Sta2023 Concepts summary. Examples.

csotuyo@ imathesis.com

Statistics → Data

Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

Population → **Sample**

Population mean is a parameter: μ

Sample mean, Statistic: $\bar{x}$

Descriptive statistics consists of the collection, organization, summarization, and presentation of data.
Inferential statistics consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.

Level of Measurement

Data

- Quantitative (numerical)
  - Continuous (measurable) — Ratio, interval
  - Discrete (countable)
- Qualitative or Categorical (names or labels) — Nominal, Ordinal

Collecting data:

Simple random sample: ***...****every possible sample of size n, has the same chance of being chosen...*

Sampling methods:

For Stratified and Cluster sampling the population is subdivided into subgroups.

- Systematics sample: *the kth element.*
- Convenience sample: easy to obtain.
- *Stratified sample: a few subjects form each subgroup.*
- *Cluster sample: select some subgroups, measure all subjects in those subgroups.*

Type of statistical studies: 1. Observational studies: *observe, measure. Do not modify.*
2. Experimental Studies: *Modify. Control. Treatment.*

Organizing and summarizing data:

McDonald's lunch service time:

Time (secs)	Frequency
75-124	11
125-174	24
175-224	10
225-274	3
275-324	2

Time (secs)	Frequency	Relative freq.
75-124	11	0.22
125-174	24	0.48
175-224	10	0.20
225-274	3	0.06
275-324	2	0.04
Total	50	

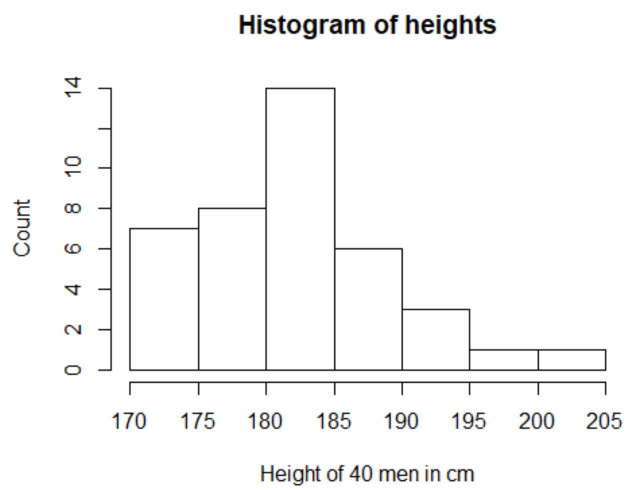
Time (secs)	Cumulative freq.
Less than 125	11
Less than 175	35
Less than 225	45
Less than 275	48
Less than 325	50

For the first class: 75-124:
*Class limits: lower limit, 75; upper limit, 124
*Class width: difference between two consecutive class lower limits; 125-75 = 50.
*Class midpoint: value in the middle: $\frac{75+124}{2} = 99.5$
*Class boundaries; values that separate the classes: 74.5 and 124.5

Heights of 40 human males in cm:

187,171,181,180,178,171,174,177,172,178,182,187, 176,179,190,185,192,184,182,178,187,173,185,184, 184,183,185,197,202,181,181,191,178,187,185,186, 174,174,182,195.

Height (cm)	Frq (count)
170-174	7
175-179	8
180-184	14
185-189	6
190-194	3
195-199	1
200-205	1

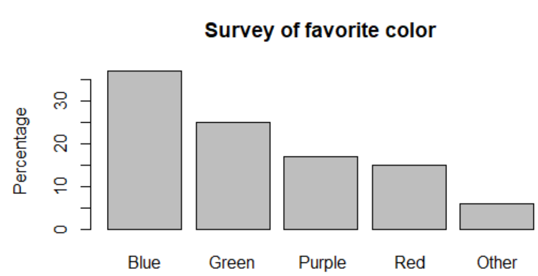


Histogram: bar plot, no gaps. Visually displays the shape of the distribution of the data.

Barplot for Categorical data:

Survey: what is your favorite color:

Colors	%
Blue	37
Green	25
Purple	17
Red	15
Other	6



Stem-leaf-plot:

Dataset (two digits numbers): 12, 23, 19, 16, 10, 17, 15, 25, 21, 12, 30, 32, 45.

The decimal point is 1 digit(s) to the right of the |

```
1 | 0225679
2 | 135
3 | 02
4 | 5
```

Stem-leaf-plot of Heights of 40 human males in cm:

```
17 | 11234446788889
18 | 01112223444555567777
19 | 01257
20 | 2
```

Standard deviation for samples:

Formula:

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

Shortcut formula:

$$s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n-1)}}$$

Standard deviation. Example.
Set of values: 0, 4, 6, 8, 10.

x	$\bar{x}$	$(x-\bar{x})^2$	=
0	5.6	$(0-5.6)^2$	31.36
4	5.6	$(4-5.6)^2$	2.56
6	5.6	$(6-5.6)^2$	0.16
8	5.6	$(8-5.6)^2$	5.76
10	5.6	$(10-5.6)^2$	19.36
		Sum	59.2

$$s = \sqrt{\frac{59.2}{4}} = 3.847$$

n = 5, there are five data values; $n-1 = 4$.

