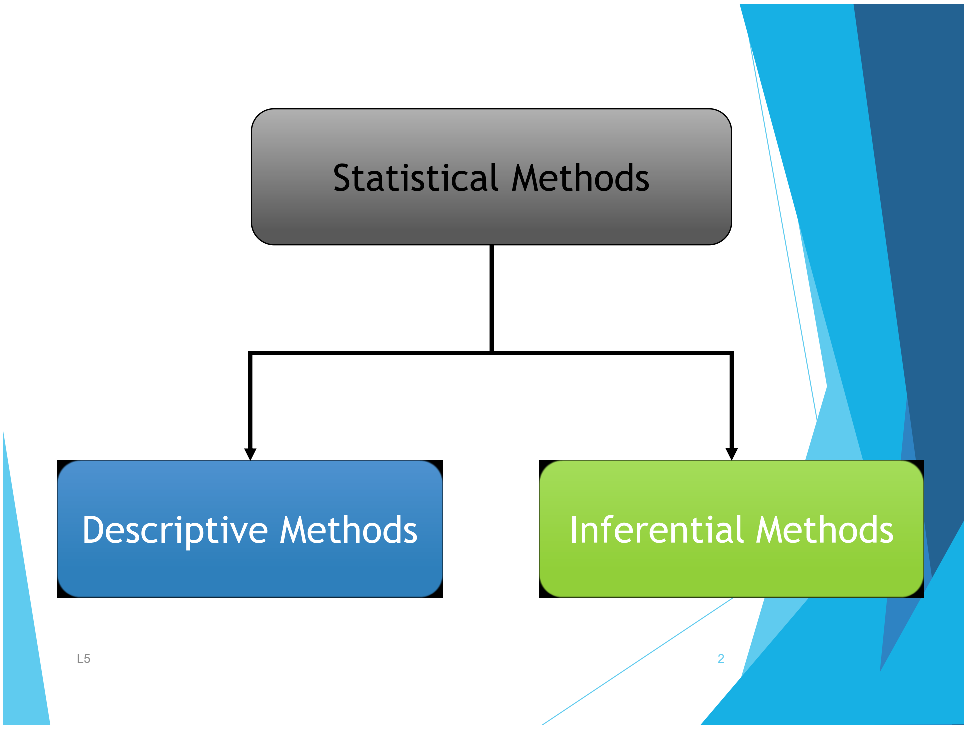
Statistical Methods

in Regional Analysis



Statistical Methods

Descriptive Methods

Inferential Methods

Descriptive Statistics

- Example: "The **average** income of the 16 employees in our company is

€4 388.75."

$$Average = \frac{\sum Income}{n}$$

ID	Income
1	€ 3 480.00
2	€ 6 600.00
3	€ 3 720.00
4	€ 5 520.00
5	€ 3 060.00
6	€ 5 040.00
7	€ 1 920.00
8	€ 3 060.00
9	€ 2 160.00
10	€ 14 400.00
11	€ 3 060.00
12	€ 5 520.00
13	€ 2 100.00
14	€ 2 100.00
15	€ 6 600.00
16	€ 1 800.00
Average	€ 4 383.75

Descriptive Statistics

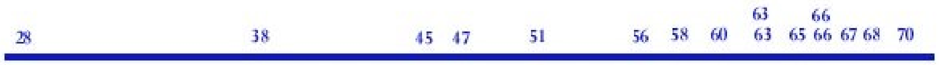
- **Descriptive statistics** - properties of a group of scores or data that we have "in hand," i.e., data that are accessible to us in that we can write them down on paper or type them into a spreadsheet.
- In descriptive statistics we are not interested in other data that were not gathered but might have been; that is the subject of **inferential statistics**.

Descriptive Statistics

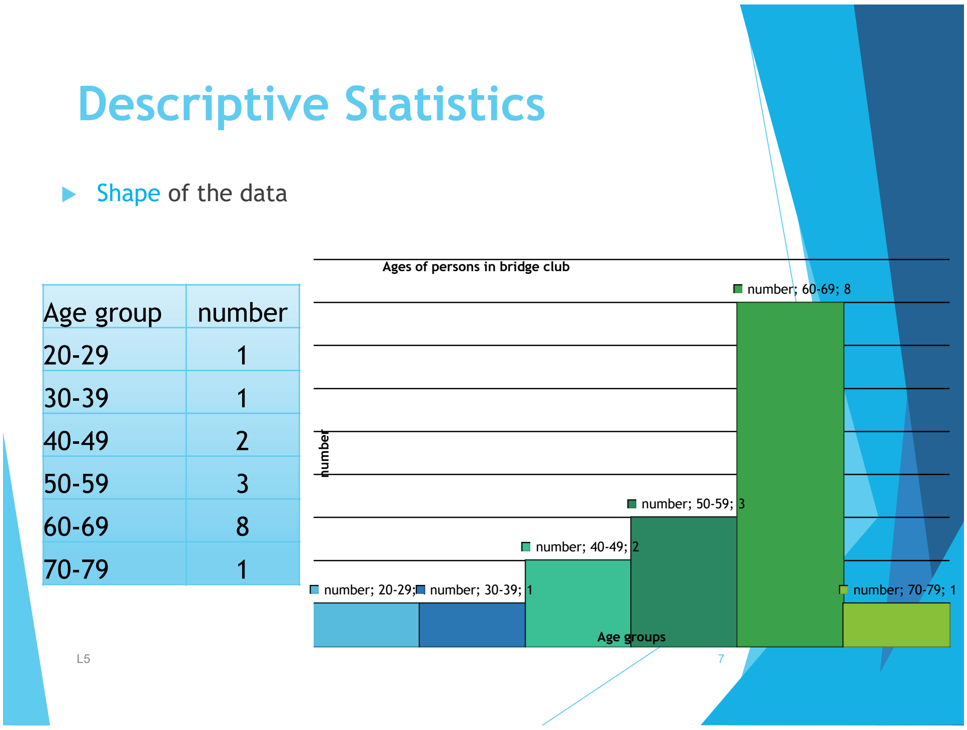
- What properties of the set of scores are we interested in? At least three: their **center**, their **spread**, and their **shape**. Consider the following set of scores, which might be ages of persons in your bridge club:
- 28, 38, 45, 47, 51, 56, 58, 60, 63, 63, 65, 66, 66, 67, 68, 70

Descriptive Statistics

- We could say of these ages that they range from 28 to 70 (**spread**), and the middle of them is somewhere around 50 (**center**). Now their shape is a property of a graph that can be drawn to depict the scores.



