Quantitative Research Methods

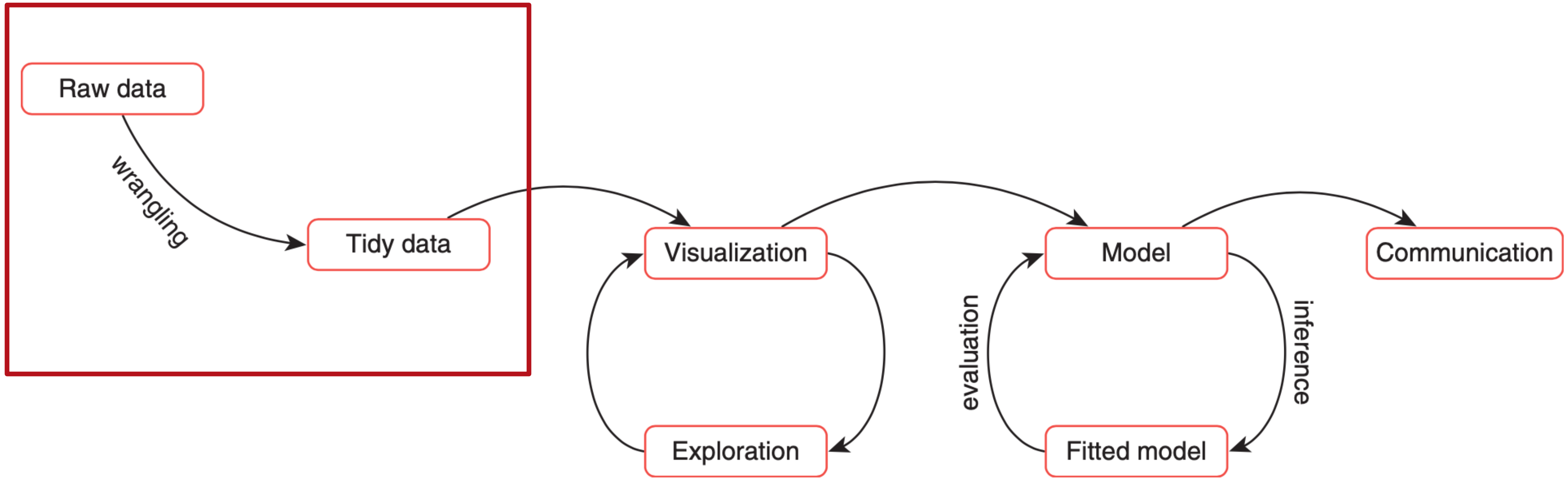
Matthew Ivory

matthew.ivory@lancaster.ac.uk

- Session 1: Introduction to quantitative research methods using R
- Session 2: Data management and data wrangling
- **Session 3: Exploratory data analysis**
- Session 4: Data visualization
- Session 5: Mid-term assignment
- Session 6: Significance tests for continuous variables
- Session 7: Tests for discrete variables: Analysing contingency tables
- Session 8: Correlation and linear regression. Tests for categorical variables.
- Session 9: ANOVA and tests for N groups
- Session 10: Multiple regression

Data wrangling: A reminder

The Data Science Workflow

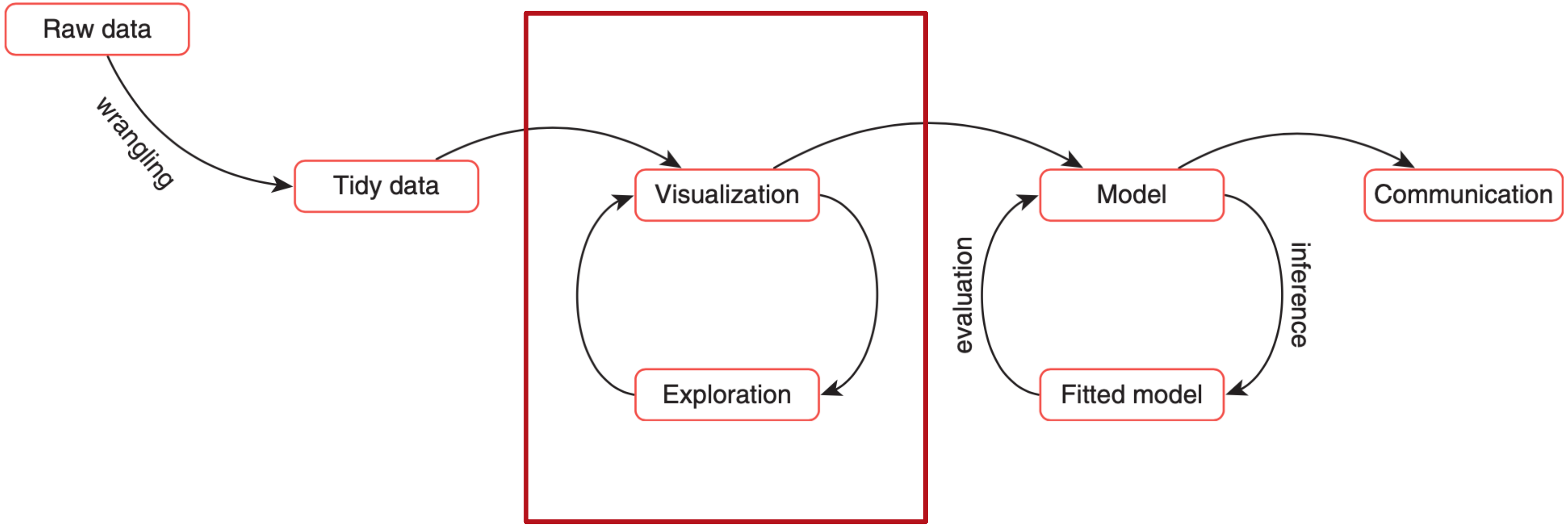


- Data science: the combined application of computational tools and statistical methods to (all aspects of) data analysis.

