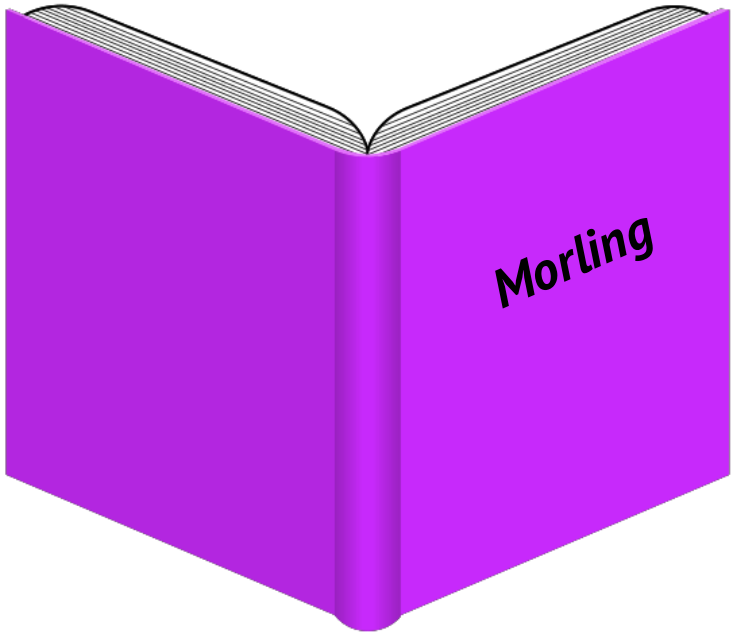


Research Methods and Statistics

Lecture 3: Why are statistics needed? + Exploring data

Johnny van Doorn

Pictures source: pixabay.org


Weihenstephaner
HEFE
WEISSBIER
ALTESTE BRAUEREI DER WELT
Weihenstephaner
HEFE
WEISSBIER
PREMIUM BAVARICUM
Jeffreys
Theory of Probability
THIRD EDITION
OXFORD
Weihenstephaner
HEFEWEISSBIER
ALKOHOLFREI
ALTESTE BRAUEREI DER WELT
Weihenstephaner
HEFEWEISSBIER
ALKOHOLFREI
PREMIUM BAVARICUM


DELL
Weihenstephaner
HEFE
WEISSBIER
Weihenstephaner
P: Alc
P: non alc


DELL

Induction

Observation

Theory

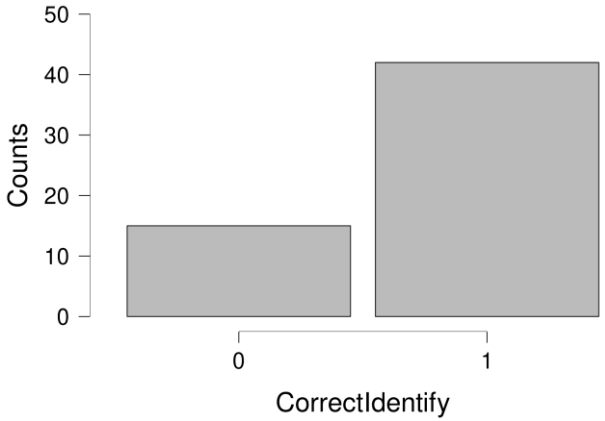
Evaluation

Deduction

42 correct (73.7%)
15 incorrect (26.3%)

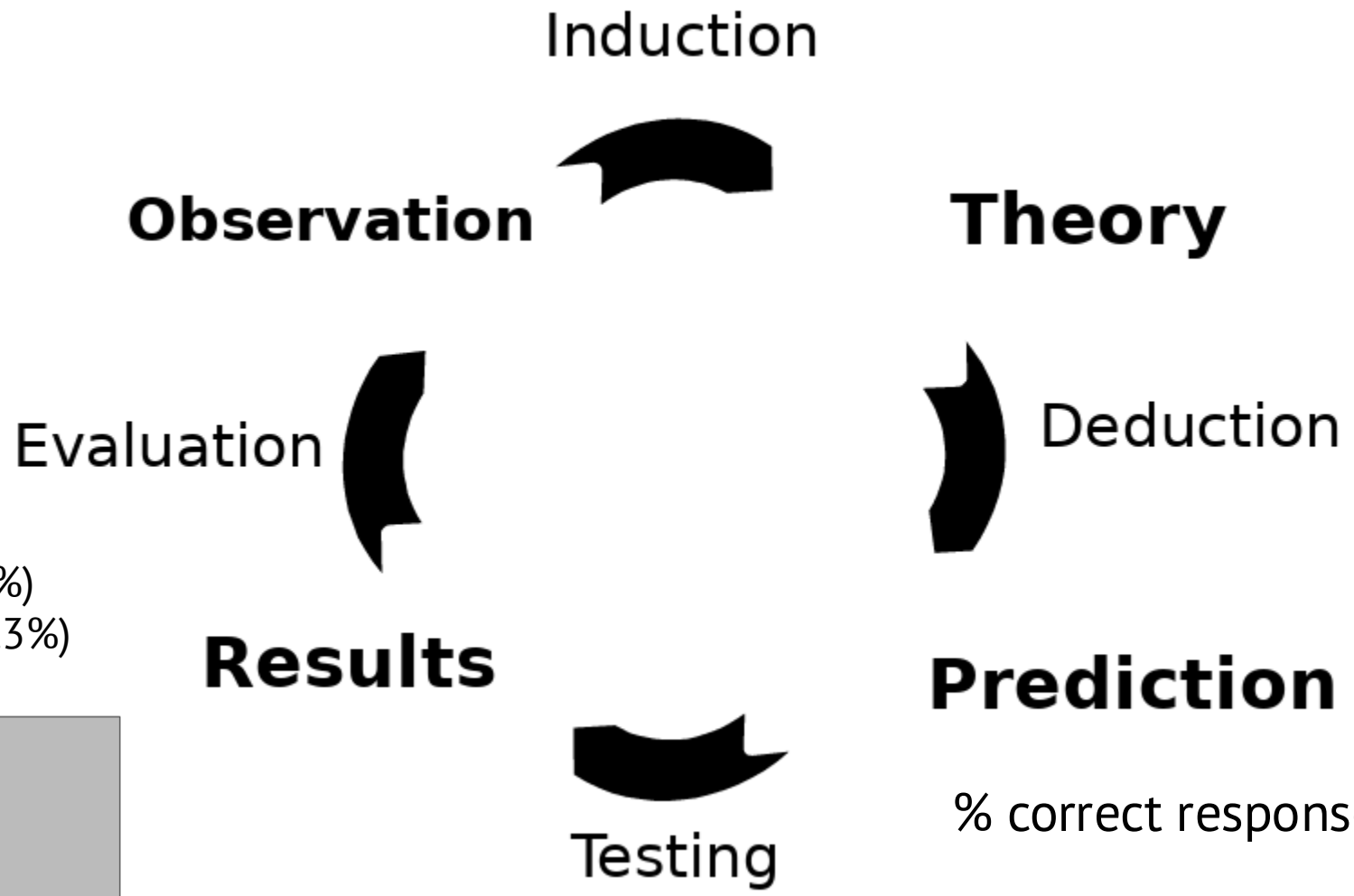
Results

Prediction

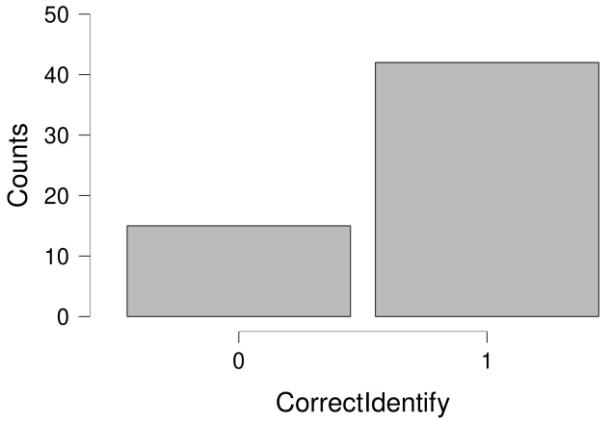


Testing

% correct responses = 50%



42 correct (73.7%)
15 incorrect (26.3%)



% correct responses > 50%

Questions that come up

- How do we know that this conclusion is correct?
- How do we know that this study is performed accurately?
- How do we know that this is not a coincidence?

Today

1. **Why are statistics needed?**
2. A little bit about the course
   - The book: Agresti & Franklin
   - Comparison to high school math
   - How to prepare
3. How can you explore data?
   - Types of data
   - Displaying data
   - Characteristics of a distribution
4. Recap
   - Next time
   - Example exam question

Why are statistics needed?

What is science?

Last week Riet said: “The search for how the world works”

- Formal vs empirical sciences
- Empirical sciences are based on experience / observation
- Inductive reasoning: specific → general

Why are statistics needed?

What is statistics?

Agresti & Franklin: "*Statistics is the art and science of learning from data*" (p. 4)

- Systematically note experiences/observations → data
- Data become overwhelming quickly

	vr11	vr12	vr13	vr14	vr15	vr16	vr17	vr18	vr19	vr20	vr21	vr22	vr23	vr24	vr25	vr26	vr27	vr28	vr29	vr30	vr31	vr32	vr33	vr34	vr35	vr36	vr37	vr38	vr39	vr40	vr41	vr42	vr43	vr44	vr45
1	0	1	1	0	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0
2	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1
3	1	1	1	0	1	1	1	0	1	0	1	1	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	0	1	1	0	0	0	0	0
4	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
5	1	1	0	0	1	1	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	1	0	1	1	1	0	0	0	0	0	0	1	1
6	1	1	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	1	1	1	0	1	1	0	0	0	0	1	1	0	1
7	0	1	1	1	0	1	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	0	0	0	1	1	1	0	1	0	1	1	1	1
8	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
12	0	1	1	1	1	1	0	0	1	0	1	1	0	1	0	1	0	1	1	0	1	0	1	1	0	1	1	1	0	1	1	0	1	1	1
13	1	1	0	1	0	1	1	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	1	1	1	1	1	1	0
14	1	1	0	1	1	1	0	1	1	0	0	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1
15	1	0	0	0	1	1	0	0	0	1	0	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	0	1	0
16	1	0	1	1	1	1	0	0	1	1	0	1	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1
17	1	1	1	1	0	1	0	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0
18	0	1	0	1	1	1	1	0	1	0	0	0	0	1	1	0	1	1	0	0	1	1	1	1	0	1	1	1	0	1	0	0	0	1	0
19	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1
20	0	1	1	0	1	1	0	0	1	0	1	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1
21	0	1	0	0	1	1	0	1	1	1	0	0	1	0	0	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1	0	1	0	1	1
22	1	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1	0
23	0	0	1	0	1	1	1	1	1	1	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	0
24	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1
25	0	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1
26	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
27	1	1	1	1	1	1	0	0	0	0	1	1	1	1	0	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0
28	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
29	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1
30	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1