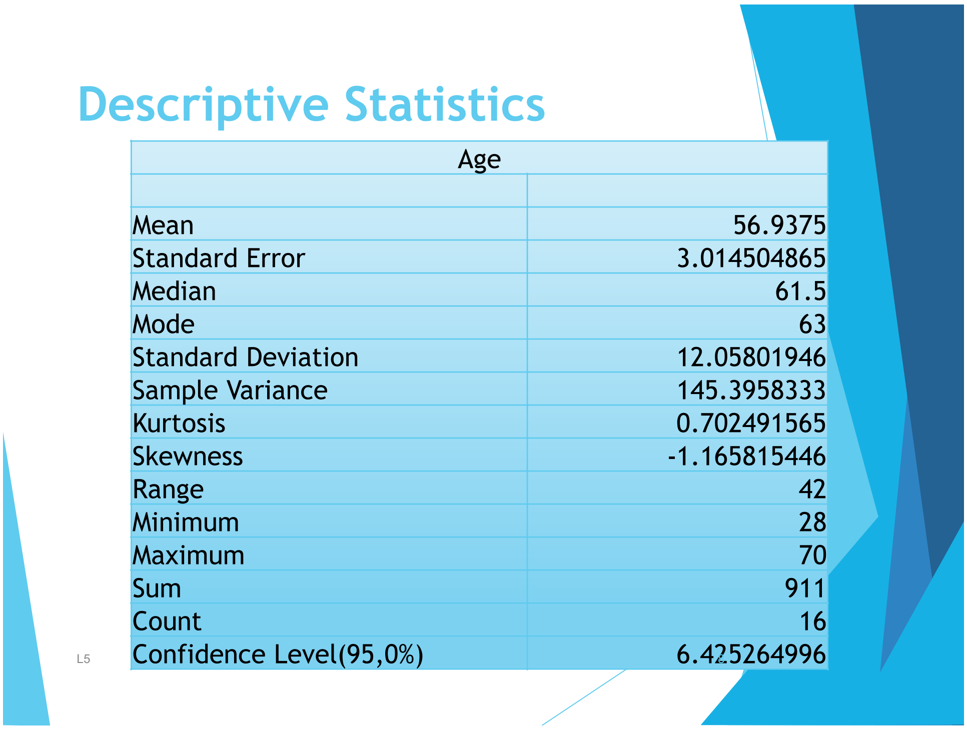
Ages of the Bridge Club Members



Descriptive Statistics

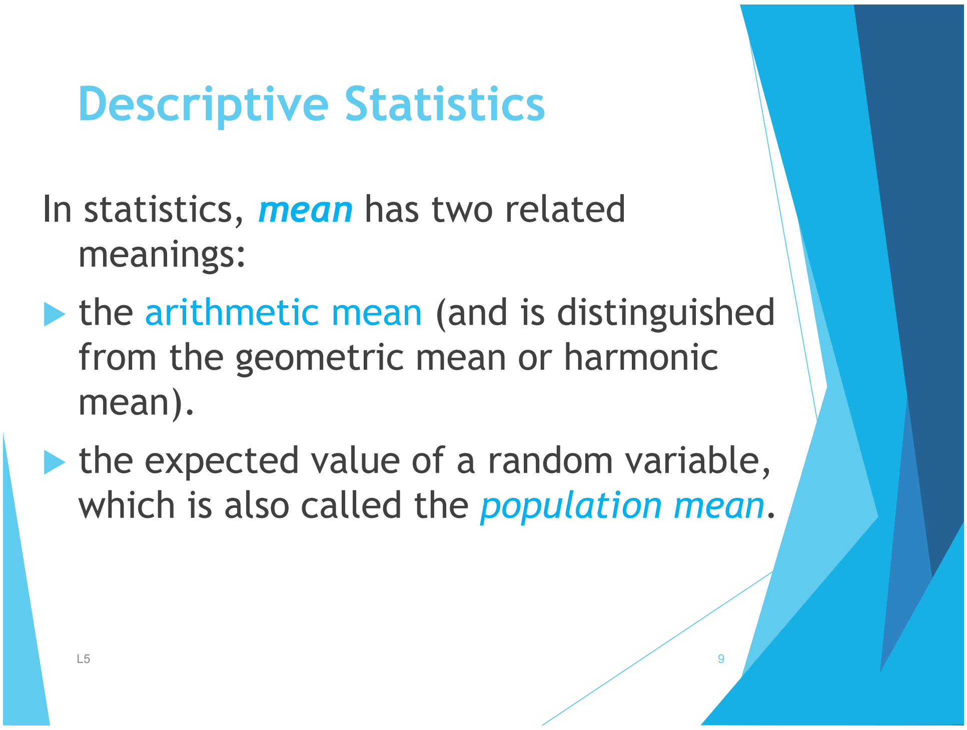
- Shape of the data

Age group	number
20-29	1
30-39	1
40-49	2
50-59	3
60-69	8
70-79	1



Descriptive Statistics

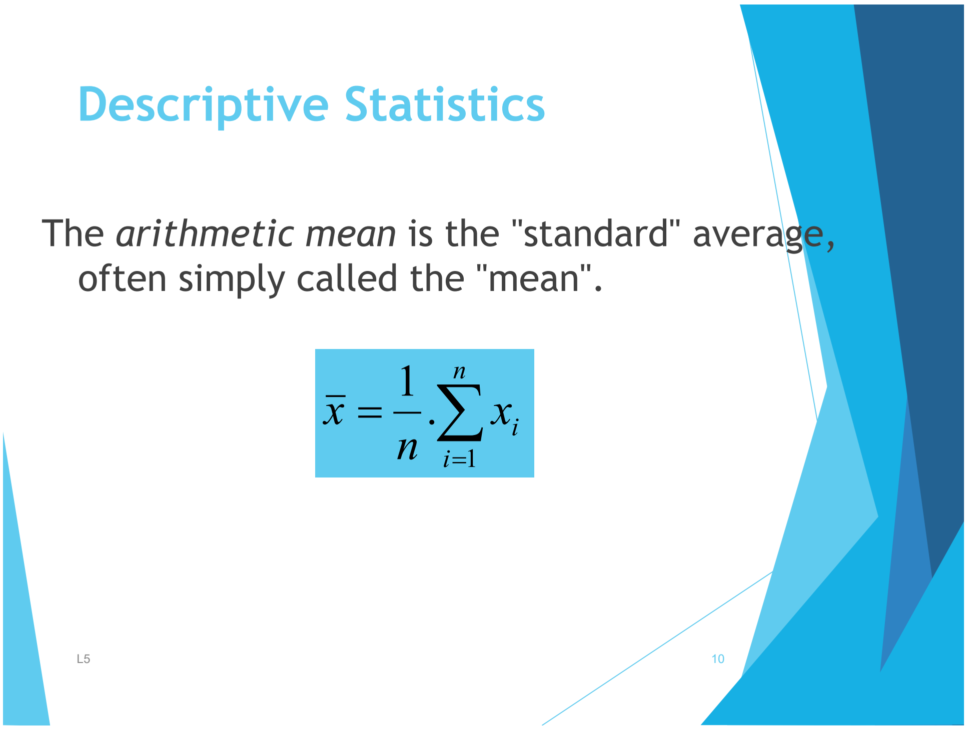
Age	
Mean	56.9375
Standard Error	3.014504865
Median	61.5
Mode	63
Standard Deviation	12.05801946
Sample Variance	145.3958333
Kurtosis	0.702491565
Skewness	-1.165815446
Range	42
Minimum	28
Maximum	70
Sum	911
Count	16
Confidence Level(95,0%)	6.425264996



Descriptive Statistics

In statistics, ***mean*** has two related meanings:

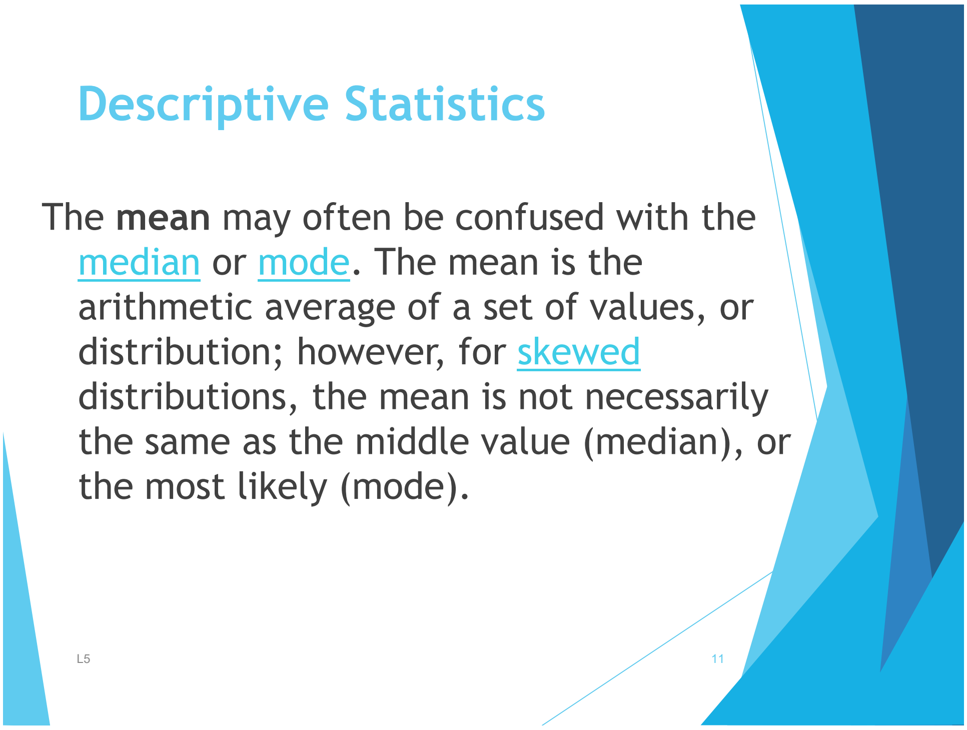
- the arithmetic mean (and is distinguished from the geometric mean or harmonic mean).
- the expected value of a random variable, which is also called the *population mean*.



