Statistics
367-1-4361
Probabilities

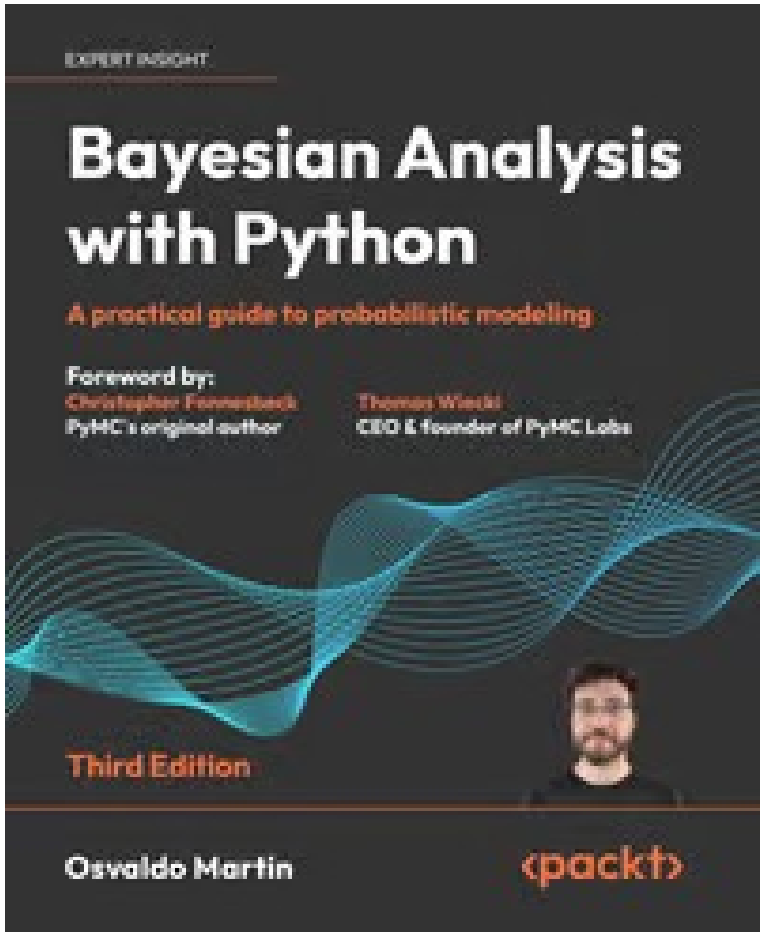
Opher Donchin

Course goals

- Understand what data is
- Create models that may explain data
- Draw conclusions about the models from the data
- Test hypotheses by comparing models

Bayesian Analysis with Python

- 3rd edition!!
- GitHub
- Publisher's website
- Full text available
  - the library
- Easy to read!



Grading

- **Pre-lecture questions**
  - Questions are worth 5% of the grade
- **Exercises**
  - Exercises are worth15% of the grade
  - 5 exercises. Each worth 3.6%
- **Midterm quiz**
  - 10% of the grade. **Can only improve grade!**
  - On computers during recitation section on 17-18/6
  - Practice for exam
- **Exam**
  - 70% of the grade
  - On computers on campus
  - Open book (material on Moodle only!)
  - 2 hours
  - Students who do not pass the exam **will not pass the course**

1A What is data

What is statistics

- Collecting
- Organizing
- Analyzing
- Interpreting

Data
Data
Data
Data

Its two main goals

- Exploratory data analysis
- Inference

From Truth to Data

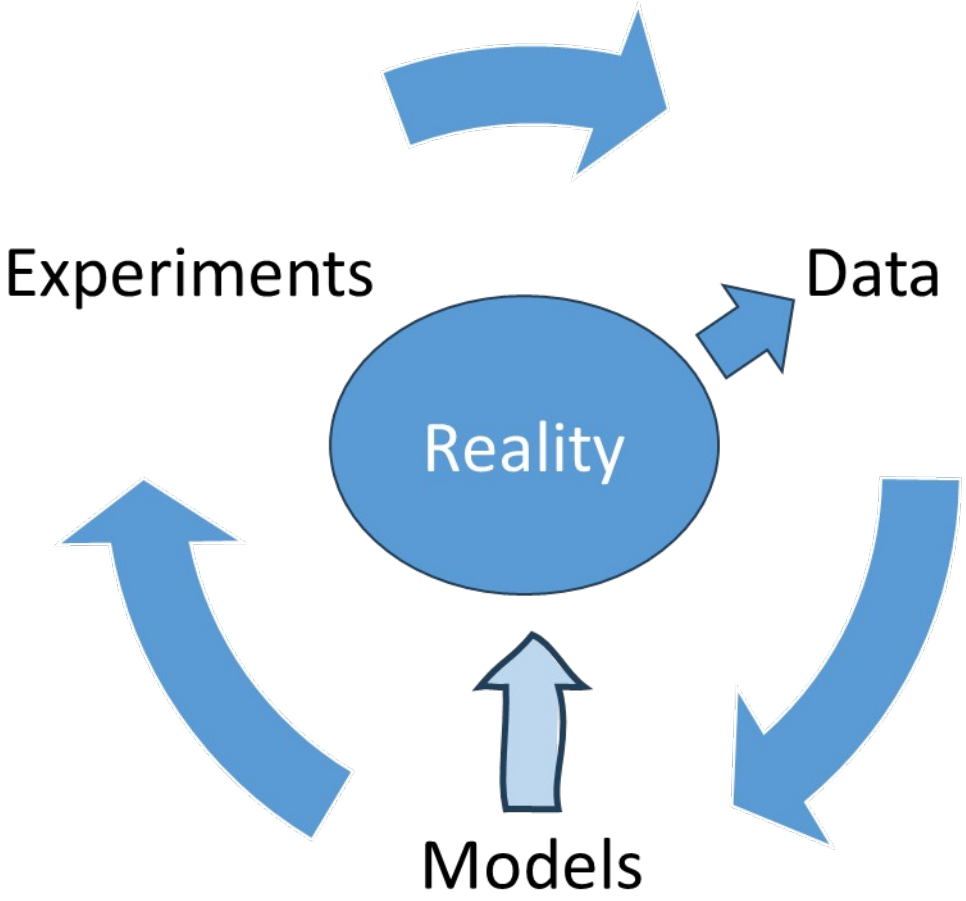
- Francis Bacon
  - *Novum Organum* 1620
  - **Facts** and **observations** give us evidence
- Pierre-Simon Laplace
  - *Theorie analytique des probabilities* 1812
  - **Observations** are **givens**, but they are uncertain
- Carl Friedrich Gauss
  - *Theoria motor corporum coelestium* 1809
  - Process **givens** to improve **estimates**
  - In Latin: **givens** is **data**

Data gives birth to statistics

- Adolphe Quetelet
  - *Sur l'homme* 1835
  - Data on heights implies an "average man"
- Francis Galton
  - *Hereditary Genius* 1869
  - Multi-variate data
  - Variability and co-variance
- Karl Pearson
  - *The Grammar of Science* 1892
  - Using data to establish scientific truth
- Ronald Fisher
  - *The Design of Experiments* 1935
  - Data is not given; it is generated by design

Reality, Data, and Models

- Experiments produce data from reality
- Data updates models of reality
- Models lead to further experiments



Sources of data in biomedical engineering

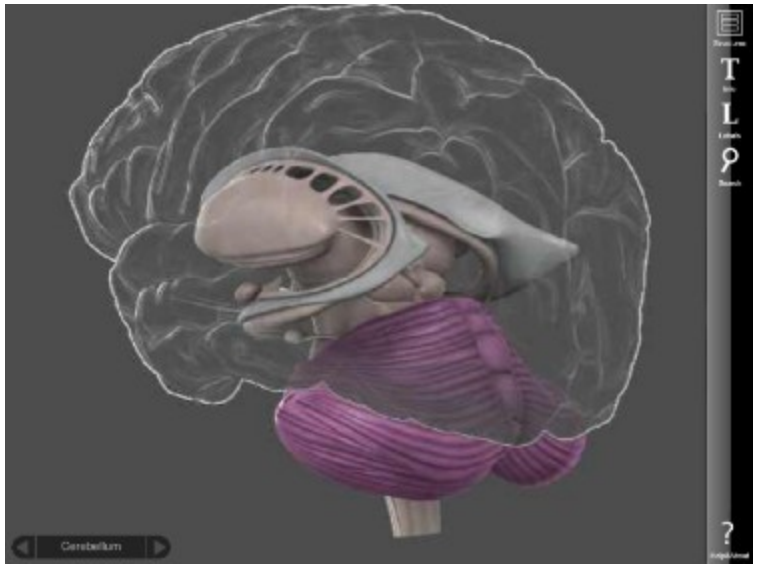
- Genomic data
- Biological data
- Physiological data
- Clinical data
- Models and simulations
- New sources:
  - Large datasets
  - Wearable technology
  - Image and video processing

