



Econometrics I

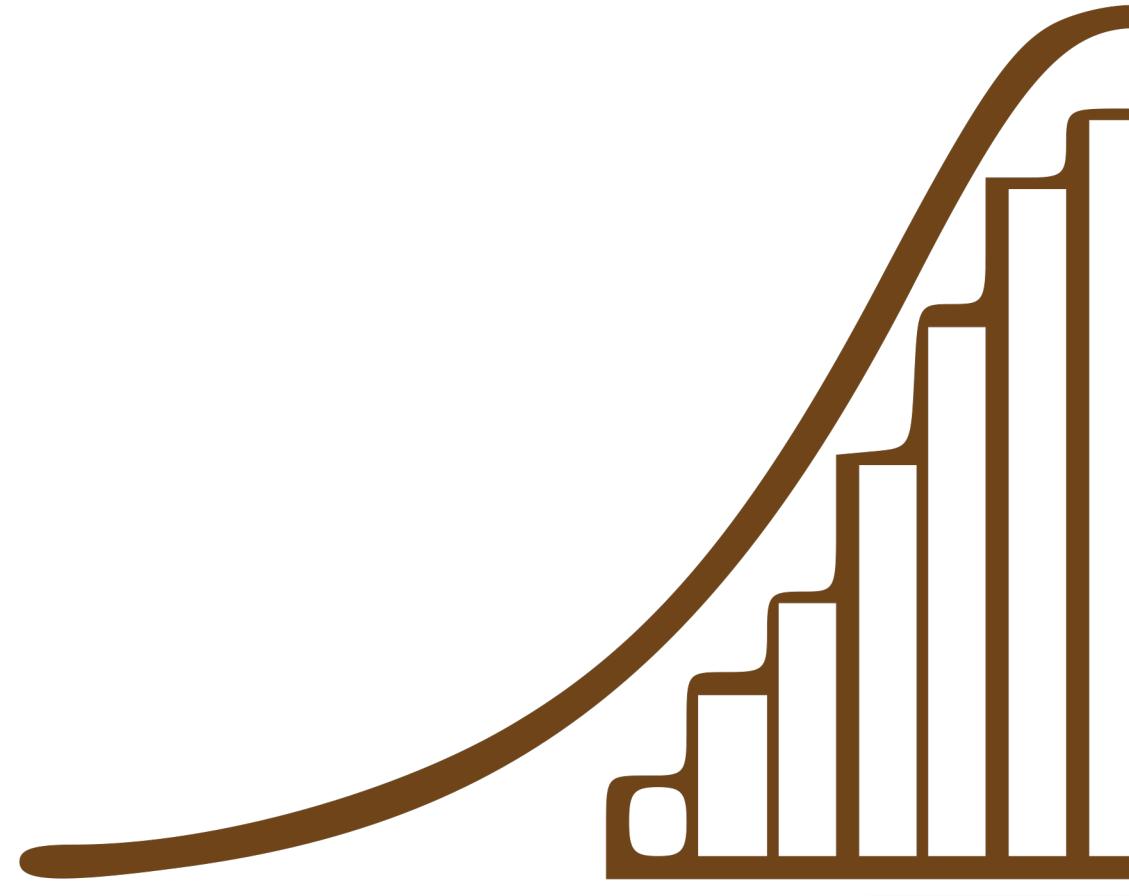
Workshop IV

Feb 21, 2023

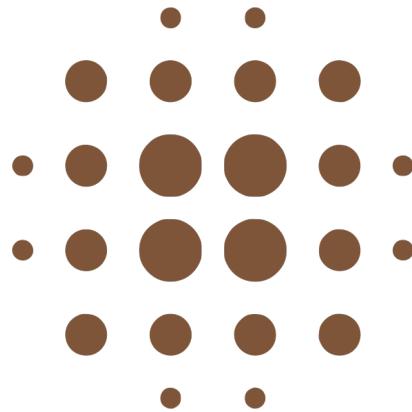
Normality



One of the objectives in statistics relies on summaries and description from numerical information through metrics that quantifies several aspects such as the distribution from a dataset.

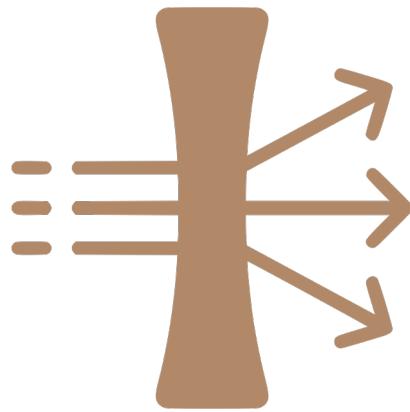


Central tendency measures.



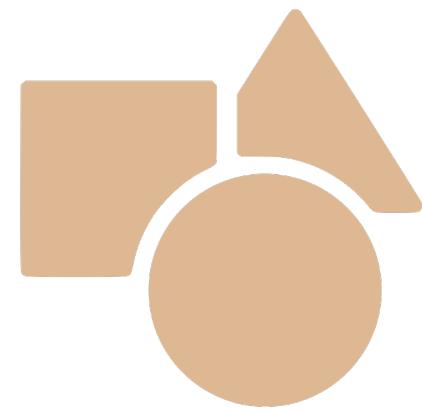
Single value that attempts to **describe** a whole dataset through the center of its distribution

Measures of dispersion



Spread of a dataset or **variability** of data
(is my data homogeneous or heterogeneous?)

