Parametric Statistics

Week 1 - Descriptive Statistics

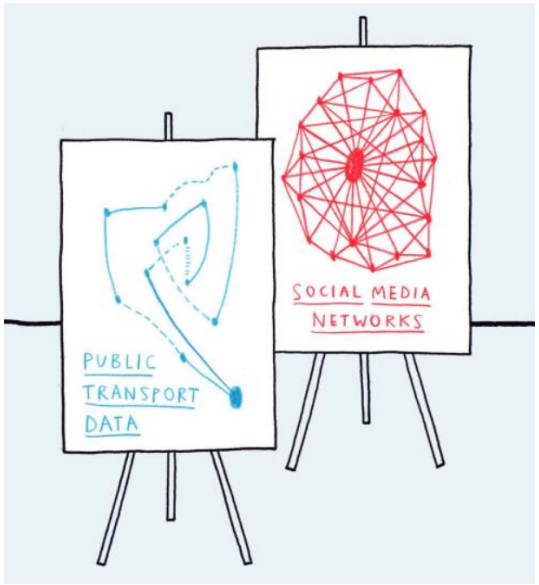
Juan Carlos Medina Serrano

Technische Universität München
Hochschule für Politik
Political Data Science

Munich, 16. October 2019

political data science

https://politicaldatascience.blogspot.de



Statistics

Statistics: Describes a set of tools and techniques to describe, organize and interpret information or data.

Descriptive Statistics: Organize and describe the characteristics of a collection of data.

Inferential Statistics: Make inferences based on a smaller group of data about a larger one. A smaller group is called a *sample*, which is a subset of a *population*.

Basis of Data Science, Machine Learning and most of Artificial Intelligence!

Parameters and Estimates

Parametric Statistics: Models the data with a fixed set of parameters.

Non-Parametric Statistics: Models the data with a set of parameters that are not fixed.

We denote any *parameter* as θ.

We normally dont know the true value of the parameter...we need to estimate it!

We denote an *estimator* of the parameter θ as $\hat{\theta}$

- Estimates of location: mean, mode, median
- Estimates of variance: range, variance, standard deviation.

Finding the Center: The Mean

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example Number of Facebook friends.
Data = {108, 103, 252, 121, 93, 57, 40, 53, 22, 116, 98}

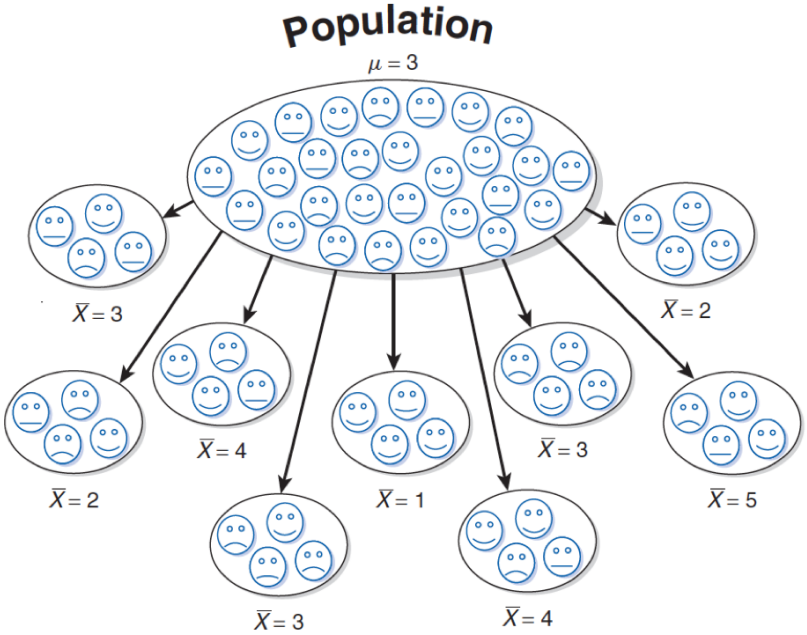
$$\sum_{i=1}^{n} x_i = 22 + 40 + 53 + 57 + 93 + 98 + 103 + 108 + 116 + 121 + 253 = 1063$$

$$\overline{X} = \frac{1063}{11} = 96.64$$

Notation: $\overline{X}$: Mean of a sample, μ : Mean of the population

- Disadvantage: It can be influenced by extreme scores
- Advantages: Takes all values in consideration, stable in different samples.