Reproduced from Andrews (2021)

Exploratory data analysis

The Data Science Workflow



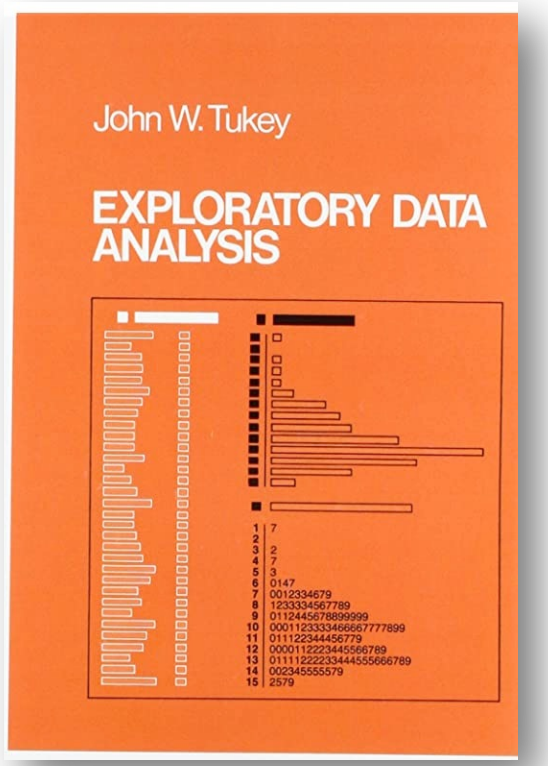
- Data science: the combined application of computational tools and statistical methods to (all aspects of) data analysis.

Reproduced from Andrews (2021)

Tukey (1977)

Exploratory data analysis

- Aim: to discover potentially interesting patterns and behaviors in the data.



Confirmatory data analysis:

- Aim: to propose and test models of our data

Tukey (1977)

Exploratory data analysis

- Detectives looking for evidence at the scene of a crime

Confirmatory data analysis:

- Courts making a prosecution case and evaluating evidence for and against

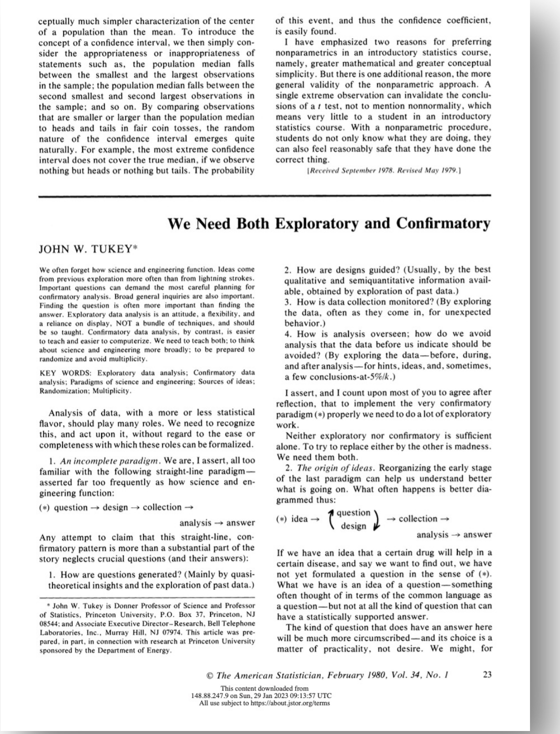
Image reproduced from here.

Tukey (1980)

We need both EDA and CDA.

EDA as an “attitude”

“Exploratory data analysis is **an attitude, a flexibility, and a reliance on display**, NOT a bundle of techniques, and should be so taught. Confirmatory data analysis, by contrast, is easier to teach and easier to computerize.” (Tukey, 1980, p. 23)



ceptually much simpler characterization of the center of a population than the mean. To introduce the concept of a confidence interval, we then simply consider the appropriateness or inappropriateness of statements such as, the population median falls between the smallest and the largest observations in the sample; the population median falls between the second smallest and second largest observations in the sample; and so on. By comparing observations that are smaller or larger than the population median to heads and tails in fair coin tosses, the random nature of the confidence interval emerges quite naturally. For example, the most extreme confidence interval does not cover the true median, if we observe nothing but heads or nothing but tails. The probability of this event, and thus the confidence coefficient, is easily found.

I have emphasized two reasons for preferring nonparametrics in an introductory statistics course, namely, greater mathematical and greater conceptual simplicity. But there is one additional reason, the more general validity of the nonparametric approach. A single extreme observation can invalidate the conclusions of a t test, not to mention nonnormality, which means very little to a student in an introductory statistics course. With a nonparametric procedure, students do not only know what they are doing, they can also feel reasonably safe that they have done the correct thing.

