be defined by combining elementary events. A={1,2};

- The probability of each of the elementary events is 1/6.

- $P(1 \cup 2 \cup 3 \cup 4 \cup 5 \cup 6) = 1$; $P(A) = 2/6$;

Random variables

- Definition: A random variable is a function from Ω to the real numbers such that every element of Ω gets one real value.

- Example

- You are interested in the color of the eyes in a population. The set of possible colors are $\{black, brown, green, blue\}$. A random variable is the function that maps this set of events to numbers.

- Let X be your random variable: "color of the eyes".

- $X = \{1, 2, 3, 4\}$. This means $X = 1$ if eyes are black, $= 2$ if brown, etc.

- The values of a random variable can be arbitrary (we could define as well: $X = \{10, 20, 30, 40\}$.)

Random variables and their realisations

- A random variable is a function;

- it represents all the possible outcomes of your random experiment.

- We can associate probabilities to each of these outcomes.

- Realisation of a random variable: Once the random experiment has taken place, we observe its realisation. This is not random anymore.

- We usually use capital letters to denote random variables (r.v.) and small letters to denote particular realisations of these variables.

X=eye color; $x = blue$.

Example

- You are about to toss a coin. $\Omega = \{heads, tails\}$; $X = \{0, 1\}$

- Each of the values of X has an associated probability (if the coin is balanced, 0.5)

- Now you toss the coin, you get heads (X=0): this is a realisation of X.

- This realisation is not random anymore.

Types of random variables

- Discrete random variables

- X is discrete if the number of distinct possible outcomes is either finite or countably infinite.

- For instance $X =$ outcome after tossing a coin; $Y =$ number of times that one should toss a coin until the first tails appears.

The assignment of probabilities in this case is done via a function, $f(x) = P(X = x)$, called the *probability mass function* (pmf), that has the properties:

- $f(x) \geq 0$

- $\sum_i f(x_i) = 1$

Continuous random variables

- X takes values in an interval.

- Examples: X: height of this class, unemployment rate, inflation rate, etc.

- The assignment of probabilities is done via the *probability density function* (pdf), $f(x)$, that has the properties:

- $f(x) \geq 0$

- $\int_i f(x_i) = 1$

- If X is continuous, then $P(X = x) = 0$ for all x.

- P$(a \leq X \leq b) = \int_a^b f(x)\, dx$.

Distribution Function

- The cumulative distribution function is defined as:

$$F(x) = P(X \leq x).$$

If X is discrete, then

$$F(x_k) = \sum_{i=x_1}^{x_k} P(X = x_i),$$

where $x_1 \leq ... \leq x_k$.

If X is continuous,

$$F(x) = \int_{-\infty}^{x} f(x)\, dx.$$

Moments of a univariate distribution

- The shape of a probability distribution can be described with the help of its moments

There are two types of moments:

$$\begin{aligned} \mu_r &= E\left(X^r\right), \text{ is the rth raw moment} \\ \mu_r^* &= E\left(X - \mu_X\right)^r \text{ is the } rth \text{ central moment} \end{aligned}$$

Each of the moments provides some information about the distribution of X. For instance μ_1 is the mean, μ_2^* is the variance.

- Exercise

- Find out what are the names of μ_3^* and μ_4^* and what aspects of the distribution of X describe.

Some important moments: Expectations

- The expected value of a random variable X, denoted as μ, is the first uncentered moment.

- It provides an idea of the central values of the distribution of X.

- Calculation.

$$E(X) = \mu_X = \begin{cases} \sum_i x_i P(X = x_i), \; if \; X \text{ is discrete} \\ \int_{-\infty}^{\infty} x f(x)\, dx, \text{ if } X \text{ is continuous} \end{cases} .$$

Expected value of a function of X

- In situations, we are interested in obtaining the mean of a function of X. Let $Z = g(X)$, be a function of X, then

$$E(Z) = \mu_Z = \begin{cases} \sum_i g(x_i)P(X = x_i), \; if \;\; X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)\,dx, \text{ if } X \text{ is continuous} \end{cases} .$$

- Why is this?

- Particular case: Expectation of a Linear transformation

- If $Z = a + bX$, then

$$E(Z) = a + bE(X). \tag{1}$$

- Because of the expression above, E(.) is called a linear operator

Variance.

- The variance of a distribution is a measure of dispersion with respect to the expected value.

$$Var\left(X\right) = \sigma_X = E\left(X - E\left(X\right)\right)^2.$$

- The variance is always larger than (or equal) to zero.

- If $Var(X) = 0$, then X is a number (has no variation).

Standard deviation

- Notice that the mean and the variance are not measured in the same units.

- Standard deviation: square root of the variance.

$$\sigma = \left(E\left(X^2\right) - E\left(X\right)^2 \right)^{1/2}$$

- The standard deviation is measured in the same units as the expected value.

Variance of a linear transformation of X

- The variance is not a linear operator.
- The variance of a linear function of X. is given by:

Let $Z = a + bX$, then

$$Var\left(Z\right) = b^2 Var\left(X\right).$$

Joint distributions

- Assume you have 2 different random experiments. It is possible to assign probabilities to the outcomes of two experiments at the same time.

- Example: for a given population, we define two variables X=age; $Y = height$. What is the probability that a person chosen at random from this population is older than 20 and taller than 1:70m? i.e., $P(X > 20, Y > 1.7)$?

- The joint distribution of X and Y will allow us to compute the probability above.

- We can also define the joint distribution of any group of variables. Let $X = (X_1, ..., X_n)'$ denote a group of random variables.

Joint distributions.

- The joint distribution completely characterizes the vector of random variables X.

- If X is discrete, then $f\left(x_1,...,x_n\right) = P\left(X_1 = x_1,...,X_n = x_n\right)$ is the joint probability mass function. It has to verify similar conditions as the univariate pmf.