104 110 101 88 101 96 105 95 90 98 98 89 95 109 83 97 75 109 108 97 99 100 100 113 98 113 110 111 105 78 105 101 114 89 96 92
107 87 101 82 107 103 103 78 114 91 110 94 101 107 88 109 111 94 100 121 109 106 105 82 98 90 100 83 106 105 94 96 104 101 117
74 95 77 113 108 81 89 94 118 106 100 114 92 93 94 102 117 97 108 99 115 118 106 98 100 84 91 100 87 104 95 104 83 113 120 108
83 98 113 103 103 98 104 114 105 124 105 100 84 98 104 112 109 105 102 106 86 100 94 91 112 117 97 105 93 103 109 103 113 101
88 95 105 96 97 120 103 98 87 102 113 91 123 102 90 103 106 114 88 92 109 93 103 93 91 102 95 104 90 115 105 88 108 99 89 90
102 74 103 105 108 89 108 113 122 103 106 114 111 101 97 92 88 78 113 110 118 115 111 74 91 87 99 103 96 113 93 97 107 102 78
103 105 126 83 92 97 90 89 95 105 94 116 105 92 105 96 91 97 103 113 75 107 86 103 115 94 114 91 103 85 100 110 102 91 126 93
97 101 95 94 102 88 103 90 92 102 104 94 107 94 92 100 103 93 96 104 101 113 111 109 103 99 95 93 95 98 92 106 103 114 106 115
105 98 92 119 85 104 86 104 129 91 105 100 103 105 111 98 104 116 99 98 77 106 96 101 118 103 101 93 101 89 101 105 121 82 84
98 121 98 104 92 106 94 113 110 102 115 87 107 106 103 103 116 85 95 104 116 107 72 113 108 93 104 102 97 71 98 101 87 111 82
88 91 109 96 110 108 90 100 97 101 102 102 109 92 88 98 102 99 93 94 100 74 90 94 89 111 87 110 96 110 89 99 114 87 101 86 113
94 103 93 101 94 90 95 98 93 99 109 115 124 90 101 95 109 98 95 101 100 91 91 115 94 97 100 81 104 99 102 106 100 103 109 100
98 98 123 97 111 103 104 95 97 90 97 110 89 96 112 107 97 104 103 88 88 95 96 99 114 102 93 98 81 93 108 106 101 92 83 80 94 99
117 95 112 99 113 121 94 95 83 88 88 88 120 107 104 102 121 84 96 74 102 111 69 78 114 87 91 89 88 87 91 91 95 99 107 95 108 99
97 94 99 84 95 106 115 86 95 104 86 102 100 94 100 103 116 101 96 100 107 91 117 94 108 104 93 114 103 95 98 95 111 107 97 94
100 94 78 96 95 115 115 106 101 96 102 102 112 93 89 104 89 103 109 87 105 97 102 106 101 106 97 112 103 89 107 100 88 93 99
99 107 105 96 84 102 109 104 86 104 90 103 108 109 115 97 97 114 105 102 108 92 100 90 112 88 105 84 91 92 96 83 107 107 87
115 116 104 111 96 113 107 84 99 120 97 100 89 98 96 93 103 99 103 89 107 92 92 90 104 107 95 106 91 83 106 102 100 91 113 95
97 78 92 87 102 96 99 93 95 100 97 109 94 105 96 100 85 93 115 103 101 92 98 106 95 82 113 113 72 102 90 97 111 99 103 102 114
108 112 111 109 111 107 90 97 94 90 102 98 92 79 108 90 96 77 104 95 95 99 111 100 94 111 107 86 98 95 102 105 87 91 93 117 99
97 125 106 94 101 85 93 133 111 97 109 104 84 103 101 109 96 98 98 96 109 95 106 111 82 114 109 108 85 106 107 103 90 95 112
98 96 88 106 93 107 89 109 95 86 85 94 104 95 88 87 101 118 86 92 84 116 102 104 102 100 106 104 91 88 127 123 93 82 98 105 108
99 98 103 90 95 98 99 107 114 95 107 113 107 104 93 92 113 83 94 105 106 110 91 104 86 98 100 93 84 104 99 95 113 98 96 96 97
100 107 103 99 102 108 118 96 116 90 112 116 115 103 103 108 106 84 86 100 110 101 90 87 113 97 111 106 85 109 110 124 85 102
106 108 106 99 114 93 100 111 122 88 95 93 107 99 111 108 120 98 104 93 105 90 96 101 104 103 95 105 98 99 89 93 105 110 113
102 100 108 96 99 73 96 104 118 97 110 84 97 106 96 109 94 95 90 111 110 88 96 118 90 122 102 110 108 90 91 103 98 95 88 104
94 99 104 85 114 98 90 98 113 89 103 102 103 89 98 99 131 108 88 94 108 108 114 112 106 107 97 97 81 108 98 82 74 105 94 106
104 96 95 105 115 106 99 90 103 123 100 111

Why are statistics needed?

What is statistics?

Agresti & Franklin: "Statistics is the art and science of learning from data" (p. 4)

- Systematically note experiences/observations → data
- Data become overwhelming quickly
  - CITO data: 159,994 children, 200 questions (picture displays 30 children x 22 questions)
  - IQ scores: +- 1,000 observations
- And we want to *learn* from the data!

Why are statistics needed?

What is statistics?

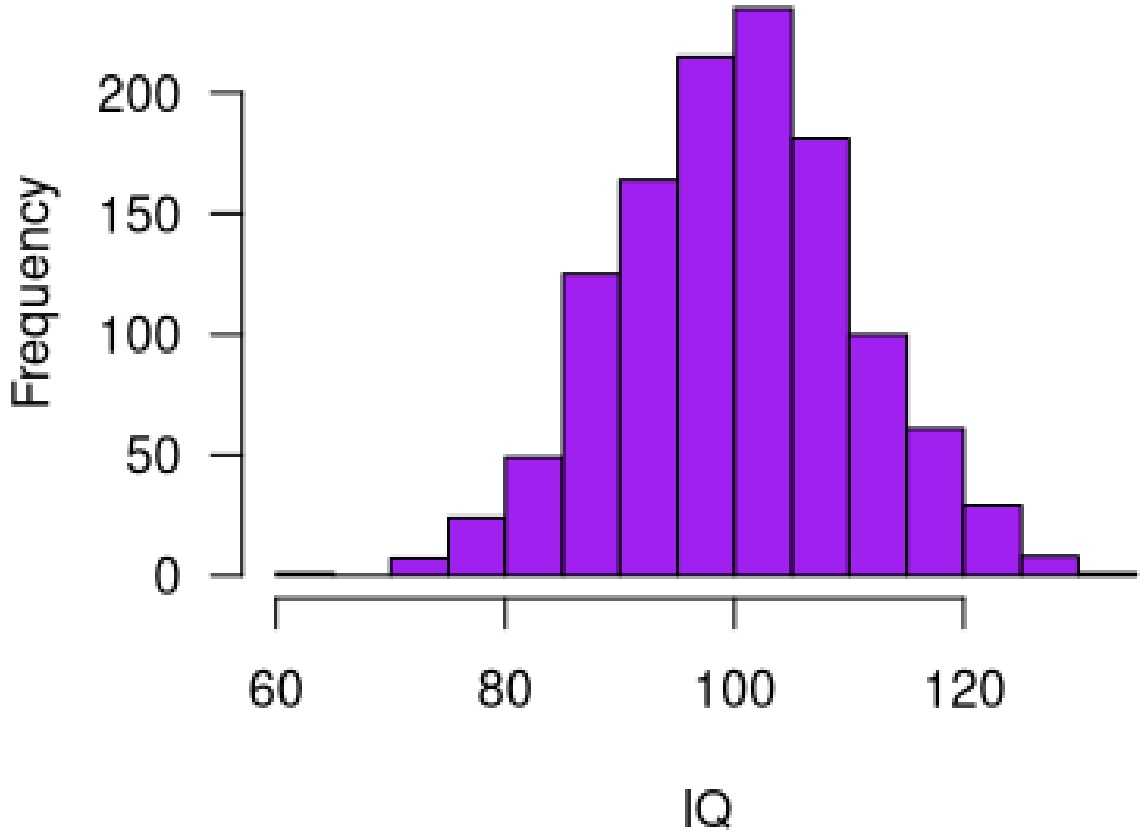
Agresti & Franklin: "Statistics is the art and science of learning from data" (p. 4)

Three components:

- *Experimental Design* (i.e., Research Methods)
- *Descriptive statistics* (summarize/compress all the data)
- *Inferential statistics* (learn from the data, generalize)

Why are statistics needed?

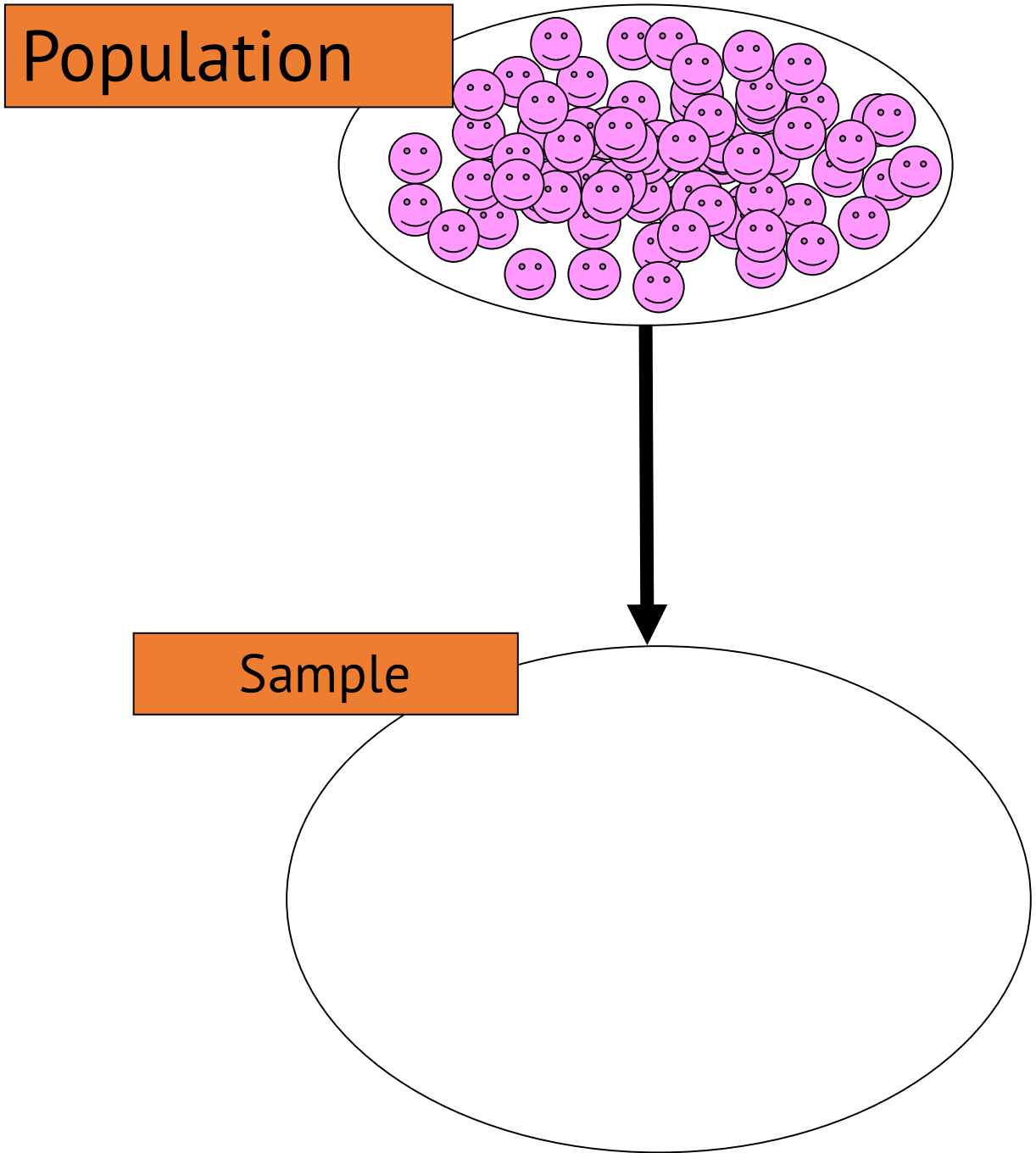
- Descriptive statistics
  - Numerically (ie, average values)
    - → Mean IQ = 101.3
  - Graphically (ie, a histogram)



A Statistic: A numerical summary of the data is called a statistic

Why are statistics needed?

- Inferential statistics
  - Making predictions and decisions about the **population**
  - Assessing the uncertainty about statements


Population
Sample

Why are statistics needed?

- Inferential statistics
  - Making predictions and decisions about the population

Population: **All** scores/data we are interested in
Sample: The part of the population that we have actually observed
Inference: Drawing a conclusion about the population, based on the sample

- Many things can go wrong when drawing inferences → Statistics

Why are statistics needed?

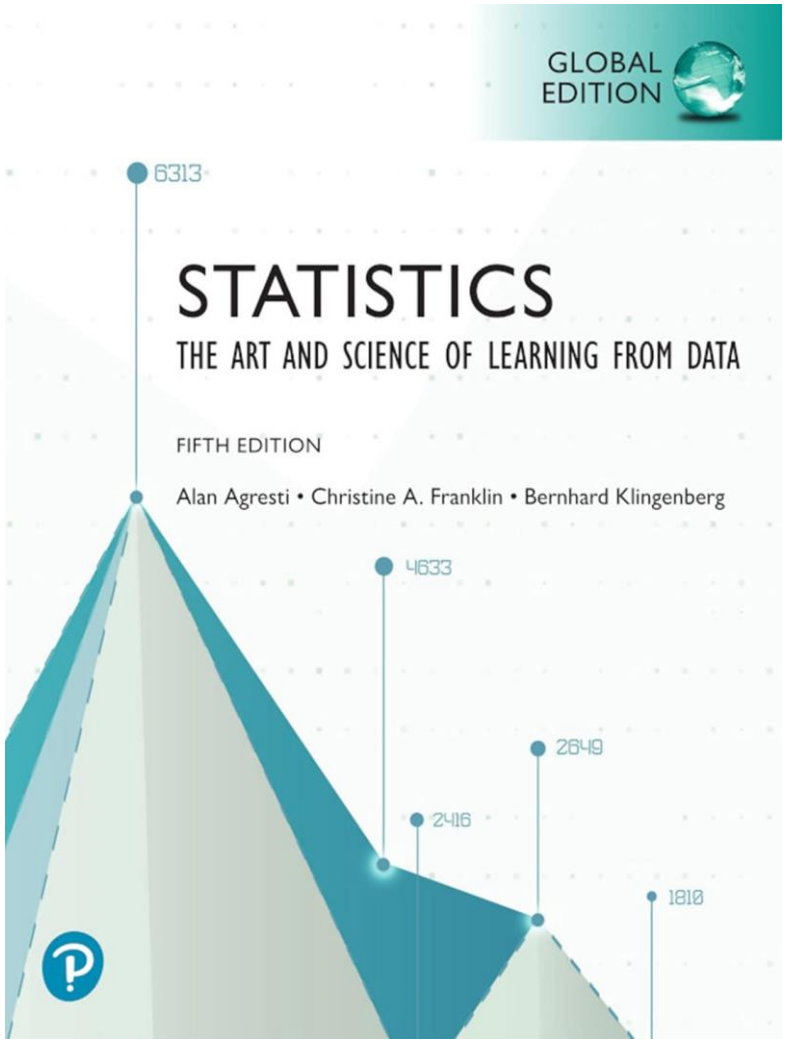
- Reproducibility crisis:
  - Center for Open Science tried to replicate 100 published psychological studies
  - Only 35 studies replicated

- https://science.sciencemag.org/content/349/6251/aac4716.full?ijkey=1xgFoCnpLswpk&keytype=ref&siteid=sci
- https://en.wikipedia.org/wiki/Reproducibility_Project

Overview of Today

1. Why are statistics needed?
2. **A little bit about the course**
   - The book: Agresti & Franklin
   - Comparison to high school math
   - How to prepare
3. How can you explore data?
   - Types of data
   - Displaying data
   - Characteristics of a distribution
4. Recap
   - Next time
   - Example exam question

Agresti & Franklin book



Do we all have the book?