Descriptive Statistics

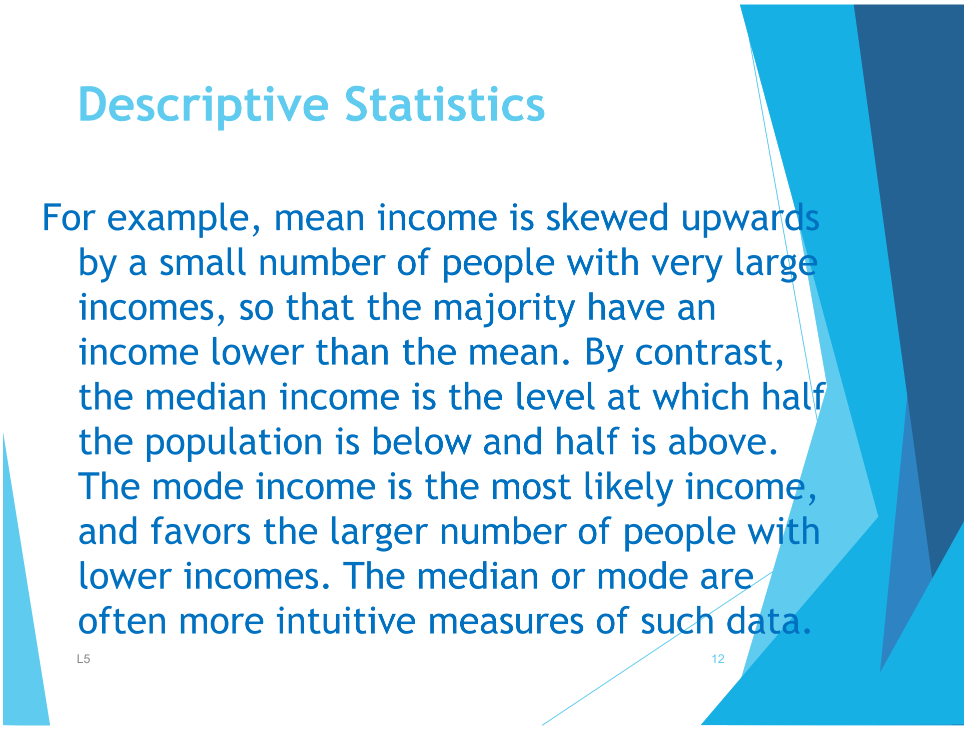
The *arithmetic mean* is the "standard" average, often simply called the "mean".

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



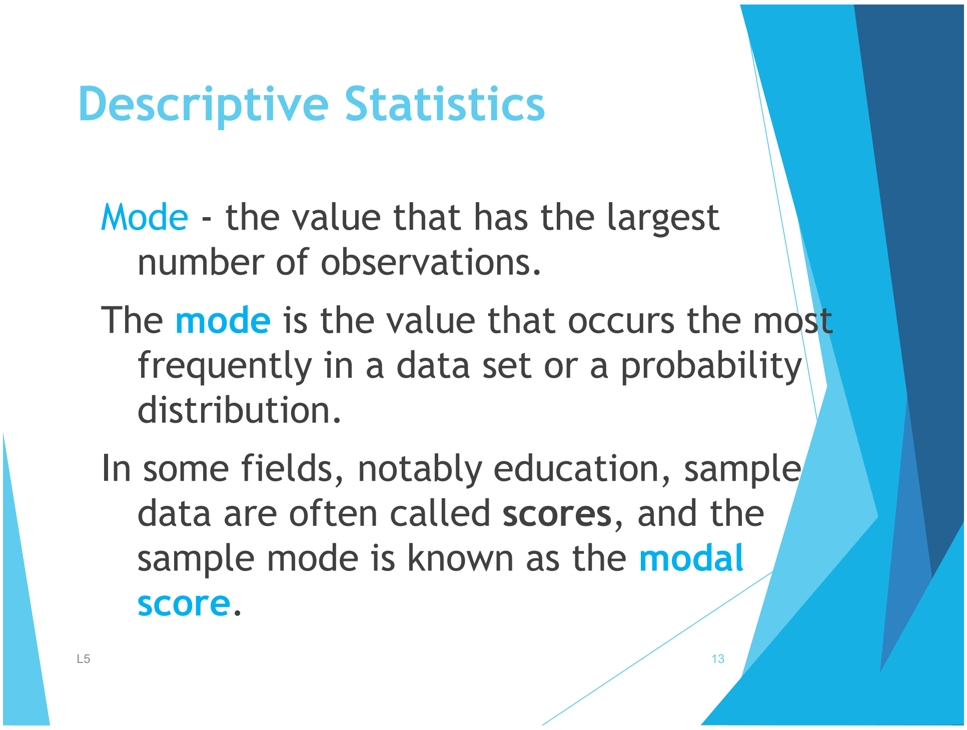
Descriptive Statistics

The **mean** may often be confused with the median or mode. The mean is the arithmetic average of a set of values, or distribution; however, for skewed distributions, the mean is not necessarily the same as the middle value (median), or the most likely (mode).



Descriptive Statistics

For example, mean income is skewed upwards by a small number of people with very large incomes, so that the majority have an income lower than the mean. By contrast, the median income is the level at which half the population is below and half is above. The mode income is the most likely income, and favors the larger number of people with lower incomes. The median or mode are often more intuitive measures of such data.

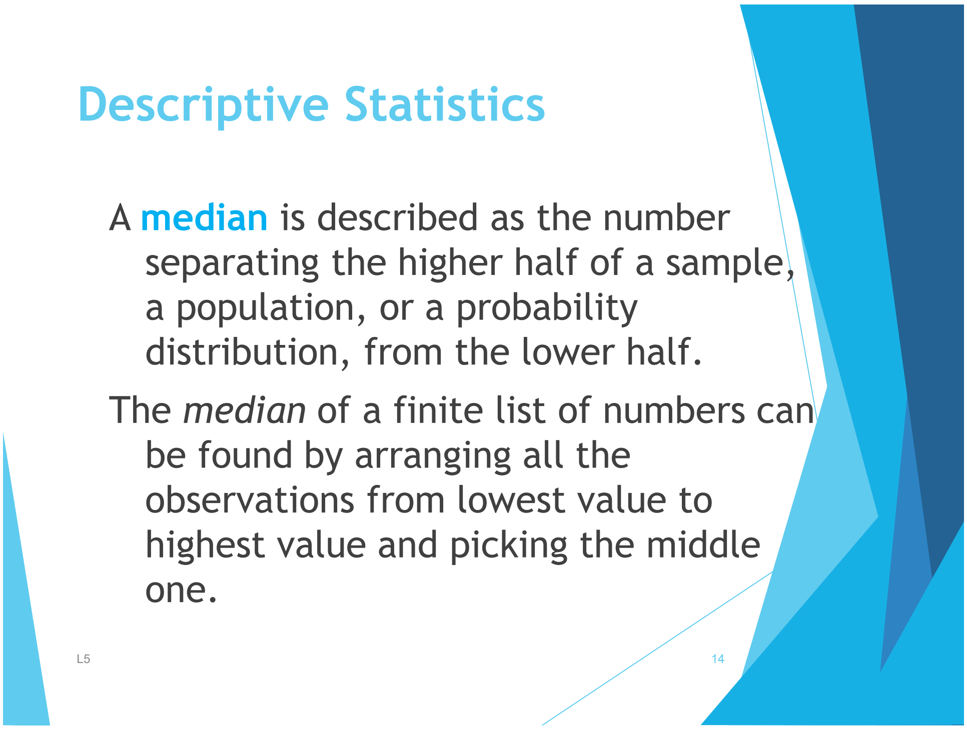


Descriptive Statistics

Mode - the value that has the largest number of observations.

The **mode** is the value that occurs the most frequently in a data set or a probability distribution.

