Introduction into Biostatistics

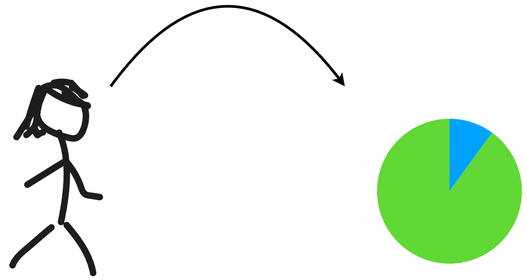
Anna Poetsch, Biotechnology Center, TU Dresden

Organisation

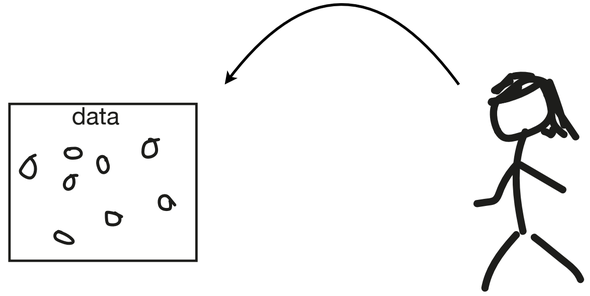
- **9.5. Introduction to biostatistics**
- **16.5. Descriptive statistics**
- **23.5. Hypothesis testing**
- 6.6. Introduction machine learning (Robert)
- 13.6. Unsupervised Machine learning (Melissa)
- 20.6. Supervised machine learning/ deep learning (Melissa)
- 21.6. Introduction into genomics data
- 4.7. Multimodal machine learning
- 11.7. Summary (all)

Recap on probability

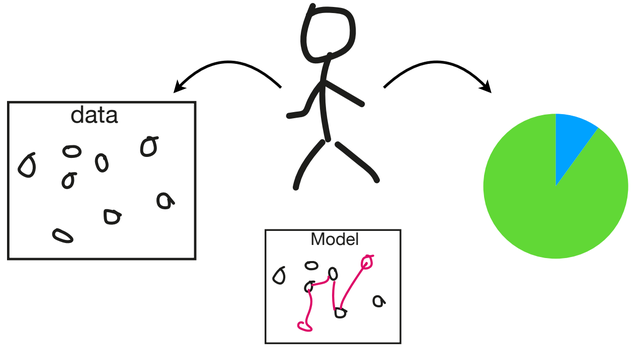
A model



Data



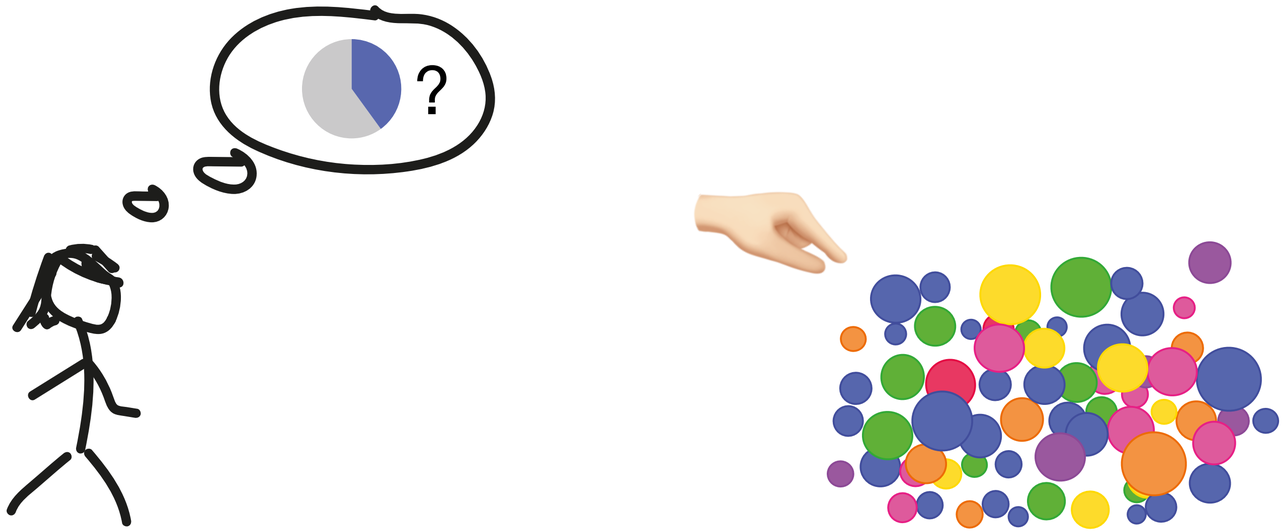
A model based on data



! **Estimating probabilities can only be as good as your assumptions/ data**

Recap on confidence of one probability measurement

- A random (or representative) sample!
- They are independent observations!
- The data are accurate!