Finding the Center: The Mean



In this case, $\theta : \mu$ and $\hat{\theta} : \overline{X}$

Finding the Center: The Median

The middle value when the data is ranked in order of magnitude.

$$22, 40, 53, 57, 93, \mathbf{98}, 103, 108, 116, 121, 252$$

With even number of data, the median is the average of the two middle values

$$22, 40, 53, 57, \textit{93, 98}, 103, 108, 116, 121$$

$$\frac{93 + 98}{2} = 95.5$$

- Advantage: Unaffected by extreme scores (**ROBUST**) and *skewed distributions*.

Finding the Center: The Mode

Value that occurs most frequently in the data set

- Disadvantage: Can often take on several values.

Finding the Center: The Weighted Mean

Each value has now a frequency attached to it. Multiply value by frequency and then calculate the mean.

Example Test results for 100 Ryanair pilots.

Value	Frequency	Value × Frequency
97	4	388
94	11	1,034
92	12	1,104
91	21	1,911
90	30	2,700
89	12	1,068
78	9	702
60 (Don't fly with this guy.)	1	60
Total	100	8,967

Mode: 90

Weighted mean : 8,967 / 100 = 89.67

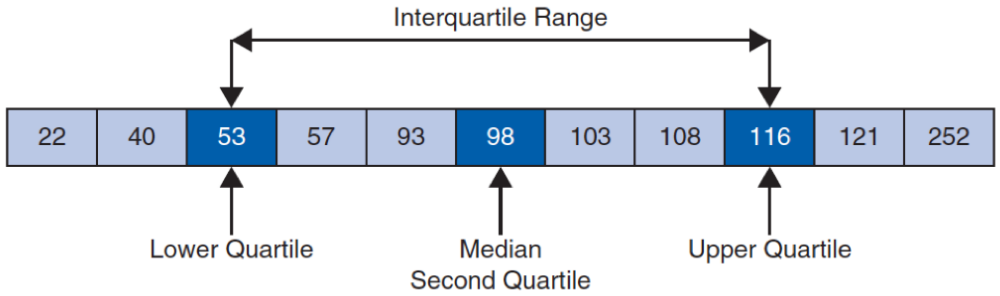
Variability of the Data

$\{3,4,4,5,4\}$ and $\{4,4,4,4,4\}$ have mean $= 4$

Variability: How different scores are from the mean.

Variability of the Data: Range and Quartiles

Range: Largest value minus the smallest value.
Quartiles: Three values that split the sorted data into four equal parts.
25% : Lower quartile
50% : Middle quartile / Median
75% : Upper quartile
Interquartile range: From the lower to upper quartile



Variability of the Data: Standard Deviation

Average amount of variability in a data set. In practical terms, it's the average distance from the mean. The larger the standard deviation, the larger the average distance each data point is from the mean.

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

Example Data $= \{5, 8, 5, 4, 6, 7, 8, 8, 3, 6\}$ $s = \sqrt{\frac{28}{9}} = 1.76$

X	$(X - \bar{X})$	$(X - \bar{X})^2$
8	+2	4
8	+2	4
8	+2	4
7	+1	1
6	0	0
6	0	0
5	−1	1
5	−1	1
4	−2	4
3	−3	9
Sum	0	28

Variability of the Data: Variance

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

Same formula as the standard deviation but without the square root bracket.

The standard deviation is stated in the original units from which it was derived (more logical and useful for description). The variance is stated in units that are squared (useful for statistical models).

- Disadvantage: Both are sensitive to extreme values, whereas quartiles are **robust**.

The numerator is called **Sum of squares**:

$$SS = \sum(x - \overline{X})^2$$

What about the denominator?

Why n-1???

Quick answer: Using only n is a biased estimator, with n - 1, the estimator is unbiased.

What???
s is an estimate of the population standard deviation
With n - 1, we force the standard deviation to be larger than it would be

Sample Size	Value of Numerator in Standard Deviation Formula	Biased Estimate of the Population Standard Deviation (dividing by *n*)	Unbiased Estimate of the Population Standard Deviation (dividing by *n* – 1)	Difference Between Biased and Unbiased Estimates
10	500	7.07	7.45	0.38
100	500	2.24	2.25	0.01
1,000	500	0.7071	0.7075	0.0004

Why n-1???

