Data Analysis 2: Fundaments of Statistics

Instructor: Arieda Muço, Fall 2017

Ways to contact me:

- mucoa@ceu.edu
- Office: 507 Nador 13
- Phone +36 1 327-3000x2865
- Office hours: Thursdays 17.00-18.00

Teaching Assistants:

- Balasz Kertész (R)
  - Kertesz_Balazs@phd.ceu.edu
- Oliver Kiss (Stata)
  - Kiss_Oliver@phd.ceu.edu

Teaching Assistants (Help):

- Luca Keresztesi
  - luca.keresztesi@gmail.com
- Daniel Baksa
  - baksa_daniel@phd.ceu.edu

What Is Data Analysis 2 ?

- Introductory course to statistics
  - Hands-on Exercizes
    - R, Stata
  - Understand the structure of data
    - How it is born, how it is stored, how it can be accessed
    - Common issues with real-life data
  - Produce, visualize and understand essentials statistics
    - Distributions of variables,
    - Statistics describing aspects of the distributions
  - Understand some fundamental statistical concepts
    - Expected value, standard deviation, etc.

Data Analysis 2 will help

- Support better decisions in
  - Business
  - Public Policy
  - And many other fields
- Formulate hypothesis and test them
- Importance of random sampling
- Understand how biases or dealing carelessly with data can misrepresent findings
- Small versus large sample size
  - Touch upon the big data revolution

Data Can Be More Useful Than Other Evidence

What stats allows you to do is not take things at face value. The idea that I trust my eyes more than the stats, I don’t buy that because I have seen magicians pull rabbits out of hats and I just know that rabbit’s not in there

Billy Beane, interview

Statistics Meaning

- The word statistics is derived from the Italian word "stato" which means "state" and refers to a person involved in the affairs of state. Therefore, statistics originally meant the collection of facts useful to the "statista".
- Statistics in this sense was used in 16-th century Italy, then spread to France, Holland, and Germany.
- Surveys of people and property actually began in ancient times.
- Oftentimes, the data are summarized, displayed in meaningful ways and analyzed.

Statistics and Information

- Today, statistics is not restricted to information about the state but exceed the realm of human endeavor.
- Neither do we restrict ourselves to merely collecting numerical information called data.
- Statistics is the science of generalizing from the data.

Information about the course

- 6 lectures
  - 150 mins each
  - in classroom
- 3 practice sessions
  - 100 mins each
  - in computer labs
- Exam during the last lecture
  - Friday, October 27

Grading

- Class participation
  - Quizzes in all coming lectures
  - Will be graded and count for the final grade
  - 3 group assignments
- Exam
  - Written examination
- Grade
  - First, you pass the written examination which counts for 60% of the course
  - 15 % will be from the quizzes
  - 25 % will be from the group assignments

Background material

- No main textbook
- Extensive support material (Handouts and codes)
  - Slides

Software

- R
  - Used by statisticians, data scientists, and economists
  - Free
  - Steep learning curve
- Stata
  - Used mainly by economists
  - Need to buy license (installed in CEU labs)
  - Easier to learn
- We will provide codes for all applications we cover
  - R and Stata as well (when possible)

Ceu Moodle

- Separate site for
  - Data Analysis 1
  - Data Analysis 2
  - Data Analysis 3
  - Etc.
- Handouts, lecture slide shows, other supporting material
- Homework assignments
  - Use the site to turn in your homework
- Register to the course
- Check the site regularly
- Read your CEU email regularly

Take away

- Data Analysis at CEU Econ
  - is an integrated sequence of many 2 credit courses
- By completing them you can
  - do meaningful data analysis on your own
  - understand other people’s data analysis
    - Unless it’s very complicated

Take away

- DA2 is a stepping stone
  - To the fundamental statistical concepts used throughout
  - You will complement your software skills with DA1
- Passing DA2 should not be hard
  - If you read the material provided
  - Do the problem sets

Classes are mandatory

- You can skip 25% of the classes
- I will know from your quizzes who participated in class
- If you don't attend lose 15 points from quizzes
- The quizz is going to be at a random time during the class
    - Begining
    - Middle
    - End

Plan for the rest of the lecture

- Measures of centrality
- Measures of dispersion
- Probability (Bayes theorem)
- Distributions

IMPORTANT!!

- Stata track with be held on computer labs
- R- track please bring your own computers
    - Install R, I recommend R-studio

Time to dive into some statistics

Summary statistics

- A statistic is a single measure of some attribute of a sample (e.g., its arithmetic mean value). It is calculated by applying a function (statistical algorithm) to the values of the items of the sample, which are known together as a set of data.
- Basic summary statistics are the most widely used statistics to describe certain aspects of data.

The mean (Example)

- Assume that you are arranging a trip to Vienna for New Years, and you are looking at hotel prices on a major hotel website.
- Suppose also that your only constraint: You are interested in staying strictly less than two kilometres away from the city center.