How can data be not accurate?

Biases!

How can data be not accurate?

Biases!

- Regression to the mean
- Sampling biases:
  - Survivorship bias
  - Volunteer bias
  - Non-response bias
  - Sampling-frame bias
- Confirmation bias
- Measurement bias
- Selection bias
- Reporting bias
- Publication bias

How can data be not accurate?

Biases!

- Regression to the mean
- Sampling biases:
  - Survivorship bias
  - Volunteer bias
  - Non-response bias
  - Sampling-frame bias
- Confirmation bias
- Measurement bias
- Selection bias
- Reporting bias
- Publication bias

How can data be not accurate?

Biases!

- Regression to the mean
- Sampling biases:
  - Survivorship bias
  - Volunteer bias
  - Non-response bias
  - Sampling-frame bias
- Confirmation bias
- Measurement bias
- Selection bias
- Reporting bias
- Publication bias

How can data be not accurate?

Biases!

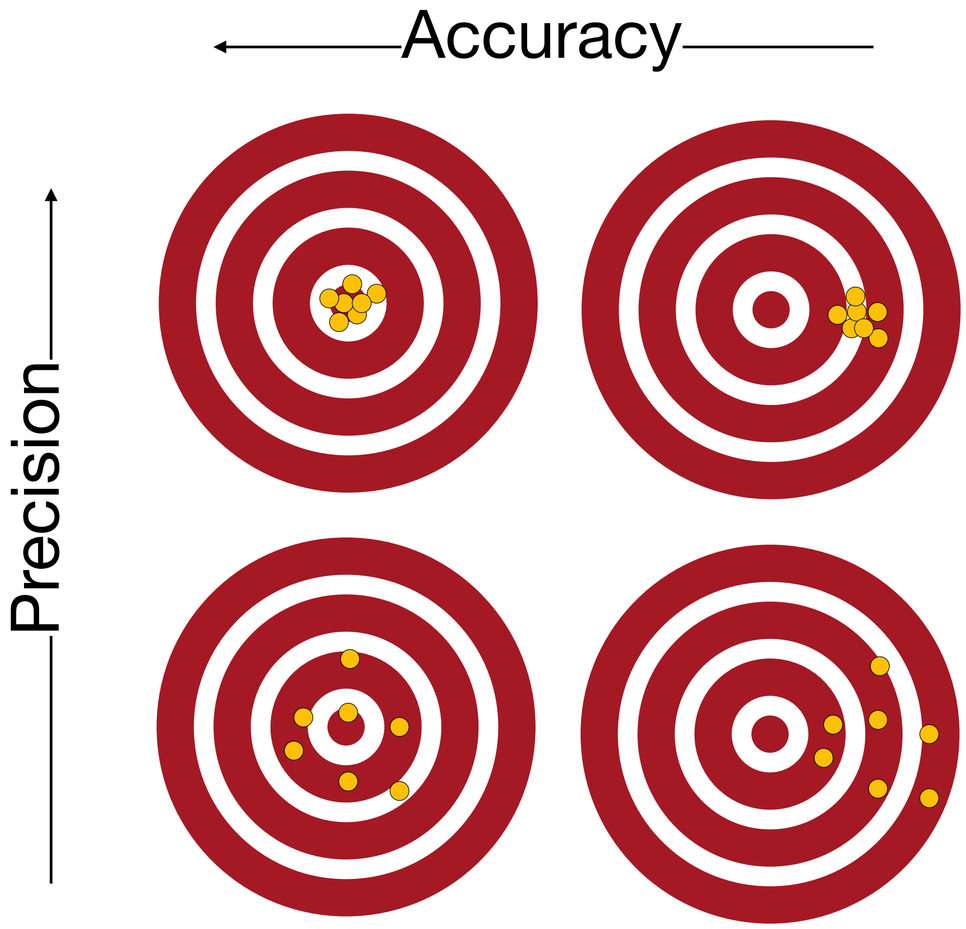
- Regression to the mean
- Sampling biases:
  - Survivorship bias
  - Volunteer bias
  - Non-response bias
  - Sampling-frame bias
- Confirmation bias
- Measurement bias
- Selection bias
- Reporting bias
- Publication bias

How can data be not accurate?