Quick answer: Using only n is a biased estimator, with n - 1, the estimator is unbiased.

What???
s is an estimate of the population standard deviation
With n - 1, we force the standard deviation to be larger than it would be

It is exactly n - 1, since they are the number of **Degrees of Freedom** (Number of observations that are free to vary). The mean of the sample is already used as an estimator of the population's mean. We can only vary n - 1 observations.

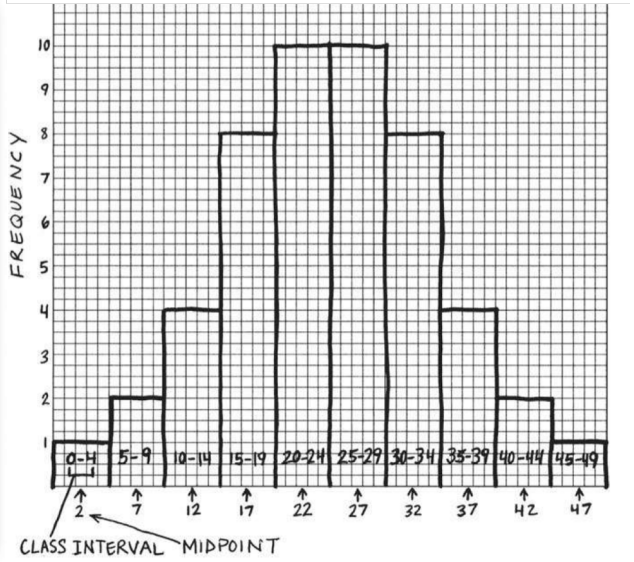
Loses importance for big datasets

Frequency Distributions

How many times each score occurs in a dataset?

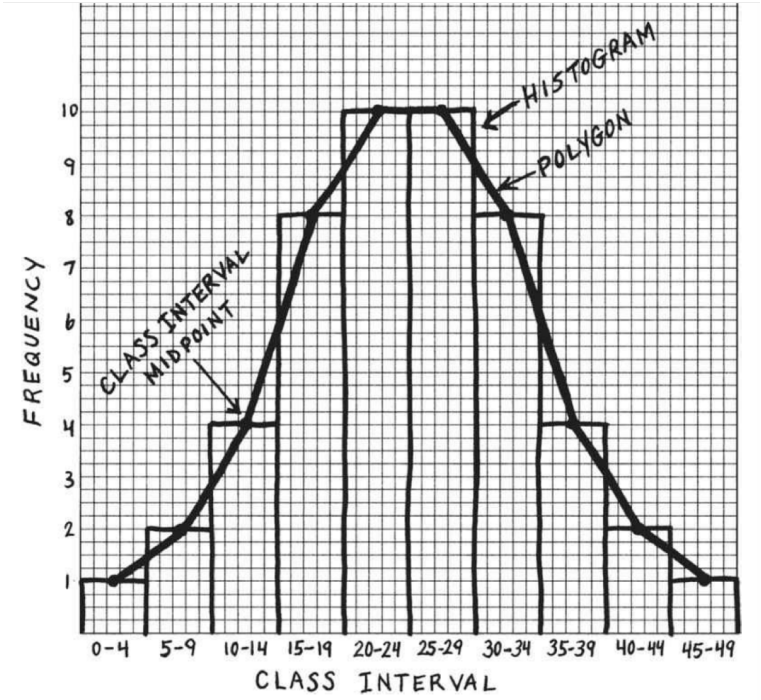
Histogram: Observations on the horizontal axis, with a bar showing how many times each value occurred in the data set.