Table 1: Hotel Prices less than 2 km from the city centre

Hotel Name	Rating	Stars	Price (huf)	Dist center km
Hotel Lamee	4.5	4	148127	.2
Hotel Topazz	4.3	4	148127	.2
CH- Wellness Apartments	3.7	3	78702	.6
Hilton Vienna am Stadtpark	4.3	4	155528	.8
Derag Livinghotel An der Oper	4.6	4	131825	.8
Le Meridien Wien	4.4	5	285718	.8
Palais Hansen Kempinski Vienna	4.8	5	168646	.9
Hilton Vienna Plaza	4.6	4	155528	.9
Das Capri - Ihr Wiener Hotel	4.5	3	133526	1.2
Citadella Residence Appartments Vienna	5	4	90883	1.2
Royal Resort Apartments Urania	4.3	3.5	117116	1.3
Ruby Sofie Hotel Vienna	4.3	3.5	54342	1.4
Royal Resort Apartments Blattgasse	3.4	3.5	59339	1.4
NH Wien City	4	4	80888	1.7
Magdas Hotel	4	2	35697	1.8

The Mean

- Is computed as the average of the values in the data at hand

$$\overline{x} = \frac{\sum x_i}{n}$$

- The mean hotel price for New Years Eve, the mean height of the class

$\overline{x}$ Sample average

$E[x]$ Expected value

The Mean (Properties)

- It changes in the same way if we transform the variable in a linear fashion.

$$\frac{\sum(x_i+a)}{n} = \overline{x+a} = \overline{x} + a$$

- If we multiply a variable with a number, say b, its mean value gets multiplied by the same number b

$$\frac{\sum(x_i b)}{n} = \overline{x \cdot b} = \overline{x} \cdot b$$

The Median

- The median is the middle value of the distribution in the sense that exactly half of the observations have lower value and the other half have higher value
- To compute the median we sort the values of our variable from the lowest to the highest
- When we have even numbers in our data, the median is the mean of the two middle values
- The median is usually less subjective to extreme values we have in our data

The Mode

- The mode is the most frequent value in your dataset

- Sometimes we might have more than a mode in our dataset.

Percentiles

- The P-th percentile of your data is the value below with lie P% of the numbers in the data. The position of the P-th percentile is given by:

$$\frac{(n+1)P}{100}$$

- where n-is the number of observations in your dataset

- Find the 50 and the 80 percentile of the following data point

    - 1, 1, 1, 1, 1, 2, 3, 3, 7

Percentiles (Answer)

- The 80 percentile is given by

$$\frac{(9+1)80}{100} = 3$$

- which is the number in the 8-th position of your data
- Find the 50 and the 80 percentile of the following data point
    - 1, 1, 1, 1,1, 2, 3, 3, 7

Quartiles

- Certain percentiles have greater importance than others as they break the distribution of data into 4 groups

- Quartiles are the percentage points that break down the dataset into quarters; first quarter, second quarter, third quarter, and fourth quarter

- The first quartile is the 25% percentile. Is that point below which lie 1/4 of your data
  - 1, 1, 1, 1, 1, 2, 3, 3, 7

Quartiles

- The median is the 50-th percentile and the second quartile

- The third quartile is called the 75th percentile. Is the point below which lie 75% of the data.

- We often call 25th percentile as the lower quartile and the 75th percentile as the upper quartile

Interquartile Range

- The interquartile range is the difference between the third and the first quartile.
- In the previous example the interquartile range is the difference between 3-1=2

Range

- The range is the difference between the largest and the smallest observation in your data point

- This is an extremely rare event as this are fake data. Most of the times range and interquartile range do not coincide.

Variance

- The variance of the observations is the average squared deviation of the data points from their mean.

$$Var(x) = \frac{\sum (x_i - \overline{x})^2}{n}$$

Standard Deviation

- The standard deviation is the (positive) squared root of the variance of the data points.

$$Std(x) = \sqrt{\frac{\sum(x_i - \overline{x})^2}{n}}$$

Variance and Standard Deviation: Properties

- When we add a number to a variable its variance and standard deviation remain the same. (Can you say why?)

- When we multiply a variable with a number the variance is multiplied by the square of the number, and the standard deviation is multiplied by the absolute value of that number. (Can you prove this?)