Uses of data in biomedical engineering

- Basic science
- Clinical testing
- Clinical trials
- Calibrating models
- Product development

An example

- Cerebellar volume



http://de.wikibooks.org/wiki/Neuroanatomie:_Kleinhirn

http://brainposts.blogspot.co.il/2011/11/neuropsychology-and-cerebellum-part-i.html

What is data

- An ordered set of ‘observations’ or ‘measurements’
  - Each observation could be vector valued
  - Data is finite!

$$\left(x_1, x_2, x_3, \cdots, x_N\right)$$

$$x_1 = \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \end{pmatrix}$$

What is data

- An ordered set of ‘observations’ or ‘measurements’
  - Each observation could be vector valued
  - Data is finite!
- Analyze the brain scans of 5 individuals and extract the size of each cerebellum
- We have to choose the right thing to measure

Cerebellar volume in liters:
0.16 0.15 0.18 0.21 0.16

Cerebellar volume in fraction of total intracranial volume (TICV):
0.13 0.13 0.12 0.13 0.13

What is data

- An ordered set of ‘observations’ or ‘measurements’
  - Each observation could be vector valued
  - Data is finite!
- Analyze brain scans of 60 subjects and extract cerebellar volumes
- We can get more data, but it will still be data

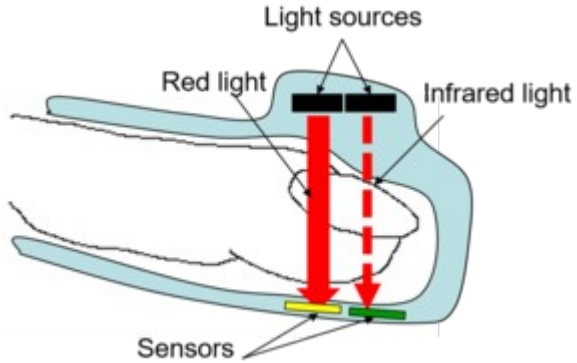
Cerebellar volume in fraction TICV for 60 subjects:

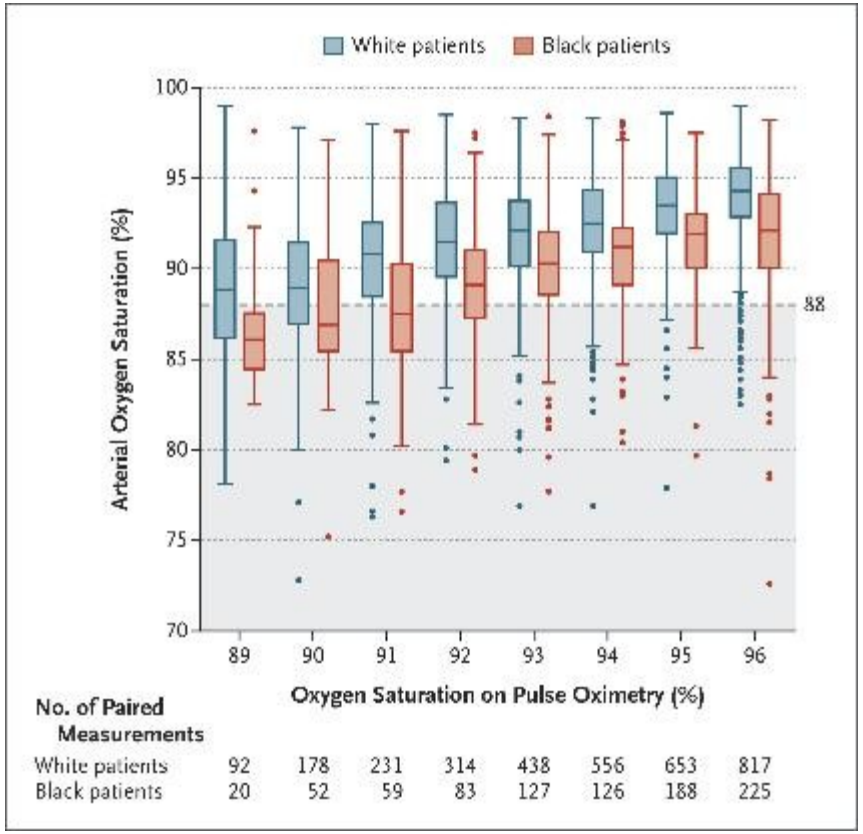
0.13 0.13 0.12 0.13 0.13 0.13 0.14 0.10 0.12 0.13
0.11 0.15 0.14 0.14 0.11 0.14 0.14 0.14 0.13 0.14
0.12 0.10 0.11 0.12 0.15 0.14 0.13 0.14 0.14 0.12
0.13 0.15 0.14 0.13 0.13 0.13 0.14 0.13 0.13 0.15
0.12 0.13 0.11 0.10 0.14 0.12 0.15 0.13 0.13 0.13
0.11 0.14 0.14 0.12 0.14 0.13 0.14 0.11 0.12 0.14

1B What is good data

Imperfect data

- Data is never perfect
  - Sampling bias
    - Pulse oximeters tested on white skin
    - Poorly calibrated for darker skin
    - During Covid-19, black patients had:
      - Lower detection
      - Delays in treatment

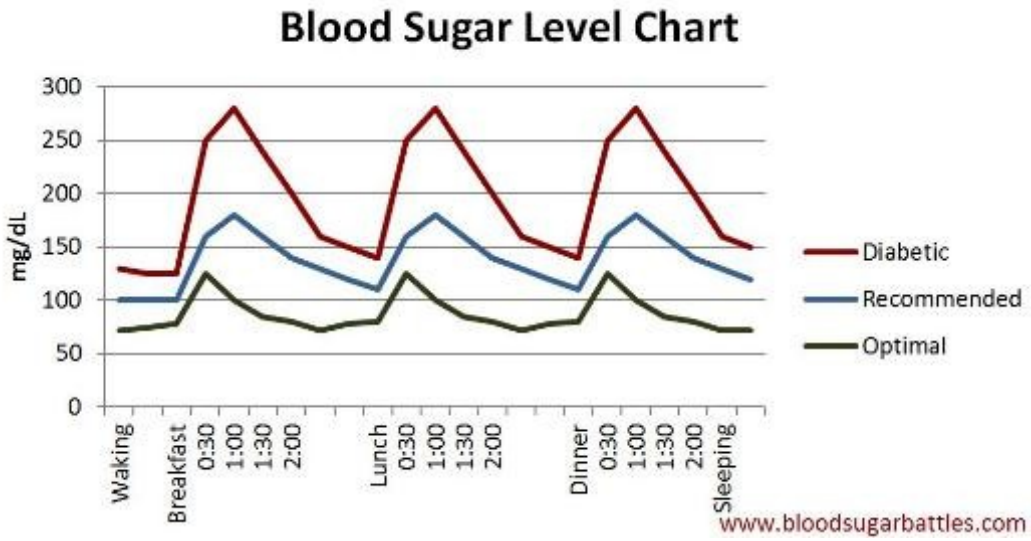




N Engl J Med 2020;383:2477-2478
DOI: 10.1056/NEJMc2029240

Imperfect data

- Data is never perfect
  - Sampling bias
  - Measurement error
    - Inaccurate glucose monitors (2018)
    - Over and underdosing insulin



Diabetes Care 1 August 2018; 41 (8): 1681–1688. https://doi.org/10.2337/dc17-1960