Biases!

- Regression to the mean
- Sampling biases:
  - Survivorship bias
  - Volunteer bias
  - Non-response bias
  - Sampling-frame bias
- Confirmation bias
- Measurement bias
- Selection bias
- Reporting bias
- Publication bias

Recap on accuracy and precision



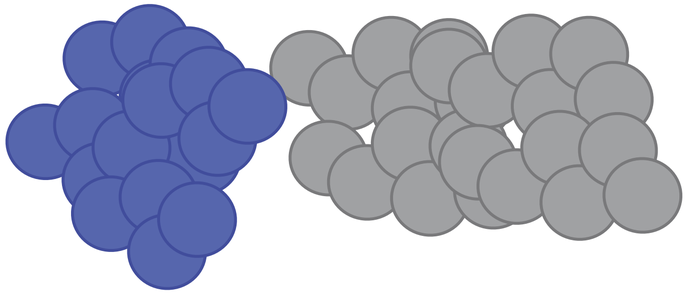
How do these relate to confidence intervals?
Does the confidence interval get bigger …
….if you increase n?
….if you increase the confidence level,
e.g. from 95% to 99%?

Know your problem, know your distribution!

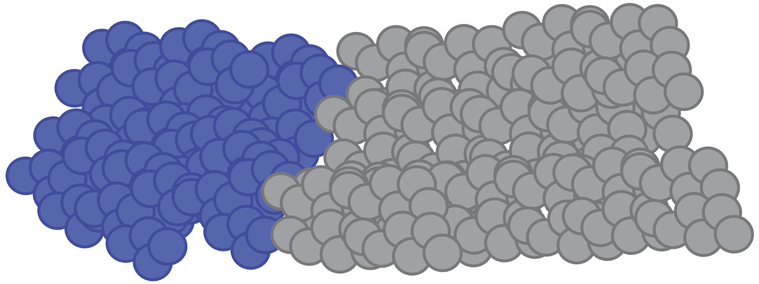
Binomial

40%

2/5



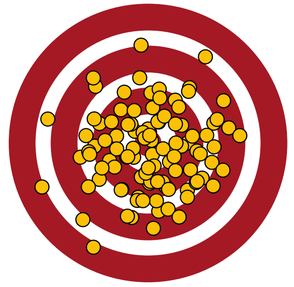
20/50