- If X is continuous, then we can assign probability through $f\left(x_1,...,x_n\right)$, the joint probability density function. It has to satisfy the conditions $f\left(x_1,...,x_n\right) \geq 0$ and

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f\left(x,y\right) dydx = 1.$$

Covariance and correlation

- The covariance between a pair of r.v. measures the degree of linear association between them. It is defined as

$$Cov(X,Y) = E\left(\left(X - E\left(X\right)\right)\left(Y - E\left(Y\right)\right)\right)$$

- Interpretation of this measure: we can only interpret the sign of the covariance.

$$\begin{aligned} Cov(X,Y) &> 0: \text{ there is a positive linear relation btw } X,Y \\ Cov(X,Y) &= 0: \text{ there is not a linear relation btw } X,Y \\ Cov(X,Y) &< 0: \text{ there is a negative relationship btw } X,Y. \end{aligned}$$

If $Cov\left(X,Y\right) = 0$ it is said that X and Y and uncorrelated.

Correlation

- The covariance depends on the units of measurement of X and Y and therefore the magnitude of the covariance is NOT informative about the strength of the linear association between X and Y.

- The correlation is a standardized version of the covariance. It is bounded between [-1,1] and therefore not only the sign but also the strength of the relationship can be assessed with it.

$$Corr(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}.$$

- Interpretation:

$$
\begin{aligned}
Corr(X,Y) &= 1: \text{ the relation btw X, Y is positive} \\
&\quad \text{ and perfectly linear} \\
1 &< Corr(X,Y) < 0: \text{ there a positive linear relation,} \\
&\quad \text{ that is higher the higher corr is to 1} \\
Cov(X,Y) &= 0: \ X,Y \text{ are uncorrelated} \\
0 &< Corr(X,Y) < -1: \text{ there a negative linear relation,} \\
&\quad \text{ that is higher the higher corr is to -1} \\
Corr(X,Y) &= -1: \text{ the relation btw X, Y is negative} \\
&\quad \text{and perfectly linear}
\end{aligned}
$$

Covariance and correlation of linear transformations of X and Y

- If $Z = a + bX$, $V = c + dY$, then,

$$Cov\left(Z, V\right) = bdCov\left(X, Y\right).$$

- If $Z = a + bX$, $V = c + dY$, then

$$Corr\left(Z, V\right) = Corr\left(X, Y\right).$$

More properties of expectations, variances and correlations

- The following relationships are very important, you should remember them!

- The expected value of a sum is the sum of expectations

$$E(\alpha_1 X_1 + \alpha_2 X_2 +, \ldots, \alpha_n X_n + c) = \alpha_1 E(X_1) + \alpha_2 E(X_2) +, \ldots, \alpha_n E(X_n) + c$$

- The variance of a sum is the sum of the variances if ONLY if the variables are uncorrelated. General case:

$$Var(\alpha_1 X_1 + \alpha_2 X_2 + c) = \alpha_1^2 Var(X_1) + \alpha_2^2 Var(X_2) + 2\alpha_1 \alpha 2 Cov(X_1, X_2)$$

- Covariance of a sum:

$$Cov(\alpha_1 X_1 + \alpha_2 X_2 + c, \alpha_3 X_3 + d) = \alpha_1 \alpha_3 Cov(X_1, X_3) + \alpha_2 \alpha_3 Cov(X_2, X_3)$$

- Combining the last two expressions, you can find out more expressions, for instance:

- Variance of n variables

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} Cov(X_i, X_j)$$

Marginal distributions

- Consider a bivariate distribution, (X, Y) with probability function $f(x, y)$.

- From $f(x, y)$, it is possible to recover the distributions of X and Y alone, i.e., distributions that do not depend on the other variable.

- These distributions are called marginal distributions.

- If (X, Y) are discrete: $f(x) = \sum_y f(x, y)$; $f(y) = \sum_x f(x, y)$;

- If (X, Y) are continuous: $f(x) = \int_y f(x, y)\, dy$; $f(y) = \int_x f((x, y)\, dx$;

Conditional distributions

- Conditional distributions play a crucial role in econometrics.

- Assume that the variables (X, Y) are related and we have some information about the variable X. Assume further that we have observed that $X = x$. We would like to update the probability of Y given the information available of X, that is, $X = x$.

- Example: Suppose that we are studing Y=height, X=weight of a population and that we have observed that the weight of a person chosen at random is 55kg. Clearly, the probability of having a particular height, say 1.90, given that we know that the person's weight is 55kg, would be different that the unconditional probability of being 1.9.

Conditional distributions

- The distribution of Y conditional to $X = x$ is defined as:

$$f(y|X = x) = \frac{f(x, y)}{f(x)}.$$

- The conditional distribution $f(y|X = x)$ is a probability function and therefore, has to verify the same conditions as any p.d.f or p.m.f (that is, is positive and has to add up –integrate– to 1).

- $f(y|X = x)$ is a function of X. That is, as X takes different values, we would obtain different $f(y|X = x)$

Independence

- In this course we will be interested in how one variable responds to the changes of a related variable.

- However, it can be the case that a variable does not react to the changes of some other variables because they are not related.

- This lack of relationship is called independence.

- The variables (X, Y) are stochastically independent iff (if and only if)

$$f(x, y) = f(x) f(y).$$

- Exercise: show that the latter result is equivalent to:

$$f(x|y) = f(x) \text{ and } f(y|x) = f(y).$$

Independence vs uncorrelation

- X, Y independent $\longrightarrow$ X, Y uncorrelated

- X, Y uncorrelated $\longrightarrow$ X, Y not necessarily independent

- Independence implies the lack of any relationship between X and Y and is a very strong condition.

- It is a much stronger condition than lack of correlation.

- Lack of correlation only means lack of linear relationship between X and Y. It can be the case that $corr\,(X, Y) = 0$ but that X and Y are not independent.

- There is an important exception: if $(X, Y)'$ follow a normal bivariate distribution and are uncorrelated, then they are also independent.

