



# Lecture 1

Descriptive Statistics  
Population & Sample  
Descriptive measures quartiles  
Percentiles & box plots

# Lecture 1

- Statistics
- Descriptive Statistics
- Statistical Inference
- Population vs Sample
- Frequency Distributions
- Cumulative Distributions
- Sample Mean
- Sample Median
- Deviations from Mean
- Variance
- Standard Deviation
- Quartiles
- Percentiles
- Box Plots



# Statistics

What is statistics?

**Statistics** is the study and manipulation of data, including ways to gather, review, analyze, and draw conclusions from data.

Why study statistics?

Answers provided by statistical analysis can provide the basis for making better decisions and choices of actions. Statistical reasoning and methods can help you become efficient at obtaining information and making useful conclusions.



# Descriptive Statistics

Interest in the field grew in 18th century

At first, descriptive statistics consisted merely of the presentation of data in tables and charts.

Nowadays, it includes the summarization of data by means of numerical descriptions and graphs.



# Statistical Inference

Statistical inference is concerned with generalizations based on sample data.

When making a statistical inference always proceed with caution.

One must decide carefully how far to go in generalizing from a given set of data.


Careful consideration must be given to determining whether such generalizations are reasonable and whether it might be wise to collect more data.




# Population Vs Sample

Population	Sample
A population is the collection of all items of interest to our study.	A sample is a subset of the population. It is representative of the population.
The measurable characteristic of the population like the mean or standard deviation is known as the parameter.	The measurable characteristic of the sample is called a statistic.
A survey done of an entire population is accurate and more precise with no margin of error except human inaccuracy in responses. However, this may not be possible always.	A survey done using a sample of the population bears accurate results, only after further factoring the margin of error and confidence interval.
All the students in the class are population.	All the students who regularly attend class is a sample.





Parameter name	Population parameter symbol	Sample statistic
Number of cases	N	n
Mean	$\mu$ (mu)	$\bar{x}$ (Sample mean)
Proportion	$\pi$ (Pi)	P (Sample proportion)
Variance	$\sigma^2$ (Sigma-square)	$s^2$ (Sample variance)
Standard deviation	$\sigma$ (Sigma)	s (sample standard deviation)
Correlation	$\rho$ (rho)	r (Sample correlation)
Regression Coefficient	$\beta$ (beta)	b (sample regression coefficient)



# Frequency Distributions

A frequency distribution is a table that divides a set of data into a suitable number of classes (categories), showing also the number of items belonging to each class.

The table sacrifices some of the information contained in the data.

Instead of knowing the exact value of each item, we only know that it belongs to a certain class.





# Example

## Data

245 333 296 304 276 336 289 234 253 292 366 323 309 284 310 338 297 314 305 330 266 391 315 305 290 300  
292 311 272 312 315 355 346 337 303 265 278 276 373 271 308 276 364 390 298 290 308 221 274 343

Height (nm)	Frequency
(205, 245]	3
(245, 285]	11
(285, 325]	23
(325, 365]	9
(365, 405]	4
Total	50

(205,245]

Note that the class limits are given to as many decimal places as the original data. Had the original data been given to one decimal place, we would have used the class limits 205.1–245.0, 245.1–285.0, ..., 365.1–405.0.



# Class Mark and Class Interval

**Class Mark:** The class marks of a frequency distribution are obtained by averaging successive class boundaries.

**Class Interval:** If the classes of a distribution are all of equal length then subtraction the lower limit from the upper limit gives the class interval.

Class mark: 225, 265, 305, 345, 385

Class Interval: 40



# Cumulative Distribution(less than or equal to variant)

Intervals	Cumulative Frequency
(205,245]	3
(245,285]	14
(285,325]	37
(325,365]	46
(365,405]	50



# Descriptive Measures: Sample Mean

N measurements/data points

$$x_1, x_2, \dots, x_i, \dots, x_n$$

Mean : the sum of the observations divided by sample size.

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



# Descriptive Measures: Sample Median

Order the  $n$  observations from smallest to largest.

sample median = observation in position  $\frac{n+1}{2}$ ,

if  $n$  odd.

= average of two observations in

positions  $\frac{n}{2}$  and  $\frac{n+2}{2}$ ,

if  $n$  even.



# When median is used over mean sometimes?

Sometimes it is preferable to use the sample median as a descriptive measure of the center, or location, of a set of data.