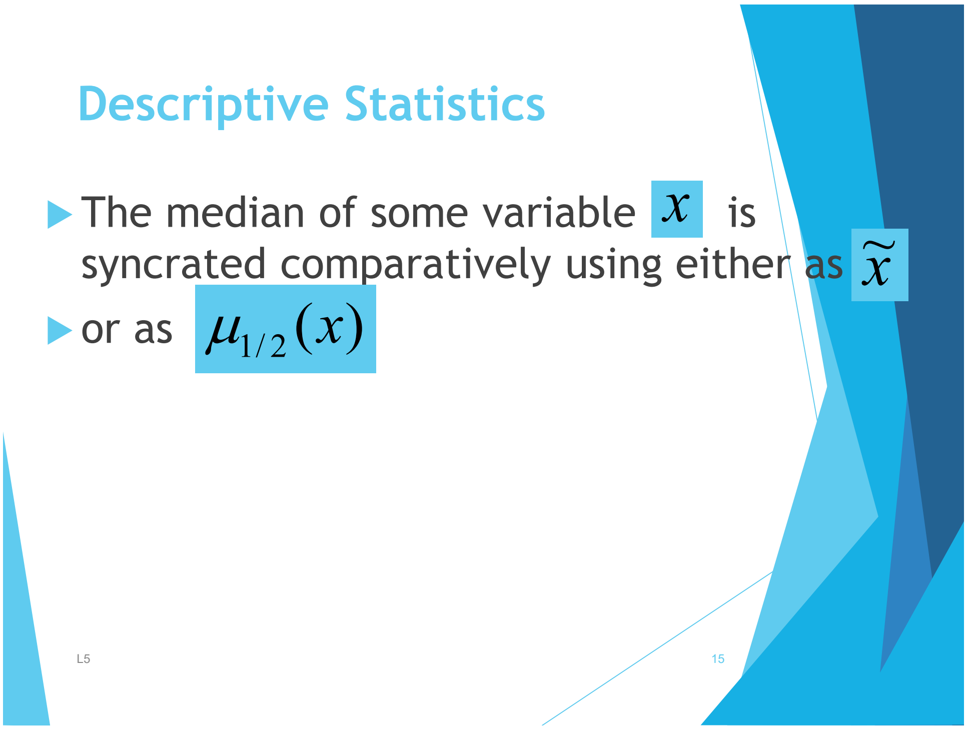
In some fields, notably education, sample data are often called **scores**, and the sample mode is known as the **modal score**.



Descriptive Statistics

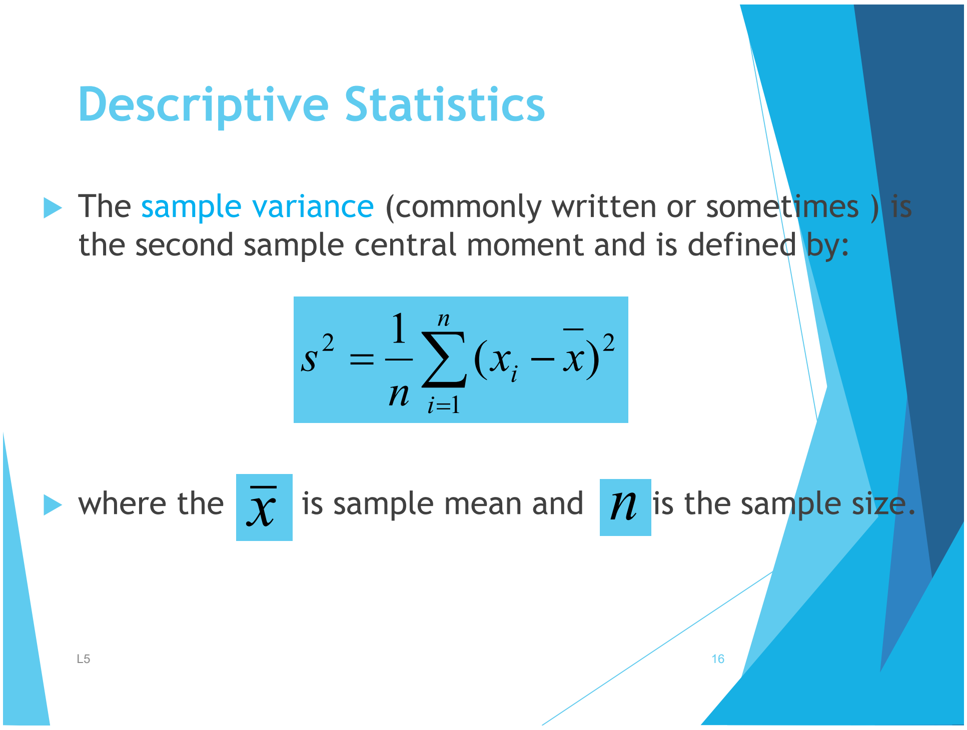
A **median** is described as the number separating the higher half of a sample, a population, or a probability distribution, from the lower half.

The *median* of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one.



Descriptive Statistics

- The median of some variable x is syncrated comparatively using either as $\tilde{x}$
- or as $\mu_{1/2}(x)$



Descriptive Statistics

- The sample variance (commonly written or sometimes) is the second sample central moment and is defined by:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

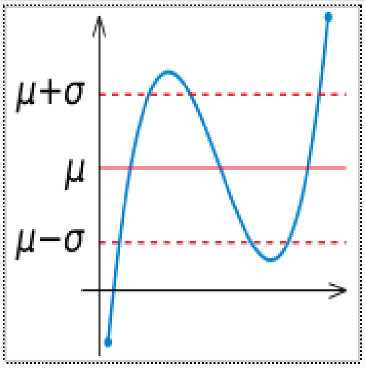
- where the $\bar{x}$ is sample mean and n is the sample size.

Descriptive Statistics

The **standard deviation** is a measure of the dispersion of a set of values. It can apply to a probability distribution, a random variable, a population or a multiset.

The standard deviation is usually denoted with the letter σ (lower case sigma). It is defined as the root-mean-square (RMS) deviation of the values from their mean, or as the square root of the variance.

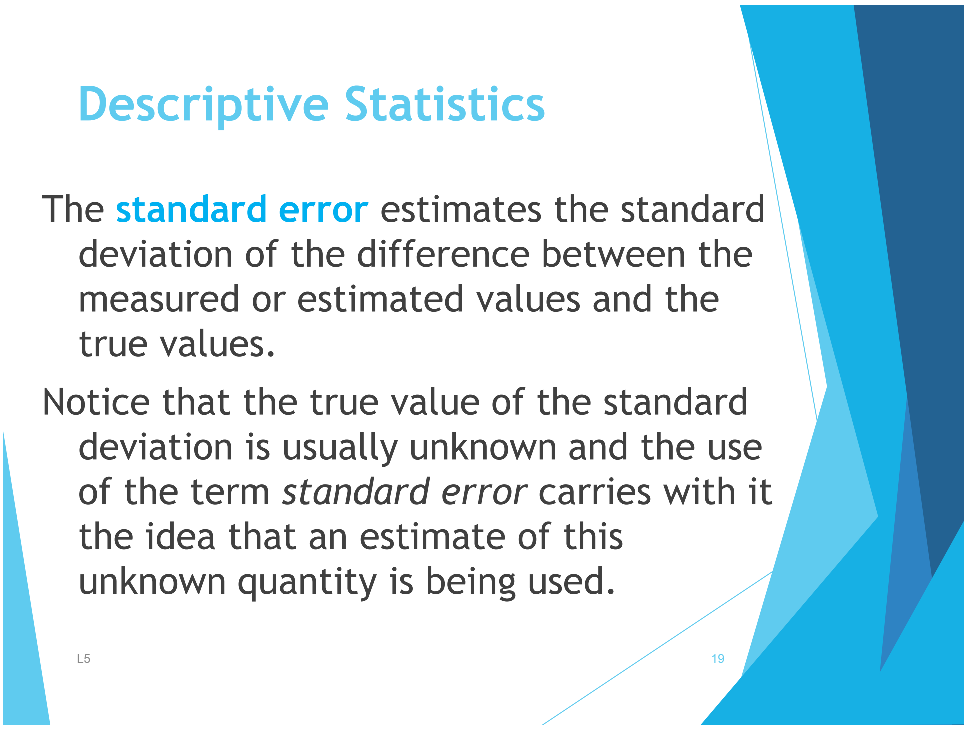
Descriptive Statistics



Given a random variable (in blue), the standard deviation σ is a measure of the spread of the values of the random variable away from its mean μ.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