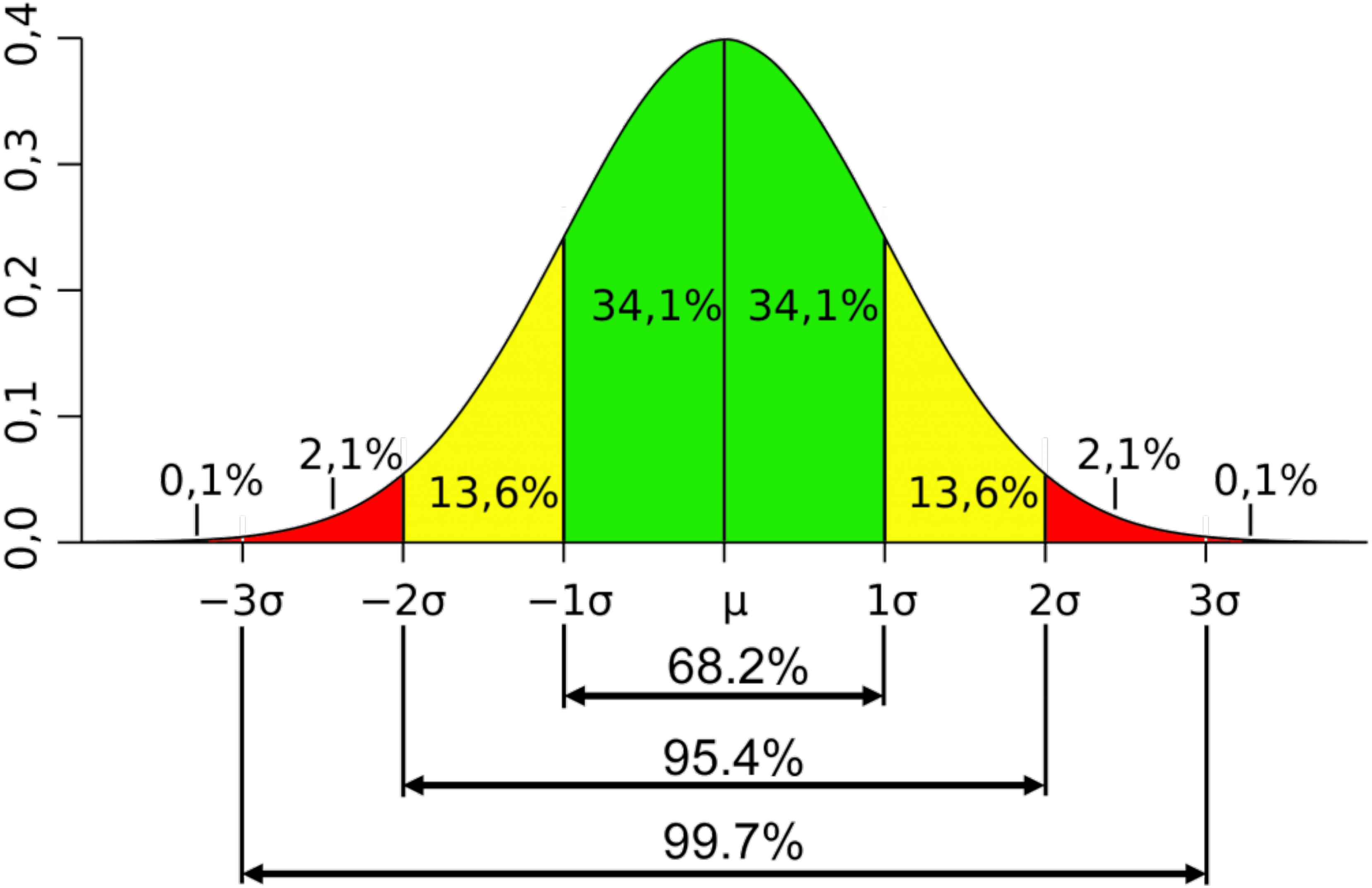
The Normal distribution.

- Univariate Normal distributions

- Let X be a continuous r.v.

- It follows a $N\left(\mu, \sigma^2\right)$ distribution (a Normal distribution with mean=μ and variance σ^2 if its pdf is given by

$$f\left(x\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(x-\mu)^2/2\sigma^2}.$$



0,4
0,3
0,2
0,1
0,0
34,1%
34,1%
13,6%
13,6%
2,1%
2,1%
0,1%
0,1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ
68.2%
95.4%
99.7%

- Standard Normal Distribution: N(0,1)
- The Normal distribution is symmetric
- To compute probabilities from a normal distribution: we use the tables (corresponding to a standard normal distribution).
- This distribution is very important in econometrics/statistics.
- Many techniques assume normality
- Many tests to check normality.
- STATA hint: use qnorm to compare the quantiles of any variable with those of a normal distribution.

The Multivariate Normal distributions

- Let $(X, Y)'$ be a pair of random variables.

- They follow a bivariate normal distribution, denoted as,

$$\left(\begin{array}{c} X \\ Y \end{array} \right) \sim N \left(\mu = \left(\begin{array}{c} \mu_x \\ \mu_y \end{array} \right), \Sigma = \left(\begin{array}{cc} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{array} \right) \right),$$

that is, a Normal distribution with vector of means μ and variance-covariance matrix Σ if for any real vector λ,

$$\lambda' \left(\begin{array}{c} X \\ Y \end{array} \right)$$

is a univariate normal distribution.

- Normal distributions are very important in statistics and econometrics for two reasons:

- They are very common (Central Limit Theorem).

- They have very good properties and then it is very convenient to work under the normality assumption.

- In particular one of this good properties is if X and Y are multivariate normal

$$Y|X \sim N(E(Y|X), var(Y|X))$$

and $E(Y|X) = a + bX$.

See Goldberger Chapter 7 for a description of the properties of these distributions.

Asymptotic theory: a quick introduction

Asymptotic Theory

- Question:

At the dinner table, your brother-in-law suggests playing heads or tails using a coin. You suspect he is cheating.

How do you prove that the coin is unbalanced?

- Context:

- Suppose you have several random variables and you combine them to produce an statistic.

- For instance, $X_1, \ldots, X_n$:

- sample mean: $\overline{X}_n = (\sum_1^n X_i)/n$

- or any other function of these variables (e.g., $T_n = \prod_1^n X_i \ \ldots$).

- How are these statistics distributed?