Class Interval	Frequency
45–49	1
40–44	2
35–39	4
30–34	8
25–29	10
20–24	10
15–19	8
10–14	4
5–9	2
0–4	1



Frequency Distributions

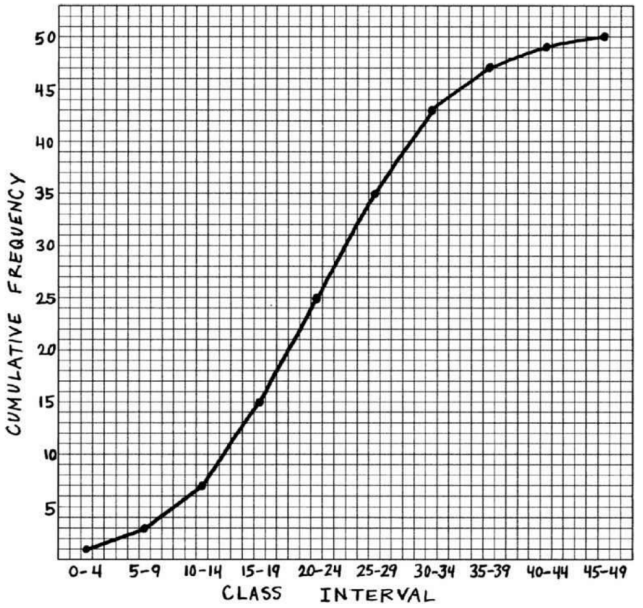
Frequency Polygon



Cumulative Distributions

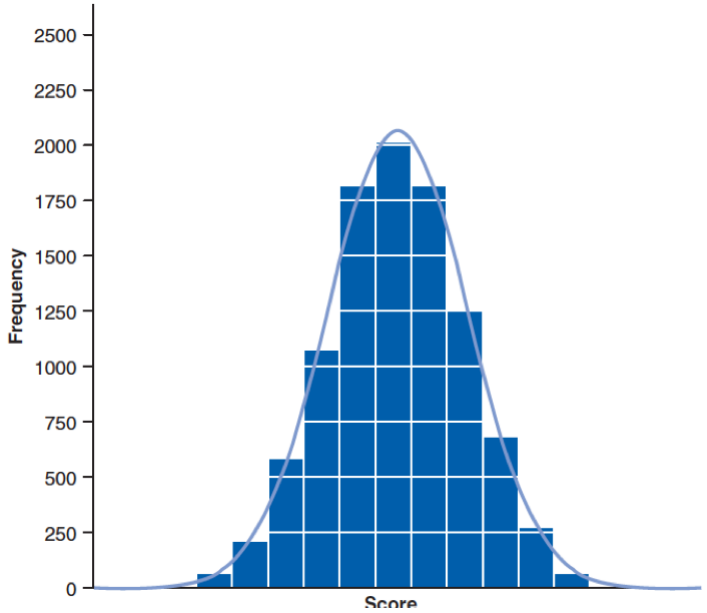
Very important to understand the difference between frequency and cumulative distributions!

Class Interval	Frequency	Cumulative Frequency
45–49	1	50
40–44	2	49
35–39	4	47
30–34	8	43
25–29	10	35
20–24	10	25
15–19	8	15
10–14	4	7
5–9	2	3
0–4	1	1



The Normal Distribution

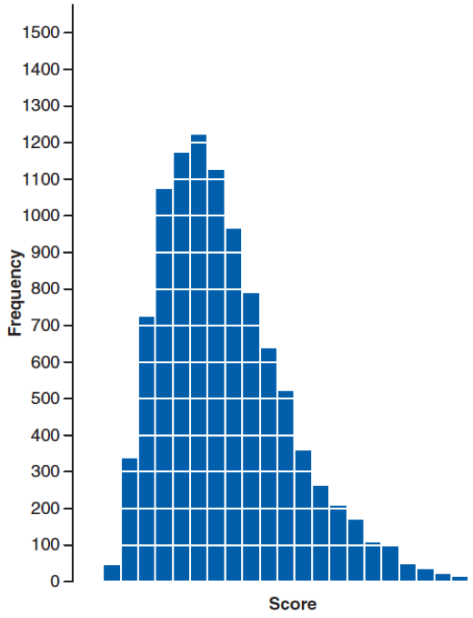
The most common distribution! Characterized by the bell-shaped curve.
The majority of the data lie around the centre of the distribution and as we get further away from the centre the frequency decreases.