Histograms

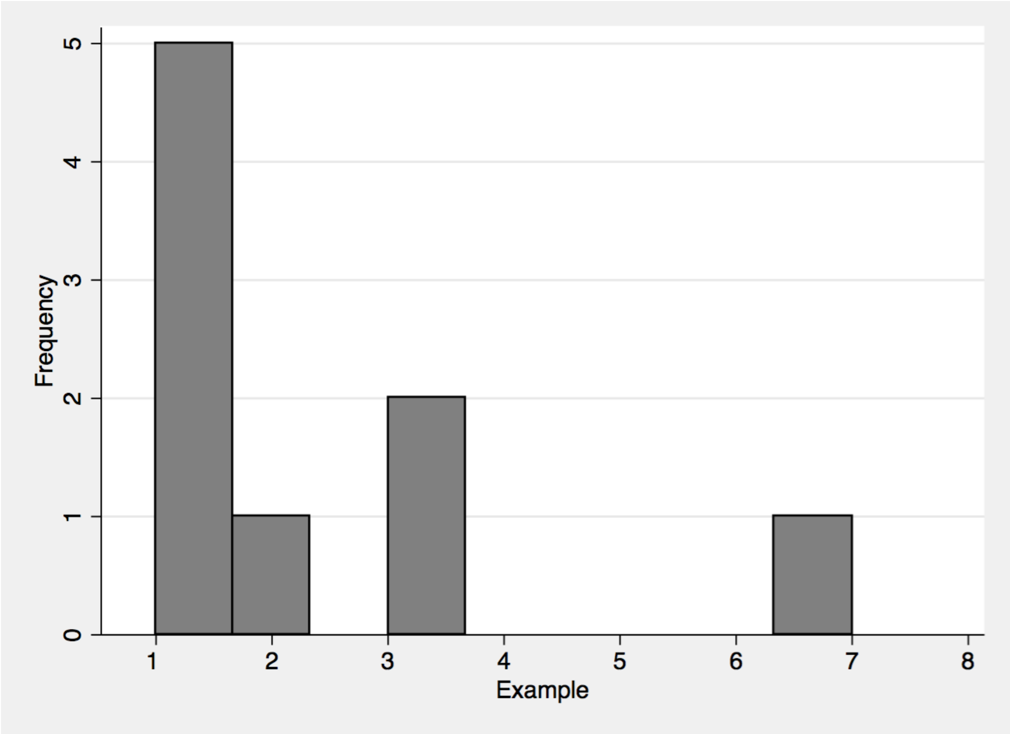
- An histogram is a plot made of bars of different heights.
- The height of each bar represents the frequency of values in each bar.
- Adjacent bars share sides.

Histograms

- In our "silly" example from earlier on, how would the histogram look like?

Histograms

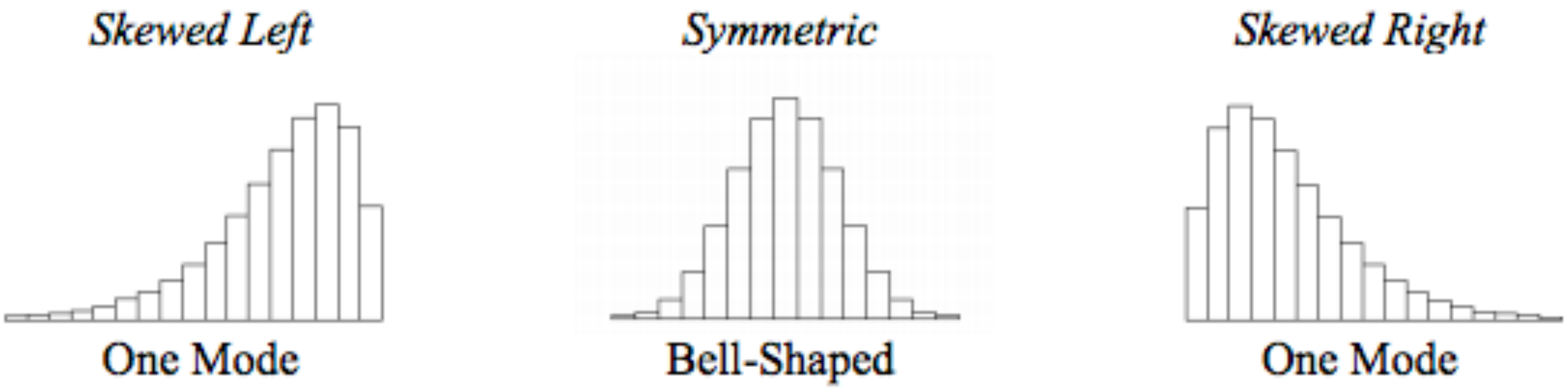
- In our "silly" example from earlier on, how would the histogram look like?



Skewness

- Your data might be skewed in two ways; having a long left tail or having a long right tail.
    - In symmetric distributions mean = median (=mode)
- When it is skewed with a long right tail the mean is larger than the median.
- When it is skewed with a long left tail the mean is smaller than the median.

Skewness


Skewed Left
One Mode
Symmetric
Bell-Shaped
Skewed Right
One Mode

The mean-median measure of skewness

- The mean -median measure of skewness captures this intuition, and it standardizes the mean-median difference by dividing it with the standard deviation.