+ Excellent explanations
+ Many examples
+ Many practice questions
+ Summary of each chapter
- Boring at times
- Difficult to find specific topics

Basis of the course: Read -> Practice

Lectures vs practice

- Learning Statistics is mainly ***doing***
  - "Like riding a bike"
- Practice, Practice, Practice

- Lectures:
  - Provide structure
  - Point to important topics
  - Extra explanation of difficult topics

- *Caveat: The lectures are not a complete summary of all topics!*
  - *Always check the Canvas modules for the interim exam information*
  - *→ Formula sheet*

Comparison to high school math

- Some topics are a repetition of high school math (Dutch system: “Wiskunde A/C”)
- Examples:
  - Some descriptive statistics (mean, median, quartiles, standard deviation) and graphs
  - Probability, probability distributions etc.
  - Population vs sample
  - Hypothesis testing

- In our course:
  - Short recap + extension
  - Deeper knowledge + relevant applications to psychological science

- *More focus on **understanding***

How to prepare for the course (stats)?!

- The different stages of confusion when learning statistics
  - Book: can keep rereading
  - Exercises: can keep (re)doing
  - Lectures: happen once

- Use the practicals
- Use the Canvas discussion board
  - Preferred to sending us a direct message
    - If you do send a direct message, please indicate who you send it to! (Riet and/or Johnny)
- Work together on the exercises

How to prepare for the exam?!

- Practice the exercises, WA, trial exam, formula sheet (see Canvas)
- Practice the software (Ans calculator, later Excel)
  → Speed is important!
- Read the book while keeping lecture slides in mind

- On the exam:
  - You don’t have to stick to the question order!
  - Focus on **fast** questions (e.g., without calculations)
  - Focus on **open** question (partial points possible)
  - Then go back to time intensive questions

You'll get there!

https://www.youtube.com/watch?v=JC82Il2cjqA

Overview of Today

1. Why are statistics needed?
2. A little bit about the course
   - The book: Agresti & Franklin
   - Comparison to high school math
   - How to prepare
3. **How can you explore data?**
   - Types of data
   - Displaying data
   - Characteristics of a distribution
4. Recap
   - Next time
   - Example exam question

Types of data

Variable (general): A thing that can take multiple "values"
Variable (book): Any characteristic observed in a study (p. 25)

- "Length": I am 184 cm, you are (most likely) something else
- "Eye Color": Some have green eyes, others have brown eyes

Categorical: Each observation belongs to one category
Quantitative: Each observation has a numerical value representing a magnitude

Discrete: Every observation is one of a specific set of values (e.g., 0, 1, 2)
Continuous: Every observation comes from a range (e.g., 0-3)

Let’s practice this

- “Race track”
  - → Categorical
- “Weight”
  - → Quantitative, continuous
- “Number of students”?
  - → Quantitative, discrete
- “Percentage of students that will pass this course”?
  - → Quantitative, continuous
- “Study year”?
  - → Quantitative, Discrete or Continuous?

Displaying Data

- Data is often overwhelming
- Data is often complex

	vr11	vr12	vr13	vr14	vr15	vr16	vr17	vr18	vr19	vr20	vr21	vr22	vr23	vr24	vr25	vr26	vr27	vr28	vr29	vr30	vr31	vr32	vr33	vr34	vr35	vr36	vr37	vr38	vr39	vr40	vr41	vr42	vr43	vr44	vr45
1	0	1	1	0	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0
2	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	1
3	1	1	1	0	1	1	1	0	1	0	1	1	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0
4	1	1	0	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
5	1	1	0	0	1	1	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	1	0	1	1	1	0	0	0	0	0	0	1	1
6	1	1	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	1	1	1	0	1	1	0	0	0	0	1	1	0	1
7	0	1	1	1	0	1	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	0	0	0	1	1	1	0	1	0	1	1	1	1
8	1	1	1	1	0	1	1	0	1	0	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
12	0	1	1	1	1	1	0	0	1	0	1	1	0	1	0	1	0	1	1	0	1	0	1	1	0	1	1	1	0	1	1	0	1	1	1
13	1	1	0	1	0	1	1	0	0	0	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	1	1	1	1	1	1	0
14	1	1	0	1	1	1	0	1	1	0	0	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	1
15	1	0	0	0	1	1	0	0	0	1	0	1	1	0	0	0	1	1	1	1	1	0	1	1	1	1	0	0	1	0	0	1	0	1	0
16	1	0	1	1	1	1	0	0	1	1	0	1	0	1	1	1	0	1	1	0	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1
17	1	1	1	1	0	1	0	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	0	0
18	0	1	0	1	1	1	1	0	1	0	0	0	0	1	1	0	1	1	0	0	1	1	1	1	0	1	1	1	0	1	0	0	0	1	0
19	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
20	0	1	1	0	1	1	0	0	1	0	1	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1
21	0	1	0	0	1	1	0	1	1	1	0	0	1	0	0	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1	0	1	0	1	1
22	1	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0	1	1	1	1	0
23	0	0	1	0	1	1	1	1	1	1	0	0	1	1	0	1	0	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0
24	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	0	1	0	1	1	1	1	1
25	0	1	1	1	1	1	0	1	1	1	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1
26	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
27	1	1	1	1	1	1	0	0	0	0	1	1	1	1	0	1	0	1	1	0	1	1	1	1	1	1	1	1	0	1	0	1	0	1	0
28	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	0	1	1	1	1	1
29	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1
30	1	1	0	1	1	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1

- Good display methods help you focus on the relevant aspects of the data

Example 1: Movie genres of the IMDB Top 250

Frequencies for Genres

Genres	Frequency
Drama	73
Comedy	21
Action	35
Crime	34
Adventure	23
Biography	22
Animation	24
Horror	5
Mystery	6
Western	5
Sci-Fi	1
Film-Noir	1
Missing	0
Total	250

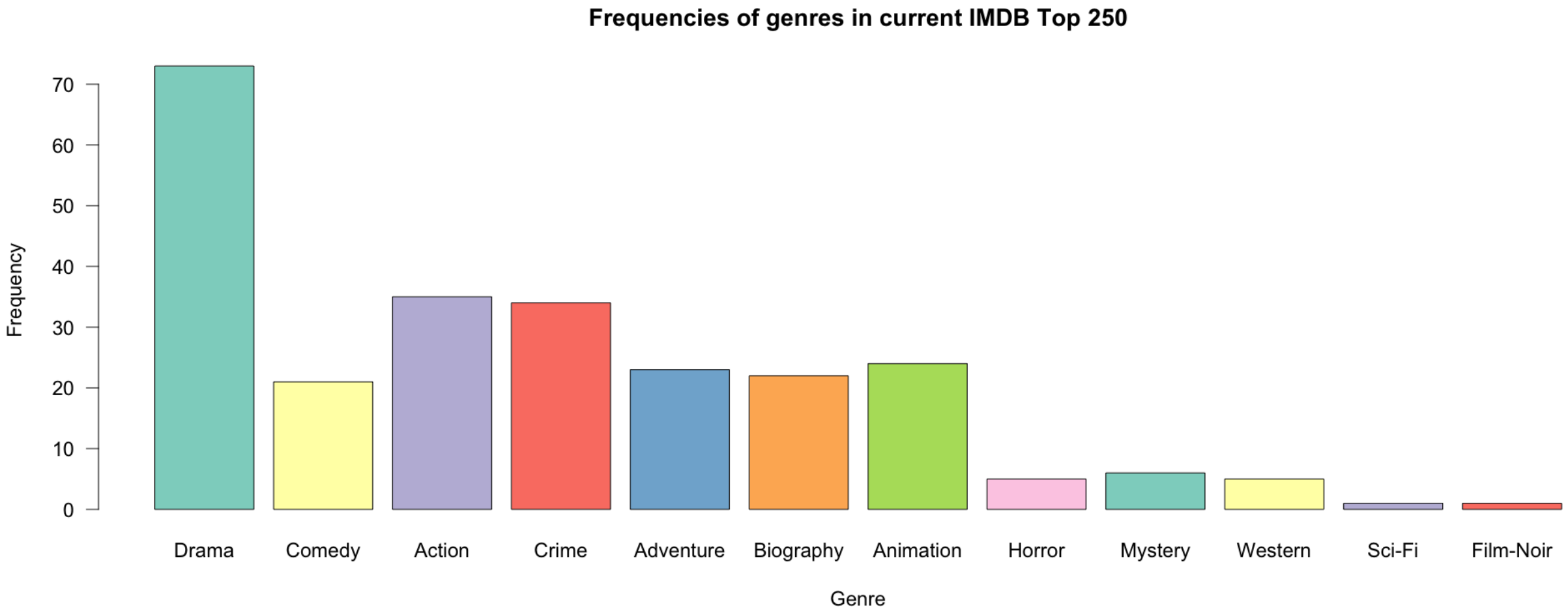
Example 1: Movie genres of the IMDB Top 250

Frequencies for Genres

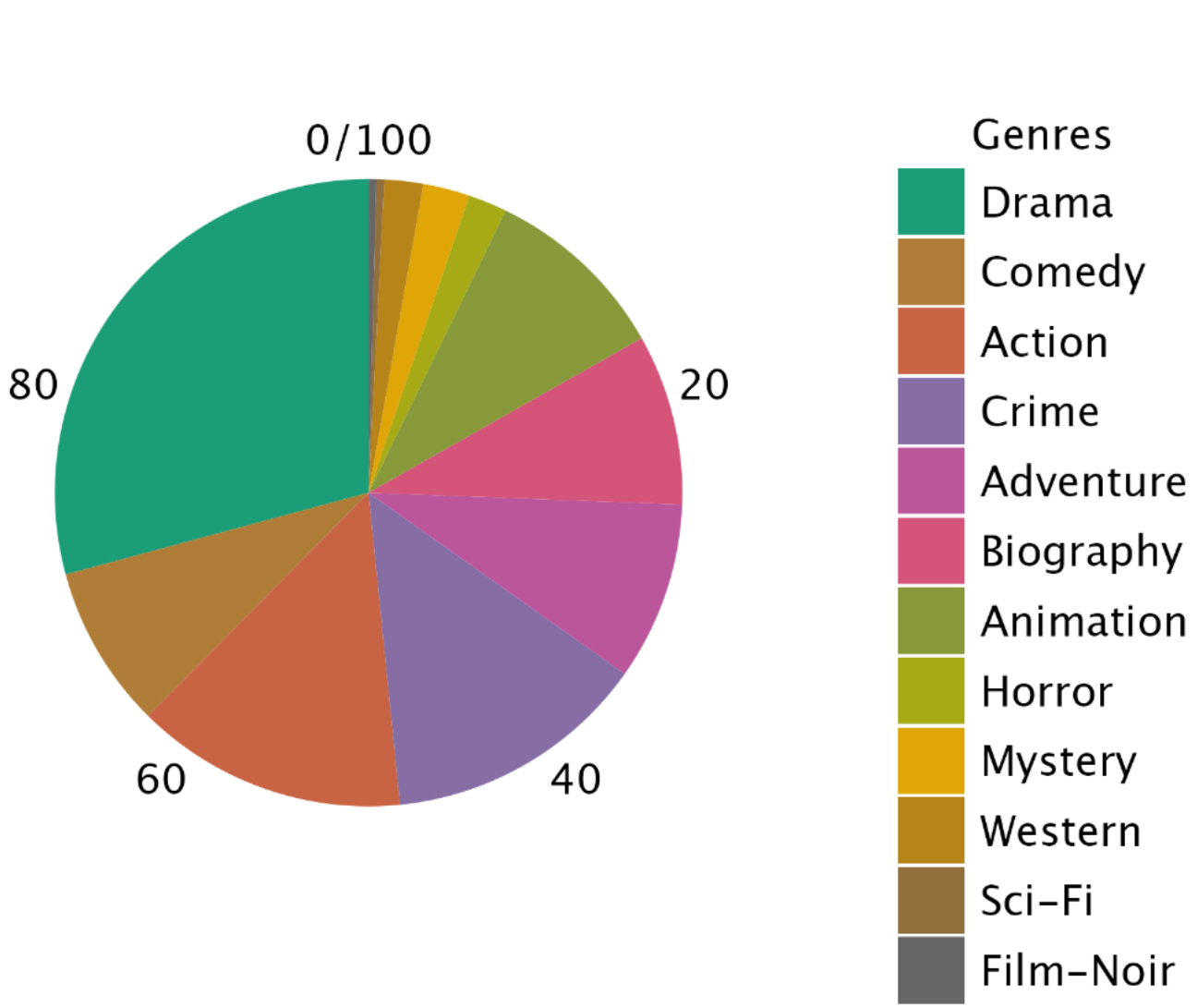
Genres	Frequency	Percent
Drama	73	29.200
Comedy	21	8.400
Action	35	14.000
Crime	34	13.600
Adventure	23	9.200
Biography	22	8.800
Animation	24	9.600
Horror	5	2.000
Mystery	6	2.400
Western	5	2.000
Sci-Fi	1	0.400
Film-Noir	1	0.400
Missing	0	0.000
Total	250	100.000