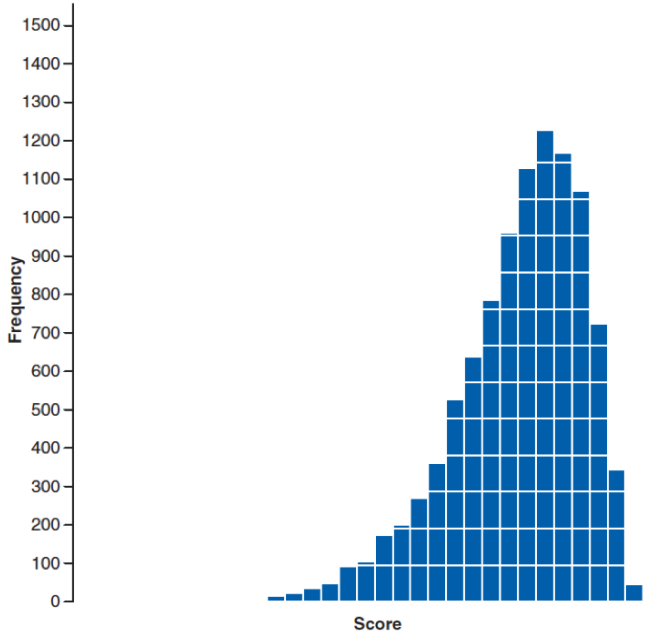


Next week we review its properties

Distributions that deviate from normal: Skew Distributions

Skewed distributions are not symmetrical and instead the most frequent values are clustered at one end of the scale. Positively (right) and negatively (left) skewed:



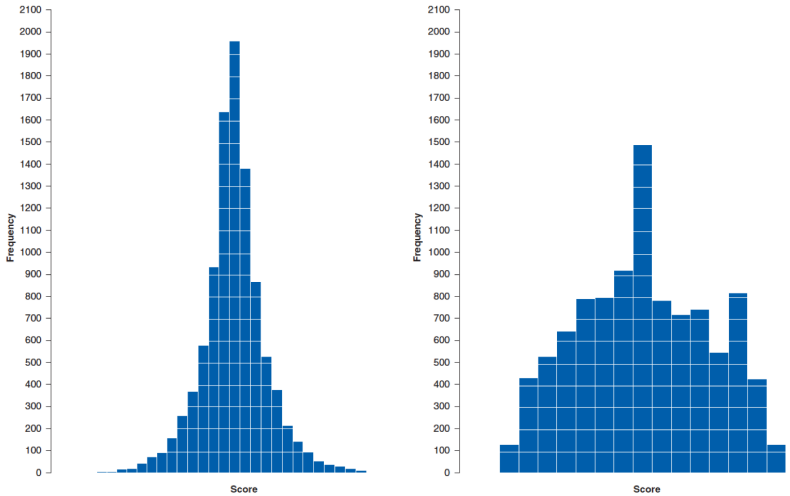


Distributions that deviate from normal: Kurtosis

What a weird name...for describing pointyness
Refers to the degree to which the data cluster at the ends of the distribution (known as the *tails*) and how pointy a distribution is.

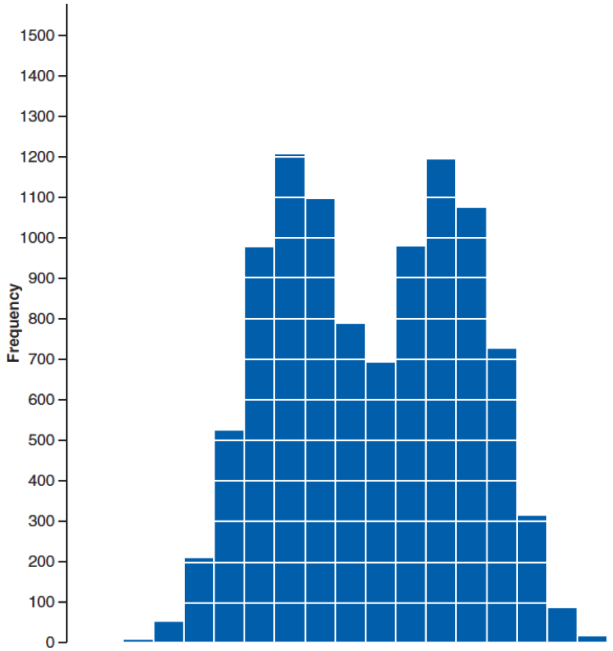
- Positive Kurtosis: Many values in the tails (heavy tails) and is pointy
- Negative Kurtosis: Thin tails and is flatter than the normal

Positive (right) and negative (left) kurtosis:



Distributions that deviate from normal: Bimodal

There are many other distributions that deviate from normality! For Example Bimodal