[*Received September 1978. Revised May 1979.*]

We Need Both Exploratory and Confirmatory

JOHN W. TUKEY*

We often forget how science and engineering function. Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. Broad general inquiries are also important. Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught. Confirmatory data analysis, by contrast, is easier to teach and easier to computerize. We need to teach both; to think about science and engineering more broadly; to be prepared to randomize and avoid multiplicity.

KEY WORDS: Exploratory data analysis; Confirmatory data analysis; Paradigms of science and engineering; Sources of ideas; Randomization; Multiplicity.

Analysis of data, with a more or less statistical flavor, should play many roles. We need to recognize this, and act upon it, without regard to the ease or completeness with which these roles can be formalized.

1. *An incomplete paradigm*. We are, I assert, all too familiar with the following straight-line paradigm—asserted far too frequently as how science and engineering function:

(∗) question → design → collection → analysis → answer

Any attempt to claim that this straight-line, confirmatory pattern is more than a substantial part of the story neglects crucial questions (and their answers):

1. How are questions generated? (Mainly by quasi-theoretical insights and the exploration of past data.)
2. How are designs guided? (Usually, by the best qualitative and semiquantitative information available, obtained by exploration of past data.)
3. How is data collection monitored? (By exploring the data, often as they come in, for unexpected behavior.)
4. How is analysis overseen; how do we avoid analysis that the data before us indicate should be avoided? (By exploring the data—before, during, and after analysis—for hints, ideas, and, sometimes, a few conclusions-at-5%/k.)

I assert, and I count upon most of you to agree after reflection, that to implement the very confirmatory paradigm (∗) properly we need to do a lot of exploratory work.

Neither exploratory nor confirmatory is sufficient alone. To try to replace either by the other is madness. We need them both.

2. *The origin of ideas*. Reorganizing the early stage of the last paradigm can help us understand better what is going on. What often happens is better diagrammed thus:

(∗) idea → question / design → collection → analysis → answer

If we have an idea that a certain drug will help in a certain disease, and say we want to find out, we have not yet formulated a question in the sense of (∗). What we have is an idea of a question—something often thought of in terms of the common language as a question—but not at all the kind of question that can have a statistically supported answer.

The kind of question that does have an answer here will be much more circumscribed—and its choice is a matter of practicality, not desire. We might, for

* John W. Tukey is Donner Professor of Science and Professor of Statistics, Princeton University, P.O. Box 37, Princeton, NJ 08544; and Associate Executive Director–Research, Bell Telephone Laboratories, Inc., Murray Hill, NJ 07974. This article was prepared, in part, in connection with research at Princeton University sponsored by the Department of Energy.

Tukey (1977, 1980)

To explore, we can use summary statistics and data visualizations.

Example: **Stem-and-leaf display**

```
stem_example <- c(12, 24, 15, 15, 12, 24, 29, 22, 21, 25, 30,
39, 45, 50, 51)
stem(stem_example)
```

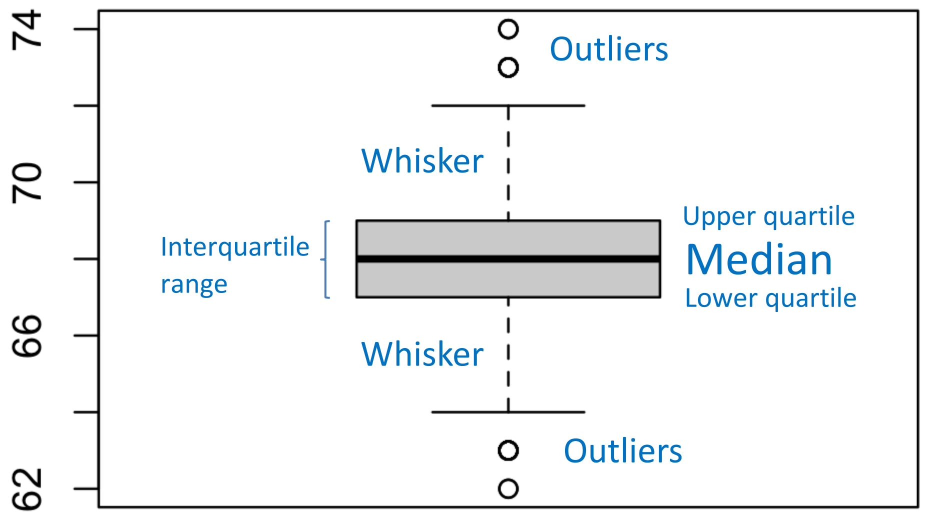
```
  The decimal point is 1 digit(s) to the right of the |

  1 | 2255
  2 | 124459
  3 | 09
  4 | 5
  5 | 01
```

Tukey (1977, 1980)

Example: **Boxplot**

- Upper quantile: 75%
- Lower quantile: 25%
- Median: 50%
- Interquartile range: Range between lower and upper quartile.
- Whisker length: 1.5*IQR from upper and lower quantiles respectively
- Outliers: Extreme values, each bubble represents one observation



Types of data

Types of data

We can classify data types in different ways.

1. Continuous data
2. Ordinal data
3. Count data
4. Categorical data

Continuous data

- Variables can take any value in a continuous metric space.
- Between any two values (e.g., 0 and 100) can exist an infinite number of other values.
- Continuous data can be ordered.
- Example: height, weight, age; speed, time, distance.