- In finite samples (i.e., when n is finite), it's often very difficult to provide an exact answer

- asymptotic theory: makes an additional assumption: n is "large" ($n \longrightarrow \infty$) and computes the "limit" of the relevant statistic under this assumption.

- In the coin example:

- Toss the coin n times. Let X_i be the random variable associated to each of these tossings, ($i = 1, \ldots, n, X_i$=1 if heads; X_i=0 if tail). If the coin is balanced, $E(X_i) = 1/2$.

- These variables are i.i.d.

- Compute the sample mean: $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$

- $\bar{X}_n$: share of heads.

- If the coin is balanced: if n is sufficiently large, $\bar{X}_n$ should be close to 1/2.

- A key result in asymptotic theory will tells us that: $\bar{X}_n$ converges to $E(X_i)$!

Key elements of asymptotic theory:

- The meaning of convergence of random variables.
- The most important convergence results
  - The Law of Large Numbers
  - The Central Limit Theorem.

Convergence of Deterministic Sequences

- Before considering convergence of random sequences, let us refresh some notions of convergence of deterministic sequences.

- A sequence of nonrandom numbers $\{a_N\}_{N=1}^{\infty}$ converges to a if for all $\varepsilon > 0$, $\exists\ N_\varepsilon > 0$ such that if $N > N_\varepsilon$, then $|a_N - a| < \varepsilon$. The constant a is called the limit of a_N.

- Example:

  - consider the sequence 1, 1/2, 1/3, . . . , 1/n
  - what is the limit of this sequence as n$\rightarrow\infty$?

- Now, we'll do the same but instead of considering numbers, we'll look at convergence of random variables.

Convergence of random sequences

- Consider a sequence of random variables: $X_1, X_2, \ldots, X_n$

- A sequence of r.v. converges to a limit if for large values of n the sequence and the limit are "close".

- But what does "close" mean when considering random variables?

Convergence of random sequences

- Consider a sequence of random variables: $X_1, X_2, \ldots, X_n$

- A sequence of r.v. converges to a limit if for large values of n the sequence and the limit are "close".

- But what does "close" mean when considering random variables?

- Defining 'closeness' in random variables is a bit more complicated than in the deterministic case.

- There are several ways to define "closeness". We will now look at two: convergence in probability and convergence in distribution.

Convergence in probability

- Consider a sequence of random variables $X_1, \ldots, X_n$ or $\{X_i\}_{i=1}^n$.

- X_n converges in probability to X, written X$_n \xrightarrow{p} X$, if for every $\varepsilon > 0$

$$P(|X_n - X| > \varepsilon) \to 0 \text{ as } n \to \infty$$

- Convergence in probability looks at the values of the variables: the probability that the distance between X_n and its limit is "large" tends to zero.

Convergence in distribution

- Consider a sequence of random variables $X_1, \ldots, X_n$ or $\{X_i\}_{i=1}^{n}$.

- X_n converges in distribution to X, written $X_n \xrightarrow{d} X$, or $X_n \Rightarrow X$, if for all $x \in C$, where C is the set of continuity points of the distribution function $F_X(.)$ of $X, then$

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

Convergence in distribution

- Consider a sequence of random variables $X_1, \ldots, X_n$ or $\{X_i\}_{i=1}^n$.

- X_n converges in distribution to X, written X$_n \xrightarrow{d} X$, or $X_n \Rightarrow X$, if for all $x \in C$, where C is the set of continuity points of the distribution function F$_X(.)$ of $X, then$

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

- Convergence in distribution looks at the distribution of the variables, which should be very "close" (in the standard –deterministic– sense, as they are not random) as n gets large.

- Notes:

- The limit X can be a random variable or a constant.

- if X is a constant, we say that the limit has a degenerate distribution (as all the probability mass is concentrated in one point)

- The two modes of convergence are related. If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$.

- The opposite result is not true in general (unless X is a constant).

Limit Theorems

- The Law of Large Numbers and the Central Limit Theorem are the most important results for computing the limits of sequences of random variables.

- There are many versions of LLN and CLT that differ on the assumptions about the dependence of the variables.

- Since we are assuming random sampling (=our data is i.i.d), then we have enough with their simplest versions: LLN and CLT for i.i.d. random variables.

Law of Large Numbers for iid sequences

Let $\{\mathsf{X}_i\}_{i=1}^n$ be an $i.i.d$ sequence of random variables with finite mean μ then

$$\bar{X}_n = n^{-1} \sum_{i=1}^n X_i \xrightarrow{p} \mu.$$

Proof. A very simple proof of this result can be provided if we further assume that var$(X_i) = \sigma^2 < \infty$. Then, by Chebychev's inequality:

$$\begin{aligned} P\left(\left| n^{-1} \sum_{i=1}^n X_i - \mu \right| > \varepsilon)\right) &\leq var(n^{-1} \sum_{i=1}^n X_i)/\varepsilon^2 \\ &= n^{-2} \sum_{i=1}^n var(X_i)/\varepsilon^2 \\ &= \frac{n\sigma^2}{n^2\varepsilon^2} \to 0. \end{aligned}$$

What does this mean?

- The STATA file handout2_LLN.do computes a small Monte Carlo simulation that shows you that this theorem is actually true. Run it so you can start experimenting with random numbers!

- The file does the following:

1. Fix n=100. Generate n random numbers using a χ^2 distribution with one degree of freedom. Notice that $E(X_i) = 1$. Compute $\bar{X}_n$ and store this value.