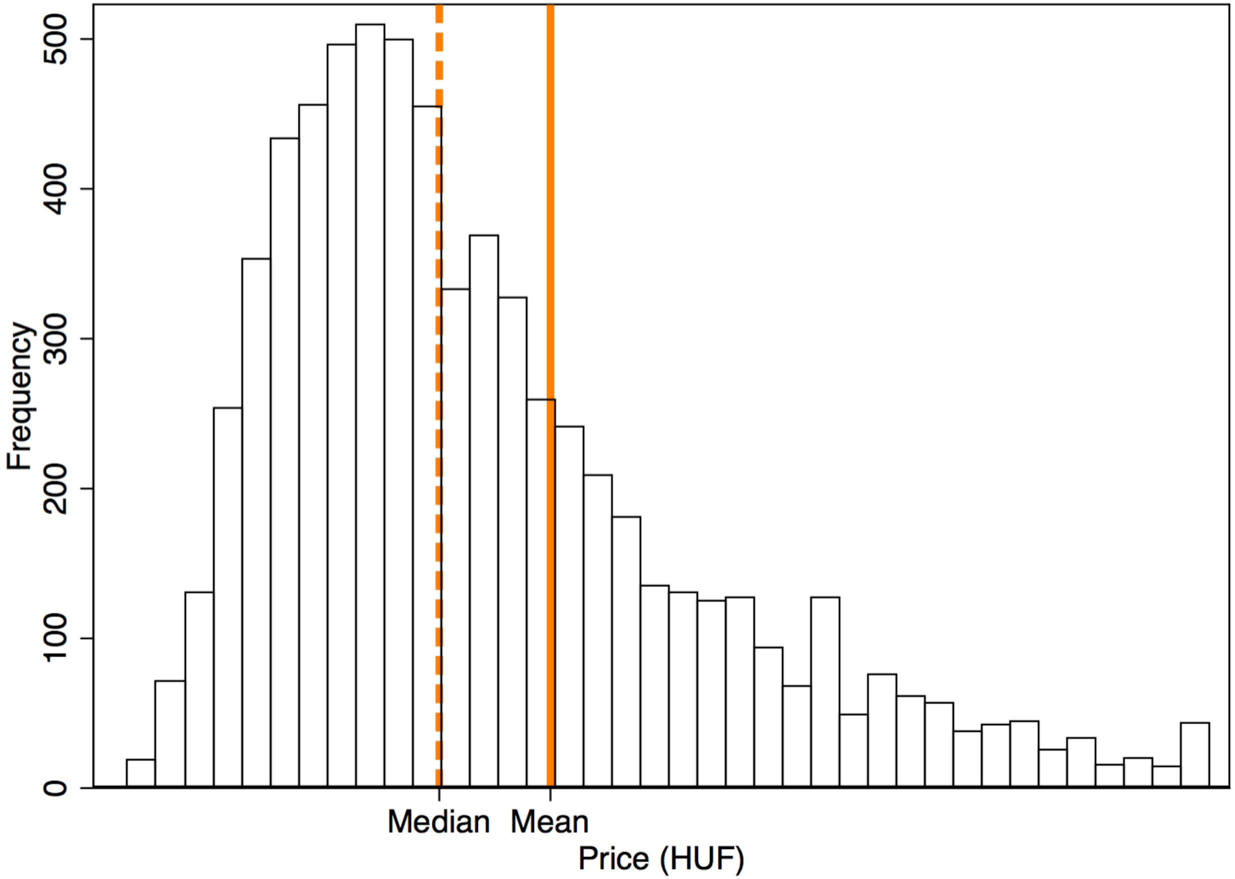
$$\frac{(\overline{x} - med(x))}{Std(x)}$$

Can you tell me the type of skewness?

- In the price data, the mean-median statistics is of 0.249. What does this imply? Skewness to the right or the left?

$$\frac{(\overline{x} - med(x))}{Std(x)} = .249$$

Skewness (Price Data)