Ordinal data

- Values of a variable that can be ordered but there is no natural sense of distance between these values.
- First, second, third… These values have a natural order, but there is no sense of distance between them.
- Example: Students scoring in first, second, third place in a test. The first might score 100, the second 45, and the third 10. Or they might score 99, 98, 97. → The data can be ordered but the distance is unknown

Count data

- Tallies of the number of times something has happened.
- Example: the number of sunny days per year, the number of restaurants in towns
- Number of crimes commited

Categorical data

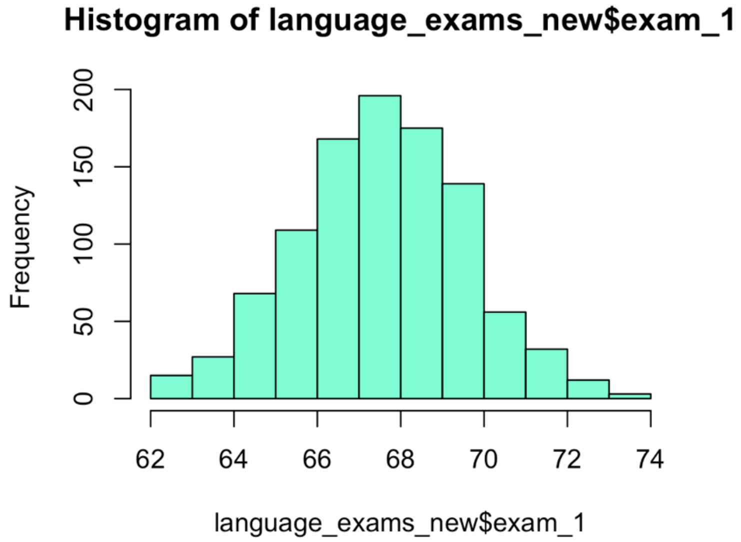
- Each value takes one of a finite number of values that are categorically distinct.
- Values of categorical variables are usually names, labels. Hence, also known as **nominal** data.
- Values of categorical variables cannot be placed in order, nor is there a natural sense of distance between them.
- Example: nationality, country, occupation, experimental conditions (experimental vs. control condition)

Characterizing distributions

Characterizing distributions

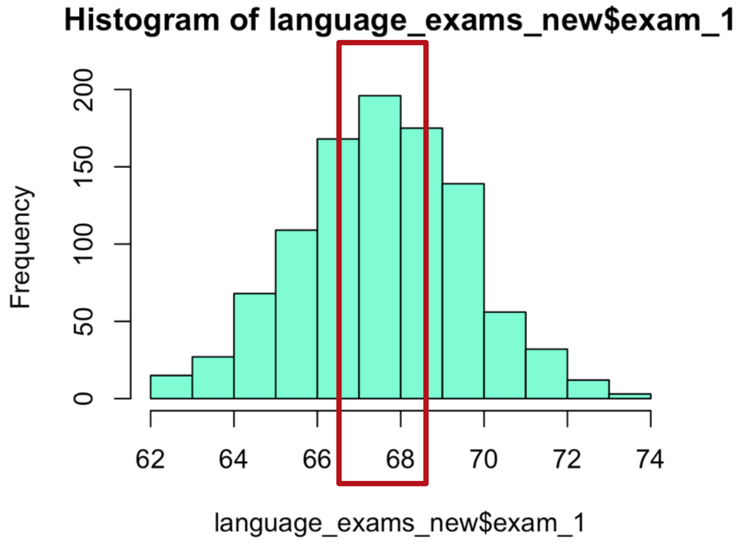
We can describe (univariate) distributions in terms of three major features:

- Location (central tendency)
- Spread (dispersion)
- Shape



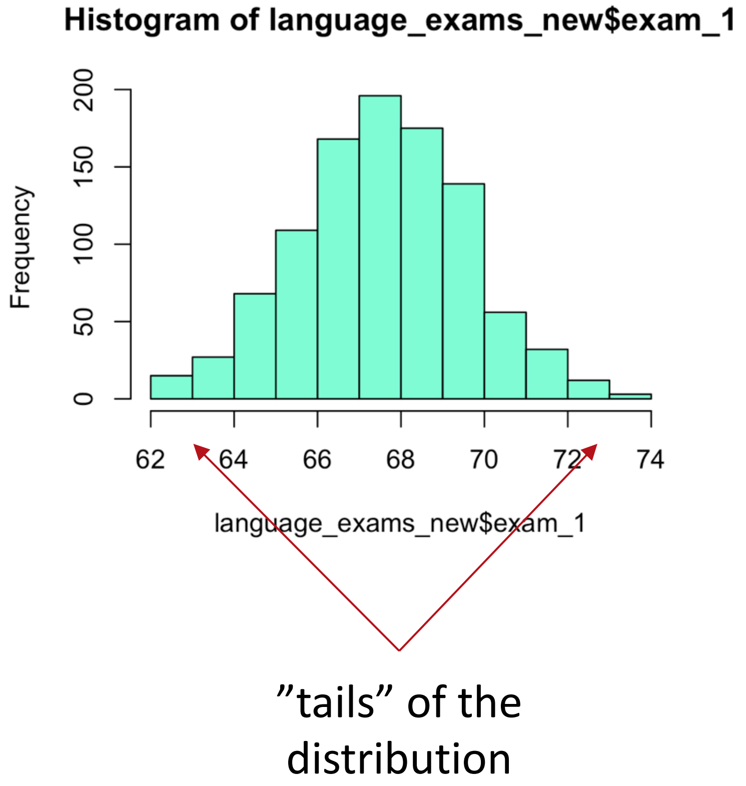
Location (central tendency)

- The location of a distribution describes where most of the value fall.
- Example: Most of the exam scores in the histogram.