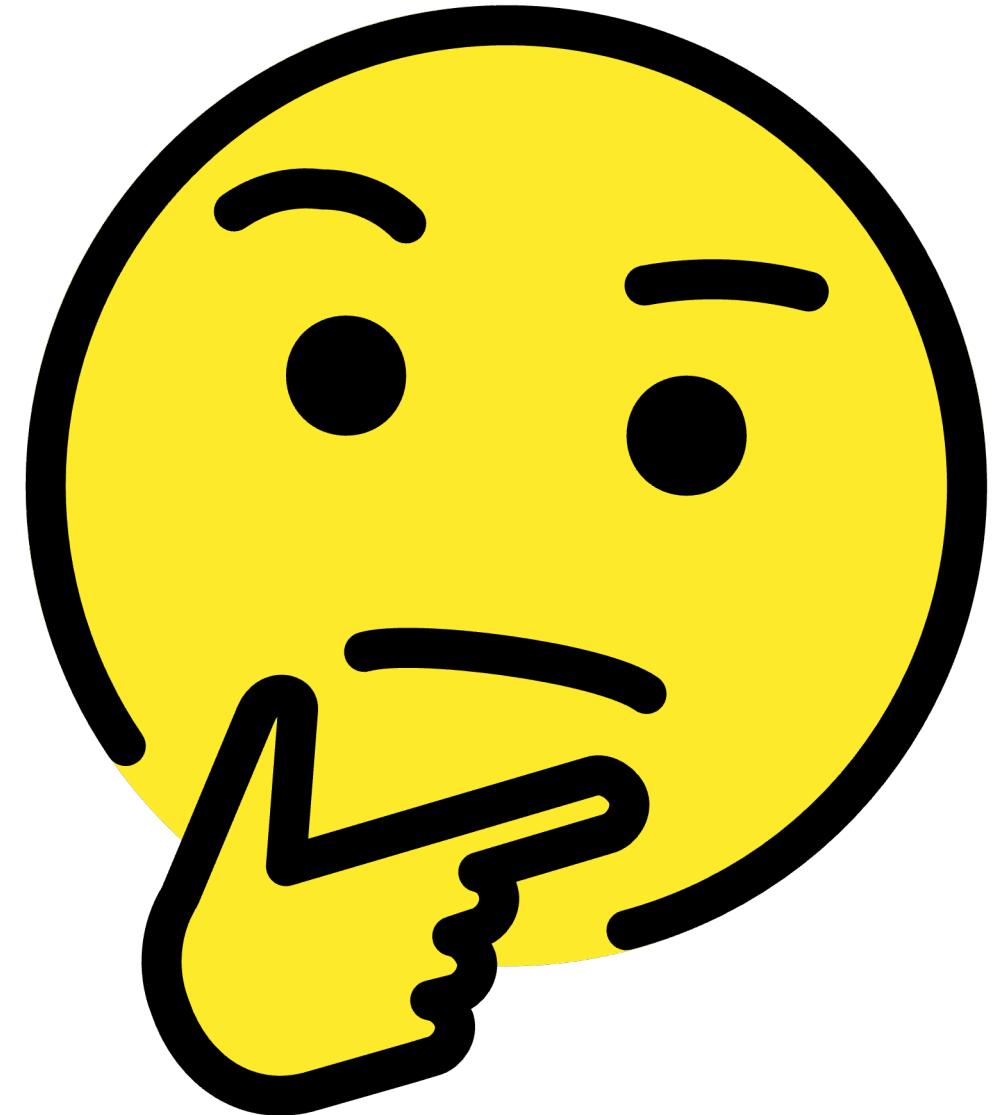
Measures of shape



Indicate how data is **grouped** according to their frequency

Okay but...How
can we measure
all this nonsense
stuff?

Hint: we use statistical
methods such as...



Position

They divide an ordered dataset in groups with the same quantity of individuals.

Quantiles, percentiles, quartiles, decil.

Centralization

Each of these measurements give us a reference value for establishing how a dataset is centered.

Mean, median, mode

Dispersion

They indicate the degree of concentration of data with respect to measures of central tendency. They are used to quantify the variability of a dataset.