Exploring the Data: Data Types

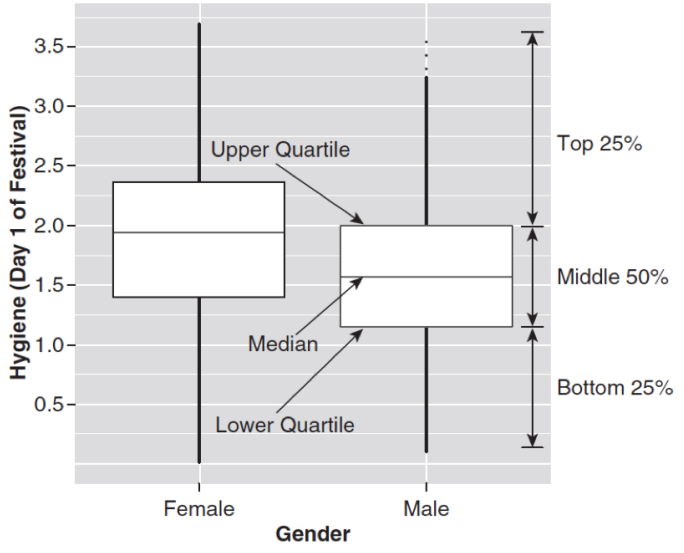
- Categorical - Can take only a specific set of values representing categories.
  - Binary - Just two categorical values.
  - Nominal - More than two values, no order
  - Ordinal - Data has explicit ordering.
- Discrete - Can only take integer values, such as counts.
- Continuous - Can take any value in an interval.
  - Interval - Equal intervals represent equal differences in the measured property
  - Ratio - Same as interval, but also requires the ratio of values to be meaningful.

Examples?

Exploring the Data: Visualization

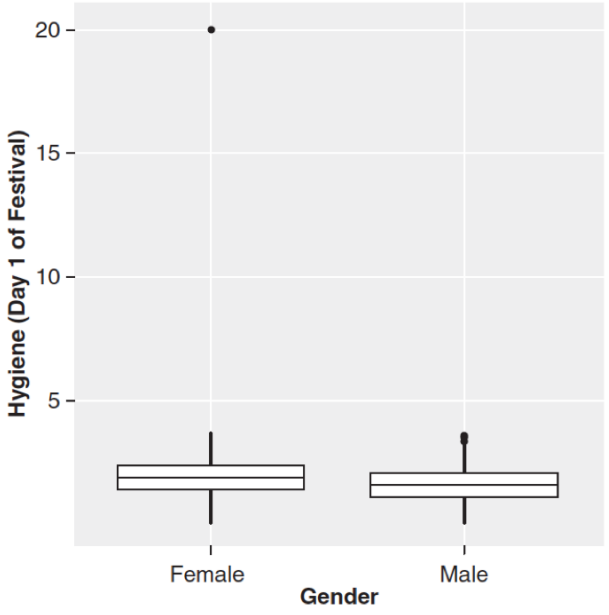
We already met the histogram, we now explore other important visualizations.

Boxplot: Quantiles and data range in a visual representation. The blacklines (whiskers) extend until 1.5 times the Interquartile range.
Example Hygiene score after one day in Tomorrowland



Exploring the Data: Visualization

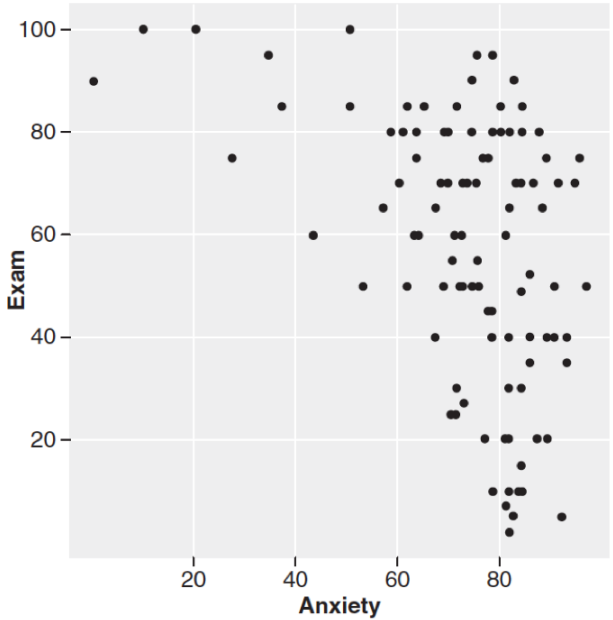
Beware of **outliers**!
Outliers are extreme values. They can either be part of a keyboard error or an interesting property of the data.
Remove them only if they are a mistake.
Outliers are a source of *bias* in statistical models. Robust methods are unaffected by outliers.