e.g., 73 / 250 = 0.292 → 29.2%

We can also visualize the frequency table using a bar chart:



We can also visualize the frequency table using a pie chart:



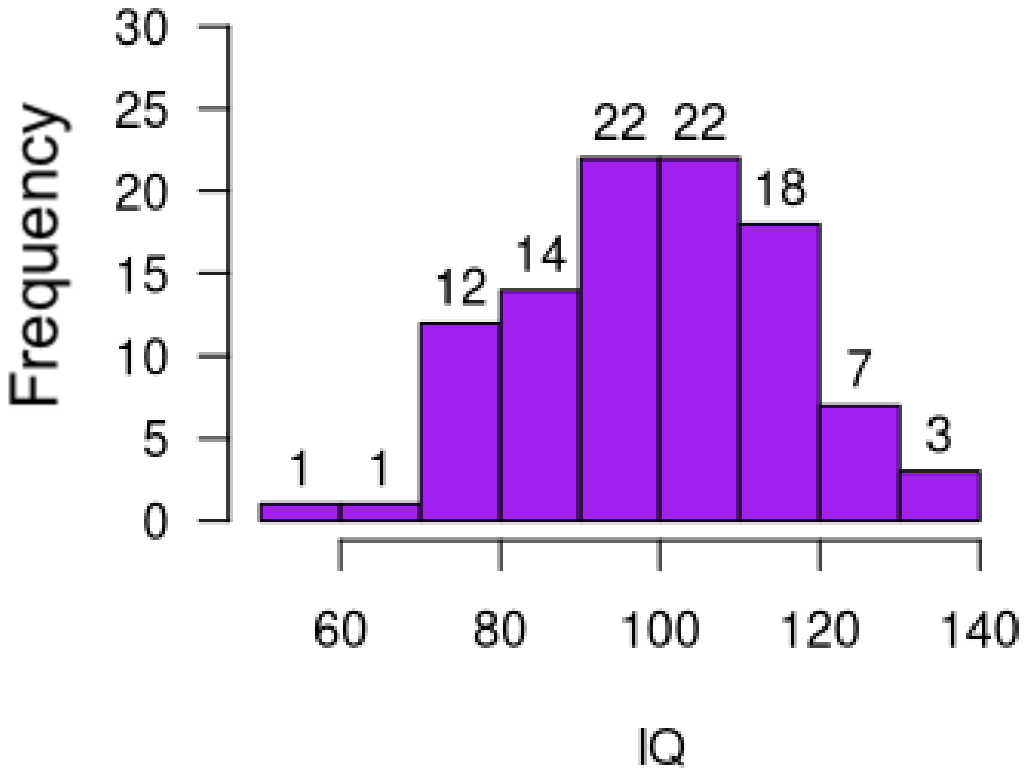
Example 2: IQ

- Standard measure of intelligence
- Set of questions/items leads to a score, such that 100 is average

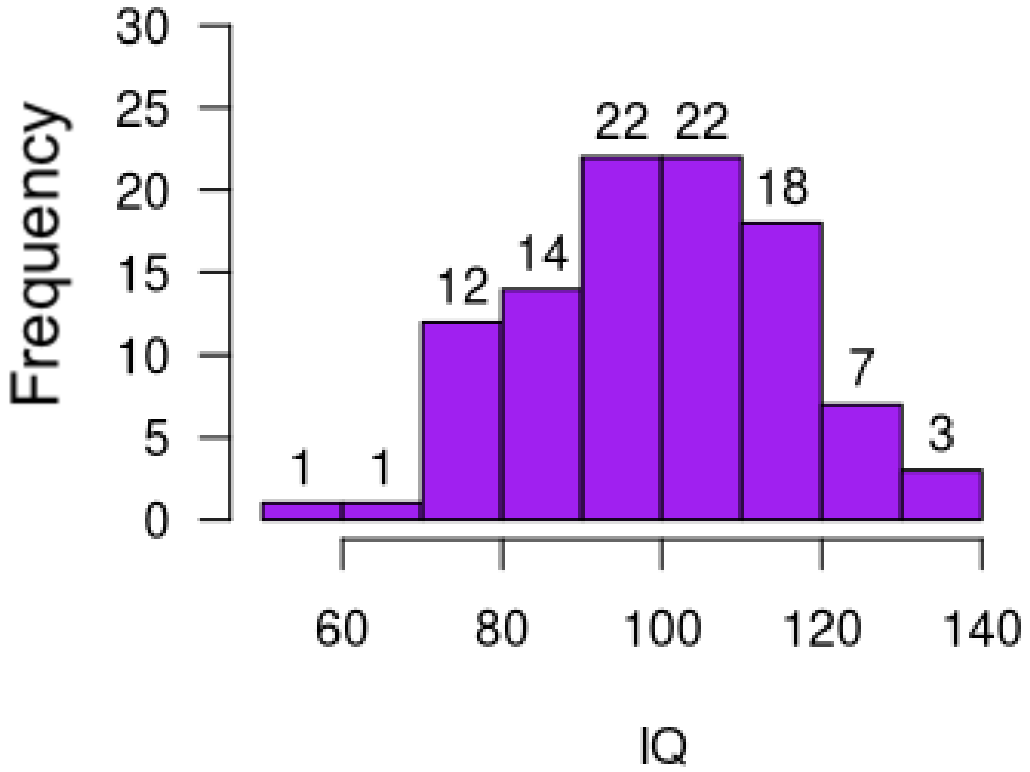
https://en.wikipedia.org/wiki/Alfred_Binet

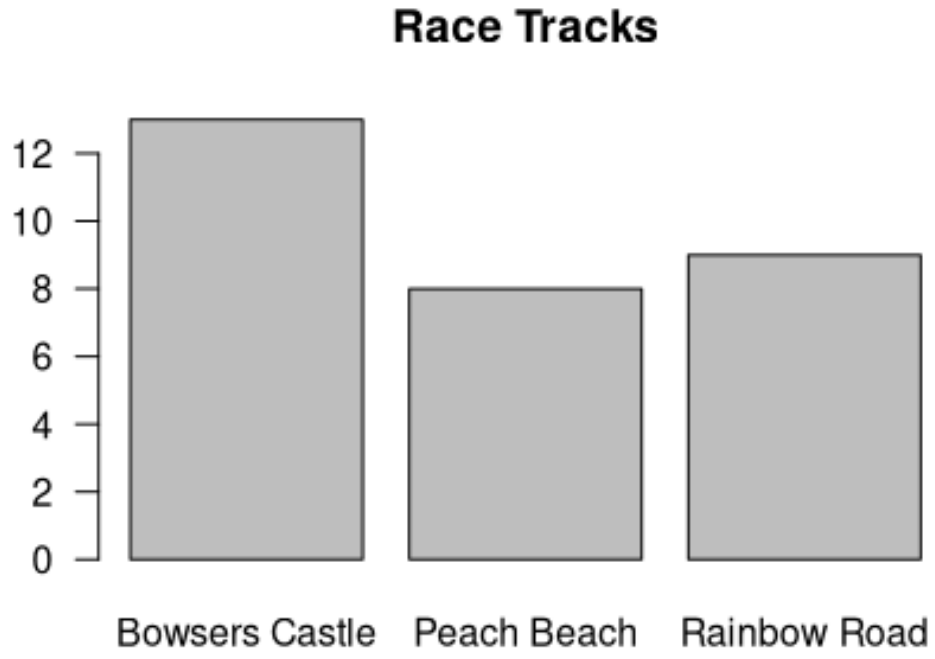
Example 2: IQ Frequencies

IQ range	Frequency
50-60	1
60-70	1
70-80	12
80-90	14
90-100	22
100-110	22
110-120	18
120-130	7
130-140	3
	***n* = 100**



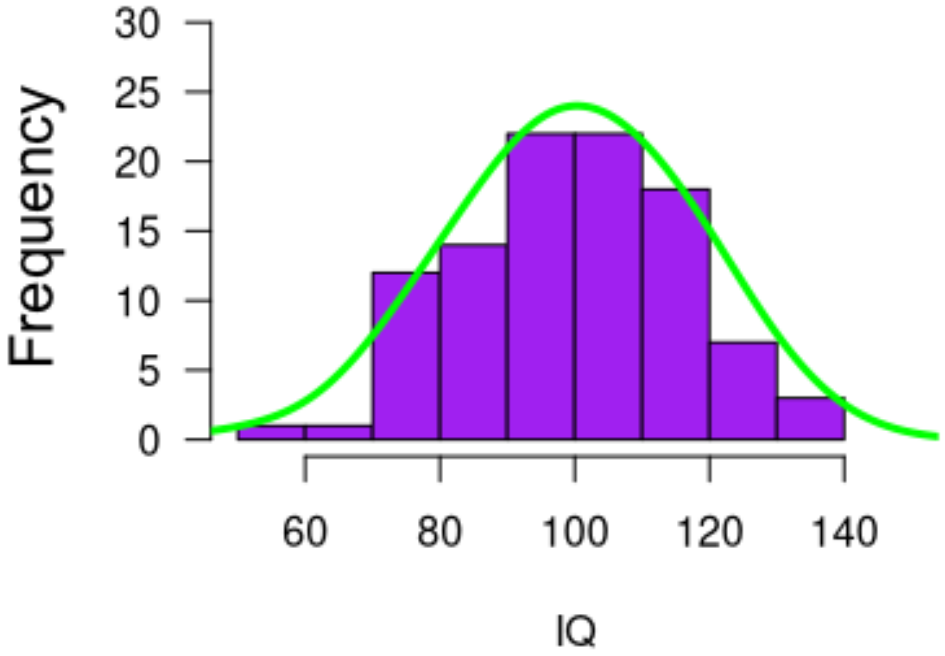
Histogram vs Bar chart



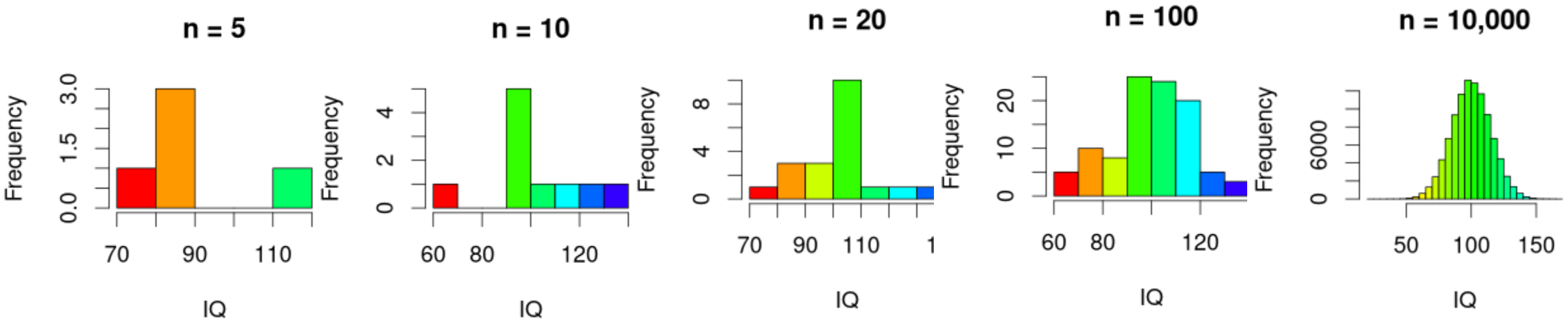


Example 2: IQ

- Really a quantitative variable
- Histogram describes the **sample**
- But the **population** is described by a continuous distribution