Spread (dispersion)

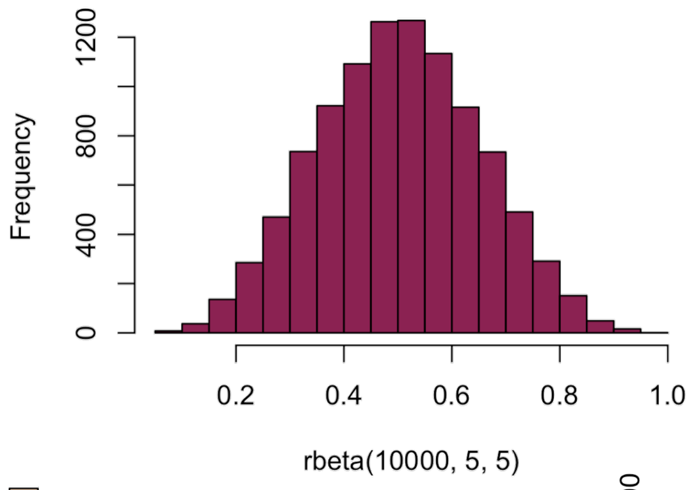
- Tells us how dispersed or spread out the distribution is.
- Are most of the scores clustered around the central value? → Short ”tails”
- Are they more spread out? → Long “tails”



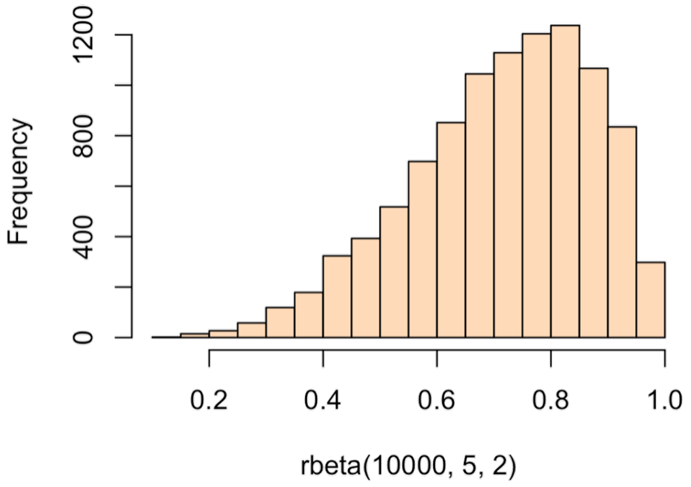
Shape: Skewness

How much (a)symmetry there is in the distribution.