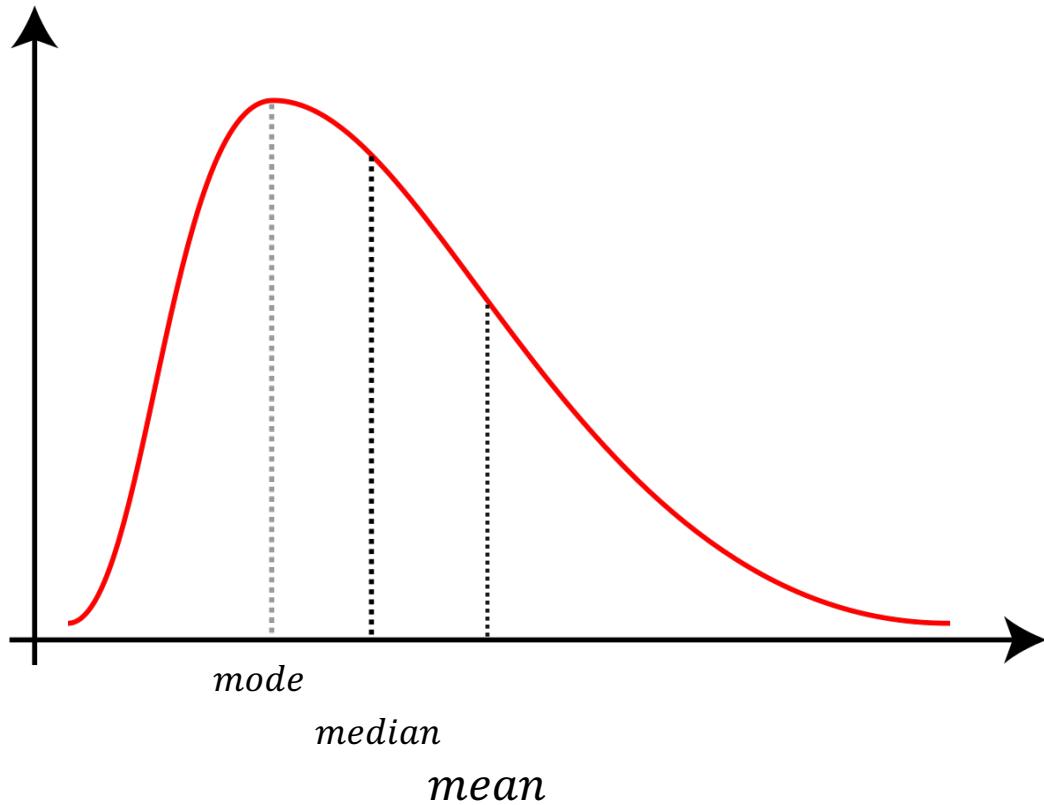
Standard Deviation, variance, coefficient of variation