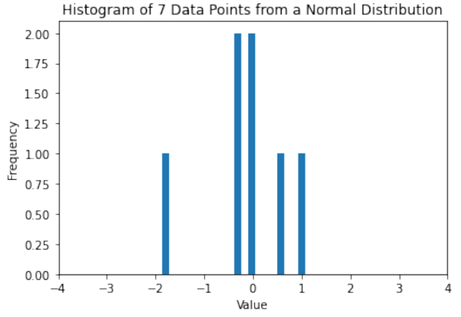


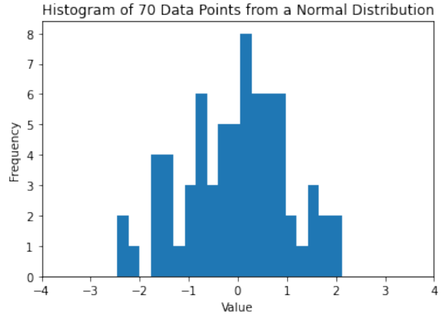
200/500

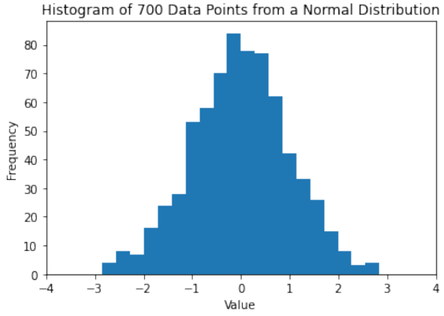
Confidence increases with n

Normal









Confidence does not increase with n

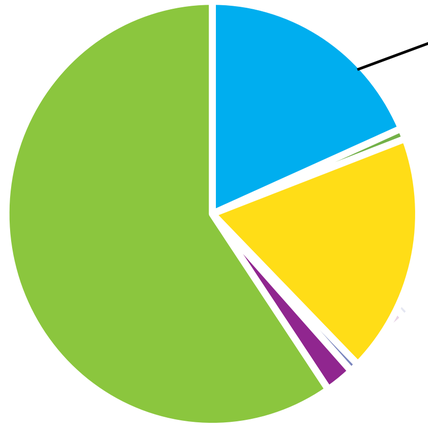
Descriptive statistics

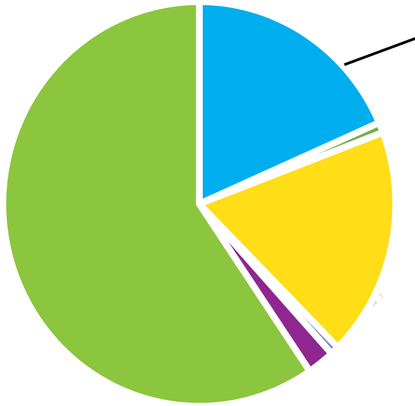
Data types and distributions

- Probability data (binomial distribution)
- Counted data (Poisson distribution)
- Normal distribution

Probability data/ Binominal distribution

Election results



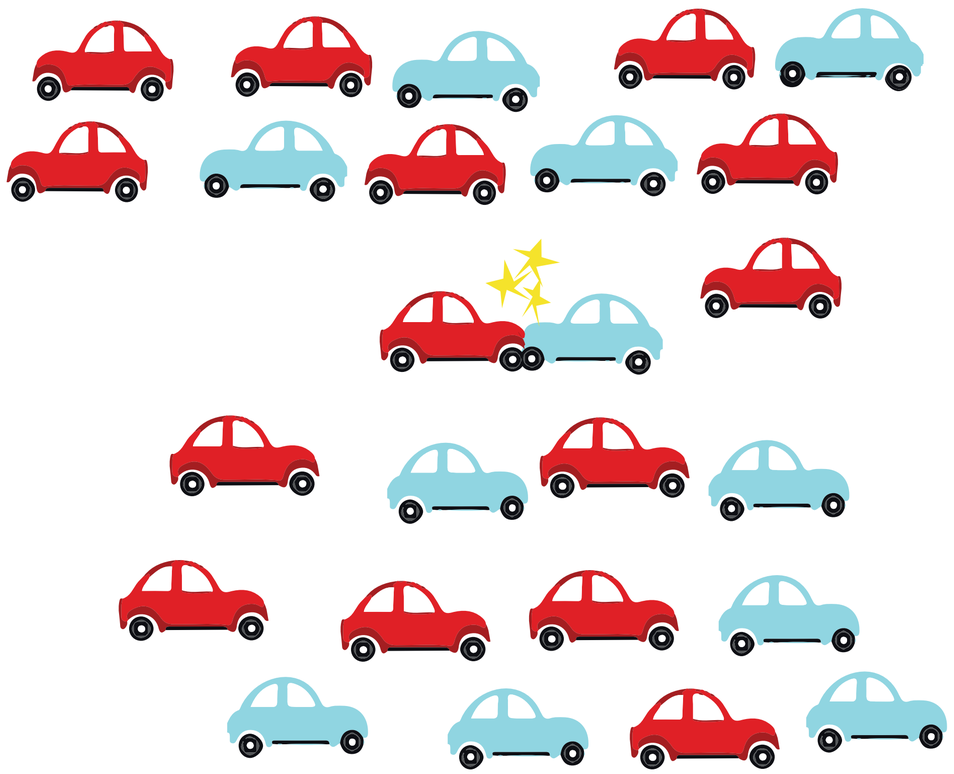


Based on “representative” studies, election outcomes can be predicted.

Why is this frequently going wrong?