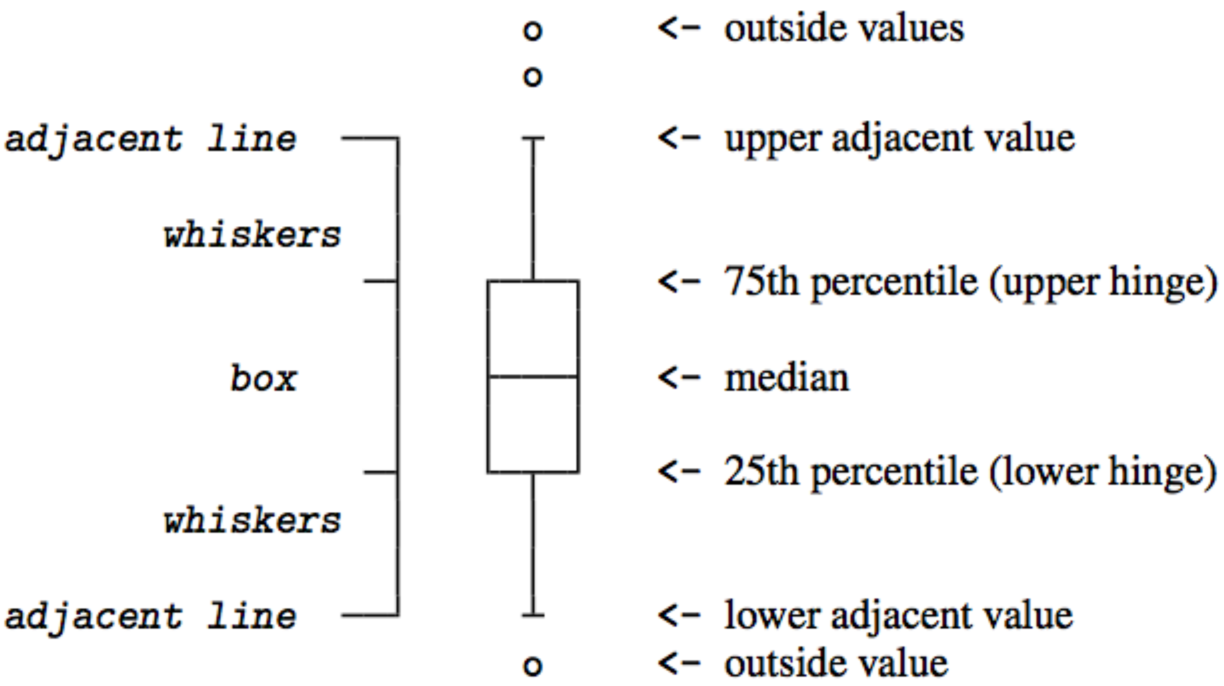


Box Plot

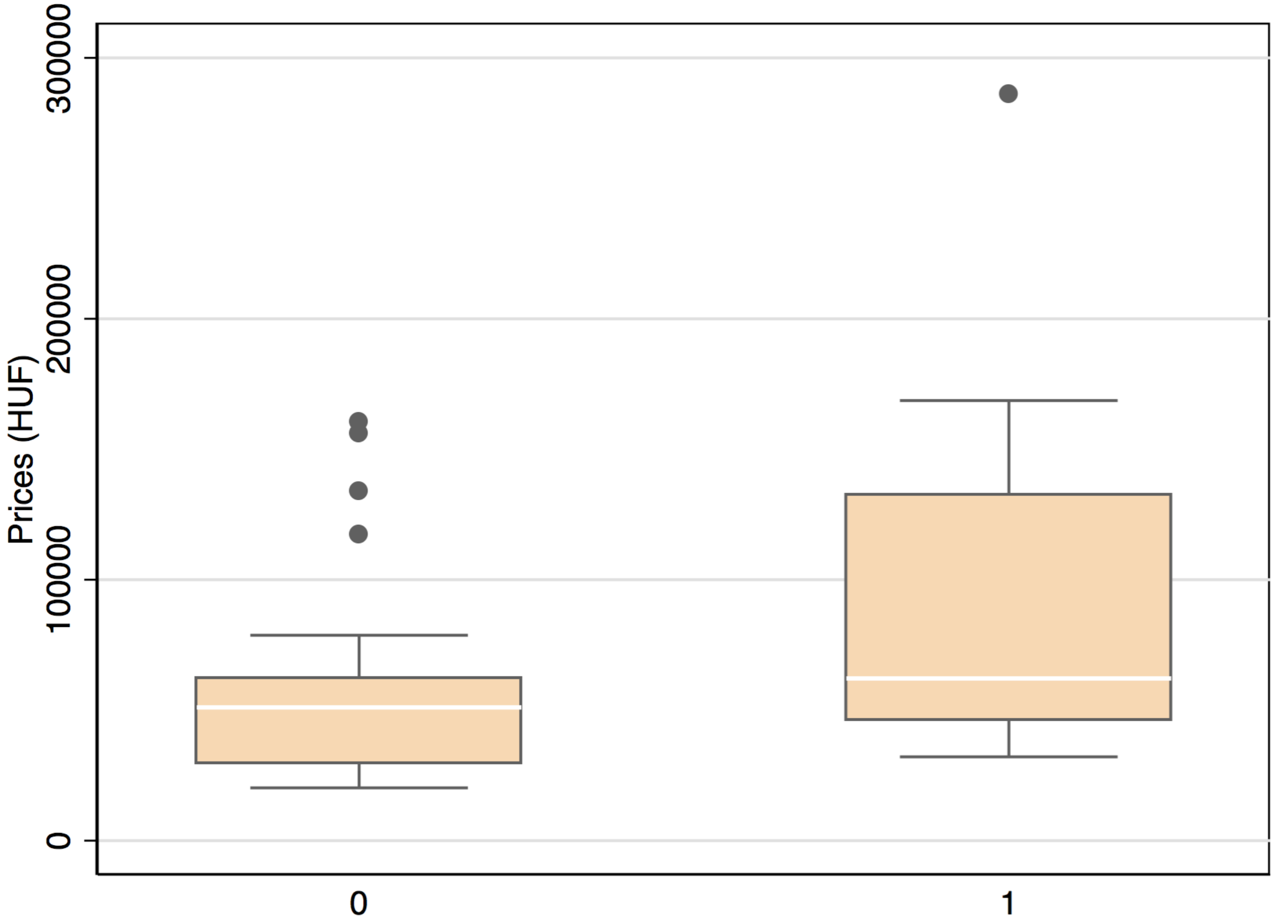
- The Box Plot is a combination of many statistics that we have already seen by now, namely:
  - The median of the data
  - The lower quartile
  - The third quartile
  - The smallest observation
  - The largest observation