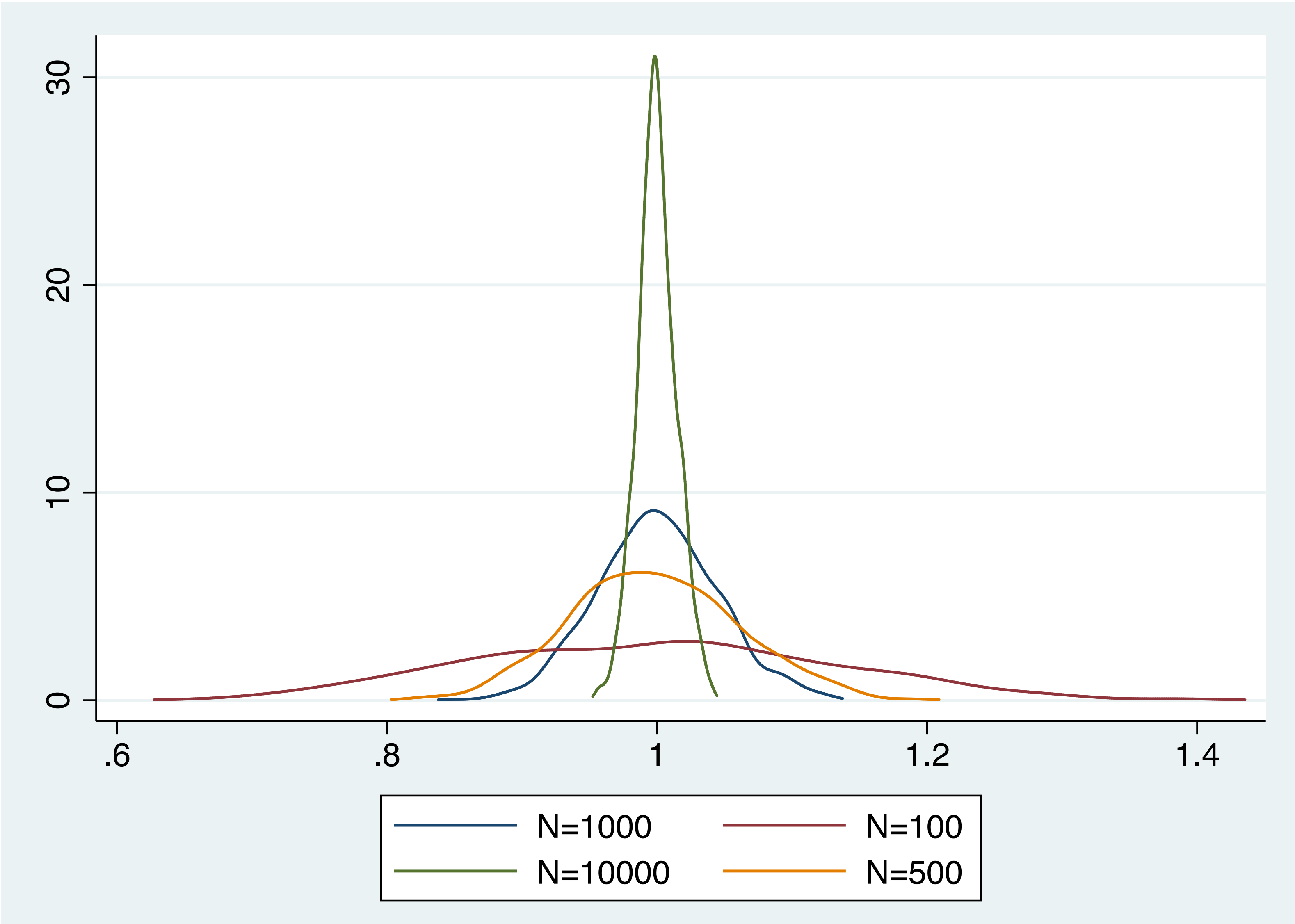
2. Repeat this R=1000 times. This allows us to see the distribution of $\bar{X}_{100}$

3. Repeat 1. and 2. for different values of $n = \{500, 1000, 10000\}$.

4. Plot the obtained distributions corresponding to $\bar{X}_{100}$, $\bar{X}_{500}$, $\bar{X}_{1000}$ and $\bar{X}_{10000}$.

The LLN in practise

This is what you get . . . what do you observe?



Central Limit theorem for $i.i.d.$ sequences

Let $\{X_i\}_{i=1}^n$ be a sequence of $i.i.d(\mu,\, \sigma^2)$ random variables.

Then

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \xrightarrow{d} N(0,1)$$

The CLT in practice

Go to the following link to see an illustration of the CLT

https://demonstrations.wolfram.com/IllustratingTheCentralLimitTheoremWithSumsOfBernoulliRandomV/

- Exercise: Try to modify the STATA file mentioned above so that it illustrates the CLT. To do that you need to plot the distribution of $\sqrt{N}\frac{(\hat{X}_N - E(X))}{\sigma}$; Recall that if X_i follows a χ^2 with 1 degree of freedom then $E(X_i) = 1$ and $Var(X_i) = 2$.

Takeaway

- Asymptotic theory: tools to approximate the distribution of (functions of) random variables.

- Why? because in most cases we won't be able to determine the exact distribution of those variables.

- We would compute limits of a sequence of random variables X_n as gets approaches infinity

- Two modes of convergence: in probability and in distribution

- Two key results: CLT and LLN

3. Estimators and basic properties

Estimators and basic properties

- Remember that our goal is to be establish causal relationships between variables.

- Often, this relationship is captured by a parameter(s) relating those variables.

- Example. Y=wages; X=years of education. Suppose that these variables are related linearly, then

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- β is our parameter of interest.

- If the parameter is identified, we can gather data to obtain an estimate of it and of its standard error (=a measure of the uncertainty of our estimator).

- As we will see shortly, in the example above the identification condition will be $E(\epsilon|X) = 0$

- An estimator is a function of the observable data that is used to estimate an unknown population parameter.

- An estimate is the result from the actual application of the function to a particular dataset,

- In general, many different estimators are possible for any given parameter.

- Estimators are random variables, thus we can (should) compute their associated distribution

- An estimate is a realisation of the corresponding estimator. Thus, it is not random.

- Let n be the size of the sample used in the computation of an estimator of the parameter θ. Thus, for each sample size we can define an estimator: $\bar{\theta}_n$.

- Let $\hat{\theta}_n$ be an estimator of the population parameter θ computed with a sample of size n. Then, $\hat{\theta}_n$ is a function that maps each sample S to its sample estimate $\hat{\theta}_n(S)$.The sequence $\{\hat{\theta}_n\}$ is an example of a sequence of random variables, so the concepts introduced above are applicable to $\{\hat{\theta}_n\}$.