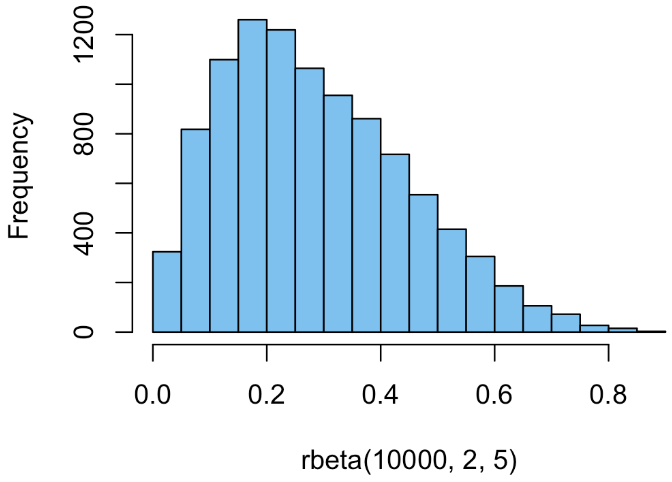
normal



negative skew

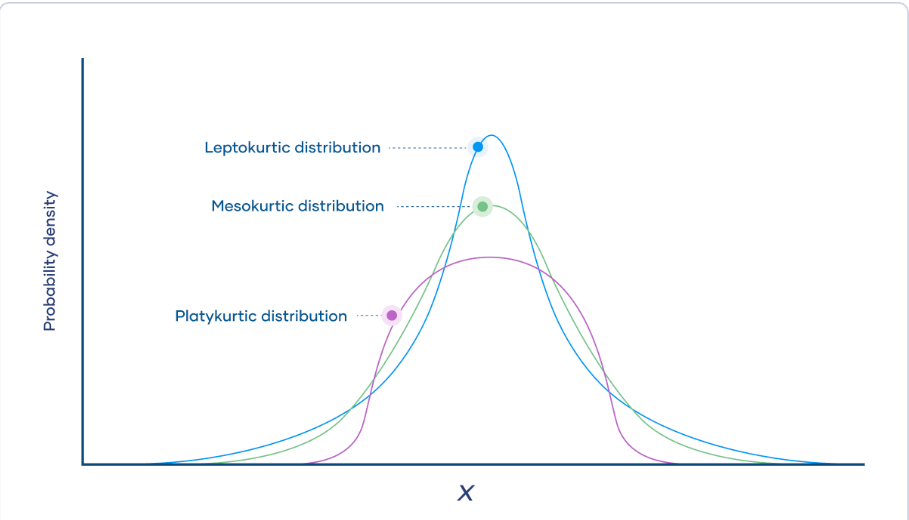


positive skew



Shape: Kurtosis

- Peaked: leptokurtic
- Flat: platykurtic



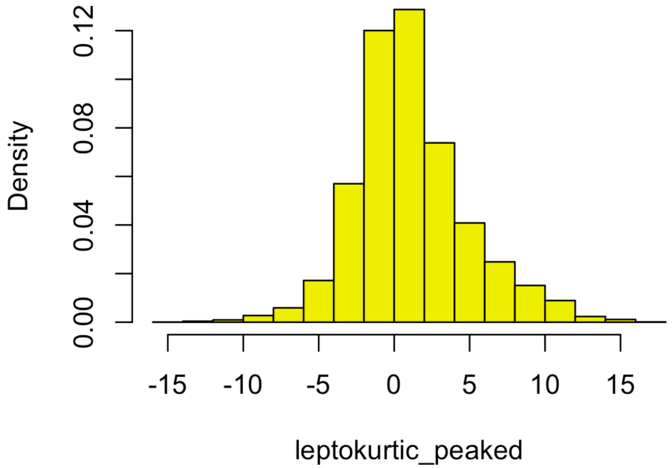


Image reference

Summary statistics

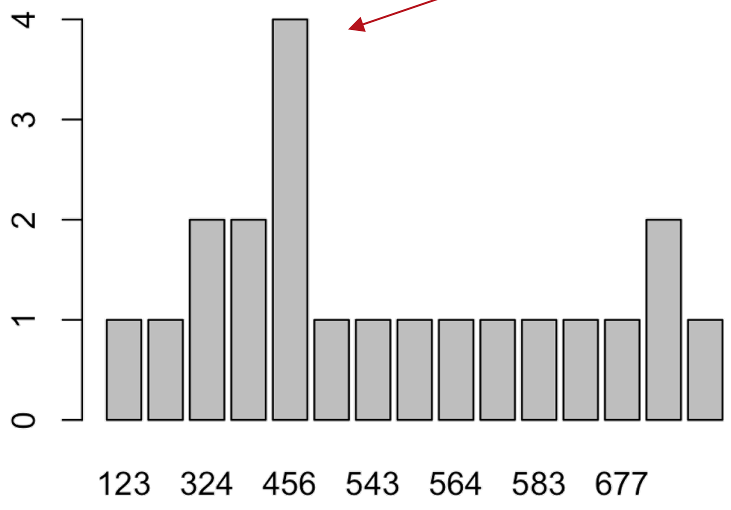
Summary statistics
(**models** of our data)

1. Measures of central tendency: mean, median, mode
2. Measures of dispersion: variance, standard deviation, range measures

Measures of central tendency

Mode (fashion in French!)

The value with the highest frequency.



	Frequency	Percent
123	1	5
322	1	5
324	2	10
345	2	10
456	4	19
465	1	5
543	1	5
546	1	5
564	1	5
567	1	5
583	1	5
663	1	5
677	1	5
876	2	10
2890	1	5
Total	21	100

Can be used for categorical, continuous, count, and ordinal data.

Arithmetic mean

- The mean: Sum of all observations ($x_1 \dots x_n$) divided by the number of observations (N)

$$\frac{x_1 + x_2 + x_3 + \dots + xn}{N}$$

- Can be calculated for continuous data and count data
- The most widely used measure of central tendency

Median

- The middle point in a sorted list of values.
- Can be used for continuous, ordinal, and count data

To calculate the median:

- First sort the values, then find the middle point.

Median

If there's an **odd** number of values:

- There is only one point in the middle of the sorted list of values. That's the median.

Data 2	Data 2 Ordered
123	123
543	324
456	456
546	489
876	543
324	546
489	876
Median	489

Median

If there's an **even** number of values:

- Find the two points in the middle of the sorted list.
- The median is the arithmetic mean of these two points.
- Here, (465 + 564) / 2 = 514.5

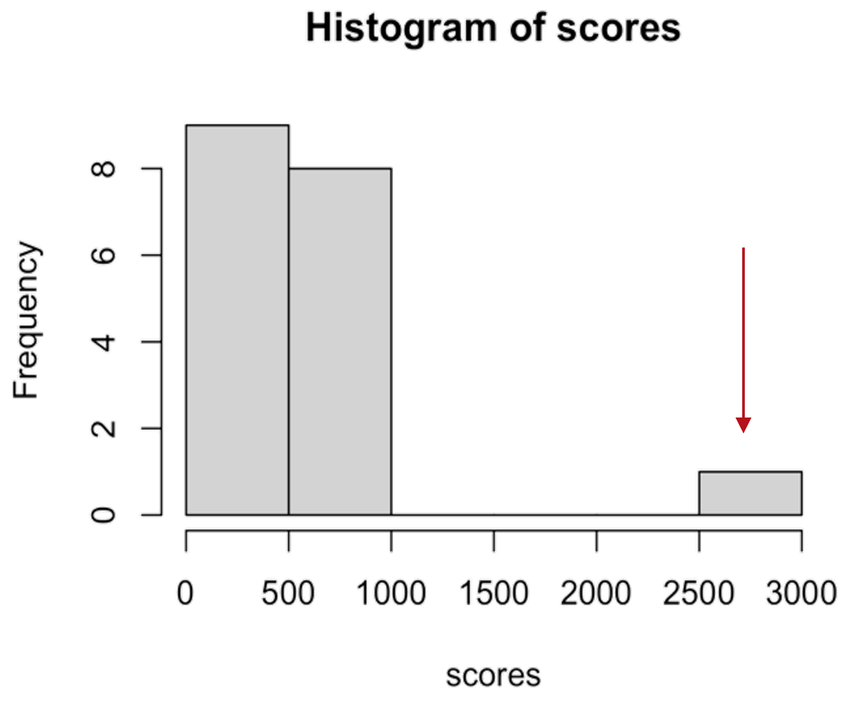
Data 1	Data 1 Ordered
564	324
677	345
345	368
465	465
368	564
583	567
324	583
567	677
Median	514.5