Probability data/ Binominal distribution

Traffic accidents

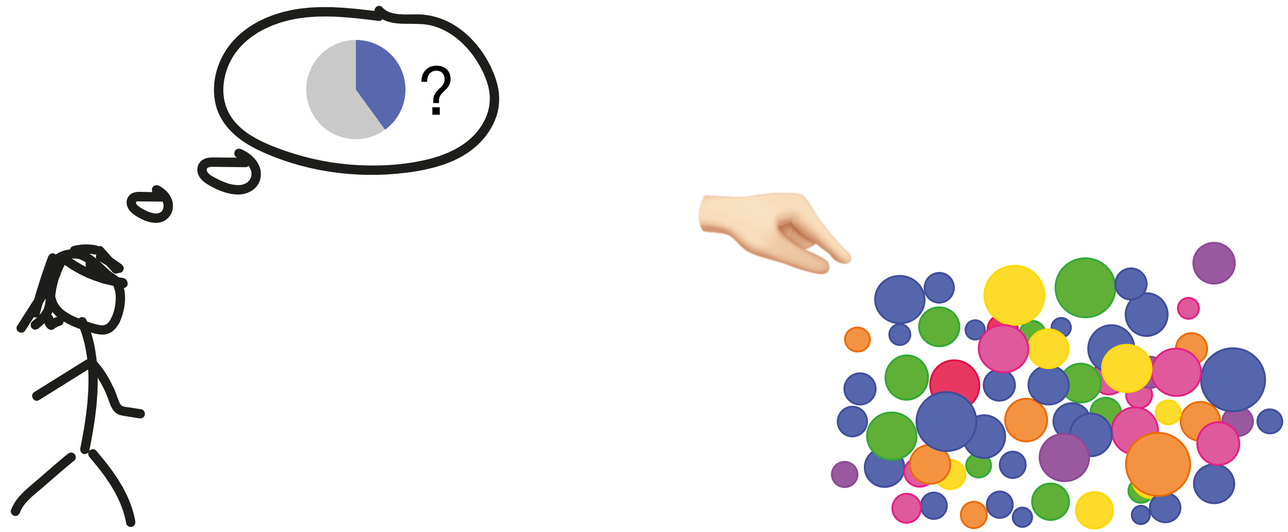


Relative to the number

of cars passing!

The probability of an accident occurring can be determined based on data
What are possible sources of sampling bias?

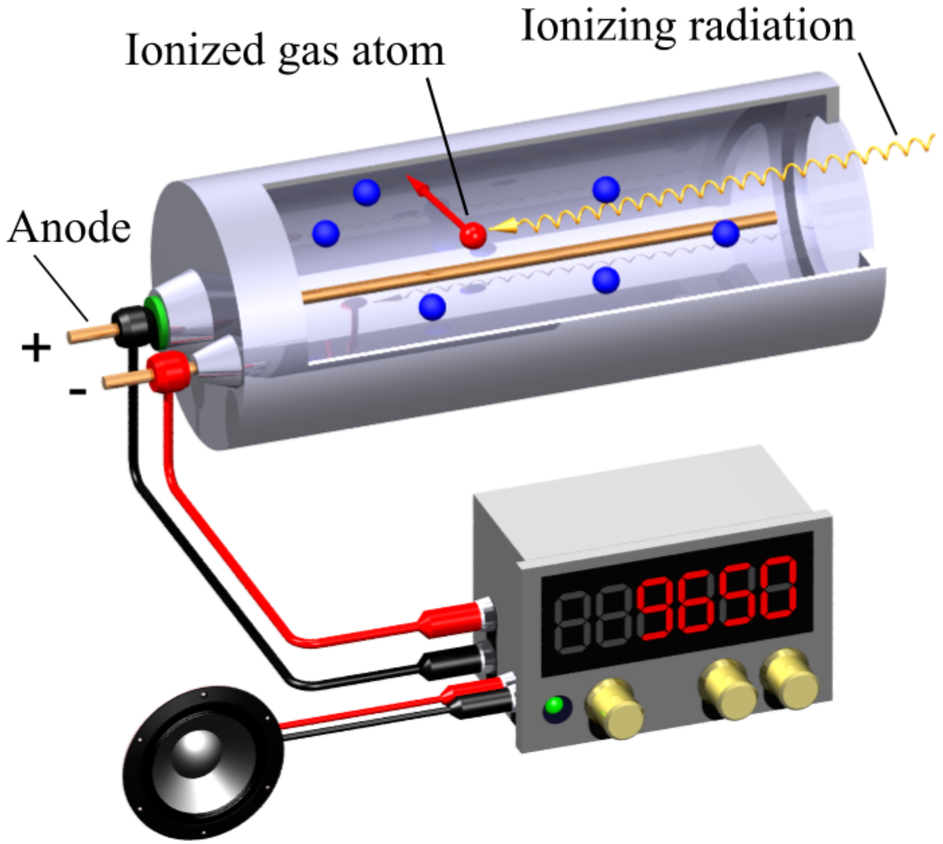
Probability data/ Binominal distribution



-> Jupyter Notebook

Counted data/ Poisson distribution

Radioactive decay



How can you view radioactive decay as a binomial distribution?