Shape

They focus on the distribution shape according to their symmetry

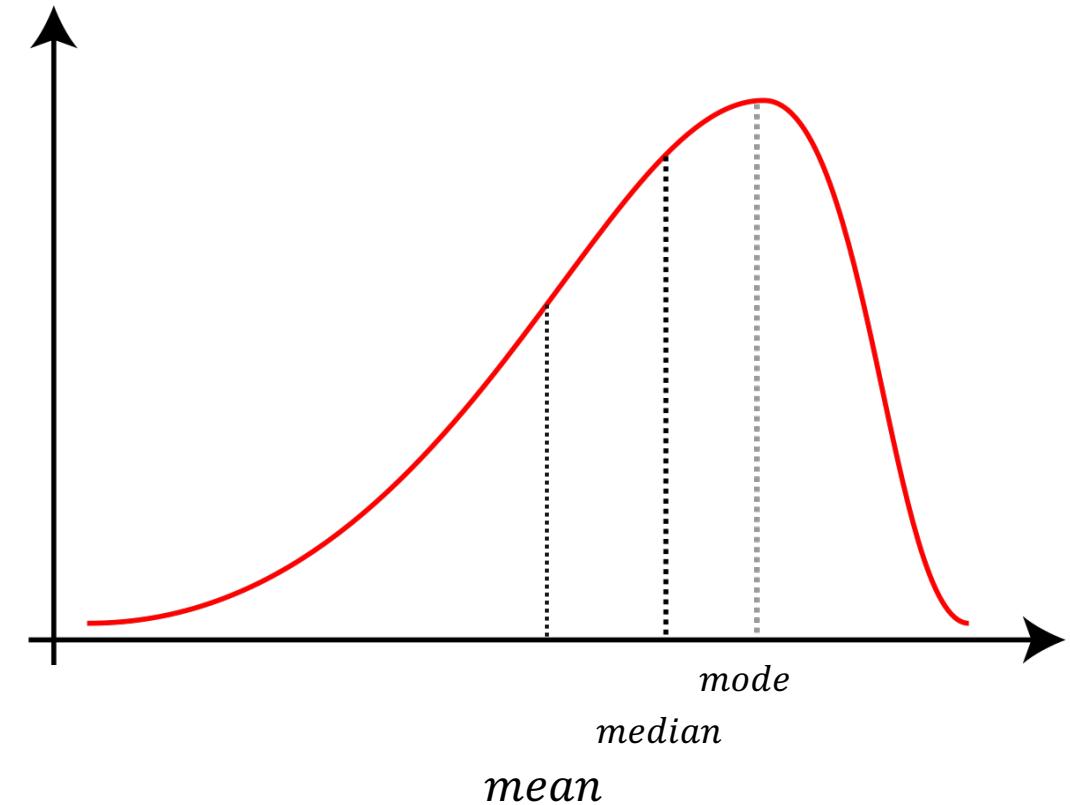
Assymmetry, kurtosis

Mean, median and mode



mode < median < mean

Positively skewed

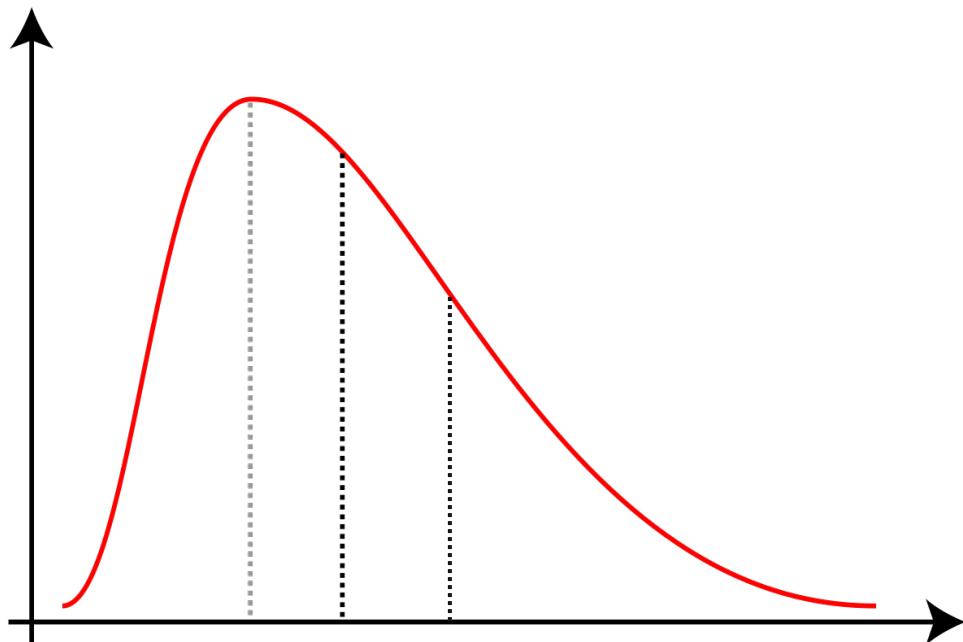


mean < median < mode

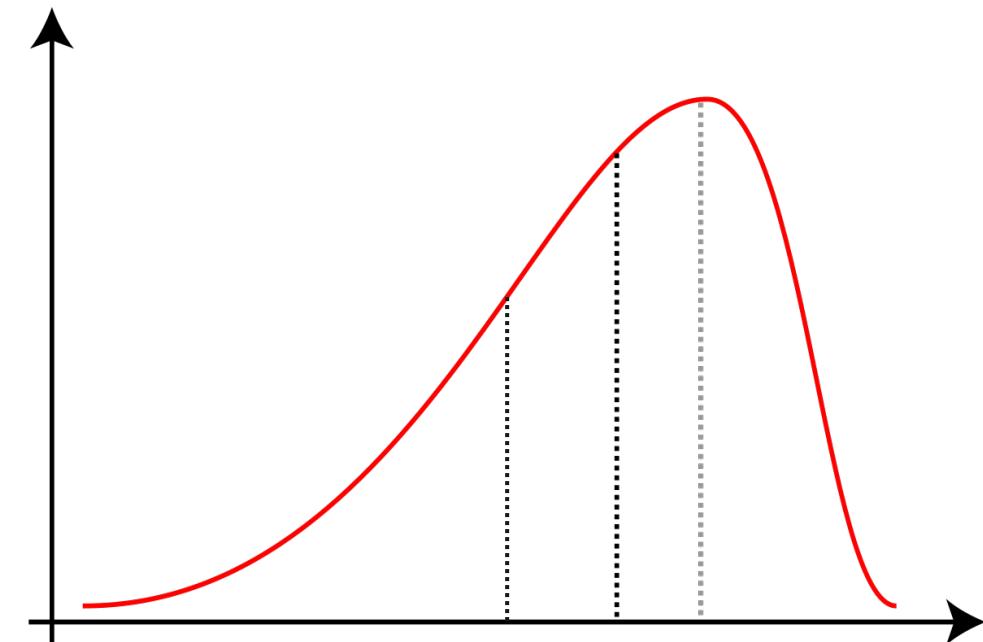
Negatively skewed

Mean, median and mode

Imagine house prices ranging from \$10k to \$1,000,000 with average being \$500,000