From histogram to continuous distribution



Statistic vs parameter

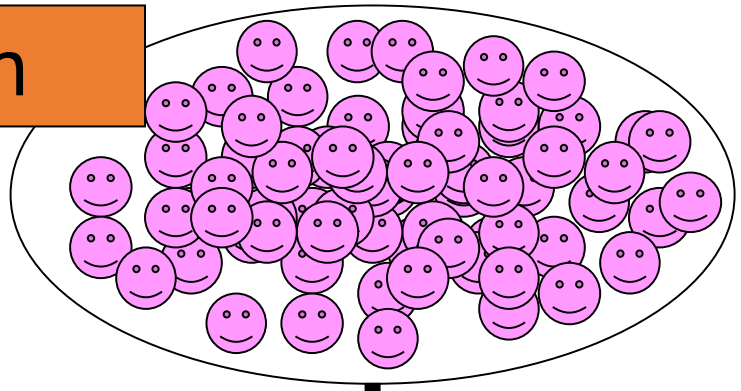
A statistic: A numerical summary of the data is called a statistic

- You can compute this! (mean, median, standard deviation, etc)

A *parameter*: A numerical summary of the population is called a parameter

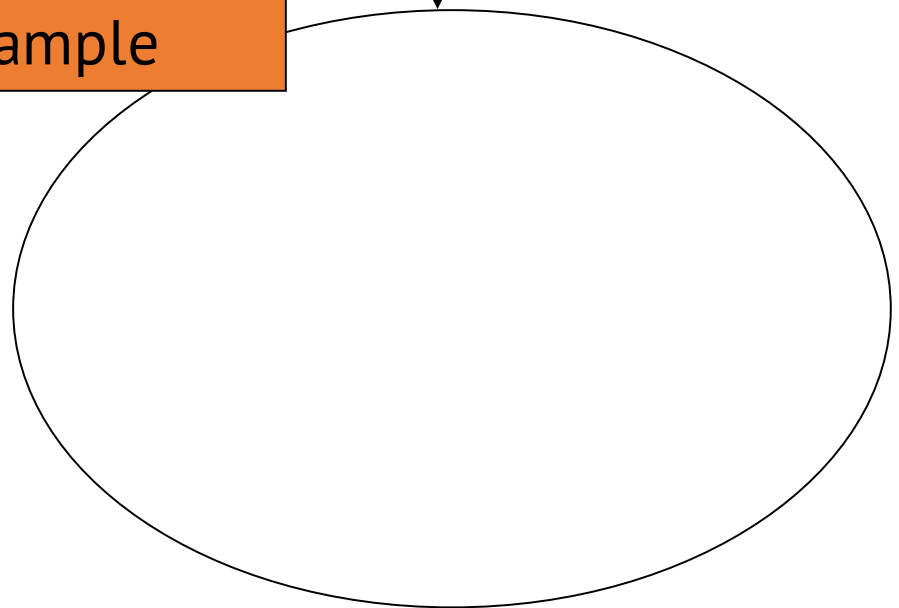
- This cannot be computed (usually)

Population



A *parameter*: A numerical summary of the population is called a parameter

Sample



A statistic: A numerical summary of the data is called a statistic

Statistic vs parameter

A statistic: A numerical summary of the data is called a statistic

- You can compute this! (mean, median, standard deviation, etc)

A *parameter*: A numerical summary of the population is called a parameter

- This cannot be computed (usually)

- ❼ We use statistics to **estimate** parameters

Example: Mean IQ

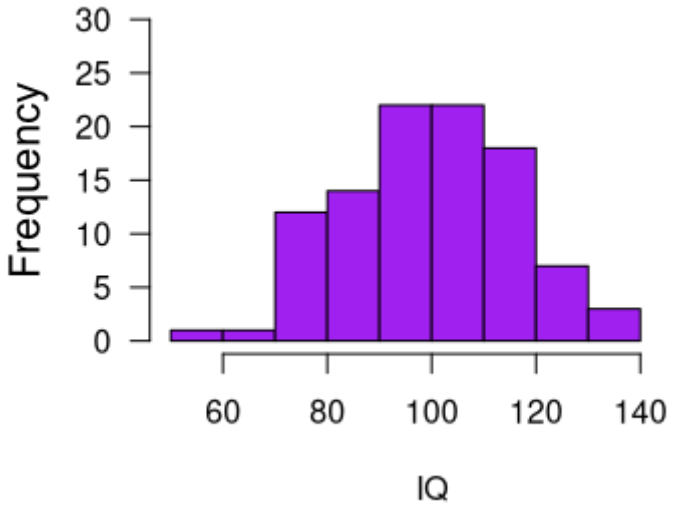
- We want to know the average IQ in particular populations
  - E.g., soccer players vs chess players

- *Location* of the distribution

Source: pixabay.com

Location of the distribution: mean

#	IQ
1	86
2	108
3	64
…	…
99	91
100	**109**

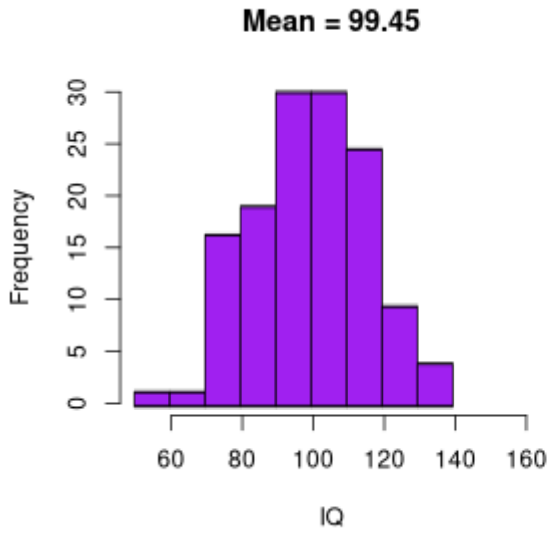


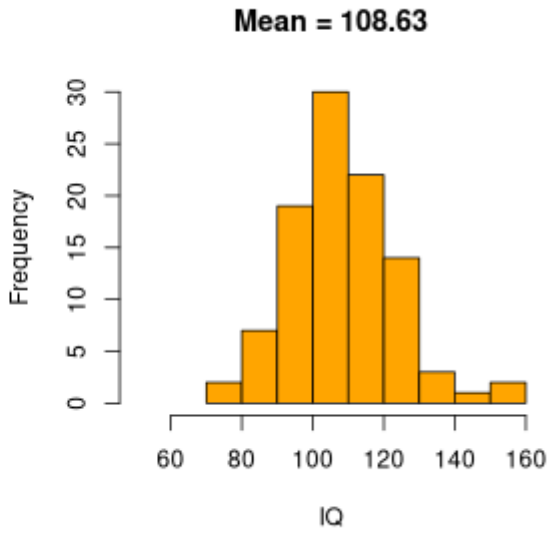
- $\bar{x} = \frac{86+108+64+\cdots+91+109}{100} = 99.45$
- $\bar{x} = \frac{\sum x}{n}$

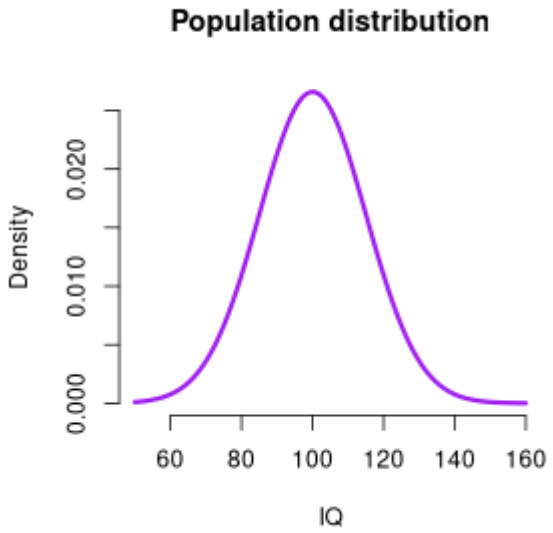
Statistic

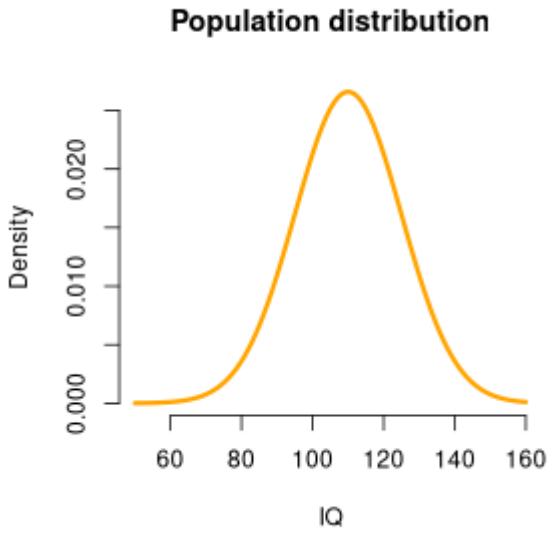
($\sum$ is the symbol for summing up all values of x)

Statistic vs parameter









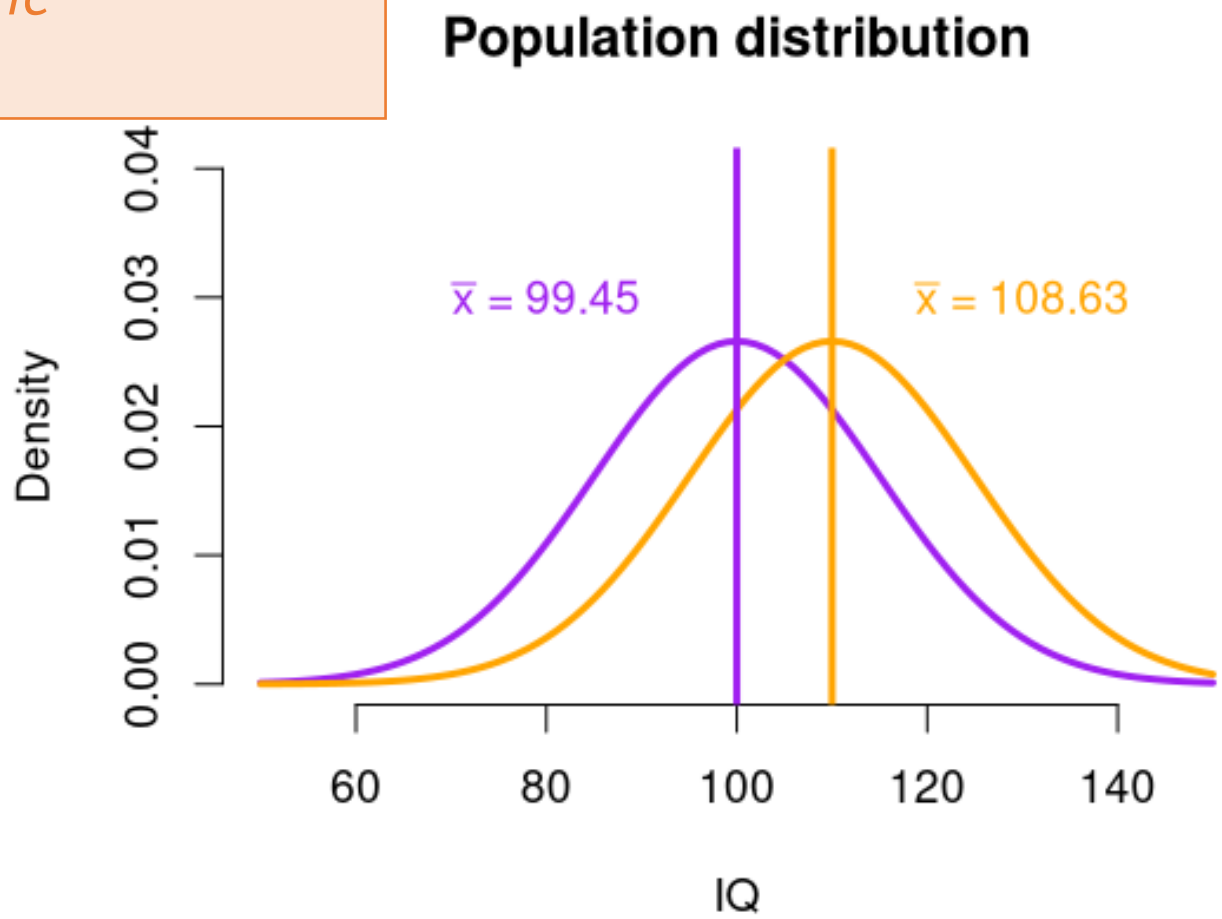
We use the sample to obtain an approximation/estimate of population distribution and parameters