Exploring the Data: Visualization

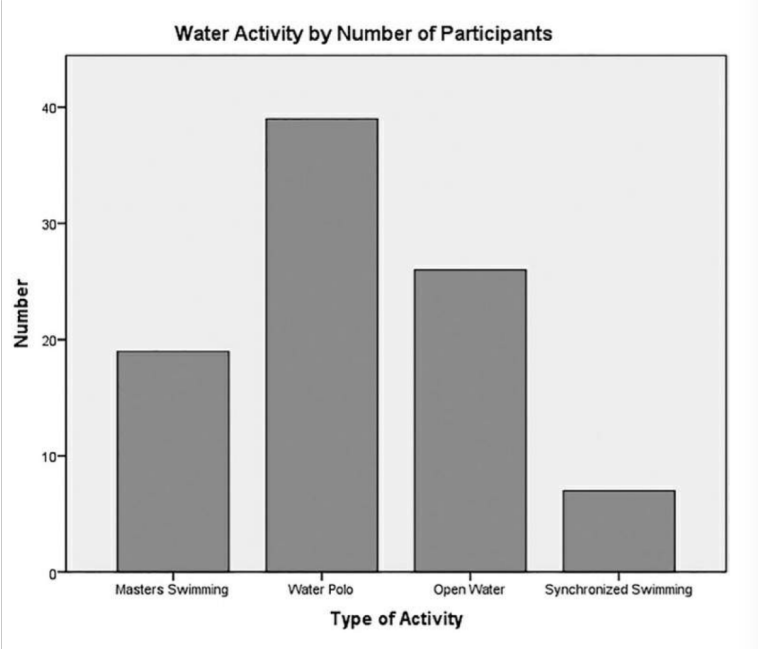
Scatterplot: Looks at the relationships between variables. Are they positively or negatively related? Or no relationship?

Example Exam anxiety vs. Exam result (Each dot is a surveyed person)



Exploring the Data: Visualization

Bar Chart: to visualize categorical data.
Example Number of participants in the water olympics



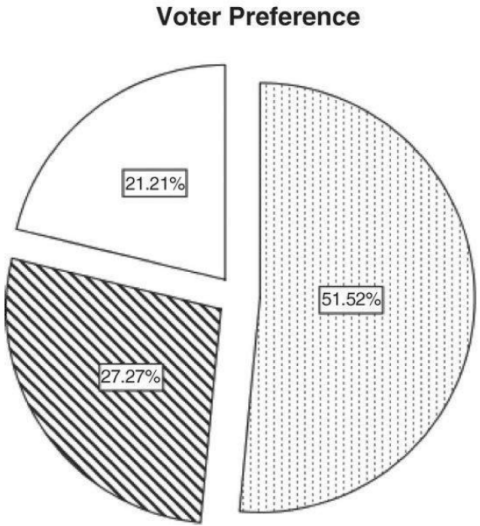
Exploring the Data: Visualization

Pie Chart: Use ONLY to show proportions that sum up to one. (Easier to distort the information...Bar charts are better)