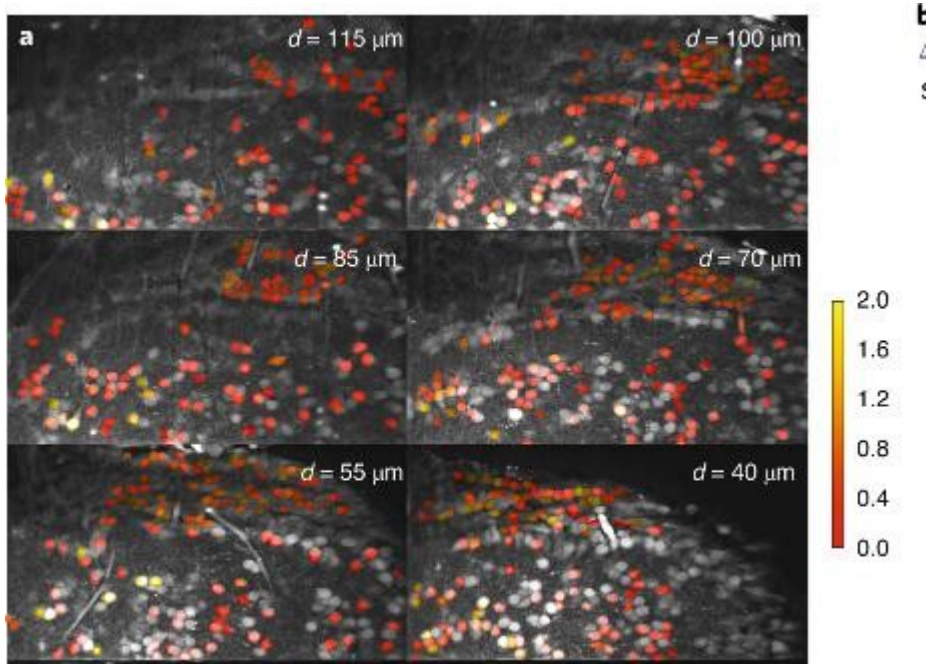
Imperfect data

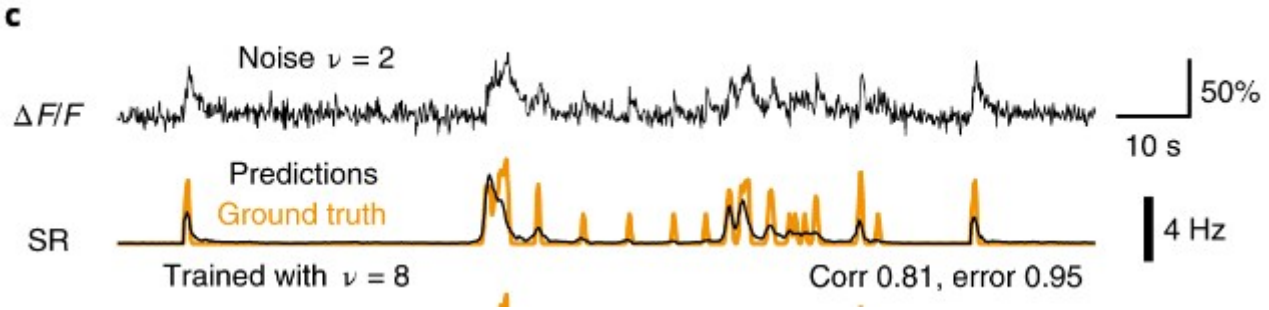
- Data is never perfect
  - Sampling bias
  - Measurement error
  - Missing data
    - Increases variance
    - May introduce bias

N Engl J Med 2020;383:2477-2478
DOI: 10.1056/NEJMc2029240

Imperfect data

- Data is never perfect
  - Sampling bias
  - Measurement error
  - Missing data
  - Proxy measures
    - Intracellular calcium
      - Used for neural activity
    - Proxies are:
      - Nonlinear
      - Incomplete
      - Context-dependent





Nat Neurosci **24**, 1324–1337 (2021).
https://doi.org/10.1038/s41593-021-00895-5

Key issues in measurement

- What are you measuring?
- What is its precision?
- Is it valid?
- Is it reliable?

Precision

- What is a reasonable precision?
- How much precision do you need to draw conclusions?
- Some things are imprecise by nature
  - Quality of life
  - Preference
  - Fitness

Validity

- Does the measure measure what you want?
  - A written test cannot measure musical ability
- Do we have a gold standard?
- Can we poll experts?

Reliability

- A measure that is precise and stable
- The value will stay the same if:
  - We measure again
  - Someone else measures
  - We measure at a different time

Data Ethics

- As data becomes increasingly important
  - We become increasingly responsible

- Privacy
- Security
- Ethical use
  - Wearable devices share health information with companies
  - Can those companies share that information?

The Information Society, *34*(1), 49–57.
https://doi.org/10.1080/01972243.2017.1391912

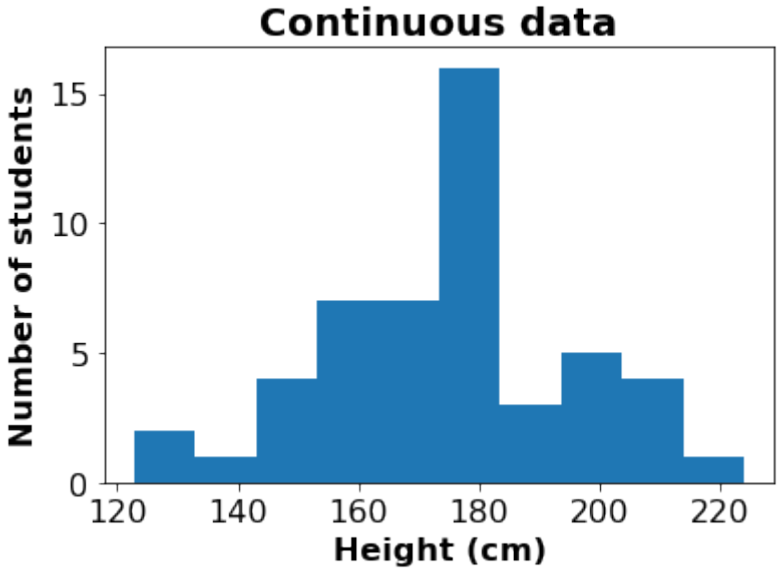
Strategies for ethical data use

- Informed consent
  - Explain how and why it is being collected
  - Allow participants to opt out
- Transparency
  - Who has access and how will it be used
- Protect vulnerable populations
  - Data should not lead to discrimination
  - Data collection should not lead to secondary harm

1C Types of data

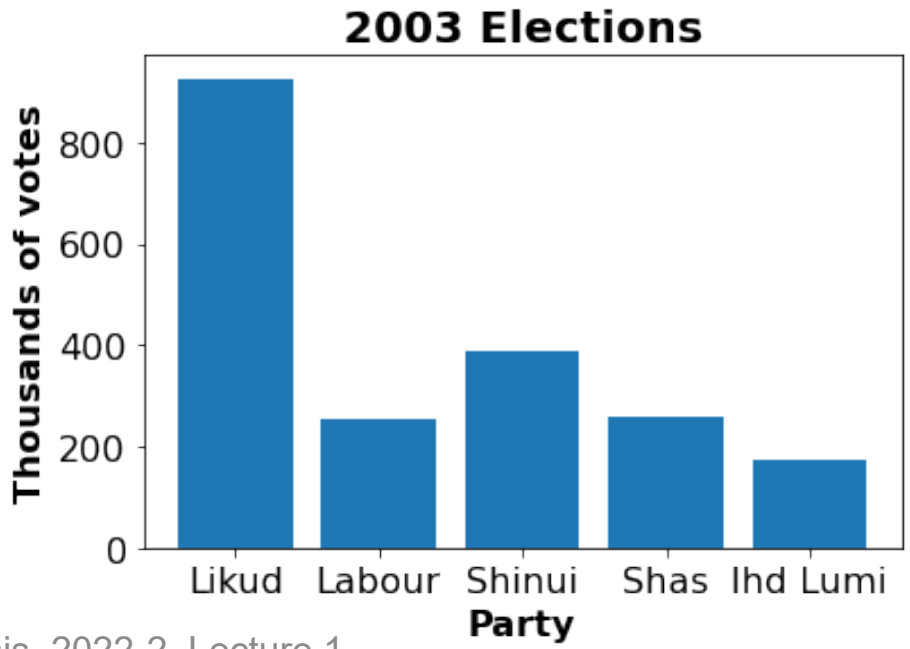
Continuous data

- The data are real numbers
  - Between any two data values there could be another data point
- Examples:
  - Rate, size, amplitude, time



Categorical data

- Values can’t really be compared or put on a scale
- Examples:
  - Gender (male / female)
  - Race
  - Personality type
  - Color



Ordinal data

- Something between continuous and categorical
- Data can be ordered but are not real numbers
- Data maps onto the integers
- Examples:
  - Child, teen, adult
  - Strongly approve, approve, don’t care, disapprove, strongly disapprove
  - Undergraduate student, graduate student, lecturer, professor

Combining different data types

- A datum could have multiple parts:
  - From a database of brain scans:
    - Gender
      - Categorical
    - Handedness
      - Categorical
    - Age
      - Ordinal / continuous
    - Education
      - Ordinal
    - Total intracranial volume
      - Continuous (ordinal)
    - Normalized brain volume
      - Continuous