- We will use the LLN and the CLT to approximate the distribution of $\hat{\theta}_n$

Properties of estimators

Some desirable properties of $\hat{\theta}_n$ are the following.

- Consistency: $\hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$.

- Unbiasedness: $\hat{\theta}_n$ is unbiased if $\mathsf{E}\left(\hat{\theta}_n\right) = \theta$ and is asymptotically unbiased if $\lim_{n\to\infty} E\left(\hat{\theta}_n\right) = \theta$.

- Asymptotic Normality. A consistent estimator $\hat{\theta}_n$ is asymptotically normal around the true parameter θ if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, V)$, where V is called the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$.

- Efficiency: An unbiased estimator $\hat{\theta}_n$ is efficient if it has the lowest possible variance among all unbiased estimators.

Takeaway

- We are interested in the value of unknown parameters

- We would use data and econometric techniques to figure out the values of those parameters

- Estimators (random variables) and estimates (the particular value that an estimator gets when a particular dataset is employed).

- Many possible estimators are available, How do we choose among them?

- We want estimators with good properties.

  - consistency, asymptotic normality, efficiency, unbiasedness. . .

4. Hypothesis Testing

Hypothesis Testing

- Consider again our previous example: the relationship between wage and years of education. You have obtained:

$$W\hat{a}ge = 2.3 + 1.2years$$

- Can you conclude from here that if you study more you will get a higher salary?

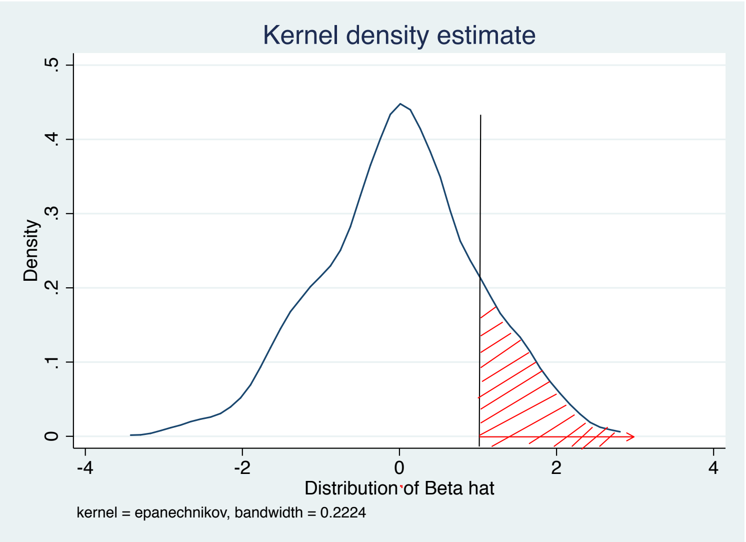
Hypothesis Testing

- Consider again our previous example: the relationship between wage and years of education. You have obtained:

$$\hat{Wage} = 2.3 + 1.2 years$$

- Can you conclude from here that if you study more you will get a higher salary?

- Not really. If $\beta = 0$ you can get positive values of $\hat{\beta}$ with some probability!



Hypothesis testing

- Solution: hypothesis testing

- Simplest context: We are interested in testing hypotheses on an unknown parameter θ

- Null hypothesis: $H_0 : \theta = \theta_0$

- Alternative hypothesis: $H_1 : \theta \neq \theta_0$; $H_1 : \theta > \theta_0$; $H_1 : \theta < \theta_0$

- H_0 and H_1 can be written in many ways but their union should cover the range of all possible values of θ

- In the example above, is θ equal to zero? You can write this as follows:

$$H_0 : \theta = 0 \text{ vs. } H_1 : \theta \neq 0$$

Decision Rule

Main idea: Reject H_0 when $\hat{\theta}$ is “far” from from θ_0

(but what is ’far’?)

Decision Rule

Main idea: Reject H_0 when $\hat{\theta}$ is "far" from from θ_0

(but what is 'far'?)

- Critical Region: values of $\hat{\theta}$ for which H_0 is rejected.

- Decision Rule: Choose a threshold that determines the critical region. Reject if $\hat{\theta}$ is in the critical region.

- The decision rule will something like: if $|\hat{\theta}| > \phi$, for some threshold ϕ then reject H_0

Decision Rule

Main idea: Reject H_0 when $\hat{\theta}$ is "far" from from θ_0

(but what is 'far'?)

- Critical Region: values of $\hat{\theta}$ for which H_0 is rejected.

- Decision Rule: Choose a threshold that determines the critical region. Reject if $\hat{\theta}$ is in the critical region.

- The decision rule will something like: if $|\hat{\theta}| > \phi$, for some threshold ϕ then reject H_0

- The shape of the critical region depends on how H_1 is specified (not determined by H_0!).

- How should we choose ϕ? (i.e., how is the critical region chosen?)

Two types of errors

- Type I error : you reject H_0 but H_0 is true.

- α=P(Rejecting $H_0|H_0$ true)

- Type II error: you don't reject H_0 but H_0 is false.