The box plot conveys some important features of your data such as their skewness and show some of the quantiles in an explicit way

Box Plot Example



Trip to Vienna (Hotel Prices)



Probability

- A probability is a measure of the likelihood of an event.

$$0 \leq p(event) \leq 1$$

- Probabilities are between zero and one. Sometimes we express them as a percentage.
- Joint probability is the probability that two events occur jointly

$$p(event1 \& event2)$$

- Probability that the event does not occur

$$p(\sim event)$$

Independent Events

- Two event are said independent if their joint probability equals the product of their individual probabilities

$$p(event1\&event2) = p(event1)p(event2)$$

- Examples?

Conditional Probability

- Conditional probability is the probability of an event if another event happens

$$p(event1|event2) = \frac{p(event1\&event2)}{p(event2)}$$

- Examples?

Conditional Probability (of independent events)

- Conditional probability is the probability of an event if another event happens

$$p(event1|event2) = \frac{p(event1\&event2)}{p(event2)}$$

- In case the events are independent then:

$$\frac{p(event1\&event2)}{p(event2)} = \frac{p(event1)p(event2)}{p(event2)}$$

Example (Trip to Vienna)

Suppose now that you allow the website to give a recommendation about a hotel in Vienna. In this case, the website will use all the hotels in the database.

	$\geq$2 km	$<$ 2 km	Total
$<$ 4 stars	24	6	30
$\geq$ 4 stars	23	9	32
Total	47	15	62

Example (Trip to Vienna)

	$\geq$2 km	$<$ 2 km	Total
$<$ 4 stars	24	6	30
$\geq$ 4 stars	23	9	32
Total	47	15	62

- What is the unconditional probability that the hotel recommended from the website is 2 km or more distant from the city centre?

- Answer: $$\frac{24+23}{24+23+6+9} = .758$$

Example (Trip to Vienna)

	$\geq$2 km	$<$ 2 km	Total
$<$ 4 stars	24	6	30
$\geq$ 4 stars	23	9	32
Total	47	15	62

- What is the conditional probability that the hotel recommended is within two km from the city center if it has at least four stars ?
- Answer: $$\frac{9}{23+9} = .281$$

Example (Trip to Vienna)

	$\geq$2 km	$<$ 2 km	Total
$<$ 4 stars	24	6	30
$\geq$ 4 stars	23	9	32
Total	47	15	62

- What is the joint probability of the website recommends a hotel less than two km from the city center and a hotel with at least 4 starts?

- Answer: $$\frac{9}{24+23+6+9} = .14$$

Bayes Theorem

- Inverse conditional probabilities are two conditional probabilities, in which the role of the conditioning event and the conditional event are switched:

$$p(doped|positive)$$

- Two inverse conditional probabilities are related

$$p(event2|event1) = \frac{p(event1|event2)p(event2)}{p(event1)}$$

Distributions

- All variables have a distribution. The distribution of a variable tells the number of times each possible value of the variable occurs in the data

- It is important to learn some theoretical distributions and their properties because it helps understand features of real data.

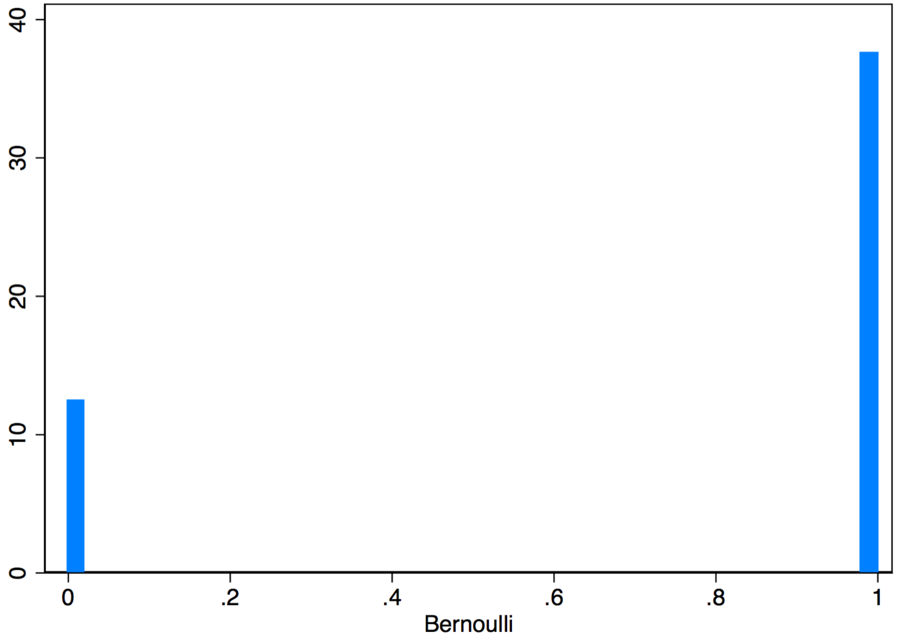
- Theoretical distributions are fully captured by a few parameters: these are statistics that determine the distributions