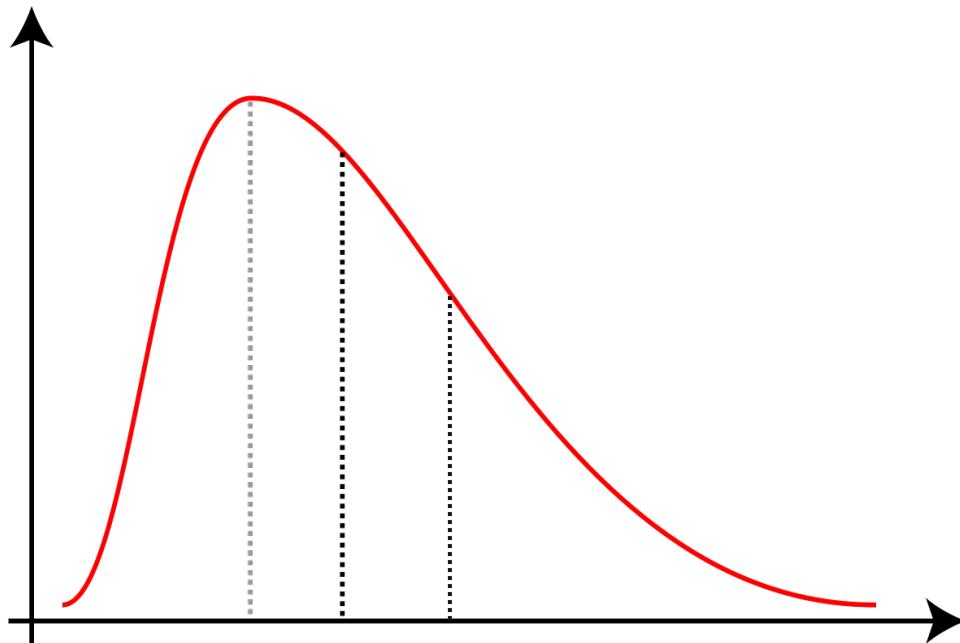


Most of our houses are being sold at a price **less than** the average

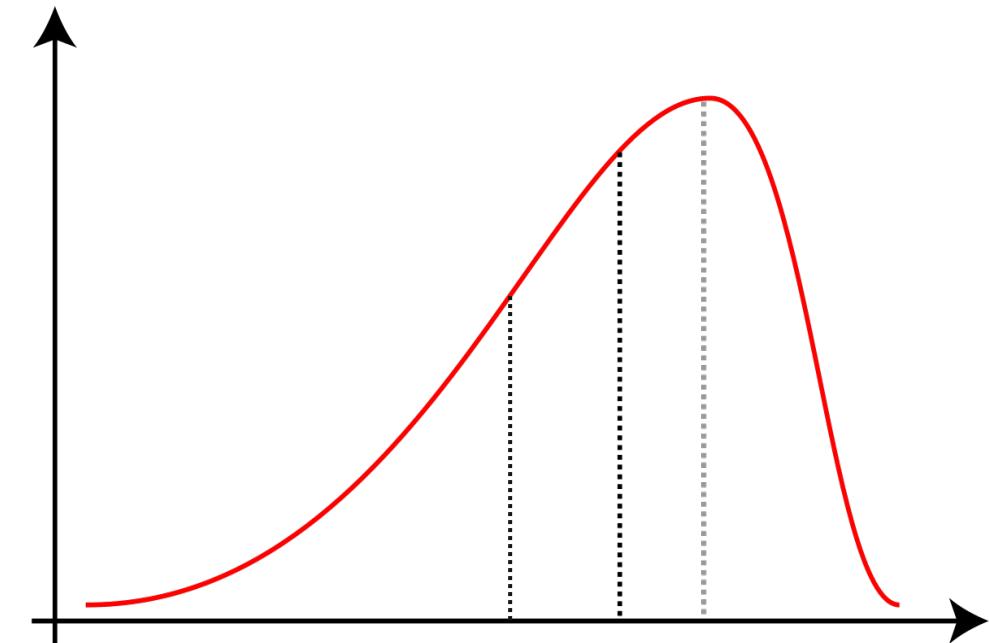


Most of our houses are being sold at a price **greater than** the average

In finance, if we want to know how the **returns** in our portfolio behaves, we state that...



Many **small losses** and
few **extreme gains**



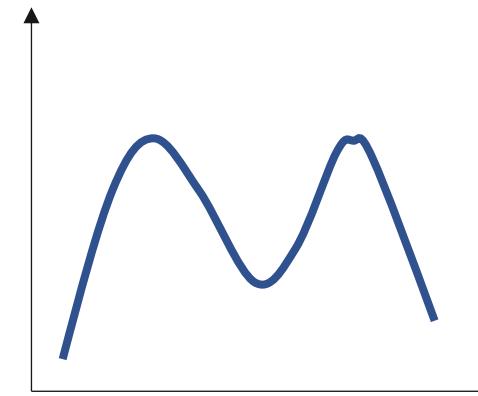
Many **small gains** and
few **extreme losses**

Distribution **function**

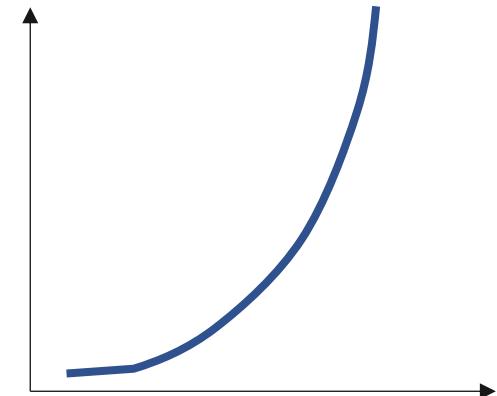
It describes the **probabilistic shape** (or behavior) of a random variable X associated with a **random experiment** which can be

$$f(x) \text{ or } Fx$$

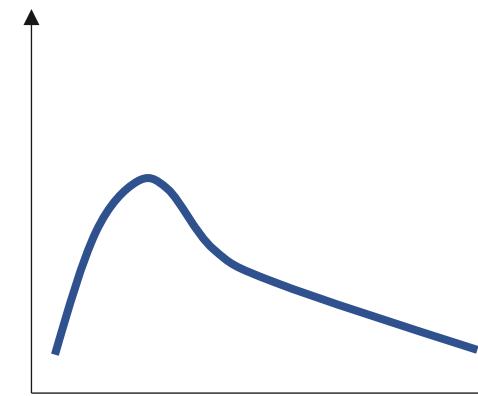
Thus, we can say that a distribution function allows to study the **behavior** of our data.