Female
Right handed
74 years
High school
1344 mm^3
0.743

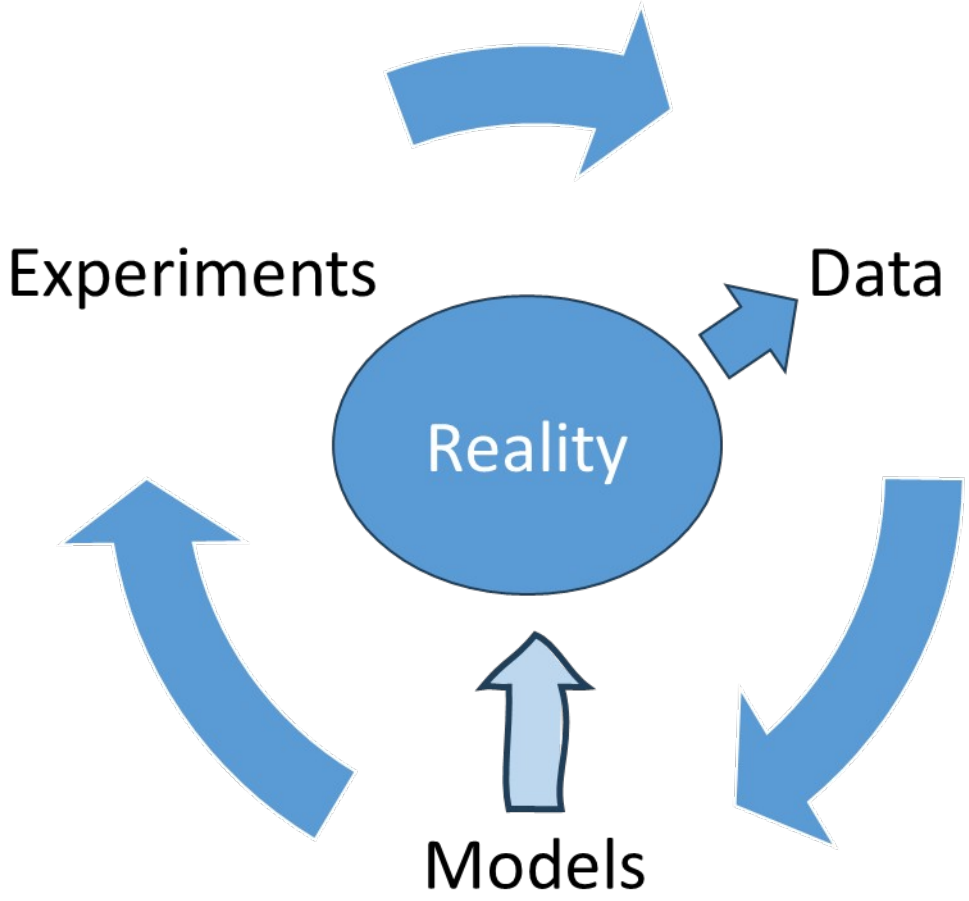
Male
Right handed
39 years
College
1636
0.739

  - Taken together, all of these would be one data point

1D What is probability

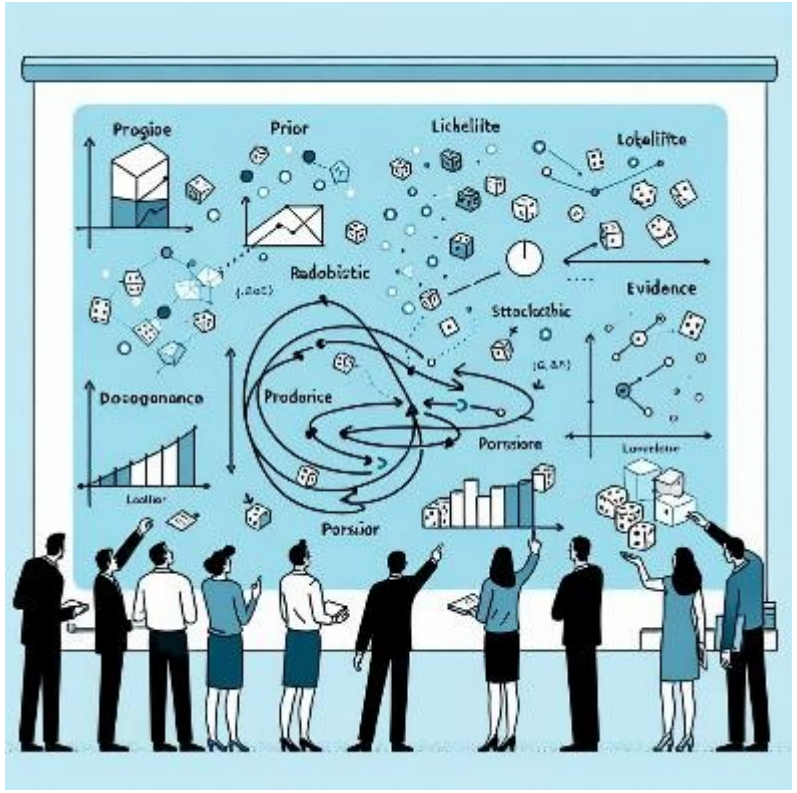
A basic approach

- Start with data
- Make model of data
- Combine data with model
  - To reach conclusions
- Evaluate model
  - Update
- Repeat



To get this basic approach to work

- Data is imperfect
- Our model will treat it as stochastic
- So:
  - We need to review probability



Definition of Probability (from Measure theory)

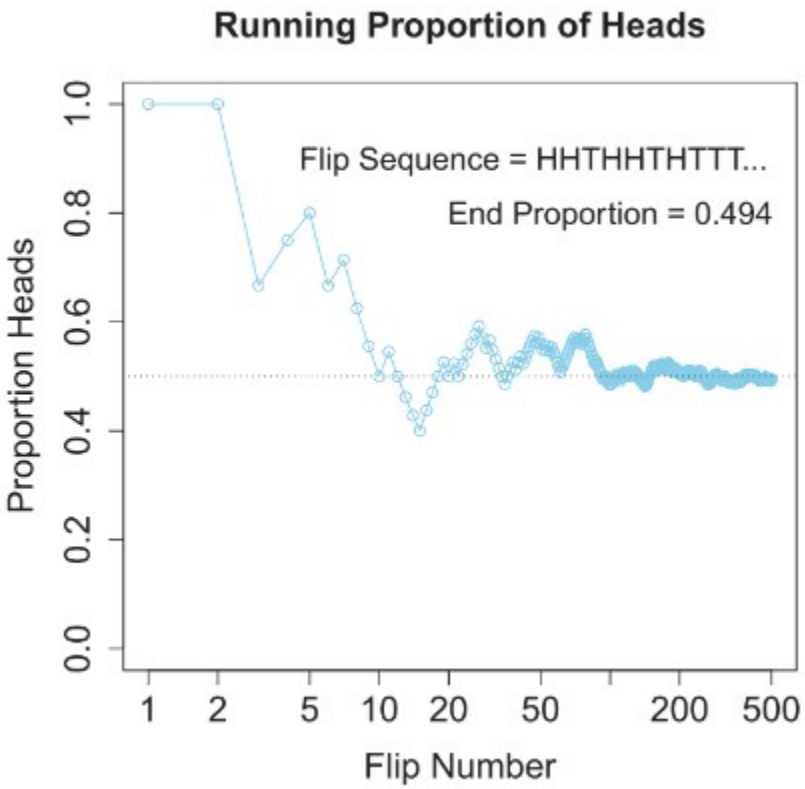
- ***Experiment***: toss a coin twice
- ***Sample space***: possible outcomes of an experiment
  - S = {HH, HT, TH, TT}
- ***Event***: a subset of possible outcomes
  - A={HH}, B={HT, TH}
- ***Probability of an event*** : an number assigned to an event Pr(A)
  - Axiom 1: $\Pr(A) \geq 0$
  - Axiom 2: $\Pr(S) = 1$
  - Axiom 3: For every sequence of disjoint events

$$\Pr(\bigcup_i A_i) = \sum_i \Pr(A_i)$$

What does P(A) mean?

- Frequentist interpretation (Out of the head)
  - If we repeat the experiment N times, the number of times A occurs (#A) converges to:

$$P(A) = \frac{\#A}{N}$$



Example: political futures

https://polymarket.com/event/trump-wins-ends-ukraine-war-in-90-days?tid=1742336531501

What does P(A) mean?

- Frequentist interpretation (Out of the head)
- Plausibility (In the head)
  - How likely is it that the coin is fair?
  - Calibrate by choosing a bet
    - Political futures markets

What does P(A) mean?