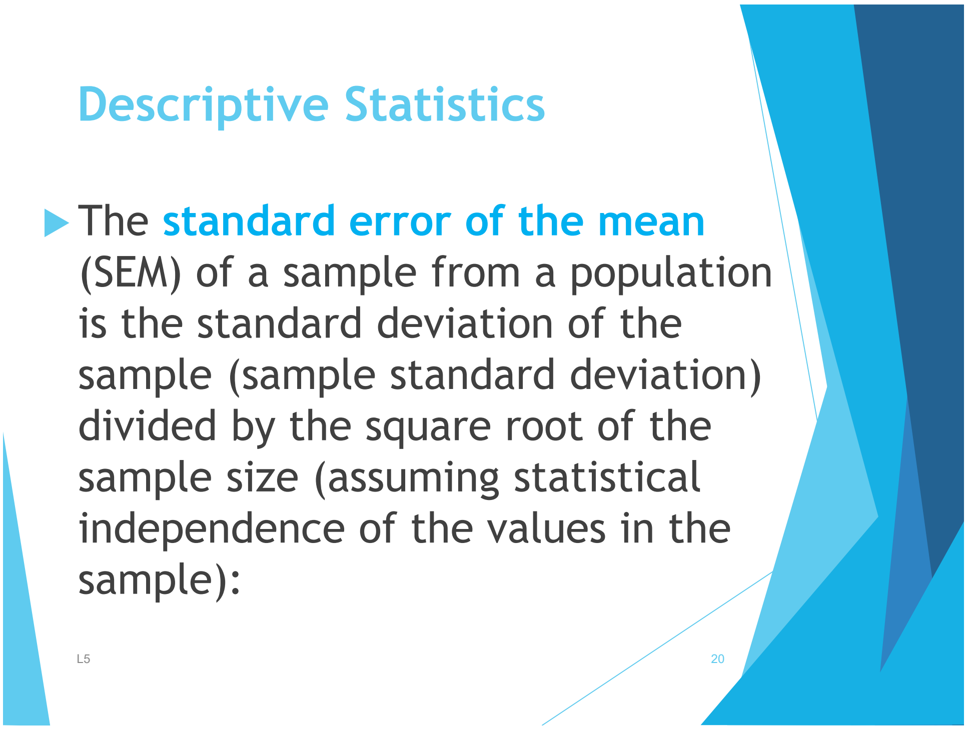
$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$



Descriptive Statistics

The **standard error** estimates the standard deviation of the difference between the measured or estimated values and the true values.

Notice that the true value of the standard deviation is usually unknown and the use of the term *standard error* carries with it the idea that an estimate of this unknown quantity is being used.



Descriptive Statistics

- The **standard error of the mean** (SEM) of a sample from a population is the standard deviation of the sample (sample standard deviation) divided by the square root of the sample size (assuming statistical independence of the values in the sample):

Descriptive Statistics

- **standard error of the mean**

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

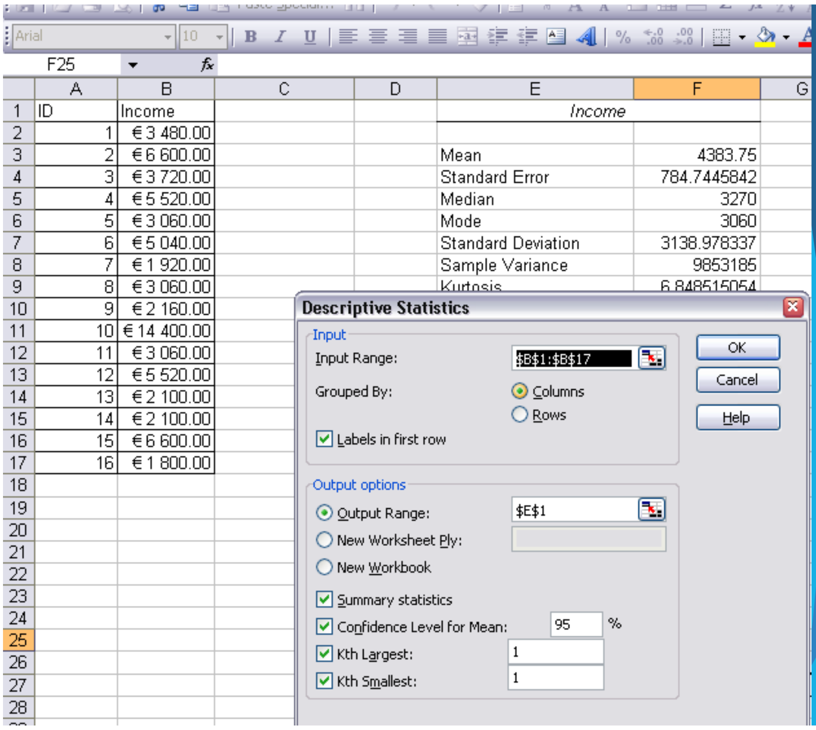
s is the sample standard deviation (ie the sample based estimate of the standard deviation of the population), and

n is the size (number of items) of the sample.

Descriptive Statistics

ID	Income
1	€ 3 480.00
2	€ 6 600.00
3	€ 3 720.00
4	€ 5 520.00
5	€ 3 060.00
6	€ 5 040.00
7	€ 1 920.00
8	€ 3 060.00
9	€ 2 160.00
10	€ 14 400.00
11	€ 3 060.00
12	€ 5 520.00
13	€ 2 100.00
14	€ 2 100.00
15	€ 6 600.00
16	€ 1 800.00
Average	€ 4 383.75

Income	
Mean	4383.75
Standard Error	784.7445842
Median	3270
Mode	3060
Standard Deviation	3138.978337
Sample Variance	9853185
Kurtosis	6.848515054
Skewness	2.35220486
Range	12600
Minimum	1800
Maximum	14400
Sum	70140
Count	16
Largest(1)	14400
Smallest(1)	1800
Confidence Level(95.0%)	1672.64348



	A	B	C	D	E	F	G
1	ID	Income			Income		
2	1	€ 3 480.00					
3	2	€ 6 600.00			Mean	4383.75	
4	3	€ 3 720.00			Standard Error	784.7445842	
5	4	€ 5 520.00			Median	3270	
6	5	€ 3 060.00			Mode	3060	
7	6	€ 5 040.00			Standard Deviation	3138.978337	
8	7	€ 1 920.00			Sample Variance	9853185	
9	8	€ 3 060.00			Kurtosis	6.848515054	
10	9	€ 2 160.00					
11	10	€ 14 400.00					
12	11	€ 3 060.00					
13	12	€ 5 520.00					
14	13	€ 2 100.00					
15	14	€ 2 100.00					
16	15	€ 6 600.00					
17	16	€ 1 800.00					

Descriptive Statistics

Input

Input Range: $B$1:$B$17

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $E$1

○ New Worksheet Ply:

○ New Workbook

☑ Summary statistics

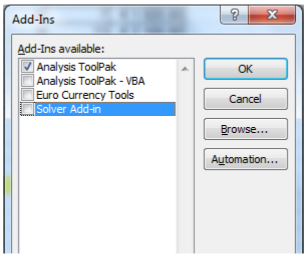
☑ Confidence Level for Mean: 95 %

☑ Kth Largest: 1

☑ Kth Smallest: 1

Activate Data Analysis

- Excel 2010:
- File - > Options - > Add-Ins - > Manage Excel Add-Ins - >Go



Add-Ins

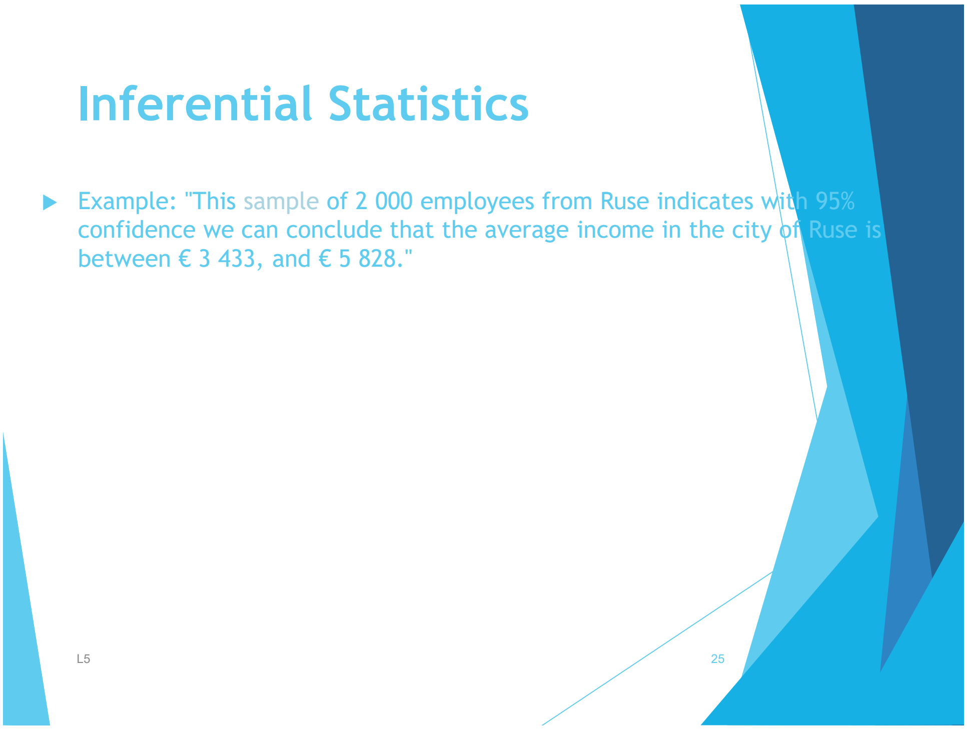
Add-Ins available:

☑ Analysis ToolPak

□ Analysis ToolPak - VBA

□ Euro Currency Tools

□ Solver Add-in

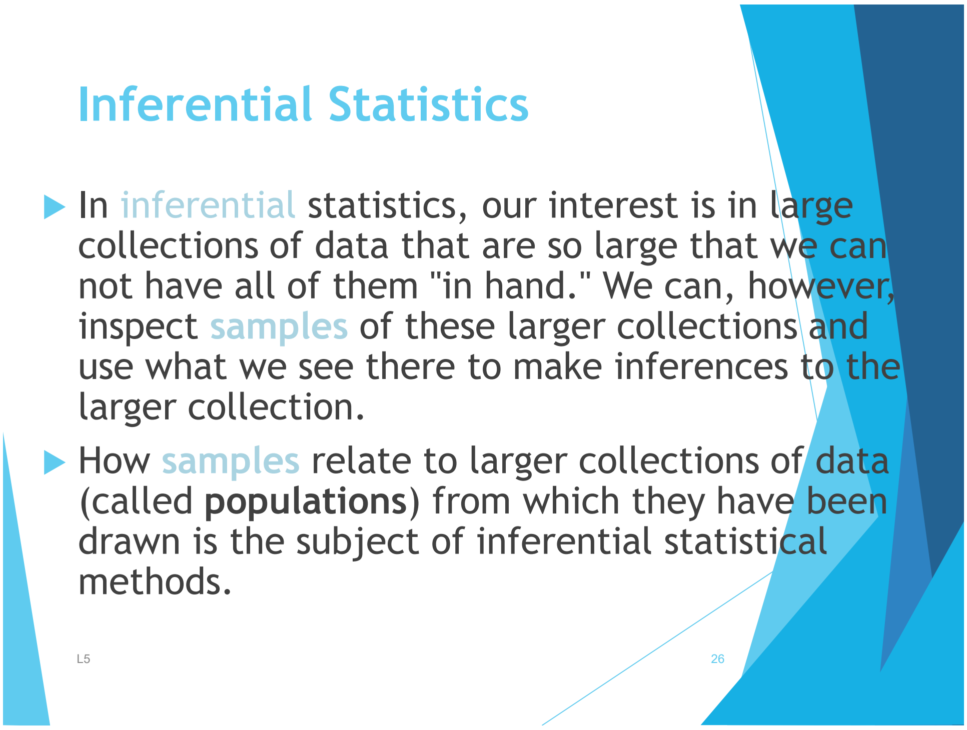


Inferential Statistics

- Example: "This sample of 2 000 employees from Ruse indicates with 95% confidence we can conclude that the average income in the city of Ruse is between € 3 433, and € 5 828."

L5

25

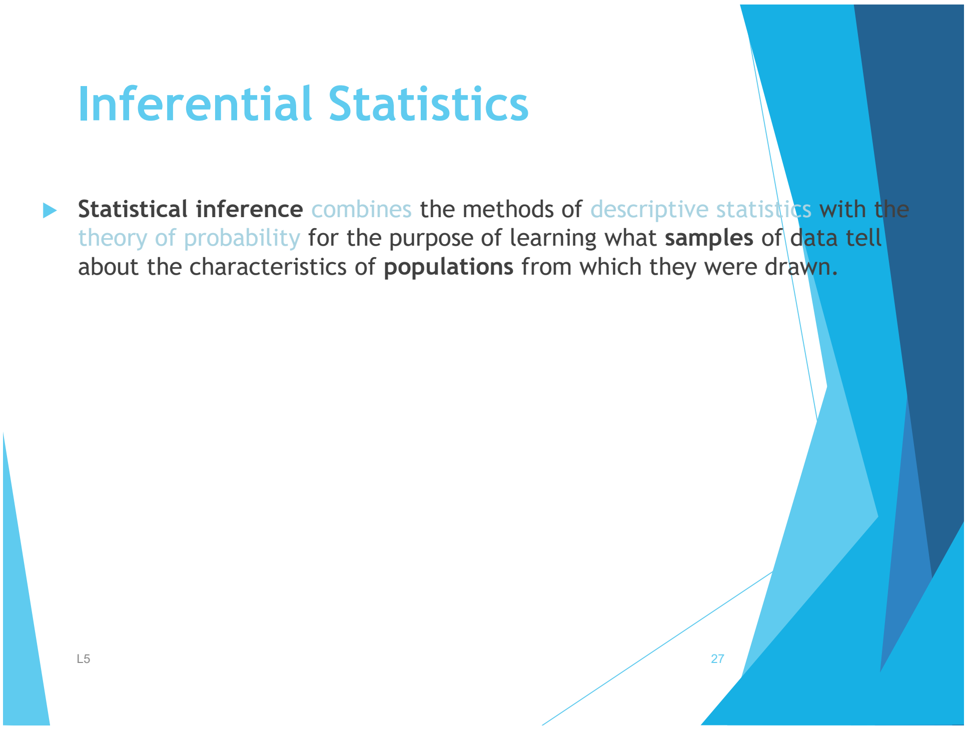


Inferential Statistics

- In inferential statistics, our interest is in large collections of data that are so large that we can not have all of them "in hand." We can, however, inspect samples of these larger collections and use what we see there to make inferences to the larger collection.
- How samples relate to larger collections of data (called **populations**) from which they have been drawn is the subject of inferential statistical methods.