Median

Data 1	Data 1 Ordered		Data 2	Data 2 Ordered
564	324		123	123
677	345		543	324
345	368		456	456
465	465		546	489
368	564		876	543
583	567		324	546
324	583		489	876
567	677			
Median	514.5		Median	489

Robust measures of central tendency

- The **mean** is the most widely used summary statistic.
- But: It's not a very "robust" statistic as it's easily affect by extreme values (outliers).



Compare:

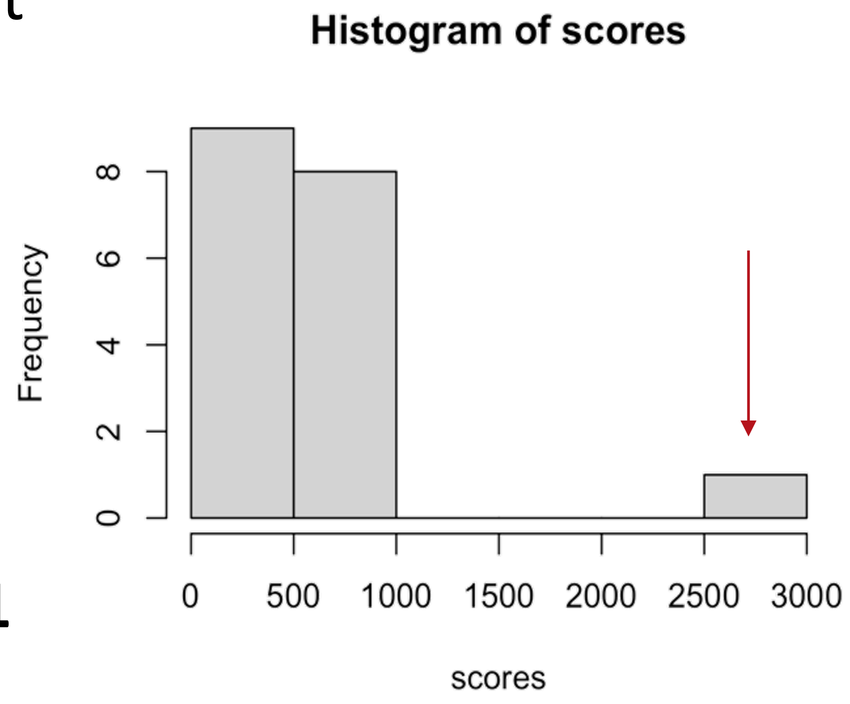
- M with outlier = **612**

Robust measures of central tendency

- The **median** is a more robust statistics; it’s not easily affected by extreme scores.

Compare:

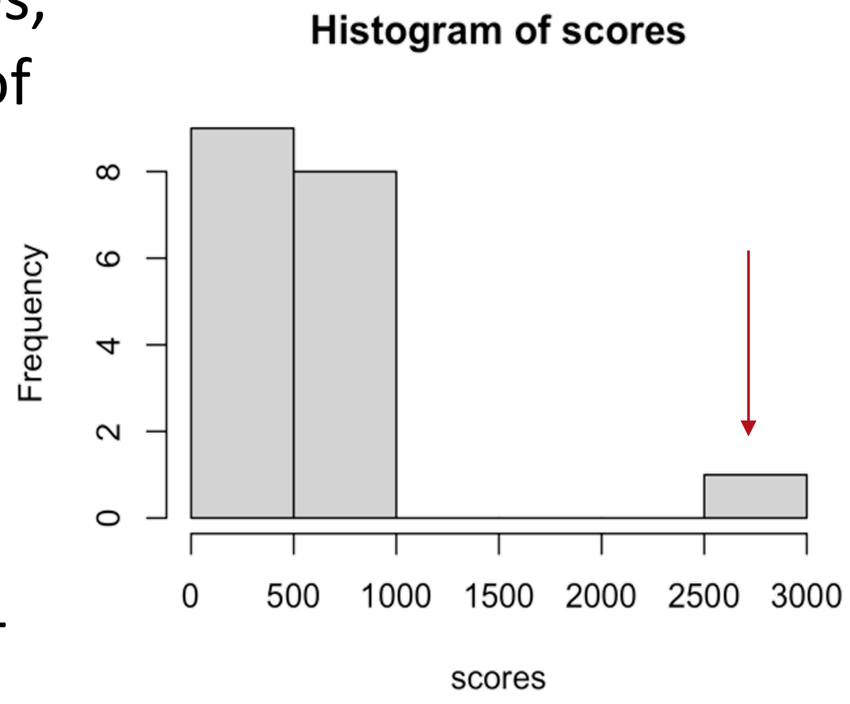
- Median with outlier: **465**
- Median without outlier: **461**



Compare mean and median

When there are extreme values, the median is a better model of our data.