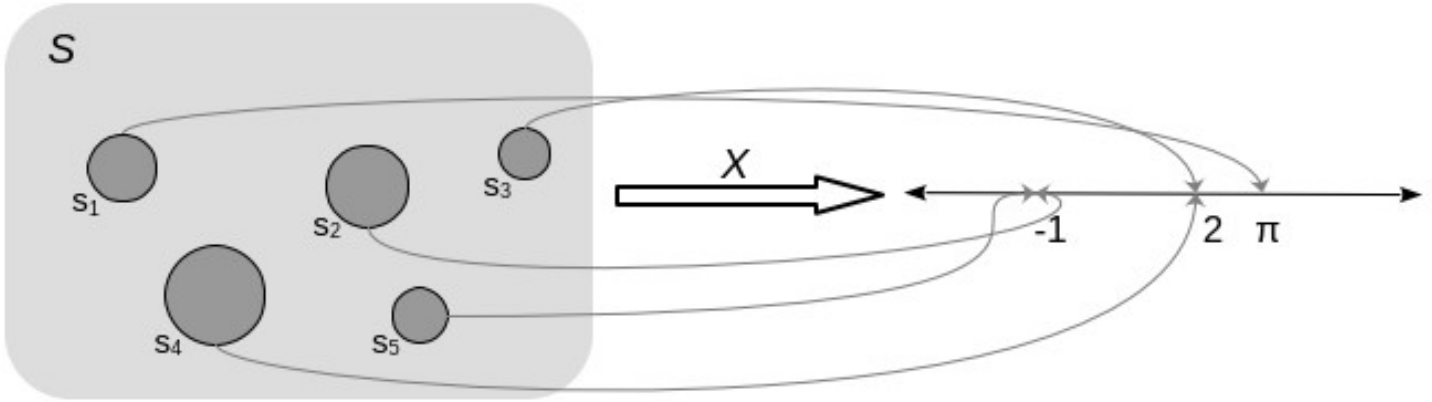
- Frequentist interpretation (Out of the head)
- Plausibility (In the head)
  - How likely is it that the coin is fair?
  - Calibrate by choosing a bet
    - Political futures markets
  - Or by assigning a distribution

de Finetti's Dutch Book (1975)

- A Dutch book is a book of bets that lets you beat the house at gambling
- de Finetti derives axioms of probability from assumption that we assign probability such that a Dutch book is impossible
- This is the idea followed by Kruschke

1E Random variables

Random variables

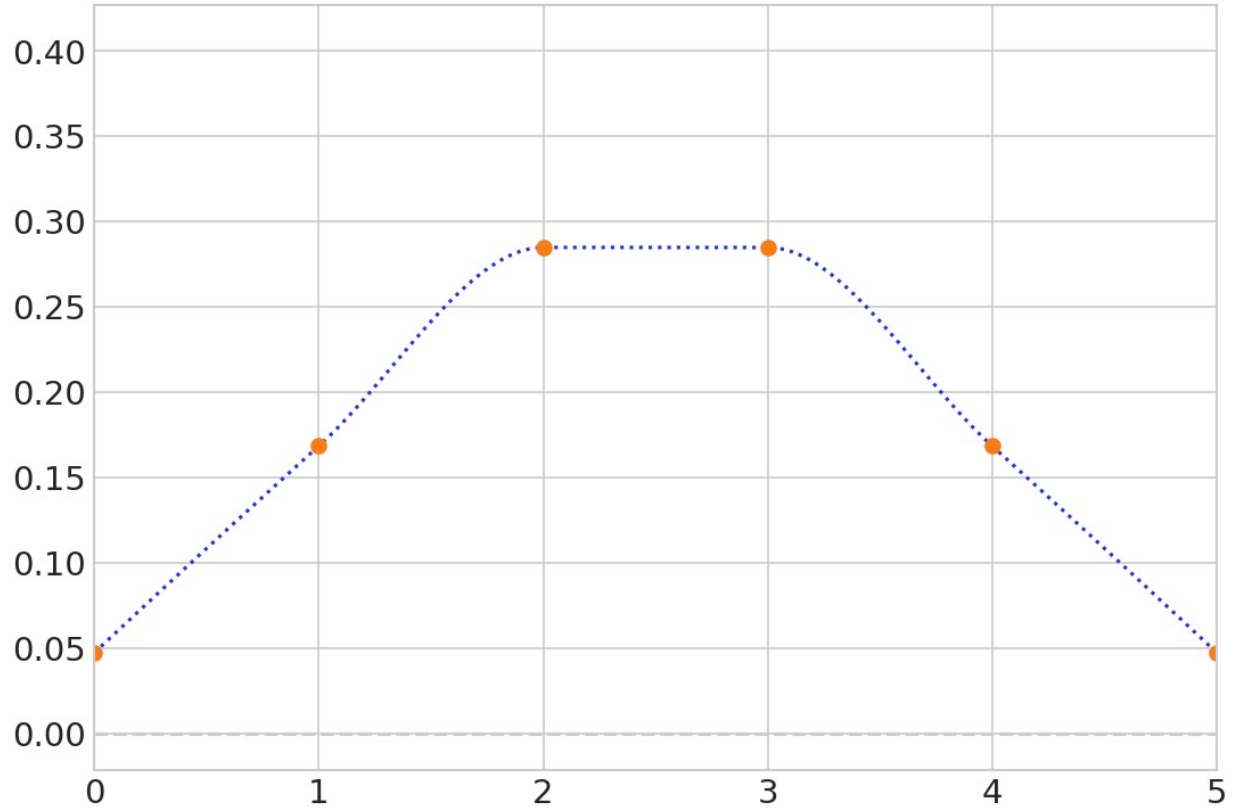


- A mapping from subsets of all possibilities to values

Exploring a distribution

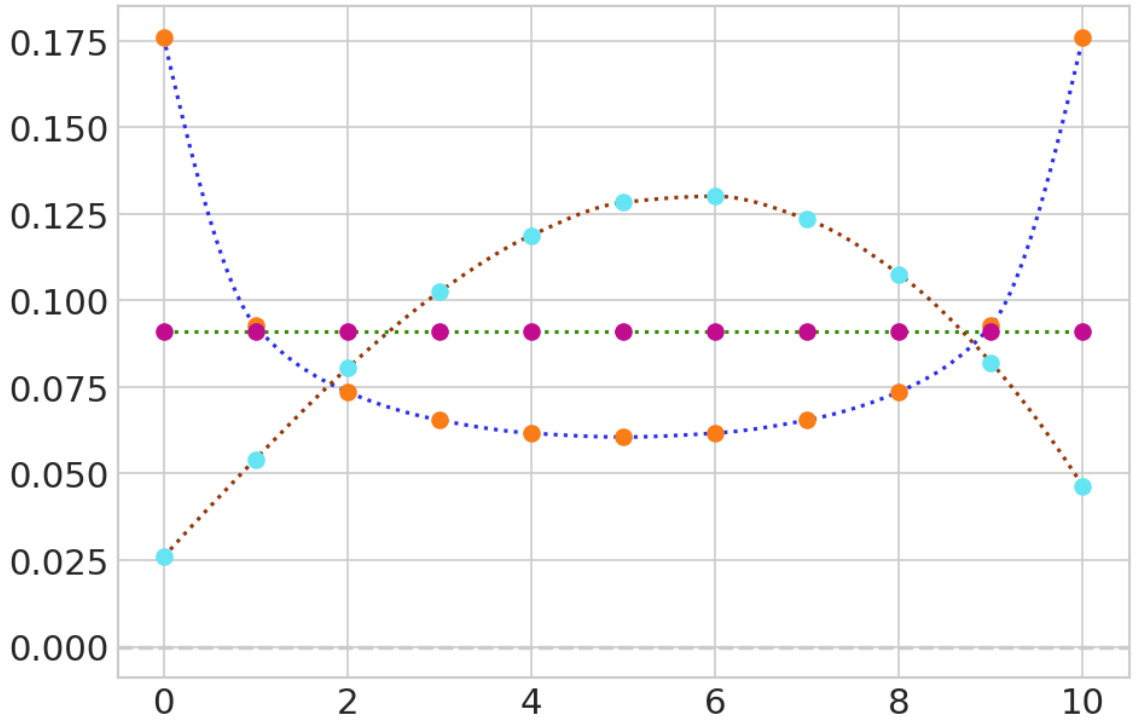
- Preliz software package has interactive plots so you can see what parameters do

```
pz.BetaBinomial(alpha=10, beta=10, n=5).plot_interactive(pointinterval=False)
```