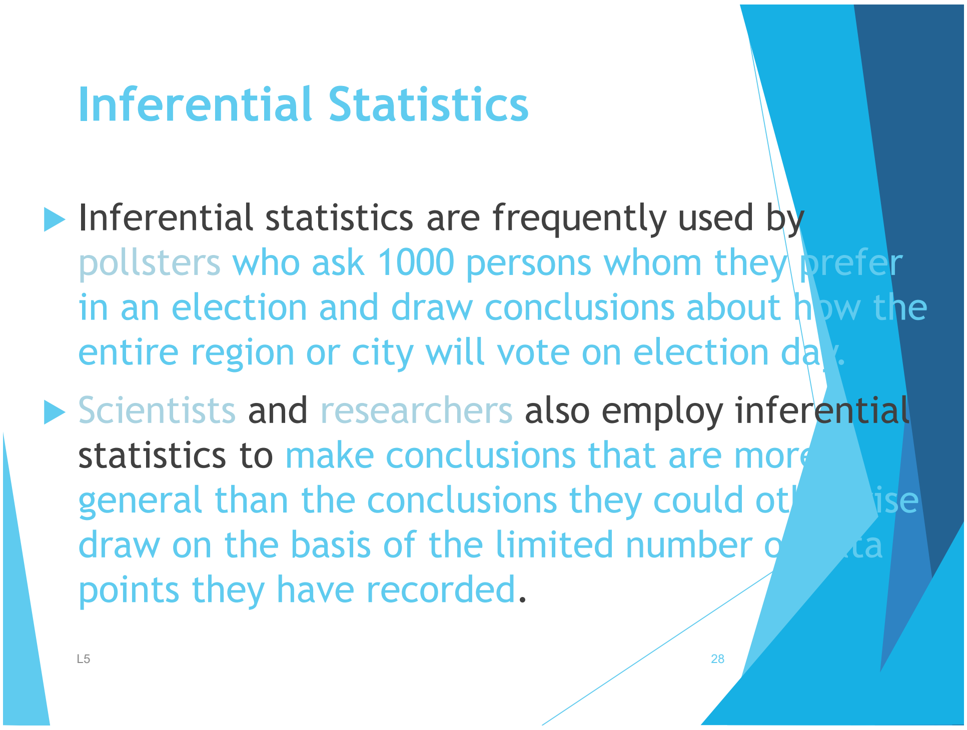
L5

26



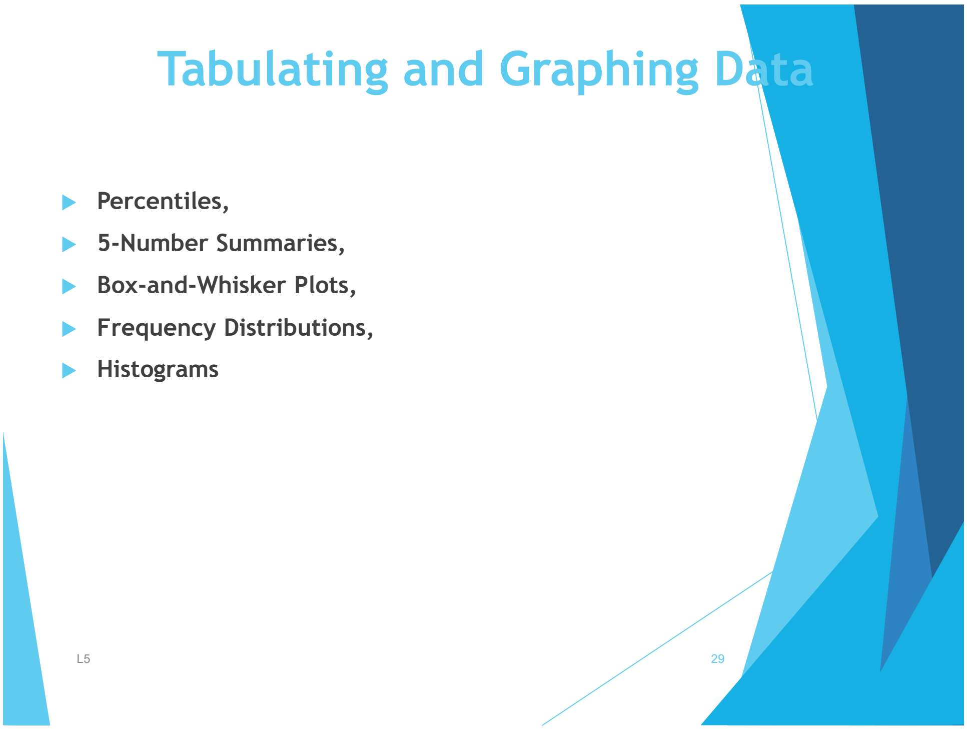
Inferential Statistics

- **Statistical inference** combines the methods of descriptive statistics with the theory of probability for the purpose of learning what **samples** of data tell about the characteristics of **populations** from which they were drawn.



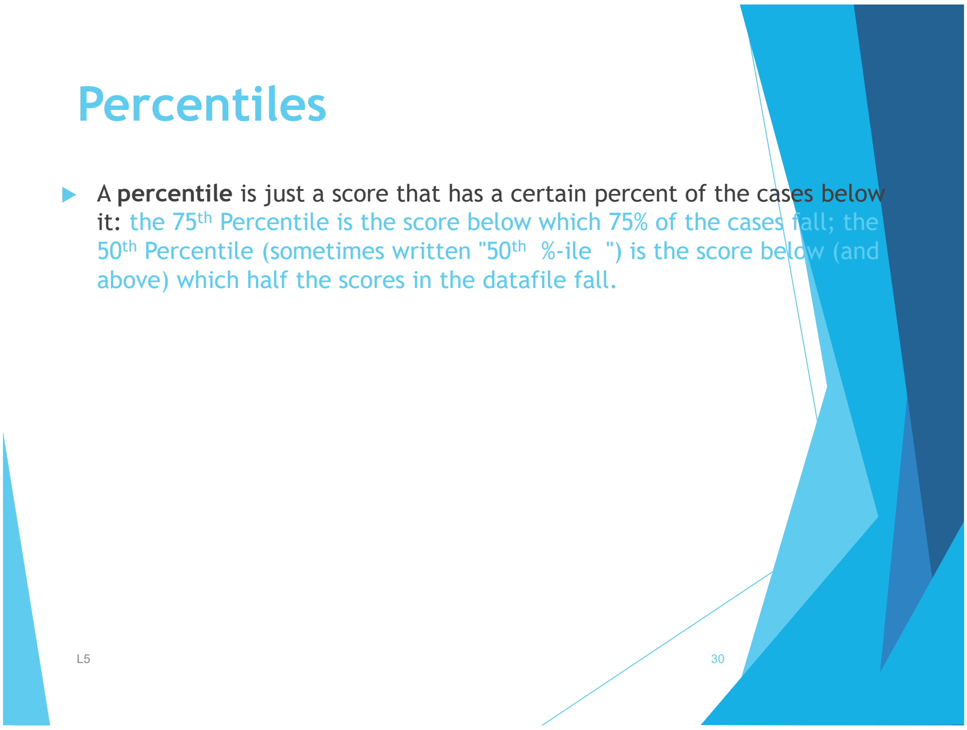
Inferential Statistics

- Inferential statistics are frequently used by pollsters who ask 1000 persons whom they prefer in an election and draw conclusions about how the entire region or city will vote on election day.
- Scientists and researchers also employ inferential statistics to make conclusions that are more general than the conclusions they could otherwise draw on the basis of the limited number of data points they have recorded.



Tabulating and Graphing Data

- Percentiles,
- 5-Number Summaries,
- Box-and-Whisker Plots,
- Frequency Distributions,
- Histograms

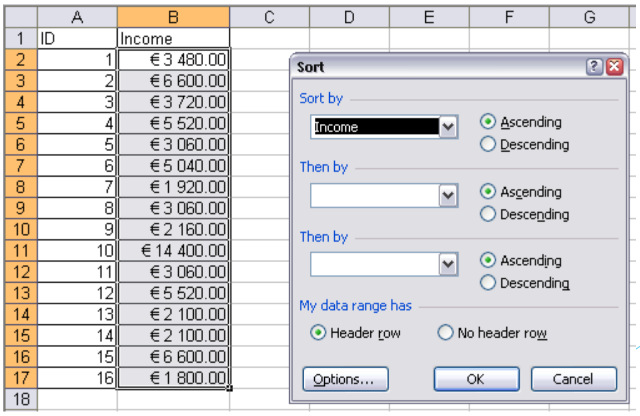


Percentiles

- A **percentile** is just a score that has a certain percent of the cases below it: the 75th Percentile is the score below which 75% of the cases fall; the 50th Percentile (sometimes written "50th %-ile ") is the score below (and above) which half the scores in the datafile fall.

Percentiles

- Data -> Sort
- 75th percentile = 16*75%=12



	A	B	C	D	E	F	G
1	ID	Income					
2	1	€ 3 480.00					
3	2	€ 6 600.00					
4	3	€ 3 720.00					
5	4	€ 5 520.00					
6	5	€ 3 060.00					
7	6	€ 5 040.00					
8	7	€ 1 920.00					
9	8	€ 3 060.00					
10	9	€ 2 160.00					
11	10	€ 14 400.00					
12	11	€ 3 060.00					
13	12	€ 5 520.00					
14	13	€ 2 100.00					
15	14	€ 2 100.00					
16	15	€ 6 600.00					
17	16	€ 1 800.00					
18							