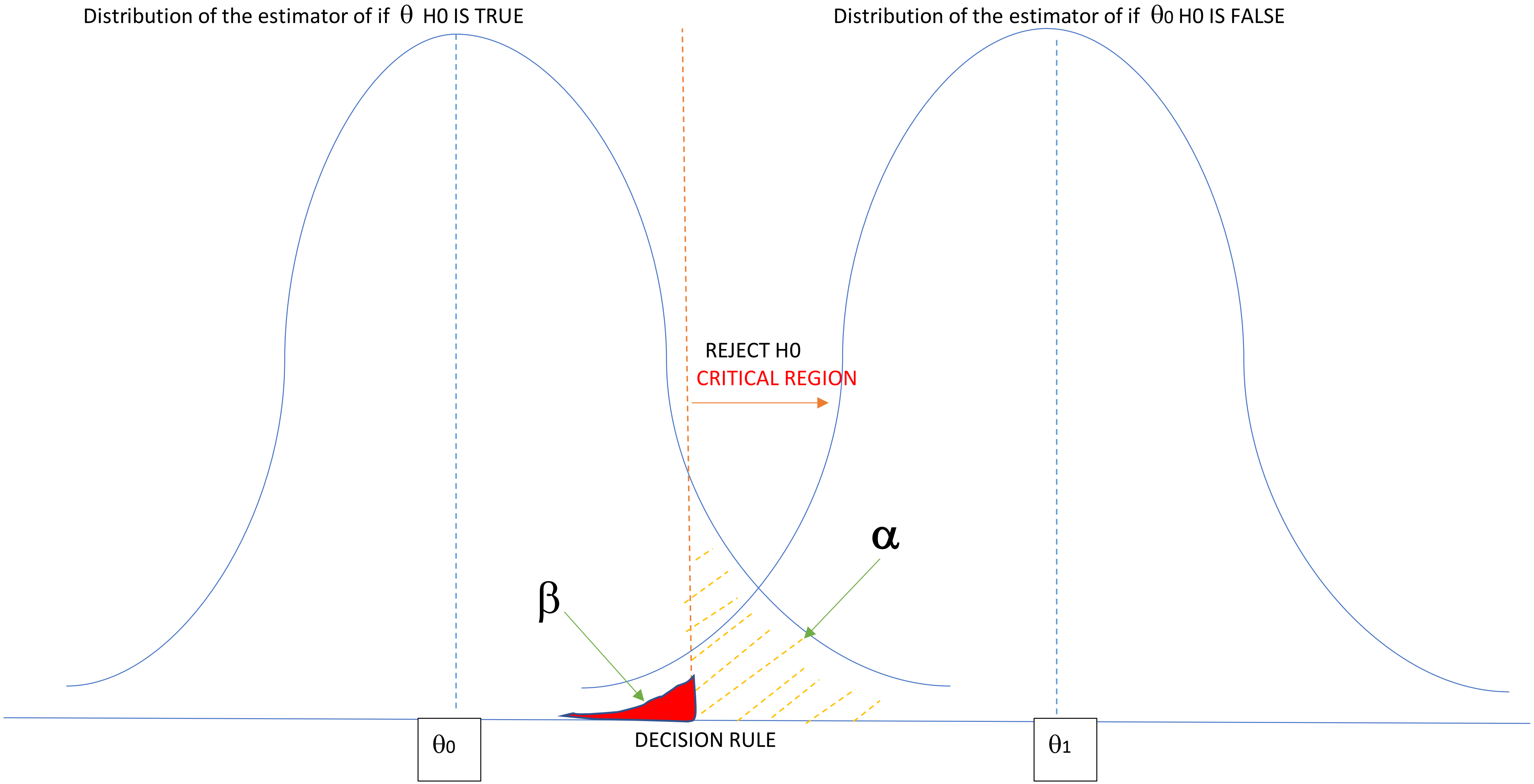
- β=P(Not Rejecting $H_0|H_0$ false)

Ideal Decision rule

- Reject H_0 if $\hat{\theta}$ is "far" from the null.

- Choose the threshold determining what "far" means by minizing both α and β.

- Is this possible?



Distribution of the estimator of if θ H0 IS TRUE
Distribution of the estimator of if θ0 H0 IS FALSE
REJECT H0
CRITICAL REGION
α
β
θ0
DECISION RULE
θ1

- You cannot minimize both α and β simultaneously
- Usual approach:
  - fix α, choose your decision rule (a threshold beyond which you would reject H_0) to minimize β
  - Fixing α: typically 0.05.
- α: size of the test;
- $1-\beta$: power of the test;

An example: t-test

- Context: We have an estimator $\hat{\theta}$ for a parameter θ. You want to test

$$H_0 : \theta = \theta_0$$

versus

$$H_1 : \theta \neq \theta_0$$

- This is a two sided test: we will reject H_0 if $\hat{\theta}$ is either much smaller (left tail) or much larger (right tail) than θ_0

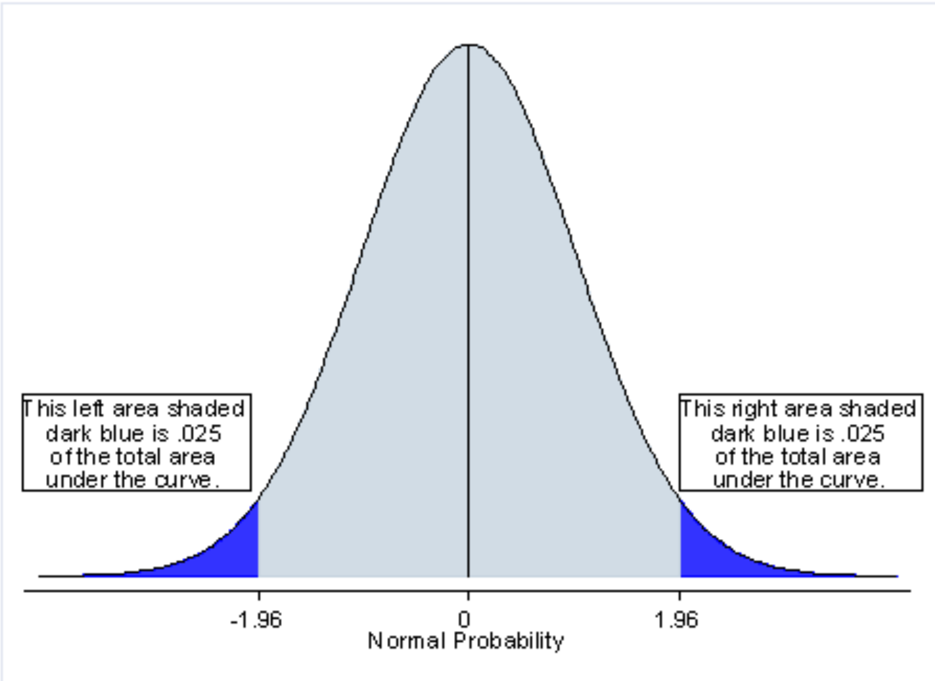
- One sided tests: $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$