Parameterized probability distributions

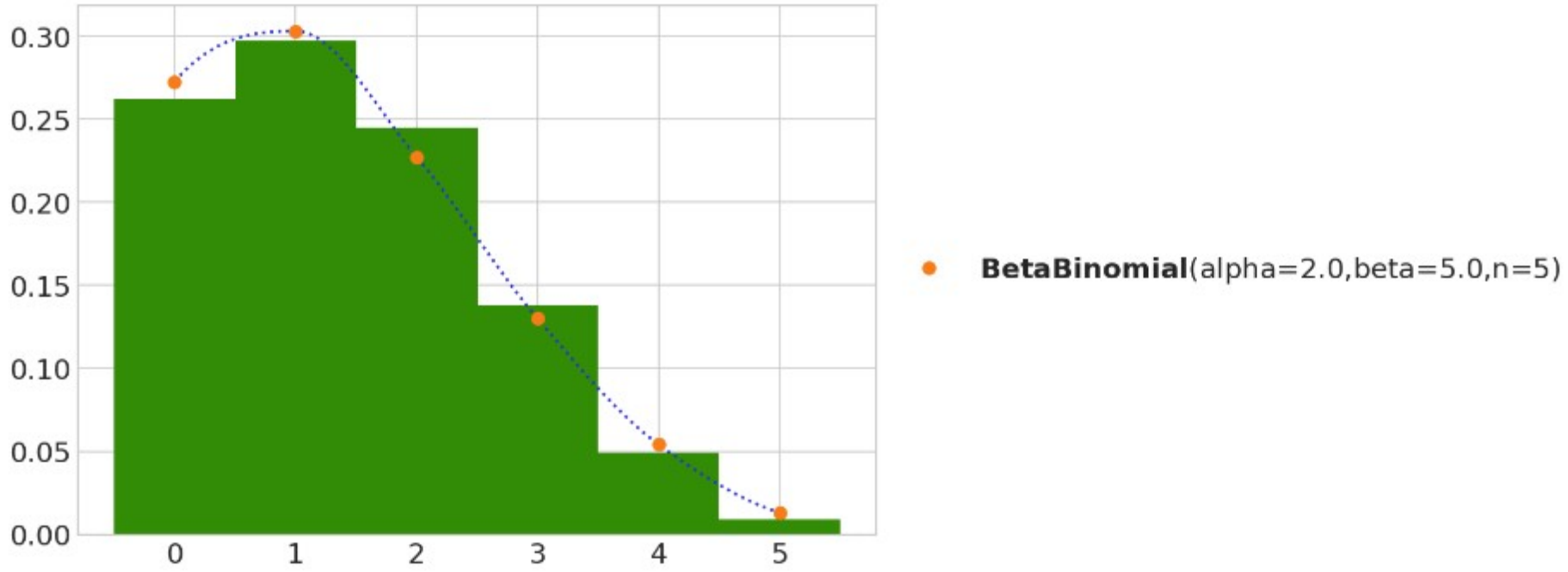
- The Beta Binomial distribution has 3 parameters



$$p(x \mid \alpha, \beta, n) = \frac{B(x + \alpha, n - x + \beta)}{B(\alpha, \beta)}$$

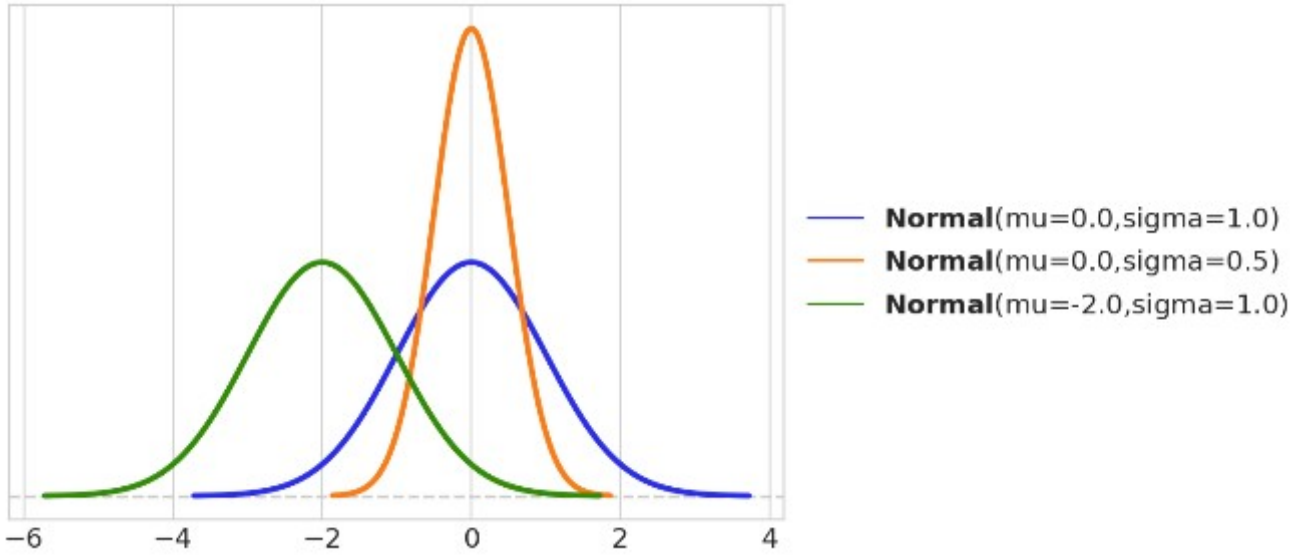
Generating samples from a distribution

```
plt.hist(pz.BetaBinomial(alpha=2, beta=5, n=5).rvs(1000),
         bins=[0, 1, 2, 3, 4, 5, 6], density=True, align="left", color="C2")
pz.BetaBinomial(alpha=2, beta=5, n=5).plot_pdf();
```

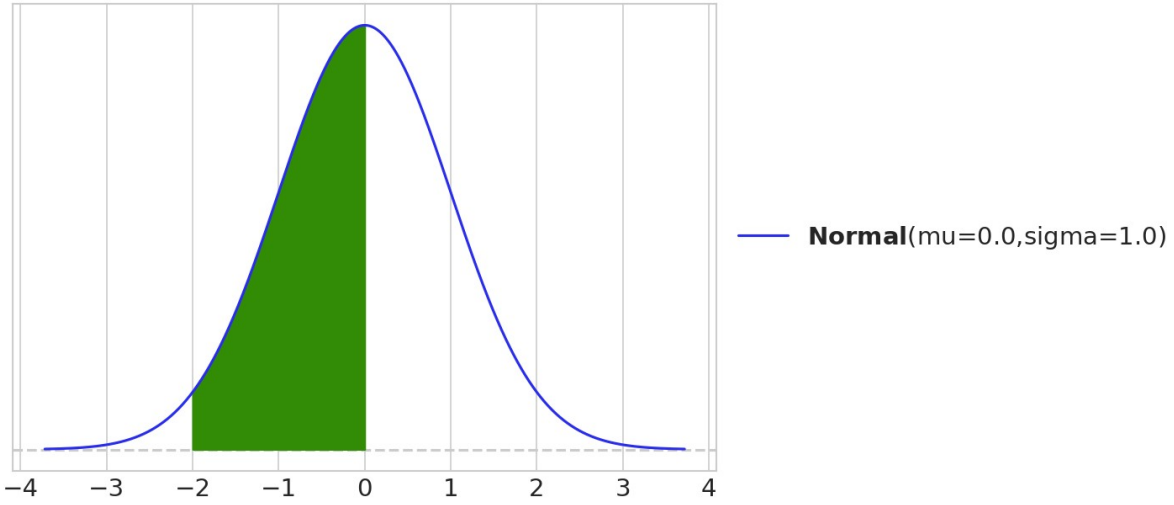


Continuous random variables

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$$



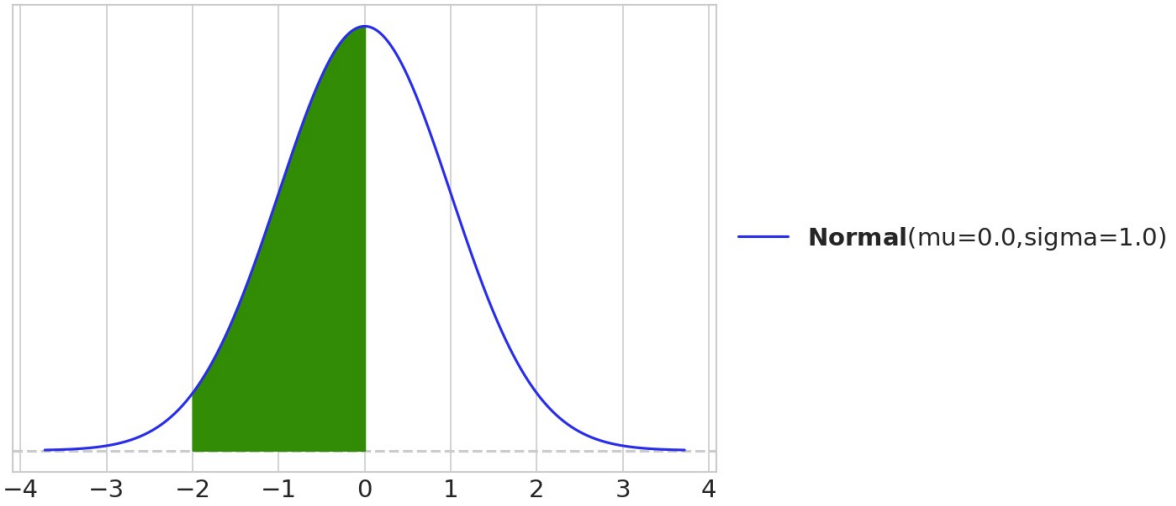
$$P\left(x < X < b\right) = \int_a^b p\left(x \mid \mu, \sigma\right) dx$$