$$Proportion : counts/totalcounts$$

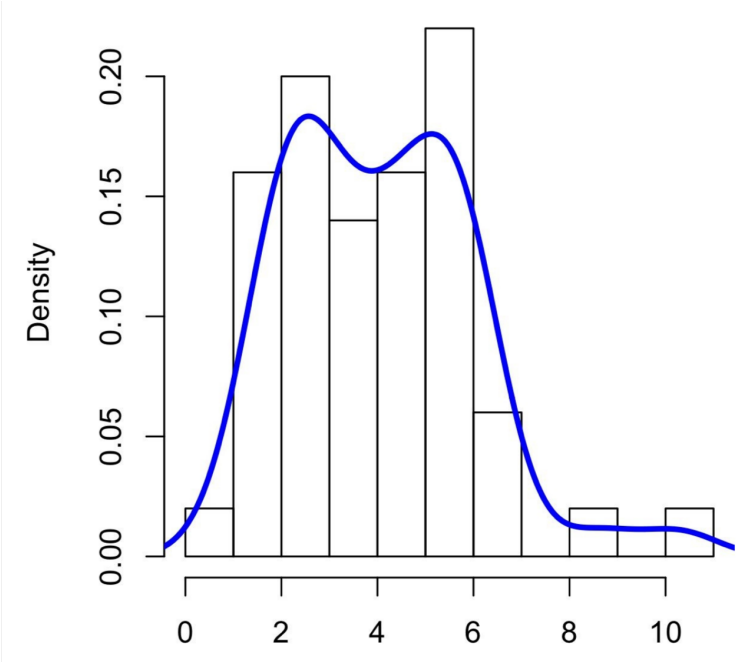
$$Percentage : Proportion * 100$$

Example Election results: Who is the best Professor in the HfP?



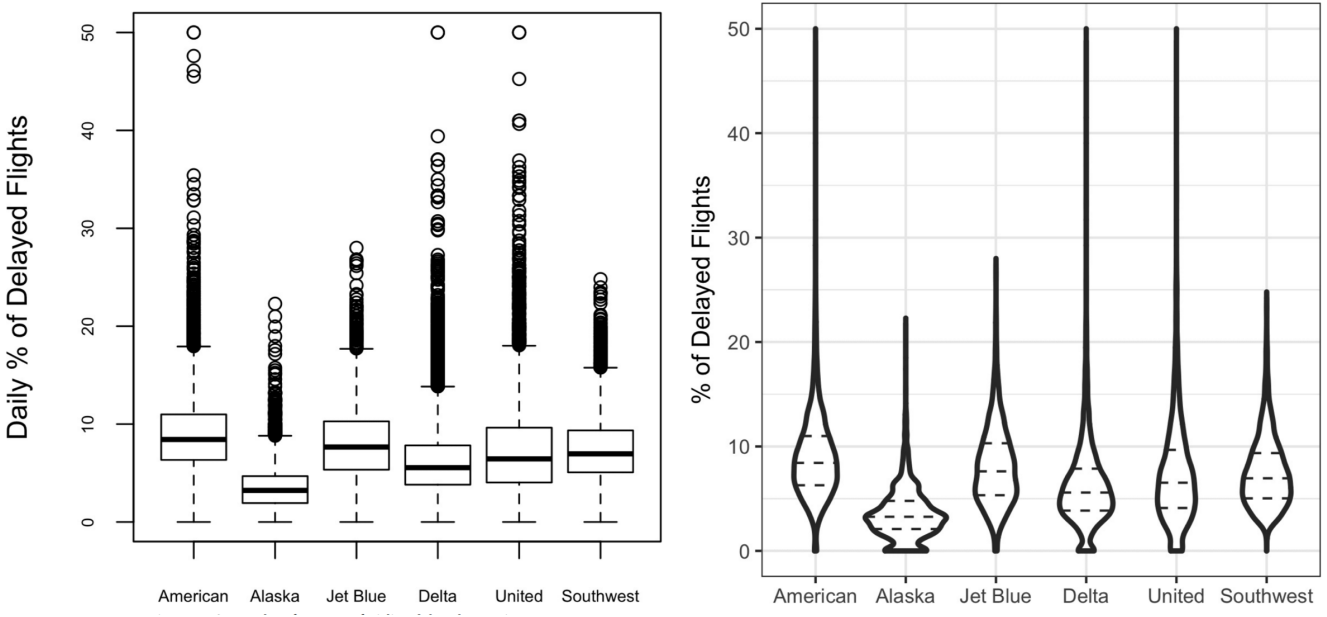
Exploring the Data: Visualization

Density plot: Related to the histogram, except that they smooth the distribution into a line. Uses proportions instead of counts in the y-axis!



Exploring the Data: Visualization

Violin plot: Boxplot + Density plot!
Example Percentage of daily delayed flights in the US



Literature

- Statistics for people who hate statistics : Chapters 2-4
- Discovering Statistics using R : Subchapter 1.7
- Practical Statistics for Data Science: Chapter 1