- M *with* outlier = 612
- Median *with* outlier: 465
- M *without* outlier = 499
- Median *without* outlier: 461



So why not just use the median?

More robust alternatives to the standard mean

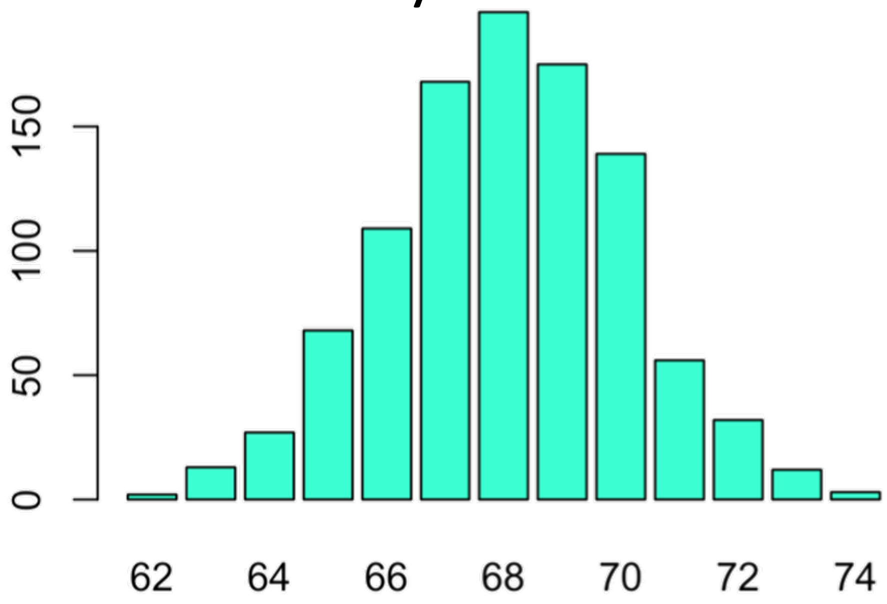
- **Trimmed mean**: Extreme values are removed before calculating the mean as normal (`trim()` function)
- **"Winsorized" mean**: Extreme values are replaced with values at the thresholds of the extremes, e.g. the 90th percentile.
- So, why not just always used these?
- Whatever you choose, justify your decision and be transparent about it in the report.

Measures of dispersion

Measures of dispersion

- Tell us how the observations in our variables are spread out.
- Provide information about the variability in our data.

Best practice:
Report a measure of central tendency and a measure of dispersion (e.g., **M** and **SD**)



Standard deviation

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

Measure of dispersion around the **mean**

To calculate the SD:

1. Calculate the mean for the sample: $\bar{x}$
2. Calculate, for each data point (x), its difference from the mean and square this value: $(x - \bar{x})$
3. Sum up the squared differences: $(x - \bar{x})^2$
4. Divide the “sum of squares” $(x - \bar{x})^2$ by the number of observations minus 1 (N – 1).
5. Take the square the root of this number.

Range

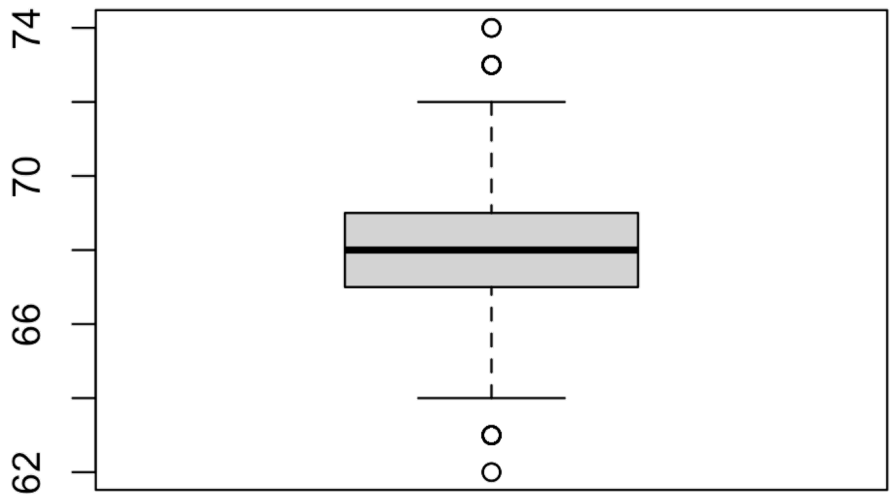
- Distance between the smallest and largest data point.
- Not very informative as it is exclusively based on the most extreme values…

Example: age range in our handout

- Minimum age = 18
- Maximum age = 25
- Range = 7

Interquartile range

- Another way of measuring dispersion is by means of the interquartile range (IQR).
- This divides the sample data into quartiles (Q1, Q2, Q3 and Q4).
- Difference above Q1 and below Q4 is called the **interquartile range =** middle 50% of values
- To find the interquartile range, subtract the value of the lower quartile (or 25%) from the value of the upper quartile (or 75%). Interquartile range = upper quartile – lower quartile.



Practical: Data wrangling and exploratory data analysis

Questions?

- I will be walking around while you work through the worksheet