How to integrate a continuous variable

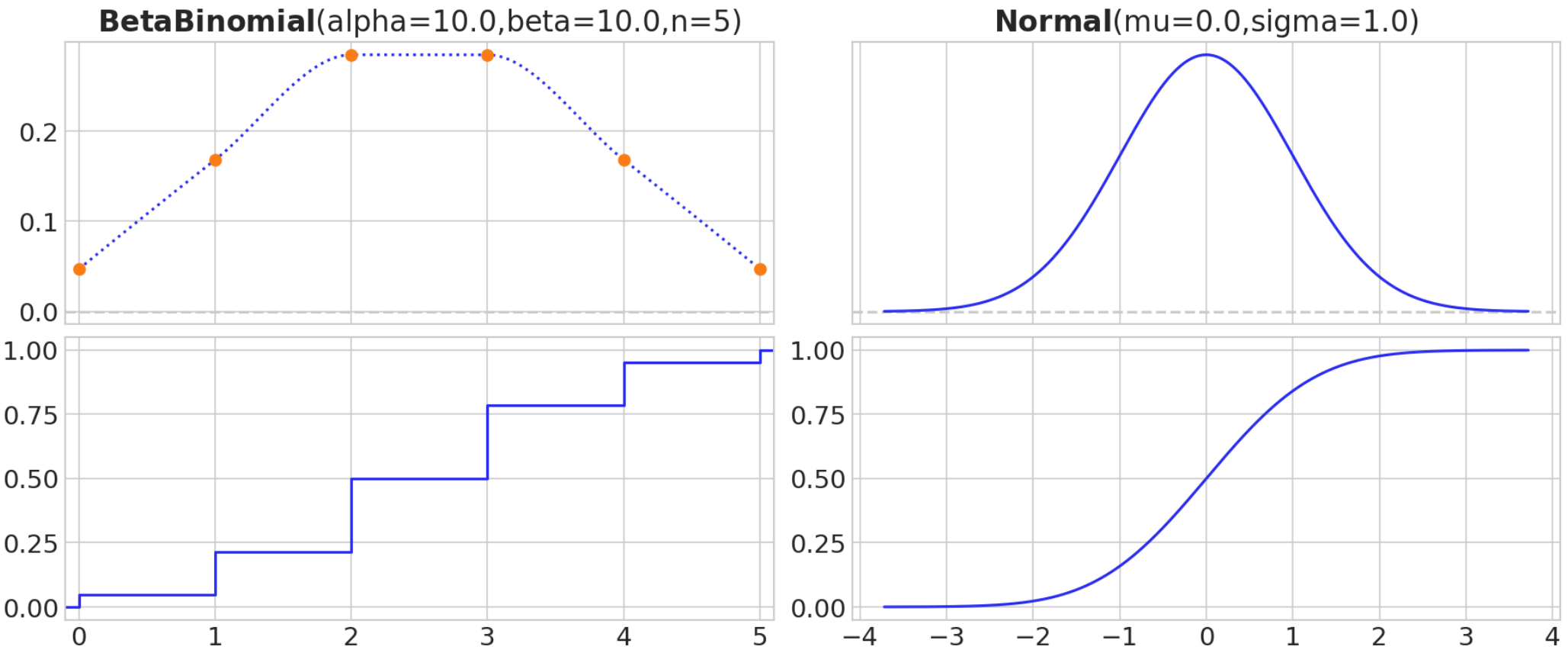
$$P\left(x < X < b\right) = \int_{a}^{b} p\left(x \mid \mu, \sigma\right) dx$$

```
dist = pz.Normal(0, 1)
ax = dist.plot_pdf()
x_s = np.linspace(-2, 0)
ax.fill_between(x_s, dist.pdf(x_s), color="C2")
dist.cdf(0) - dist.cdf(-2)
```



Cumulative distribution functions

- Always go from 0 on the left to 1 on the right
- Y axis is probability



1F Characterizing distributions

Describing the distribution with one number

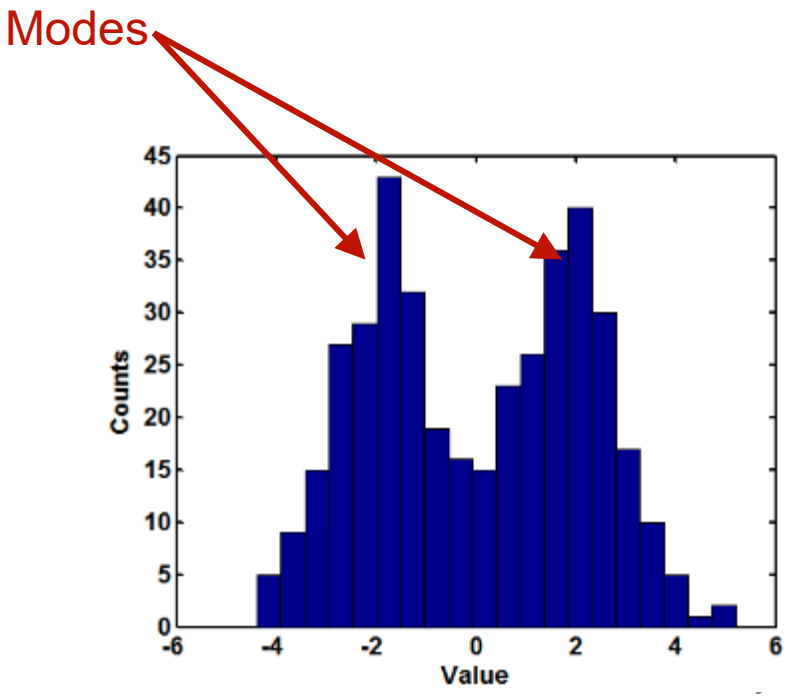
- We want to use a single number to describe the distribution
- We can choose it
- Where is the ‘middle’ of the data?

Describing the distribution with one number

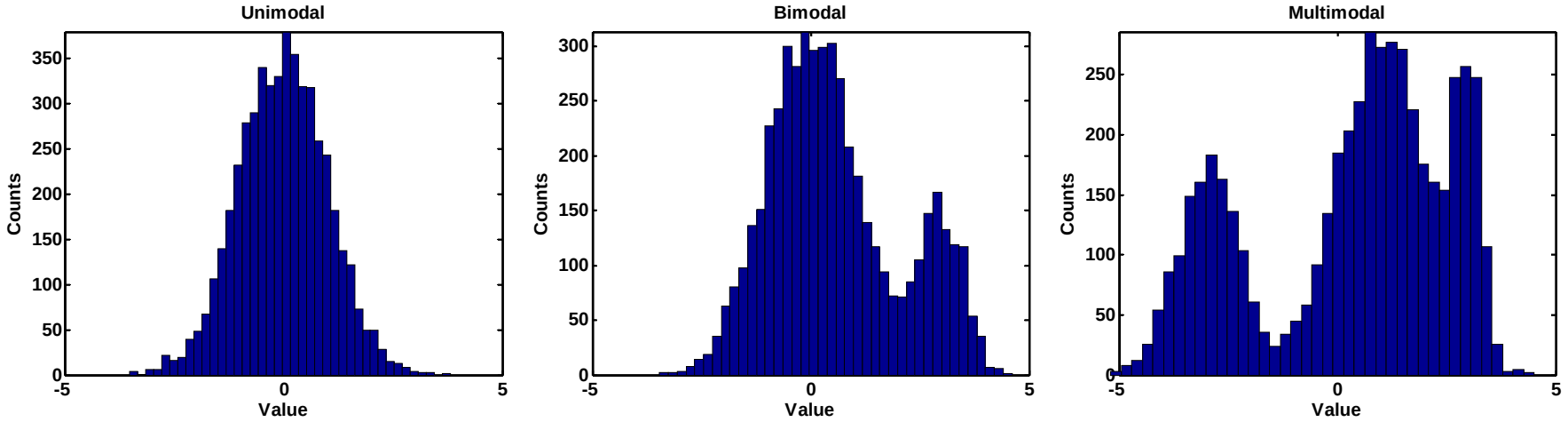
- Where is the ‘middle’ of the distribution?
- Three common measures
  - Median
  - Mode
  - Mean
- There are others
  - We use a different one to calculate your homework score