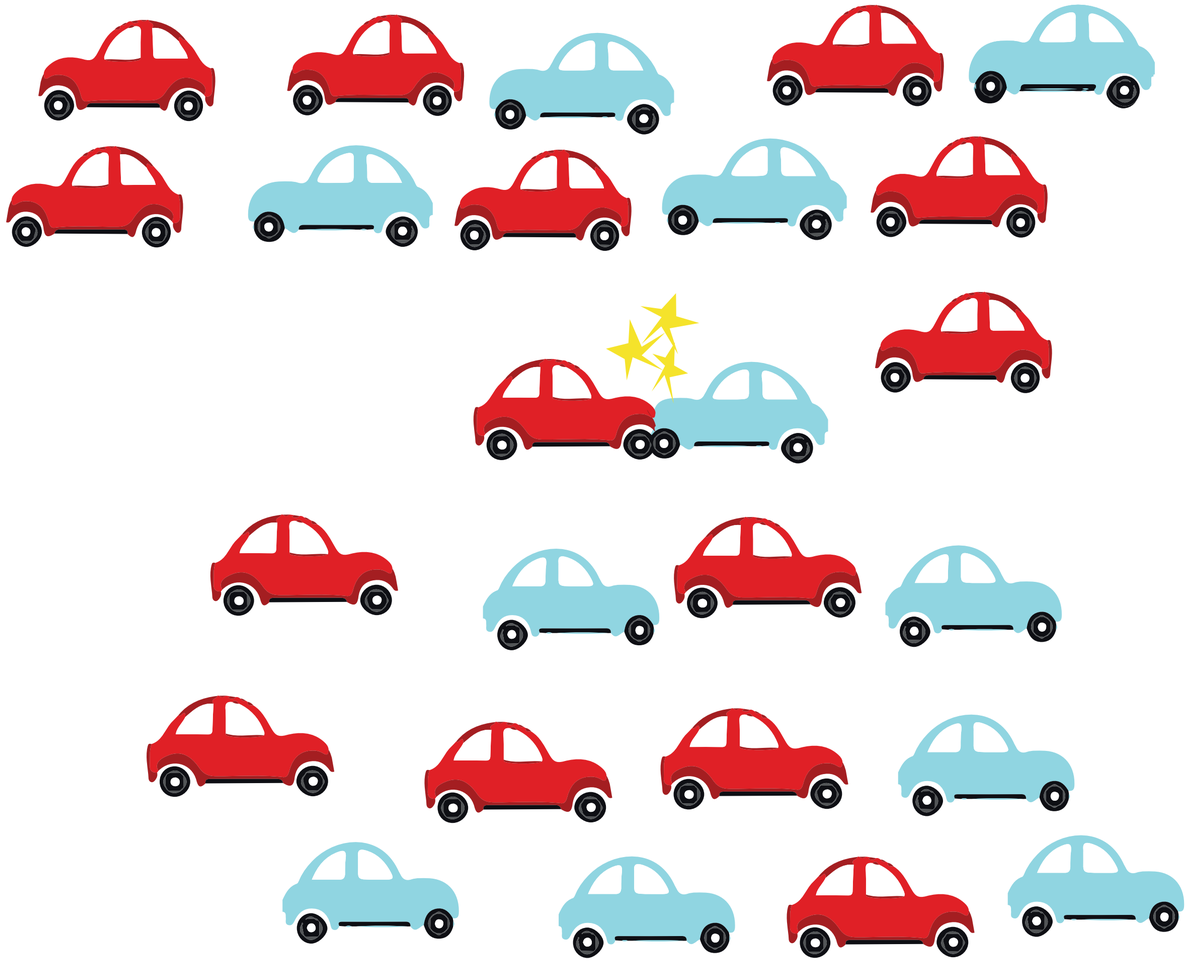
How many raisins are in a Christstollen?

How can you view raisins as a binomial distribution?

How many mutations are in a cancer genome?

How can you view mutations as a binomial distribution?

Is the traffic problem really for a binomial distribution?