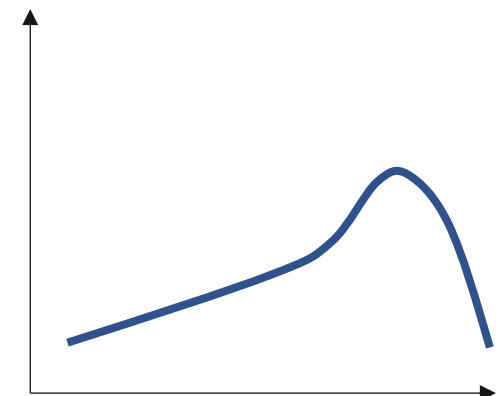
Bimodal distribution



J distribution

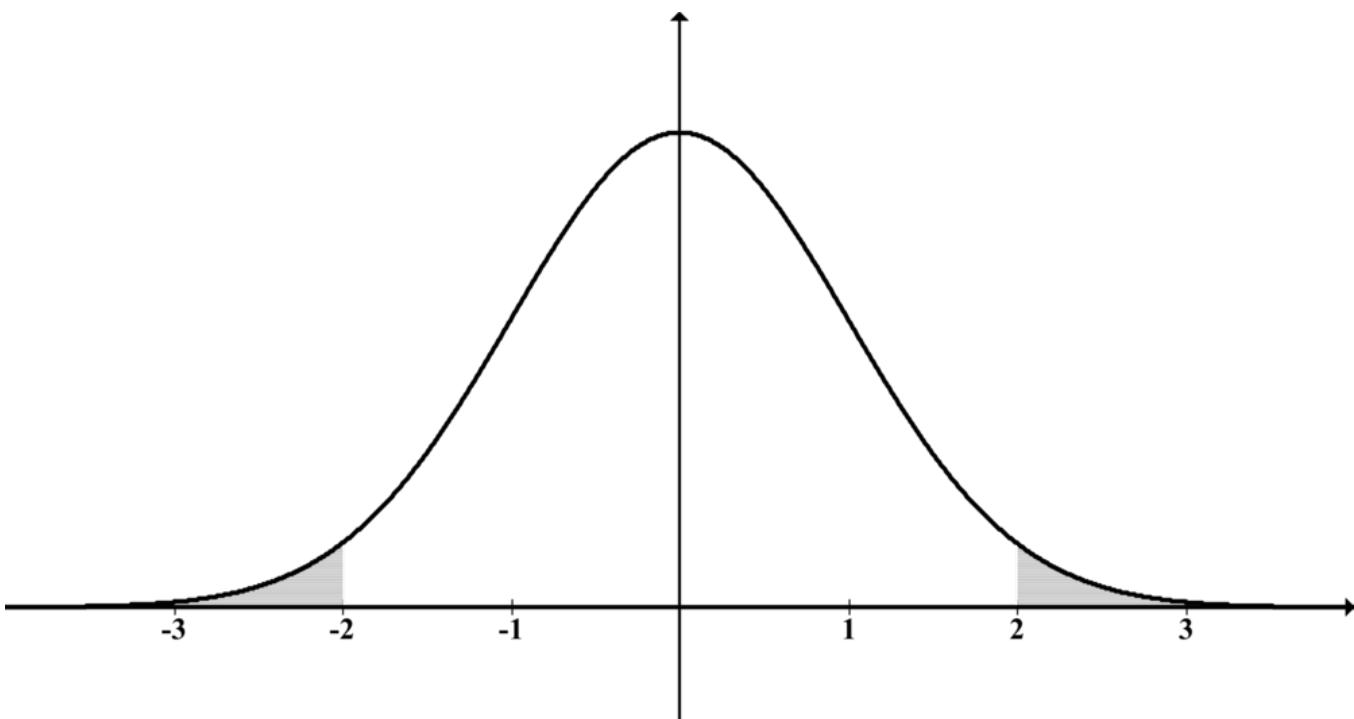


Positively skewed distribution



Negatively skewed distribution

Distribution **function**



Undoubtedly, the normal distribution is the **most well-known probability function**. It was discovered as an approximation of the a **binomial dist.** by Abraham de Moivre (1667-1754)

Two parameters: mean and standard deviation.

When normal distribution has *mean* = 0 and *variance* = 1 it is called **standard normal distribution**.

CLT is a **statistical theorem** which states that, given a large enough sample of a population, the distribution of sample means will lead to a normal distribution.

Furthermore CLT establishes **that the larger the sample**, the more the sample mean **approaches the population mean**.

Due to CTL, it has been proven that **if** there are many independent random variables with identical distributions the distribution of its sum tends to be normal.