This is particularly true if it is desired to minimize the calculations

or

If it is desired to eliminate the effect of extreme (very large or very small) values.



# Question

A sample of five university students responded to the question “How much time, in minutes, did you spend on the social network site yesterday?”

100 45 60 130 30 35

Find the mean and median.



# Question

A sample of five university students responded to the question “How much time, in minutes, did you spend on the social network site yesterday?”

100 45 60 130 30 35

Find the mean and median.

Mean: 66.67

Median: 52.5





# Descriptive Measures: Deviations from Mean

If a set of numbers  $x_1, x_2, \dots, x_n$  has mean  $\bar{x}$ , the differences

$$x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x}$$

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

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$



# Descriptive Measures: Deviations from Mean

Data: 1 2 3 4 5                      Mean 3

Data: -7 -3 3 10 12                  Mean 3

We observe that the dispersion of a set of data is small if the values are closely bunched about their mean, and that it is large if the values are scattered widely about their mean.

It would seem reasonable, therefore, to measure the variation of a set of data in terms of the amounts by which the values deviate from their mean.



# Descriptive Measures: Deviations from Mean

The sum of the deviations about mean is always zero.

Because the deviations sum to zero, we need to remove their signs. Absolute value and square are two natural choices.

If we take their absolute value, so each negative deviation is treated as positive, we would obtain a measure of variation.

However, to obtain the most common measure of variation, we square each deviation.



# Descriptive Measures: Variance

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

Reason for dividing by  $n-1$  instead of  $n$  is that there are only  $n-1$  independent deviations  $x_i - \bar{x}$ .

Because their sum is always zero, the value of any particular one is always equal to the negative of the sum of the other  $n - 1$  deviations.

If many of the deviations are large in magnitude, either positive or negative, their squares will be large and  $s^2$  will be large. When all the deviations are small,  $s^2$  will be small.



# Example

The delay times (handling, setting, and positioning the tools) for cutting 6 parts on an engine lathe are 0.6, 1.2, 0.9, 1.0, 0.6, and 0.8 minutes. Calculate  $s^2$ .

$$\bar{x} = \frac{0.6 + 1.2 + 0.9 + 1.0 + 0.6 + 0.8}{6} = 0.85$$

$$s^2 = \frac{0.2750}{5} = 0.055 \text{ (minute)}^2$$

$x_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0.6	-0.25	0.0625
1.2	0.35	0.1225
0.9	0.05	0.0025
1.0	0.15	0.0225
0.6	-0.25	0.0625
0.8	-0.05	0.0025
5.1	0.00	0.2750



# Descriptive Measures: Standard Deviation

Notice that the units of  $s^2$  are not those of the original observations.

In previous question the data are delay times in minutes, but  $s^2$  has the unit (minute)<sup>2</sup>

Consequently, we define the standard deviation of  $n$  observations  $x_1, x_2, \dots, x_n$  as the square root of their variance.

The standard deviation is by far the most generally useful measure of variation. Its advantage over the variance is that it is expressed in the same units as the observations.

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



# Descriptive Measures: Quartiles

In addition to the median, which divides a set of data into halves, we can consider other division points.

When an ordered data set is divided into quarters, the resulting division points are called sample quartiles.

The first quartile,  $Q_1$ , is a value that has one-fourth, or 25%, of the observations below its value. The first quartile is also the sample 25th percentile  $P_{0.25}$ .



# Descriptive Measures: Percentile

More generally, we define the sample 100  $p$ th percentile as :

The sample 100  $p$ th percentile is a value such that at least 100 $p$ % of the observations are at or below this value, and at least 100(1 –  $p$ )% are at or above this value.

Calculating the sample 100  $p$ th percentile:

1. Order the  $n$  observations from smallest to largest.
2. Determine the product  $np$ .

If  $np$  is not an integer, round it up to the next integer and find the corresponding ordered value.

If  $np$  is an integer, say  $k$ , calculate the mean of the  $k$ th and  $(k + 1)$ st ordered observations.





# Descriptive Measures: Percentile

<b>first quartile</b>	$Q_1 = 25\text{th percentile}$
<b>second quartile</b>	$Q_2 = 50\text{th percentile}$
<b>third quartile</b>	$Q_3 = 75\text{th percentile}$

Calculating the sample 100  $p$ th percentile:

1. Order the  $n$  observations from smallest to largest.
2. Determine the product  $np$ .

If  $np$  is not an integer, round it up to the next integer and find the corresponding ordered value.