Bernoulli Distribution

- The Bernoulli distribution is a theoretical distribution that we observe over an over: all zero-one variables are distributed Bernoulli.

$$mean = p$$

$$var = p(1 - p)$$

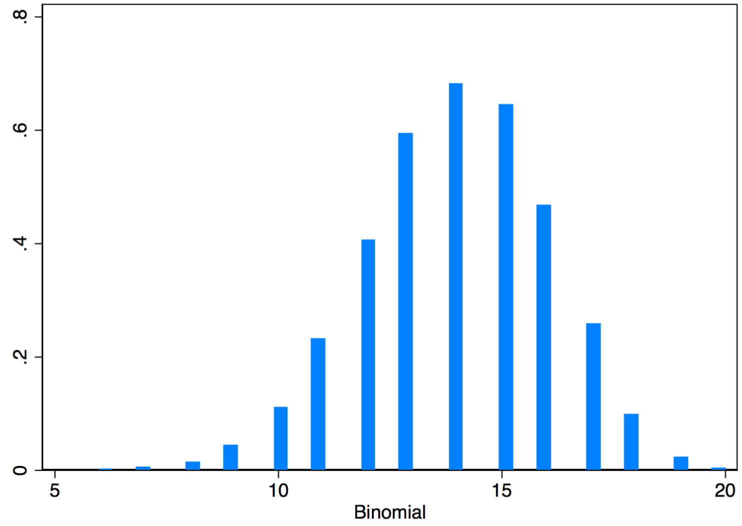


Binomial Distribution

- The Binomial distribution is based on the Bernoulli distribution. A variable is distributed Binomial if it can be viewed as the sum of many independent Bernoulli variables with the same p parameter.

$$mean = np$$

$$var = np(1-p)$$

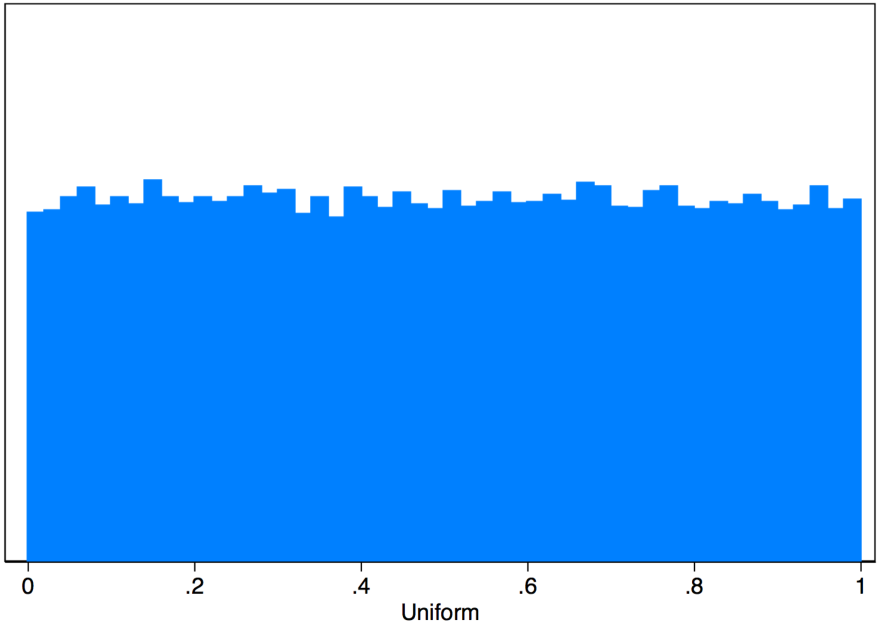


Uniform

- The uniform distribution characterizes continuous variables with values that are equally likely to occur within a minimum value and a maximum value.

$$mean = \frac{a+b}{2}$$

$$var = \frac{(b-a)^2}{12}$$



Normal Distribution

- It can be thought of as a generalization of the binomial with infinitely many Bernoulli variables added up