CLT is the ground for the assumption of normality

Normality

Classic linear regression model assumes that every U_j is normally distributed with

$$\text{Mean: } E(u_i) = 0$$

$$\text{Variance: } E[u_i - E(u_i)]^2 = E(u_i^2) = \sigma^2$$

$$\text{Cov: } E\{[u_i - E(u_i)][u_j - E(u_j)]\} = E(u_i u_j) = 0; \quad i \neq j$$

They are expressed in a compact form as:

$$u_i \sim N(0, \sigma^2)$$

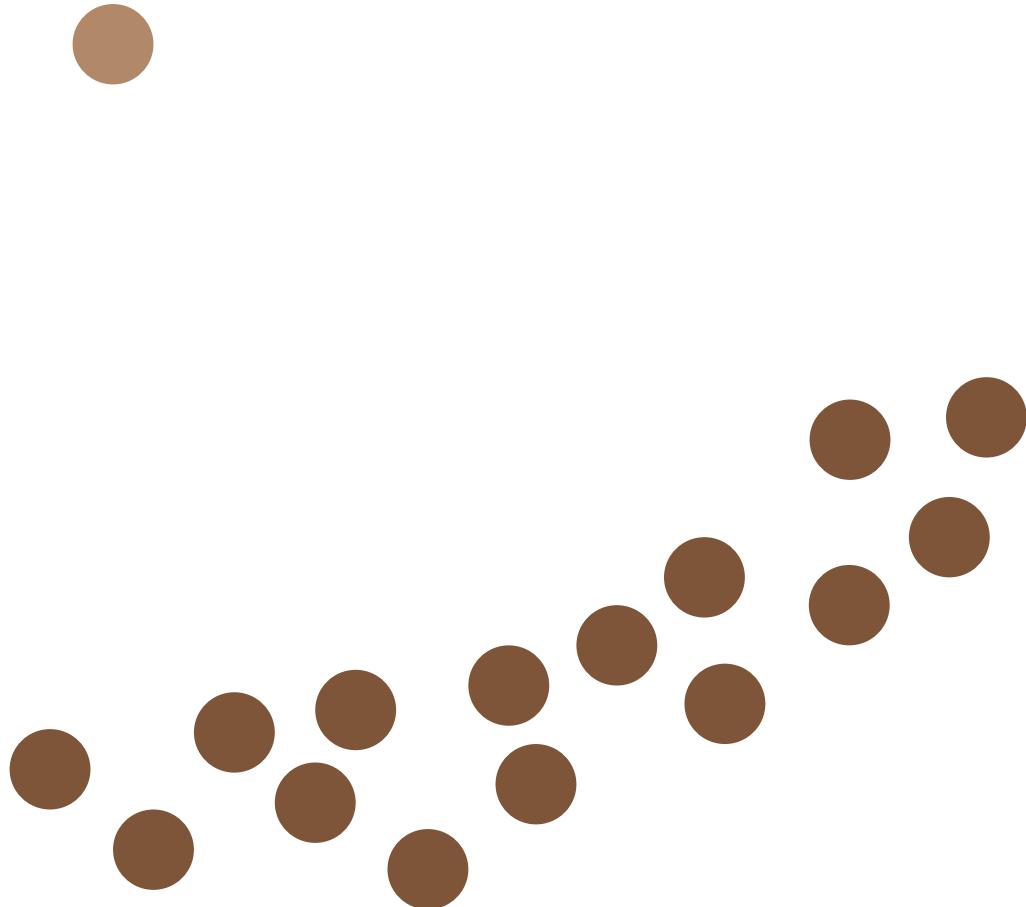
Why is it important the assumption of normality?

Normality in residuals ensure that estimator from ordinary least squares is consistent and efficient.

Tests such as **T** and **F** are calculated from the assumption of normal distribution.



Practice is much harder than an academic stone engraving sentence.



Why is it important the assumption of normality?

If there is a highly abnormal data point, caused for a special situation that is out of model's reach, it can lead to a modification of residuals' distribution

A data point is considered as an outlier, if the value of a data point for a variable differs substantially from the pattern from the rest of variables

What can we do
if we find an
outlier?

We use **elemapi2.dta** which contains information about basic academic performance in USA

Let's prove that (**api00**) depends on free meals percentage given to students (**meals**), students currently learning English (**ell**) and percentage professors with recent accreditations (**emer**)



```
// Database: elemapi2.dta
// Run regression
regress api00 meals ell emer

// Use predict to generate residuals from regression
predict r, resid
```

Kernel density is a **non-parametric** estimation method for a random variable density