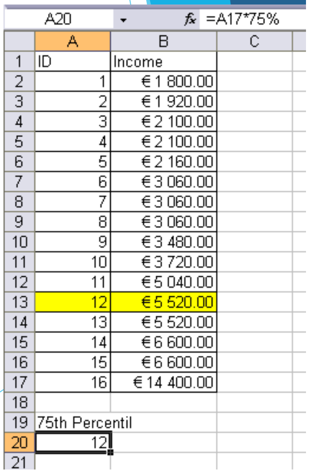
Sort

Sort by: Income — ● Ascending ○ Descending

Then by: — ● Ascending ○ Descending

Then by: — ● Ascending ○ Descending

My data range has: ● Header row ○ No header row

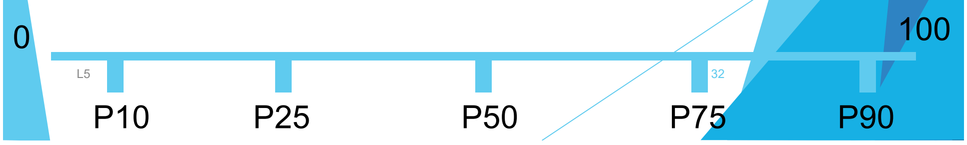


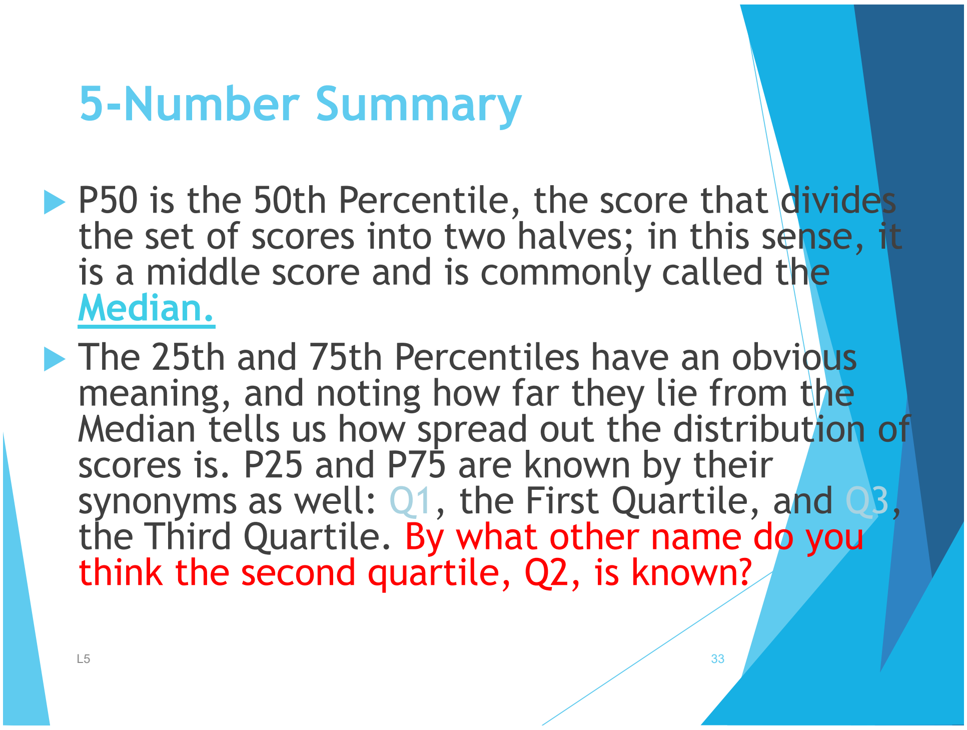
A20 fx =A17*75%

	A	B	C
1	ID	Income	
2	1	€ 1 800.00	
3	2	€ 1 920.00	
4	3	€ 2 100.00	
5	4	€ 2 100.00	
6	5	€ 2 160.00	
7	6	€ 3 060.00	
8	7	€ 3 060.00	
9	8	€ 3 060.00	
10	9	€ 3 480.00	
11	10	€ 3 720.00	
12	11	€ 5 040.00	
13	12	€ 5 520.00	
14	13	€ 5 520.00	
15	14	€ 6 600.00	
16	15	€ 6 600.00	
17	16	€ 14 400.00	
18			
19	75th Percentil		
20	12		
21			

5-Number Summary

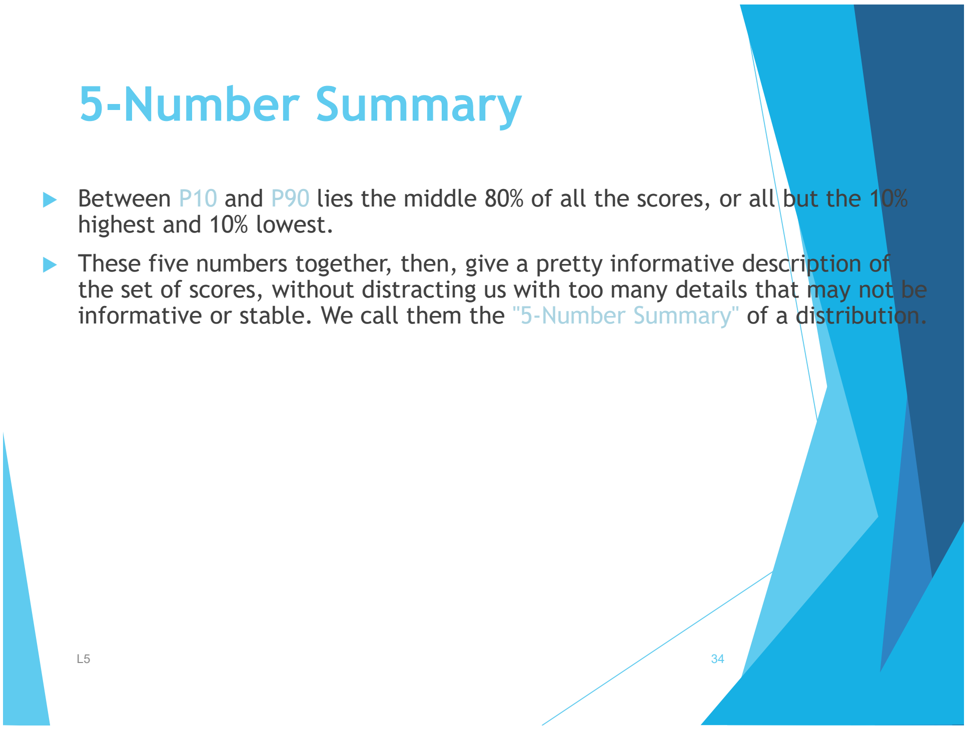
- **It** is a very concise way to describe the major features of a set of scores without getting bogged down in details.
- The 5 numbers in question are the 10th, 25th, 50th, 75th and 90th percentiles.
- In mathematical notation, these are denoted as follows: P10, P25, P50, P75, and P90.





5-Number Summary

- P50 is the 50th Percentile, the score that divides the set of scores into two halves; in this sense, it is a middle score and is commonly called the Median.
- The 25th and 75th Percentiles have an obvious meaning, and noting how far they lie from the Median tells us how spread out the distribution of scores is. P25 and P75 are known by their synonyms as well: Q1, the First Quartile, and Q3, the Third Quartile. By what other name do you think the second quartile, Q2, is known?



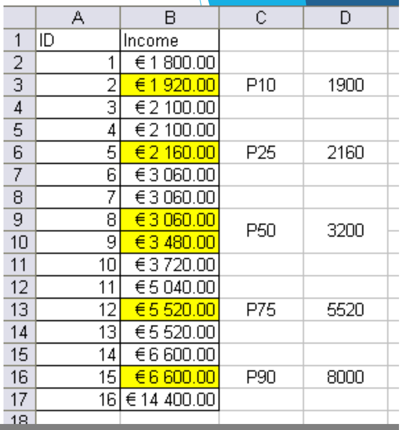
5-Number Summary

- Between P10 and P90 lies the middle 80% of all the scores, or all but the 10% highest and 10% lowest.
- These five numbers together, then, give a pretty informative description of the set of scores, without distracting us with too many details that may not be informative or stable. We call them the "5-Number Summary" of a distribution.

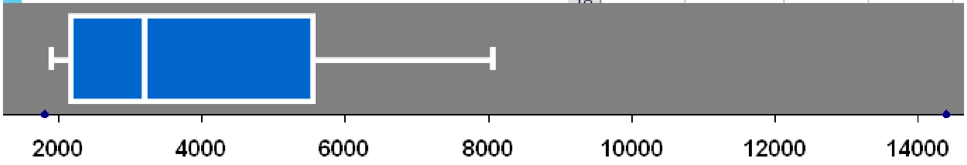
Box and Whisker Plot results

- We establish a ruled line horizontally across the page and mark off the full range of scores that we see in the set of scores we are analyzing.
- Then we draw a rectangle above the ruled scale such that the right edge is above the point on the scale corresponding to P75 and the left edge of the rectangle is above the 25th Percentile.
- We draw a line inside the rectangle at the Median. Then we draw "whiskers” that extend out from the sides of the rectangle in each direction until they reach the 10th and 90th Percentiles.

Box and Whisker Plot results



	A	B	C	D
1	ID	Income		
2	1	€ 1 800.00		
3	2	€ 1 920.00	P10	1900
4	3	€ 2 100.00		
5	4	€ 2 100.00		
6	5	€ 2 160.00	P25	2160
7	6	€ 3 060.00		
8	7	€ 3 060.00		
9	8	€ 3 060.00	P50	3200
10	9	€ 3 480.00		
11	10	€ 3 720.00		
12	11	€ 5 040.00		
13	12	€ 5 520.00	P75	5520
14	13	€ 5 520.00		
15	14	€ 6 600.00		
16	15	€ 6 600.00	P90	8000
17	16	€ 14 400.00		
18				


2000
4000
6000
8000
10000
12000
14000