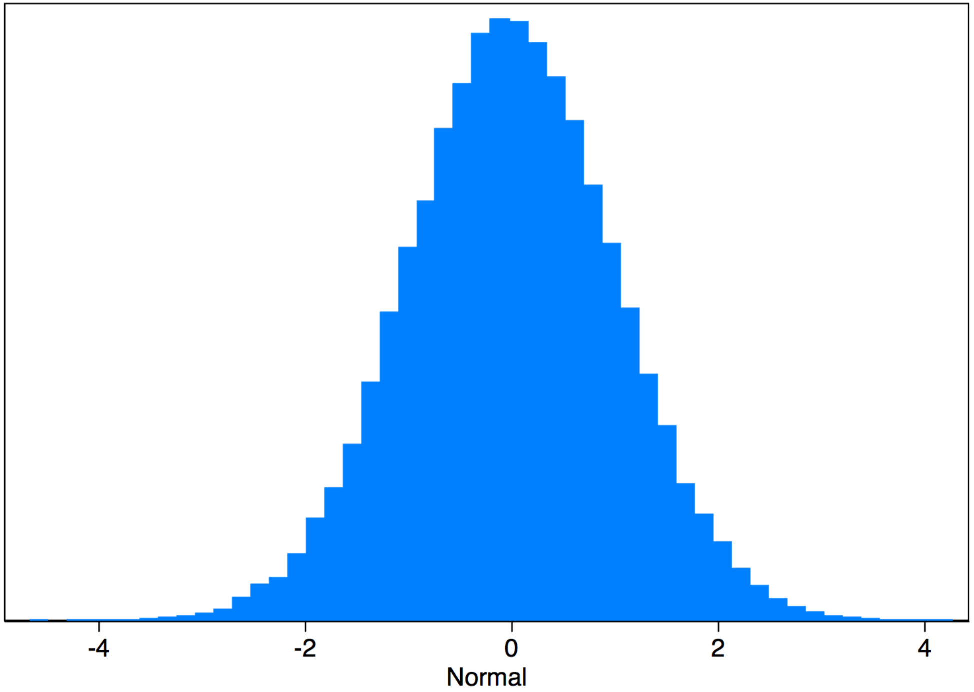
$$mean = \mu$$

$$var = \sigma^2$$

Standard Normal

$$mean = \mu = 0$$

$$var = \sigma^2 = 1$$

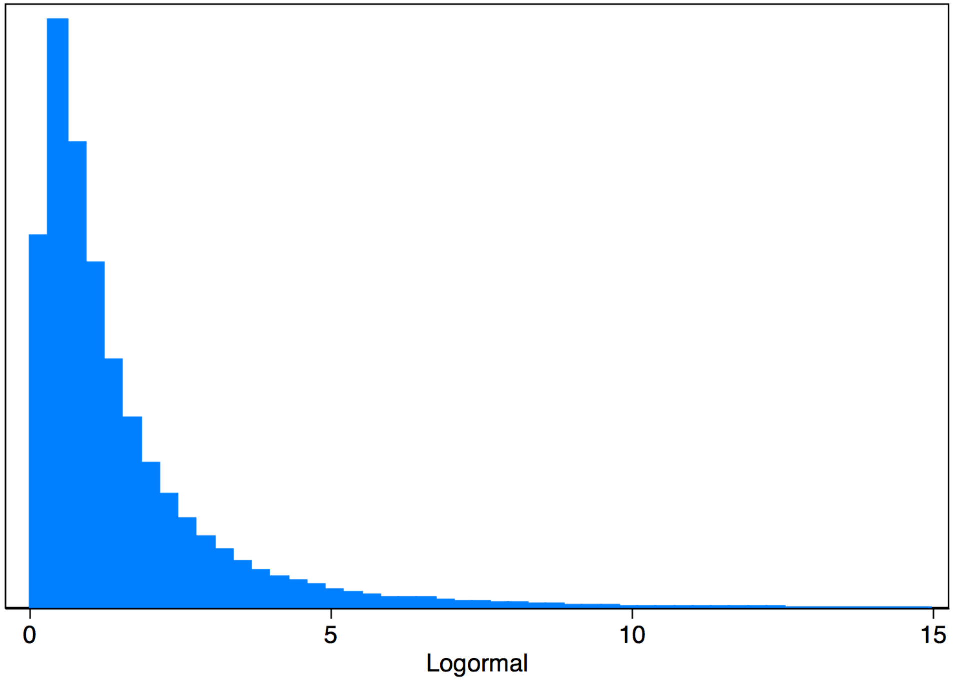


Lognormal Distribution

- If we take a variable that is distributed normal x and have the following transformation e^x

$$mean = e^{(\mu + \frac{\sigma^2}{2})}$$

$$var = e^{(\mu + \frac{\sigma^2}{2})} e^{(\sigma^2 - 1)}$$



See for yourself

http://students.brown.edu/seeing-theory/distributions/

Codes in R and Stata (which will be available in Moodle), change the parameters and see changes in the distribution

It's fun!!