Counted data/ Poisson distribution

-> Jupyter Notebook

Break

Types of variables

Discrete variables

Ordinal variables

- limited set of discrete values with order

e.g. scale from 1-10

Discrete variables

Ordinal variables

- limited set of discrete values with order

e.g. scale from 1-10

Nominal, binomial variables

- limited set of discrete values without order

e.g. responder <-> non responder

Discrete variables

Ordinal variables

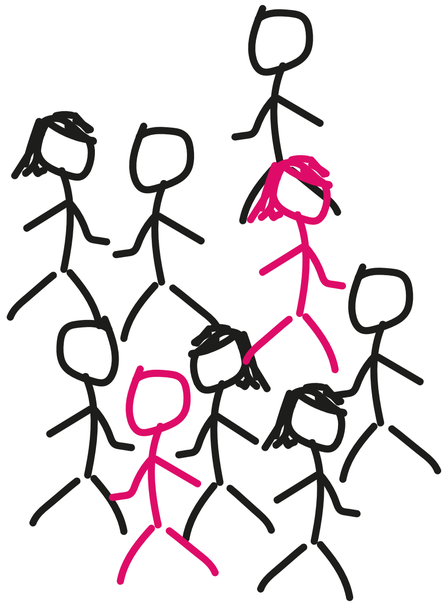
- limited set of discrete values with order

e.g. scale from 1-10

Nominal, binomial variables

- limited set of discrete values without order

e.g. responder <-> non responder



Discrete variables

Ordinal variables

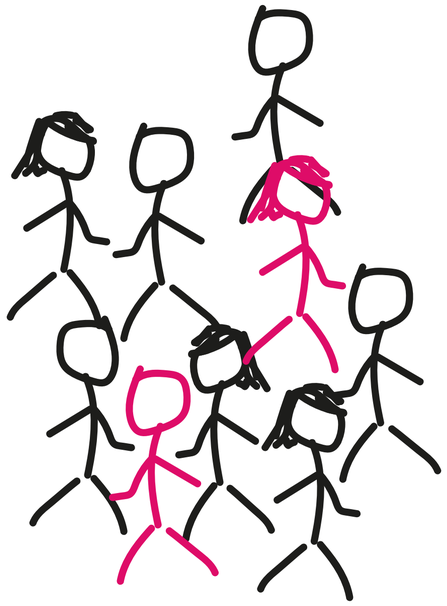
- limited set of discrete values with order

e.g. scale from 1-10

Nominal, binomial variables

- limited set of discrete values without order

e.g. responder <-> non responder



What colours would you use to visualise such variables?

Continuous variables

Interval variables

- continuous value, for which intervals make sense, but no ratios

e.g. °C

Ratio variables

- continuous value, for which ratios make sense

e.g. height, weight, enzyme activity, Kelvin

Continuous variables

Interval variables

- continuous value, for which intervals make sense, but no ratios

e.g. °C

Ratio variables

- continuous value, for which ratios make sense

e.g. height, weight, enzyme activity, Kelvin

What colours would you use to visualise such variables?

Descriptive statistics

Summary statistics

- Min, max, mean
- Mode
- Median and quartiles
- Confidence intervals

Summary parameters

1 2 2 5 5 5 10 30

Summary parameters

Min value: 1

Max value: 30

1 2 2 5 5 5 10 30

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

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

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean (μ): (1+2+2+5+5+5+10+30)/8 = 7.5

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

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

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean (μ): (1+2+2+5+5+5+10+30)/8 = 7.5

Variance: $\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

((1-7.5)²+(2-7.5)²+(2-7.5)²+(5-7.5)²+(5-7.5)²+(5-7.5)²+(10-7.5)²+(30-7.5)²)/8 = 90.57143

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

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean (μ): (1+2+2+5+5+5+10+30)/8 = 7.5

Variance: $\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

((1-7.5)²+(2-7.5)²+(2-7.5)²+(5-7.5)²+(5-7.5)²+(5-7.5)²+(10-7.5)²+(30-7.5)²)/8 = 90.57143

SD: square_root (variance) = 9.516902

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

Summary parameters

1 2 2 5 5 5 10 30

Min value: 1

Max value: 30

Parametric measures

Mean (μ): (1+2+2+5+5+5+10+30)/8 = 7.5

Variance: $\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

((1-7.5)²+(2-7.5)²+(2-7.5)²+(5-7.5)²+(5-7.5)²+(5-7.5)²+(10-7.5)²+(30-7.5)²)/8 = 90.57143

SD: square_root (variance) = 9.516902

SD = standard deviation = sigma

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

non-parametric measures:

	1	2	2	5	5	5	10	30
Ranks:	1	2	2	4	4	4	7	8

non-parametric measures:

	1	2	2	5	5	5	10	30
Ranks:	1	2	2	4	4	4	7	8

Median: the central value: 5

non-parametric measures:

	1	2	2	5	5	5	10	30
Ranks:	1	2	2	4	4	4	7	8

Median: the central value: 5

Quartiles: the value of the lower and upper quarter: 2, 6.25

non-parametric measures:

	1	2	2	5	5	5	10	30
Ranks:	1	2	2	4	4	4	7	8

Median: the central value: 5

Quartiles: the value of the lower and upper quarter: 2, 6.25

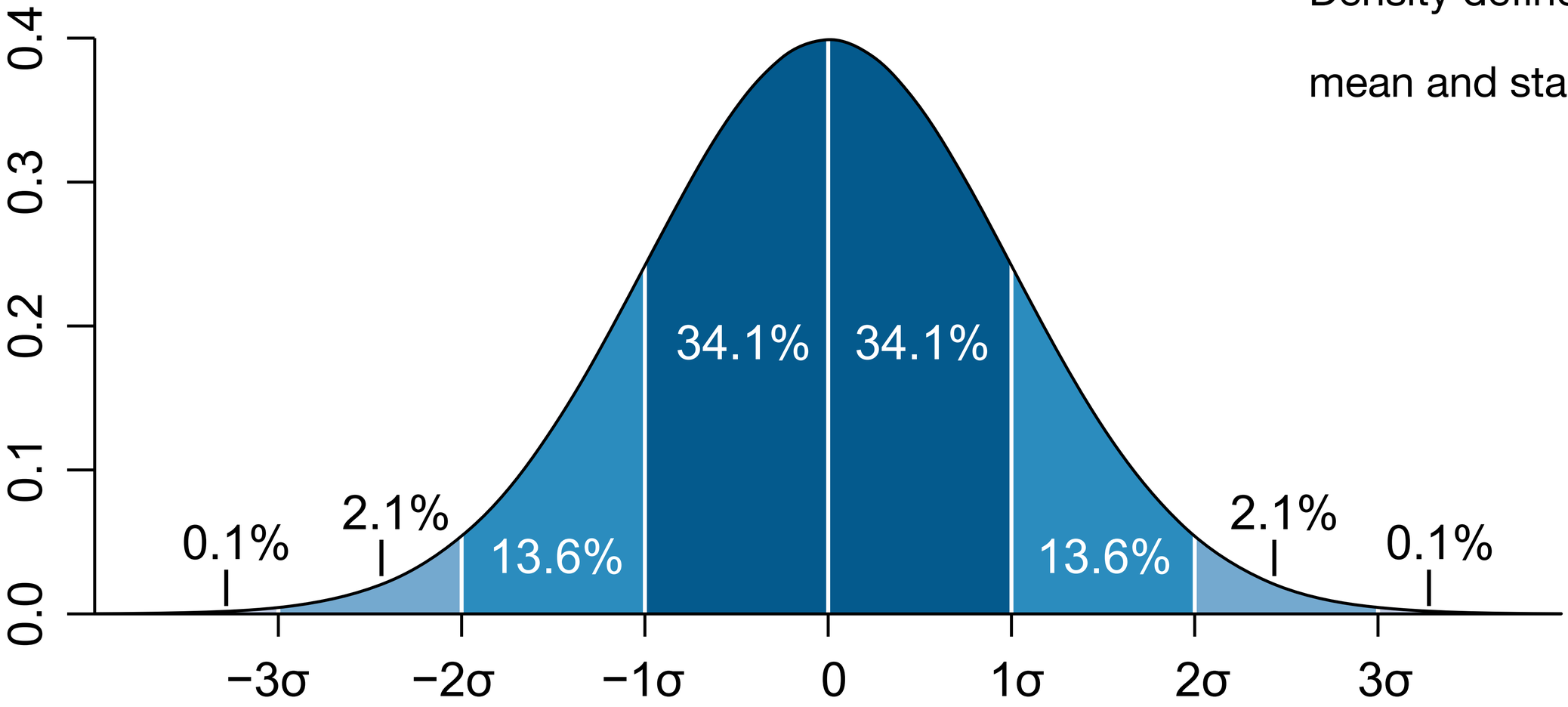
Inter quartile range (IQR): 6.25-2

Normal distribution

Gaussian distribution, bell-shaped distribution

Density defined by

mean and standard deviation



The graph is adapted from: M. W. Toews - Own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

Normal distribution

Gaussian distribution, bell-shaped distribution

The result of general imprecision: weighing, pipetting, randomness

Therefore also: height, weight, or is it?

What else?

-> Jupiter Notebook

Summary

- Probability data (Binomial distribution)
- Count data (Poisson distribution)
- Categorical and continuous data types
- Normal distributions
- Describing a distribution (mean, median, standard deviation, mode, error)