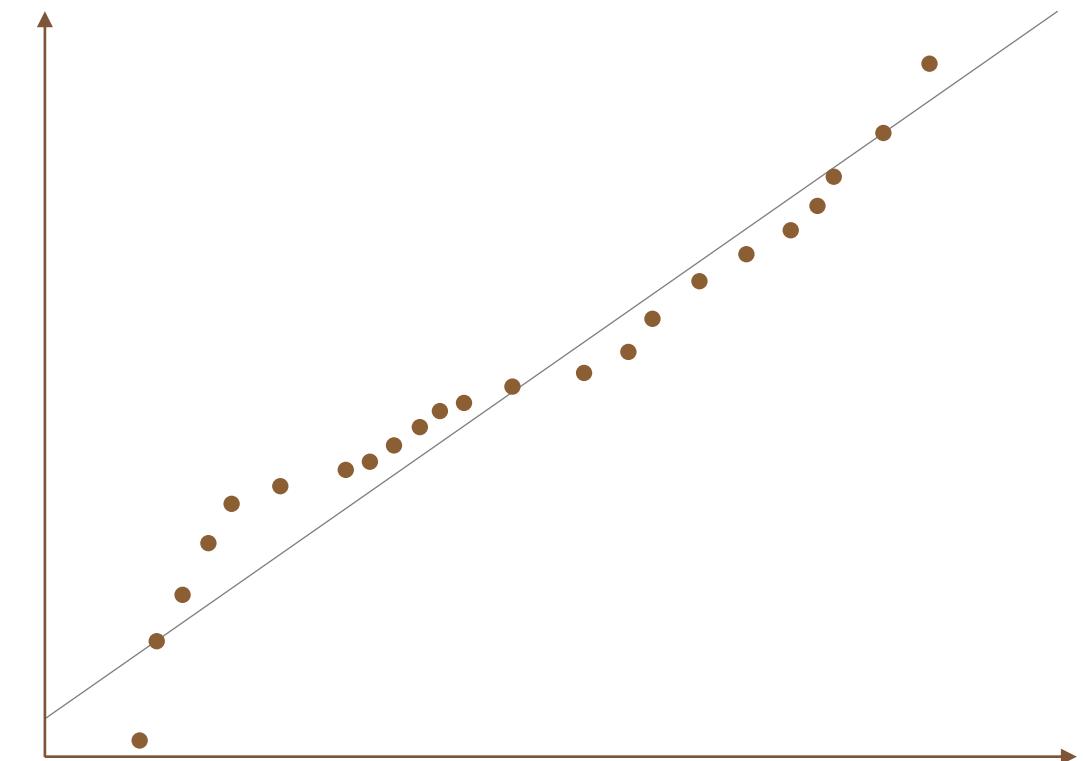


```
// Obtain kdensity with normal distribution  
kdensity r, normal
```

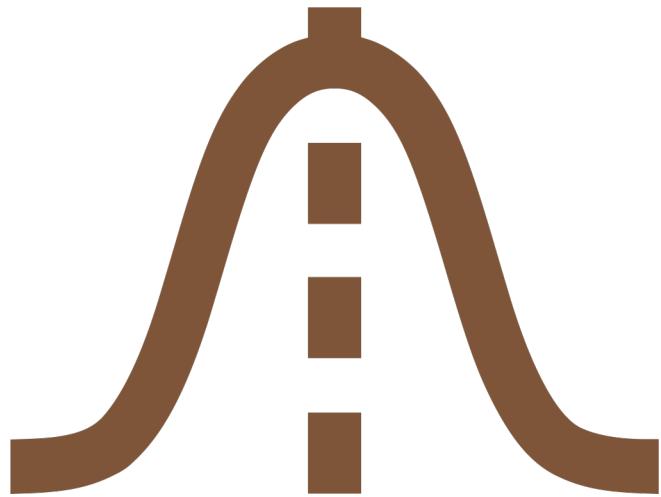
The Standardized Normal Probability Diagram is a graph technique to prove normality.

It indicates if a dataset has an approximated normal bell shape

How to read it?: if distribution is normal, then points will be close to the straight line



On the other hand, the **Normal Probability Graph** is used to answer questions such as:



Is data normally distributed?



What is the nature of standard deviation of normality (long tails, shorter tails than expected)?



```
// Show a standardized normal probability diagram  
pnorm r
```

```
// Test means of residuals from the regression  
sktest r
```

Null hypothesis indicates sample skewed and sample kurtosis distributes as a normal shape

Non-parametric test do not assume the shape of data

They are also known as distribution free.

Kolmogorov-Smirnov is one example. It is used to determine goodness-of-fit from two probability distributions

This test is sensitive to value that are closed to the median rather than extreme tail values



In varname we put data that will be proven and on the right side we evaluate with an accumulated normal distribution



```
// Equation (do not code following code line)
ksmirnov varname = normal( (varname-(mean)) / varname(sd) )

// Application
sum r
ksmirnov r = normal( (r-7.37e-08) / 57.60224)
```

Null hypothesis states that distributions are equal. In our example:

- 0.426
- 0.797
- 0.787

Then, we **cannot reject null hypothesis**

Practice (**Normality**)

On the other hand, Shapiro-Francia test shows the squared correlation among ordered values in a sample and ordered approximated quartiles that are expected in a normal distribution.

Null hypothesis states that sample approximates to a normal distribution.

What can we conclude from this test?



```
// Obtain Shapiro-Francia test  
sfrancia r
```

References

- **Salvatore, D., & Sarmiento, J. C.** (1983). *Econometría* (No. HB141 S39). McGraw-Hill.
- **Gujarati, D. N.** (2009). Basic econometrics. Tata McGraw-Hill Education.
- **Wooldridge, J.M.** (2016). *Introductory Econometrics*, Cengage Learning, 6th edition.
- **CFA Institute** (2020), “Level I, Volume 1, 2020, Ethical and Professional Standards and Quantitative Methods; Reading 7: Statistical Concepts and Market Returns”, pp. 422-430