In this case, our best estimate of the population mean, is the sample mean

A difference in means reflects a difference in distributions

Statistic vs parameter

Estimates of the parameters:



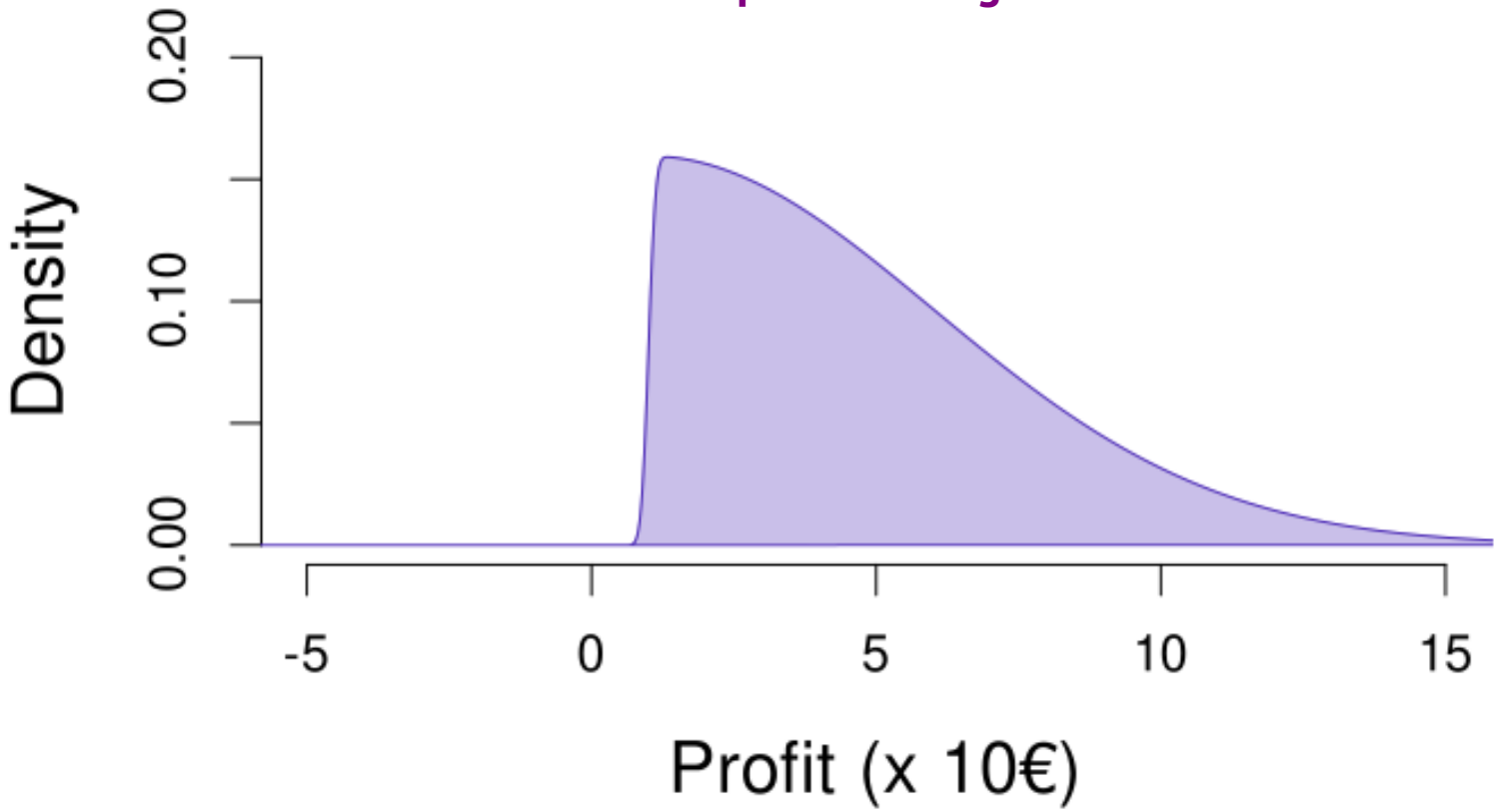
In this case, our best estimate of the population mean, is the sample mean

A difference in means reflects a difference in distributions

Allows us to conclude that there is a difference between the two populations?
→ inferential statistics (later in this course)

Location vs shape of Continuous distributions

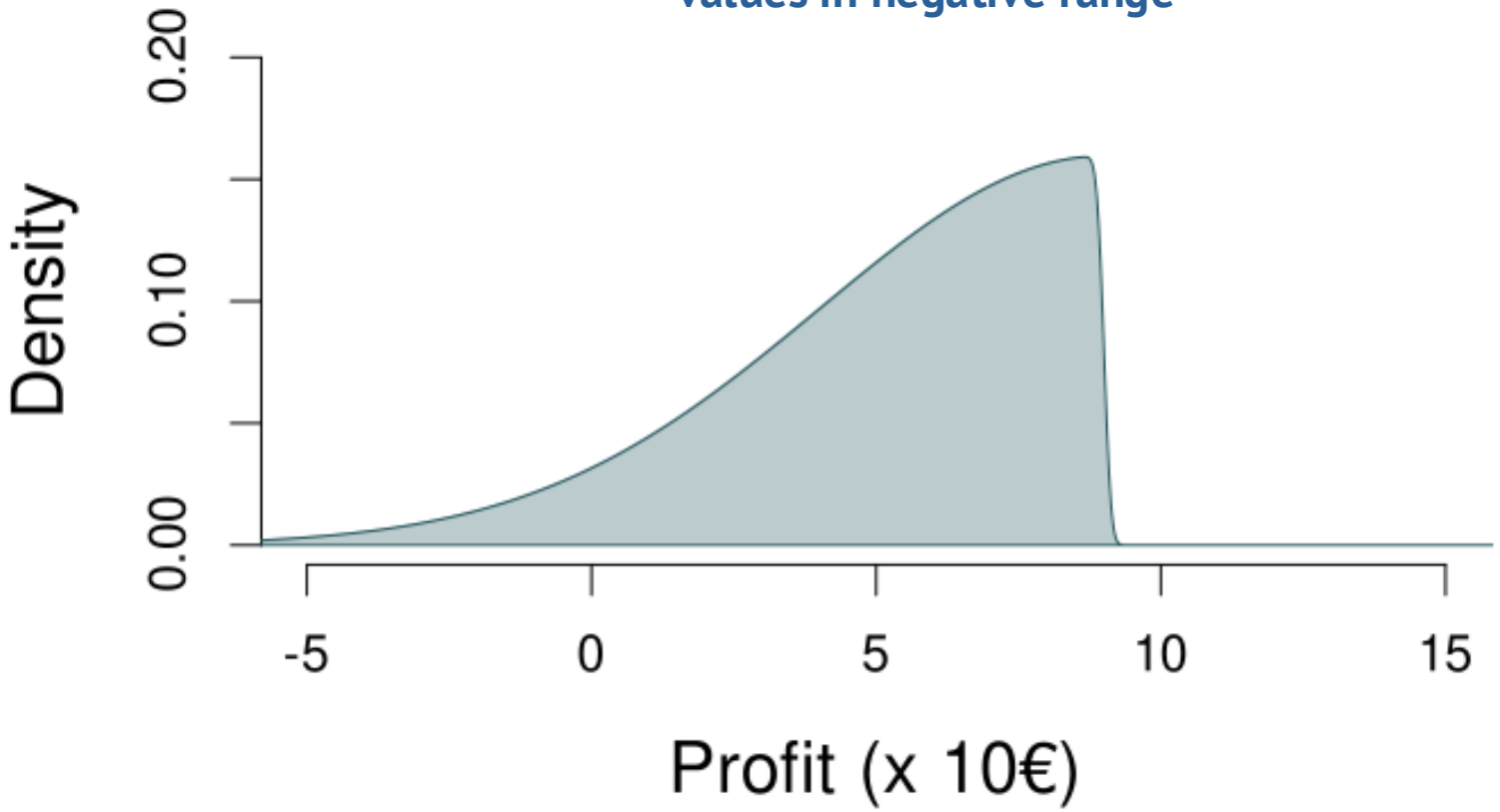
Always a profit, extreme results in the positive range



Source: pixabay.com

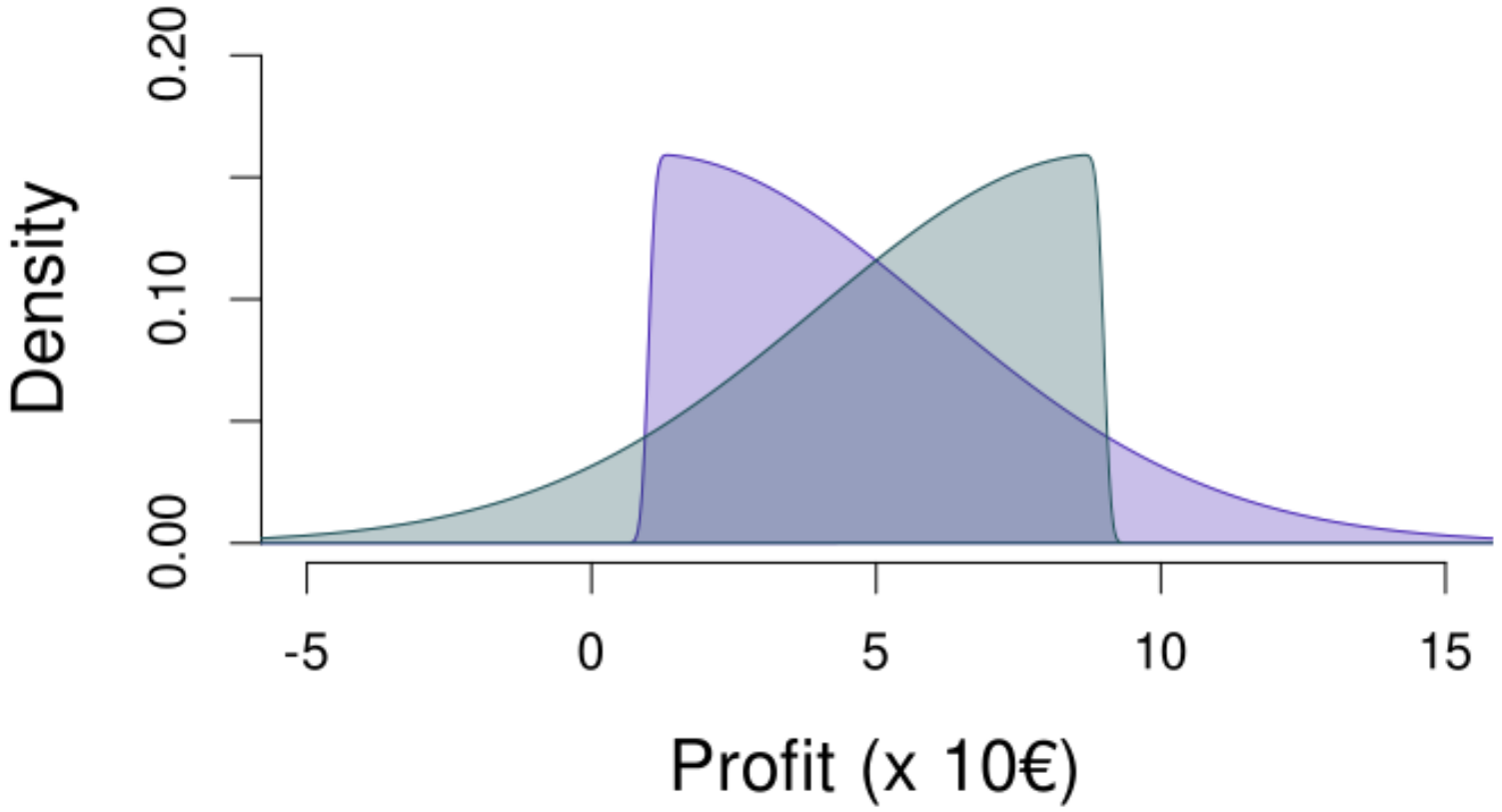
Location vs shape of Continuous distributions