Shortcut formula. Table:

	x	x^2
	0	0
	4	16
	6	36
	8	64
	10	100
sum	28	216

$$n = 5;\quad n-1 = 4;\quad \sum x = 28;\quad \sum x^2 = 216$$

Using the shortcut formula:

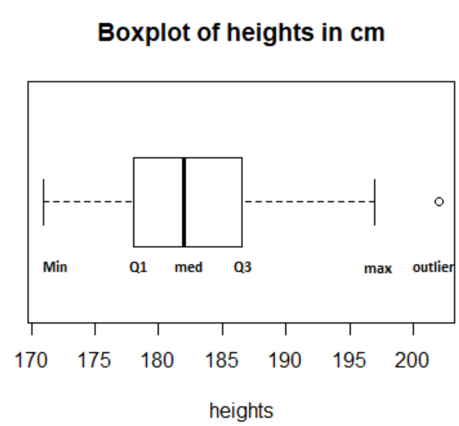
$$s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n-1)}}$$

$$s = \sqrt{\frac{5(216) - (28)^2}{5(5-1)}} = 3.847$$

Heights of 40 human males in cm:
187,171,181,180,178,171,174,177,172,178,182,187, 176,179,190,185,192,184,182,178,187,173,185,184, 184,183,185,197,202,181,181,191,178,187,185,186, 174,174,182,195.

Five data summary:

Min.	Q1	Median	Q3	Max.
171.0	178.0	182.0	186.2	202.0

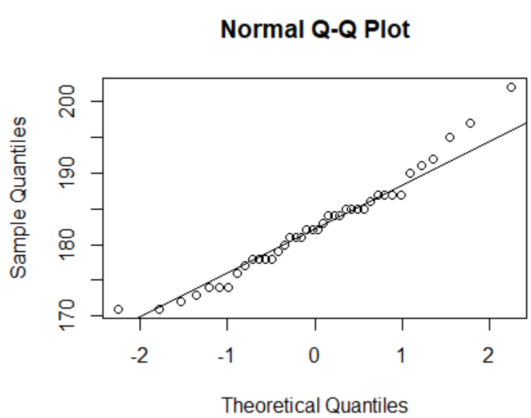


Z scores, sample data:

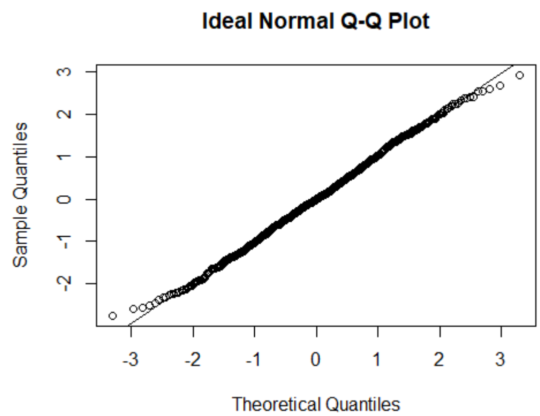
$$z = \frac{x - \bar{x}}{s}$$

```
Sort the dataset height of 40 human
males:
171 171 172 173 174 174 174 176 177
178 178 178 178 179 180 181 181 181
182 182 182 183 184 184 184 185 185
185 185 186 187 187 187 187 190 191
192 195 197 202

Mean: 182.45; sd=7.074512

The sample quantiles are: (first
five values and z scores):
1 171 -1.62
2 171 -1.62
3 172 -1.48
4 173 -1.34
5 174 -1.19

Last five values and z scores:
36 191 1.21
37 192 1.35
38 195 1.77
39 197 2.06
40 202 2.76
```

A normal qq plot is a plot of the sample quantiles versus the theoretical quantiles. If the data is normally distributed, the points will fall on the 45-degree reference line. If the data is not normally distributed, the points will deviate from the reference line.



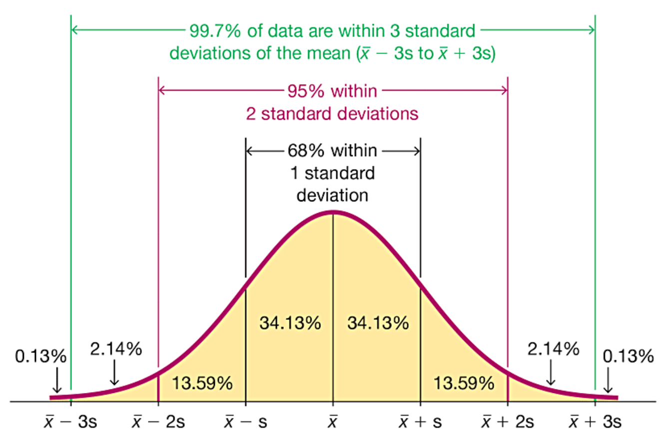
The normal qq plot is just one method to assess normality. No dataset is ideally normal. The following graph shows what we may accept as close to *ideally* normal.



The Empirical Rule

The empirical rule, also referred to as the three-sigma rule or 68-95-99.7 rule, is a statistical rule which states that for a normal distribution, almost all data falls within three standard deviations (denoted by σ) of the mean (denoted by µ). Broken down, the empirical rule shows that 68% falls within the first standard deviation (µ ± σ), 95% within the first two standard deviations (µ ± 2σ), and 99.7% within the first three standard deviations (µ ± 3σ).
(reference: Investopedia)

The Empirical rule, graphically:



Graph by Pearson Education, 2014.