If  $np$  is an integer, say  $k$ , calculate the mean of the  $k$ th and  $(k + 1)$ st ordered observations.



# Question

Given the data

136 143 147 151 158 160 161 163 165 167 173 174 181 181 185 188 190 205

Obtain the quartiles and the 10th percentile.

$n = 18$

First quartile:  $18 \times (0.25) = 4.5$  (round up to 5)

$Q_1 = 5\text{th observation} = 158$

Number of observations below or equal to 158 = 5 (atleast 4.5 required acc to definition)

Number of observations equal to or above 158 = 14 (atleast 13.5 required acc to definition)



# Question

Given the data

136 143 147 151 158 160 161 163 165 167 173 174 181 181 185 188 190 205

Obtain the quartiles and the 10th percentile.

$n = 18$

Second:  $18 \cdot (0.5) = 9$  Therefore, we average the 9th and 10th ordered values

$Q_2 = \text{average the 9th and 10th ordered values} = (165+167)/2 = 166$

$Q_3 = 181$   $P_{0.10} = 143$



# Descriptive Measures: Range & Interquartile Range

The minimum and maximum observations also convey information concerning the amount of variability present in a set of data. Together, they describe the interval containing all of the observed values.

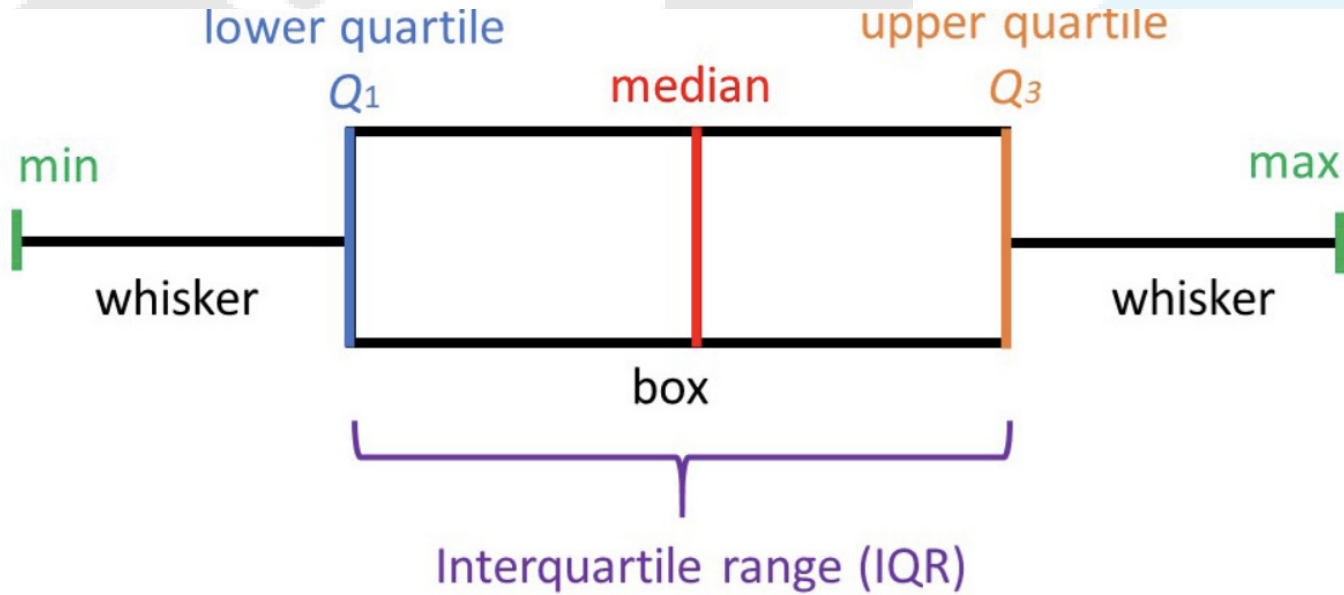
range = maximum – minimum

The amount of variation in the middle half of the data is described by the interquartile range.

interquartile range = third quartile – first quartile =  $Q_3 - Q_1$



# Descriptive Measures: Box Plots



# References

Probability and statistics for engineers RA Johnson, I Miller, JE Freund - 2000 - 117.239.47.98

Statistics for business & economics DR Anderson, DJ Sweeney, TA Williams, JD Camm

Probability and statistics for engineering and science J Deovre