Can suffer loss, extreme values in negative range



Source: pixabay.com

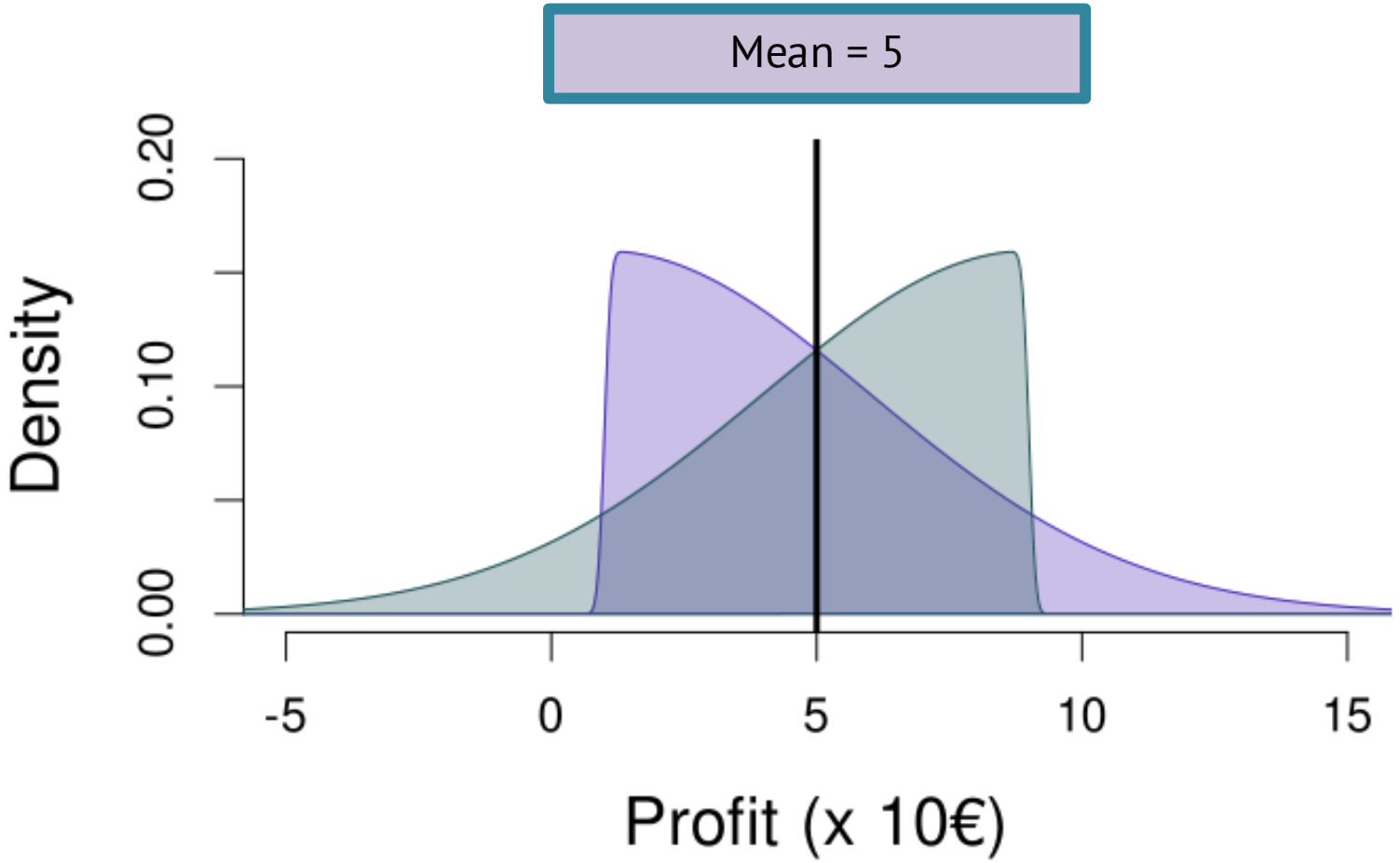
Location vs shape of Continuous distributions



Always a profit, extreme results in the positive range

Can suffer loss, extreme values in negative range

Location vs shape of Continuous distributions



Always a profit, extreme results in the positive range

Can suffer loss, extreme values in negative range

Another location measure: median

- The value that divides the lower 50% of the observations

1. Order all observations
   - If odd *n*: Middle observation

Median = **4**

n = 9

x: 1,9,2,4,5,3,3,7,12

x: 1,2,3,3,**4**,5,7,9,12 (ordered)

   - If even *n*: Mean of middle two observations

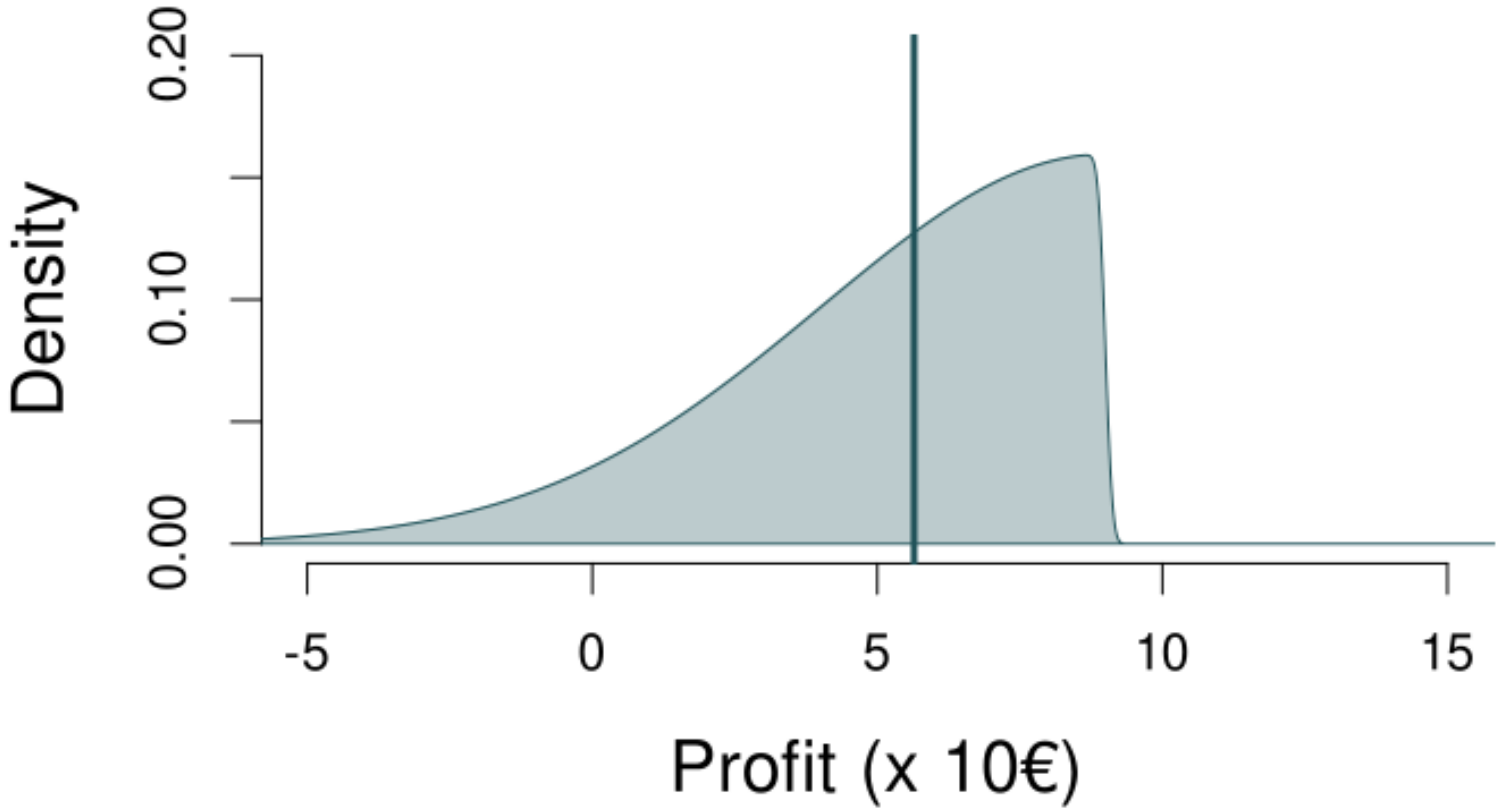
Median = (5+7)/2 = **6**

n = 8

x: 7,9,4,5,-1,8,9,2

x: -1,2,4,**5,7**,8,9,9 (ordered)

The Median of a Continuous Distribution



Median ≈ 5.63

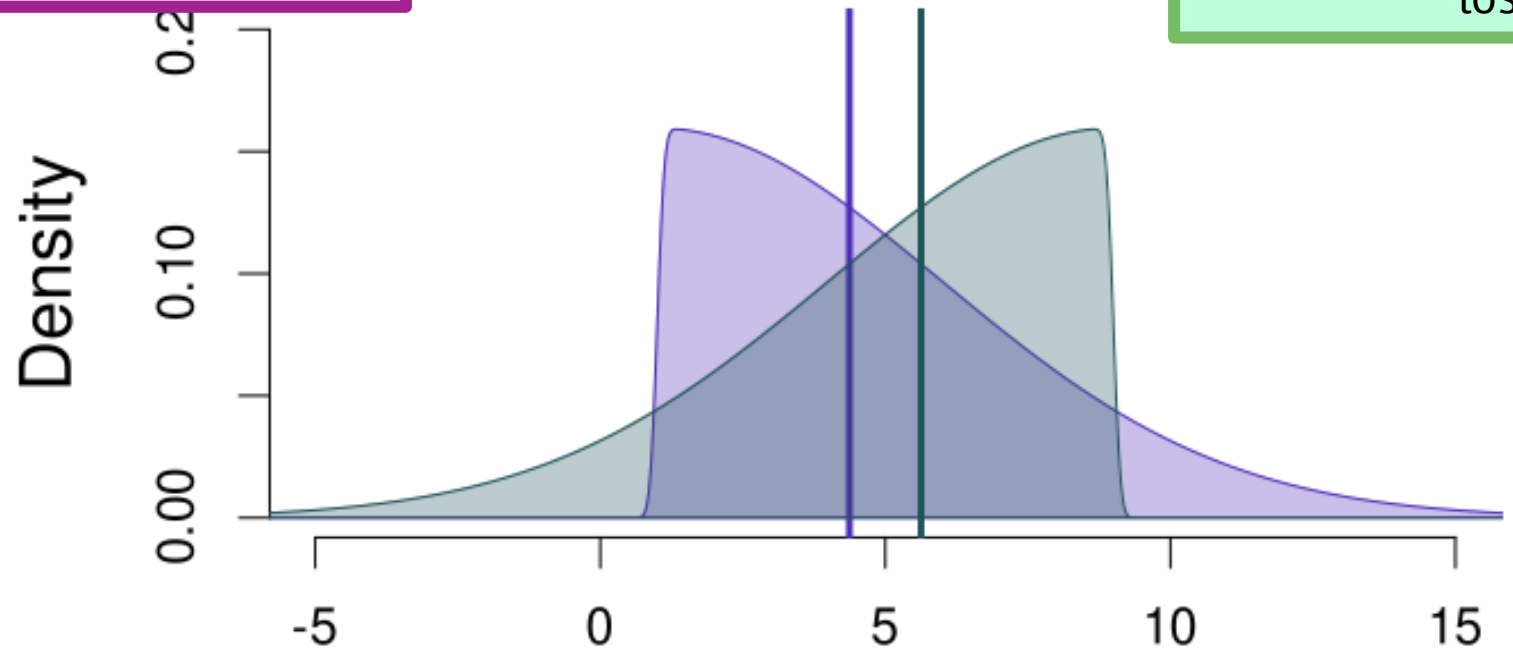
If you play this game, there is a 50% probability that your profit is 5.63 or **lower**

If you play this game, there is a 50% probability that your profit is 5.63 or **higher**

The Median of a Continuous Distribution

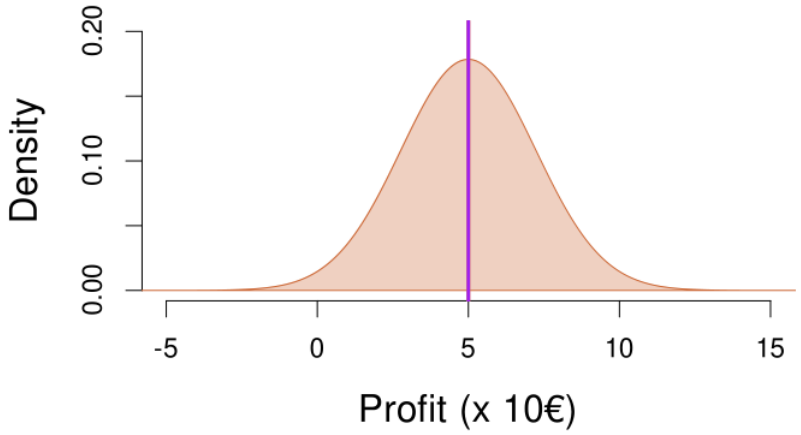
Median ≈ 4.37
Most sessions pay less than average, with a rare big win