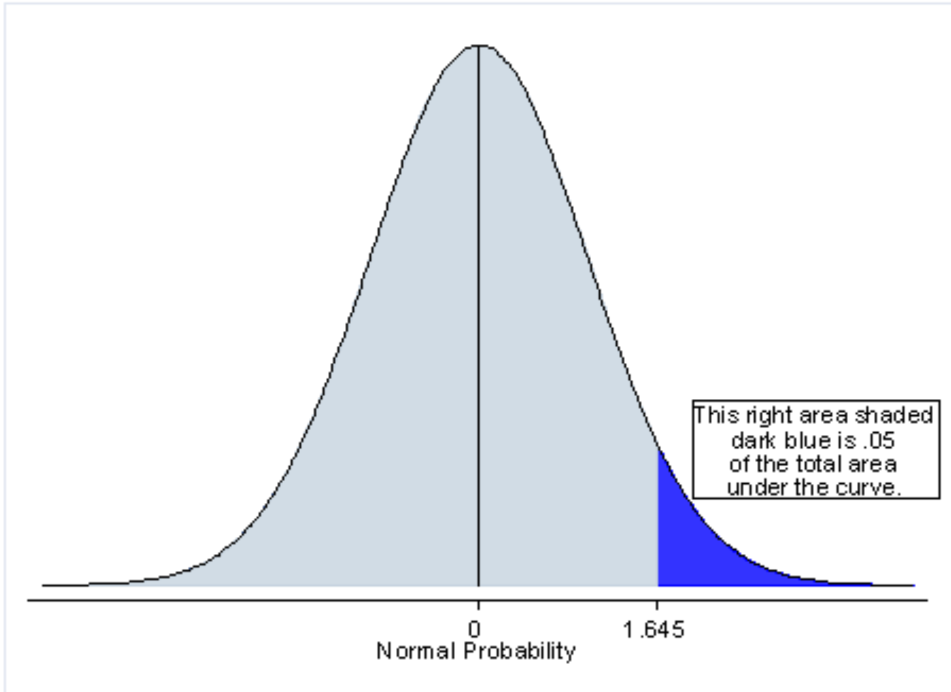
- Assume that $\hat{\theta}$ is asymptotically normally distributed. Under H_0, $\theta = \theta_0$, then

$$\frac{(\hat{\beta} - \beta_0)}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

- But σ is unknown! Solution: replace it by an estimate of the standard error of $\hat{\theta}$. By the continuous mapping theorem

$$\frac{(\hat{\beta} - \beta_0)}{\hat{\sigma}/\sqrt{n}} \xrightarrow{d} N(0,1)$$


This left area shaded
dark blue is .025
of the total area
under the curve.
This right area shaded
dark blue is .025
of the total area
under the curve.
-1.96
0
1.96
Normal Probability


This right area shaded
dark blue is .05
of the total area
under the curve.
0
1.645
Normal Probability

- Two sided test: Reject H_0 if $\hat{\theta}$ is too large ...

$$\frac{(\hat{\beta} - \beta_0)}{\hat{\sigma}/\sqrt{n}} > 1.96$$

- ... or too small

$$\frac{(\hat{\beta} - \beta_0)}{\hat{\sigma}/\sqrt{n}} < -1.96$$

- That is, reject if:

$$|\frac{(\hat{\beta} - \beta_0)}{\hat{\sigma}/\sqrt{n}}| < -1.96$$

Going back to the coin question. . .

- Consider again the problem of the coin formulated a few slides ago.

- Suppose you have toss the coin one hundred times and you have obtained that the share of heads is .53.

- Is the coin balanced?

Solution

- Define the random variable X_i: heads when you toss once the coin.

- Bernoulli random variable, probability p.

- $E(X_i) = p$; $Var(X_i) = p(1-p)$;

- $\bar{X}_n = \sum_{i=1}^{100} X_i$: share of heads in 100 tosses.

- H_0: the coin is balanced (p=.5)

- X_i are i.i.d.: the CLT applies!

- Under H_0

$$\frac{(\bar{X}_n - .5)}{\hat{\sigma}/\sqrt{n}} \xrightarrow{d} N(0,1)$$

- We need an estimate of σ!

- Replace p by $\hat{p}$: $\hat{\sigma} = \sqrt{(}.53 \times .47) = .495$

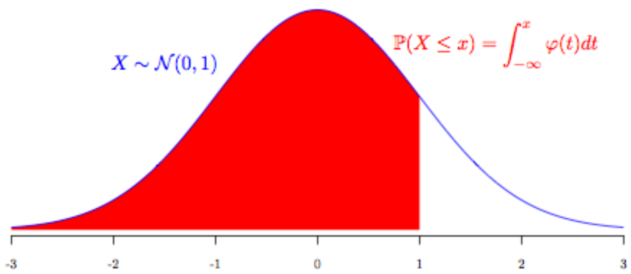
- $\frac{(\bar{X}_n - .5)}{.49/10} = .61$

- Shall we reject H_0 if α=0.05?

- What is the minimum value of α for which we would reject H_0?

- In other words: assuming that the null hypothesis is true, what's the probability of obtaining test results that are more extreme than .61 (in either direction)?

- (The figure you've just computed has a name, do you remember it?)



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Takeaway

- Estimators are random variables. Inferential techniques are employed to get a better idea of the uncertainty involved in our estimation.

- Hypotheses testing (but there other inferential techniques, for instance, confidence intervals).

- Key concepts (you need to remember what they mean!)

  - Null and alternative hypothesis

  - Type I (α) and Type II (β) errors

  - Critical region

- A very important example: t-test (testing 1 parameter). (But how would you test more than 1 restriction?)

Takeaway

- Estimators are random variables. Inferential techniques are employed to get a better idea of the uncertainty involved in our estimation.

- Hypotheses testing (but there other inferential techniques, for instance, confidence intervals).

- Key concepts (you need to remember what they mean!)

  - Null and alternative hypothesis

  - Type I (α) and Type II (β) errors

  - Critical region

- A very important example: t-test (testing 1 parameter). (But how would you test more than 1 restriction?)

  - F-tests

References:

Greene, Appendix B, C, D

Wooldridge, Chapter 2, Appendix 2A, Chapter 3.