Median ≈ 5.63
Most sessions pay more than average, with a rare heavy loss



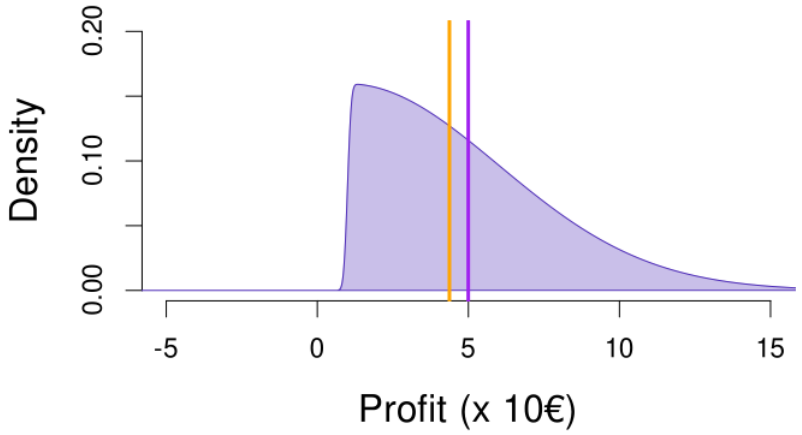
On average (over the course of many sessions), it will not matter for your wallet, but each machine provides a different experience:
Purple: chasing a jackpot, while never losing
Green: getting many bigger wins but suffering the occasional loss

Mean vs Median tells something about shape of a distribution



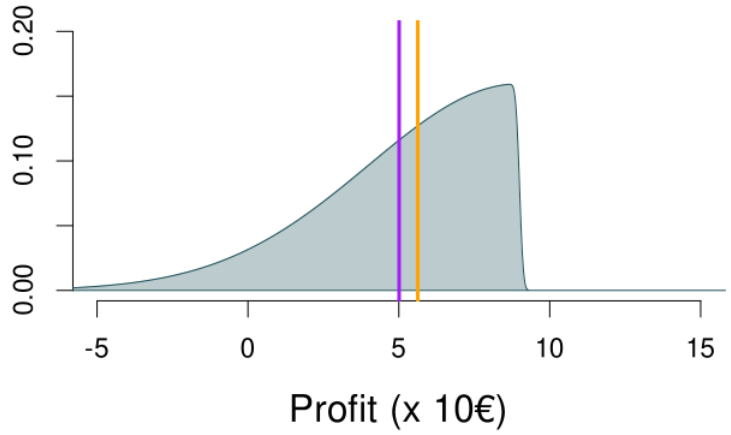
Symmetric distribution

Mean = *Median*



Skewed to the right

Mean > *Median*



Skewed to the left

Mean < *Median*

Why is that?

- x = 3, 5, 4, 8, 10
  - ➔ Mean: (3 + 5 + 4 + 8 + 10) / 5 = 6
  - ➔ Median: n is odd, so middle number of sorted x ⟶ 5
- x = 3, 5, 4, 8, 9000
  - ➔ Mean: (3 + 5 + 4 + 8 + 9000) / 5 = 1804
  - ➔ Median: n is odd, so middle number of sorted x ⟶ 5

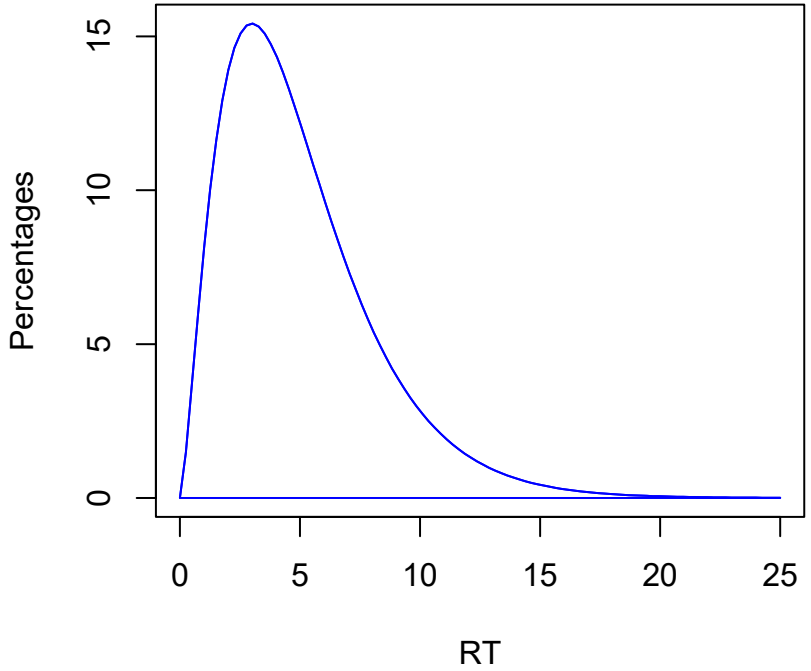
The median ignores extreme values, so is less affected by those (compared to the mean)

Overview of Today

1. Why are statistics needed?
2. A little bit about the course
   - The book: Agresti & Franklin
   - Comparison to high school math
   - How to prepare
3. How can you explore data?
   - Types of data
   - Displaying data
   - Characteristics of a distribution
4. **Recap**
   - Next time
   - Example exam question

Example question

- The figure displays a skewed distribution



- Which statement about the mean and median of this distribution is true?

a. median < mean

b. median = mean

c. median > mean

Recap of Today

- Why statistics?
  - Get an overview of data, numerically or graphically → descriptive statistics
  - Make statements about the whole population, based on a sample from the population → inferential statistics
- Different types of variables exist (e.g., discrete vs. continuous)

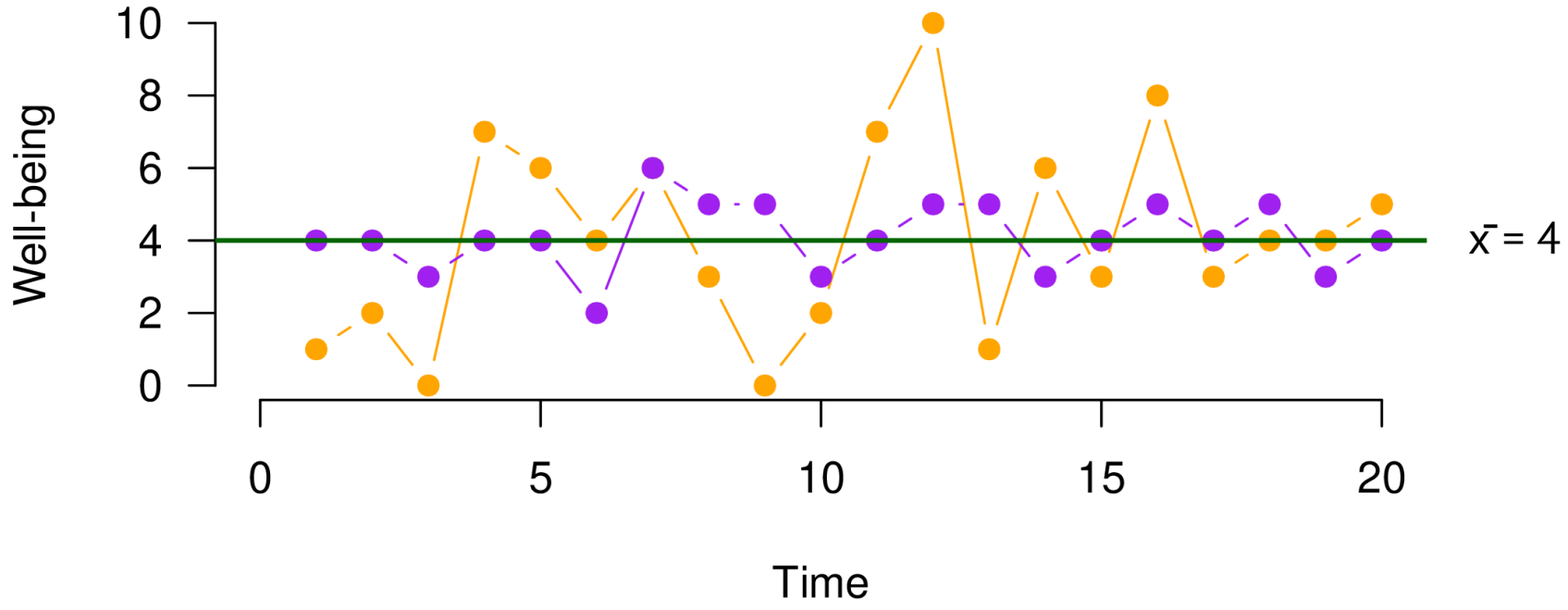
- Science quality = methods + statistics

Design experiments & gather data

Analyze the data and draw conclusions

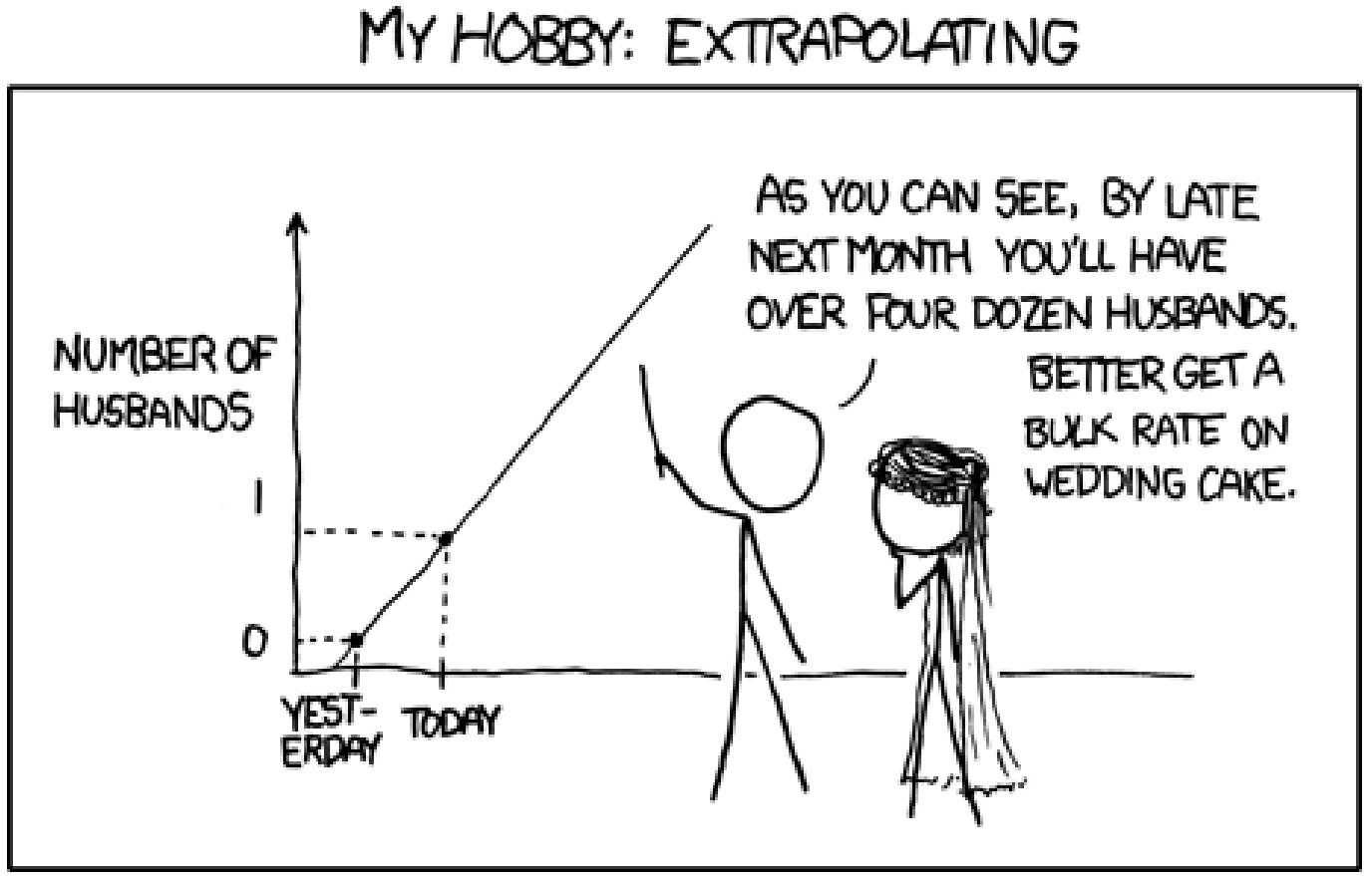
Next time

- Variability in behavior



Questions?

Thank you for your attention



Source: https://www.xkcd.com/605/

Bonus Video

Hans Rosling on Data Visualization

https://www.youtube.com/watch?v=jbkSRLYSojo

"Having the data is not enough – I have to show it in ways people enjoy and understand"