More than one mode

- It is possible to have more than one mode
- This often means that the distribution combines two different sources



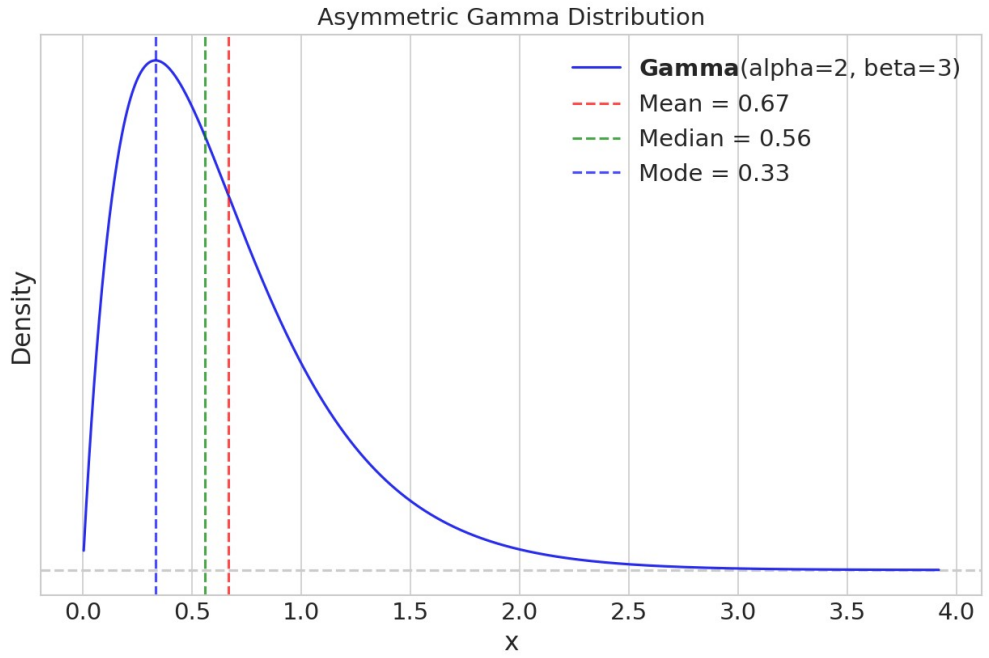
Unimodal and multimodal distributions



- The mode is the point with the largest probability
- We usually take multi-modal data to be selected from different unimodal distributions
- So **we** will usually only deal with unimodal data

Measures of central tendency

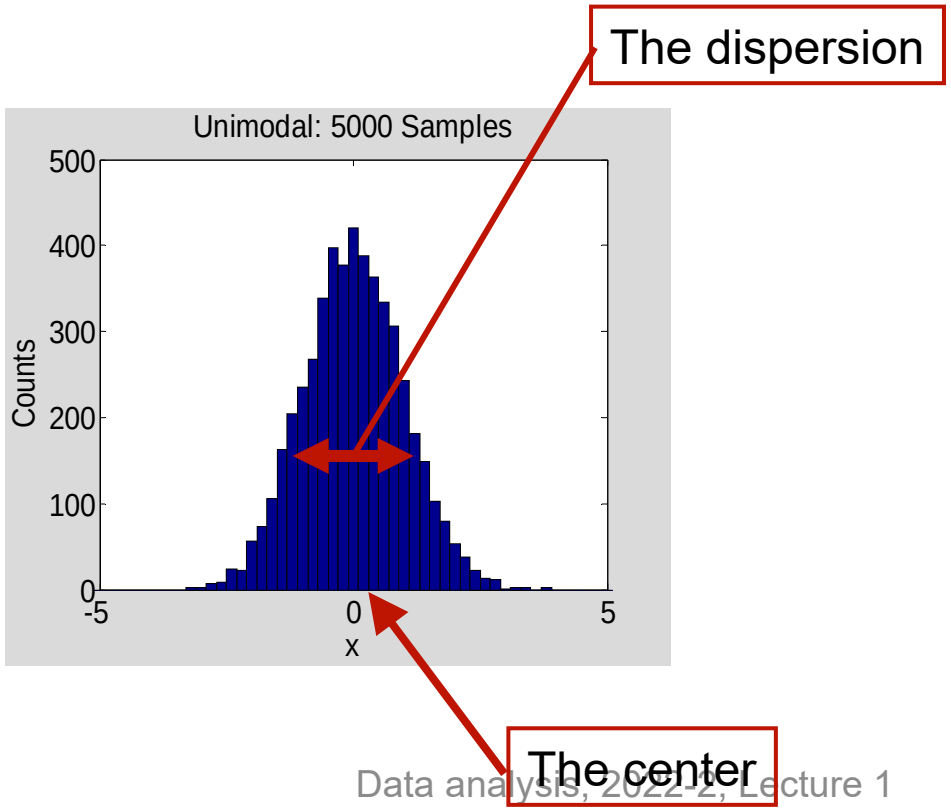
- Notice the difference between the mean, the median, and the mode



Which is better?

Representing a distribution with two numbers

- The first number: where is the center?
- The second number: how close to the center are we?



Representing the distribution with two numbers

- Center and spread
- Again, we have different measures
  - The standard deviation is the most commonly used

The variance: $\sigma^2 = \int_{-\infty}^{\infty} p(x)(x-\mu)^2\,dx = E\left((x-\mu)^2\right)$

The standard deviation: $\sigma = \sqrt{\sigma^2}$

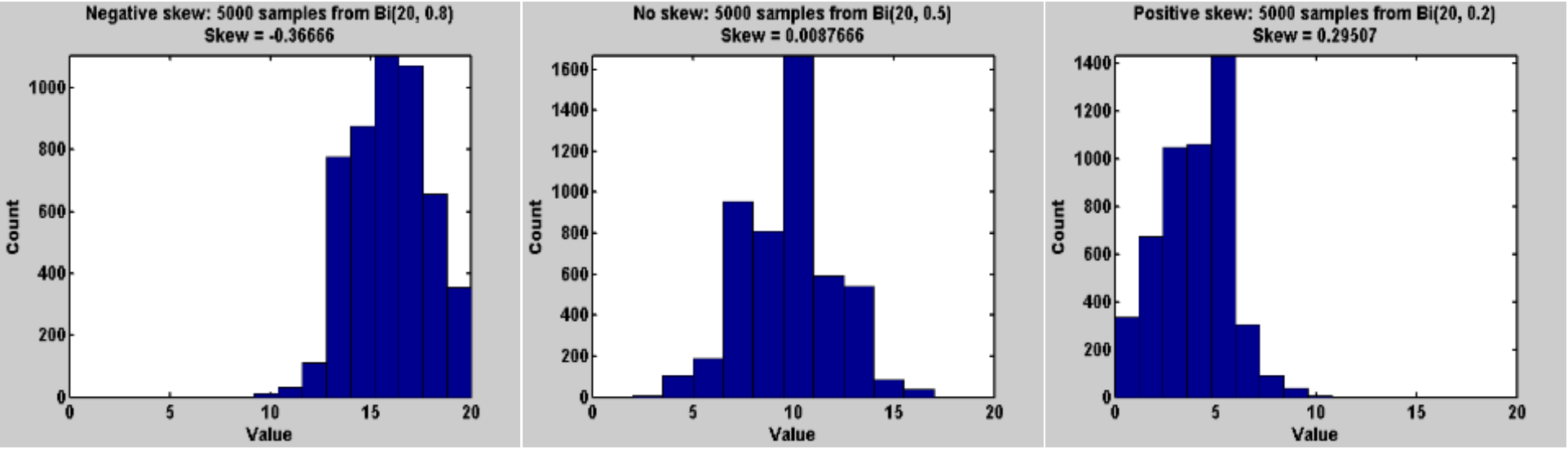
The absolute deviation: $\mathrm{MAD} = \int_{-\infty}^{\infty} p(x)|x-\mu|dx$

The inter-quartile interval: The difference between 0.75 and 0.25 of the distribution

Representing more details about the distribution

- What should we use?
- Skewness
  - How symmetric is the data?
- Kurtosis
  - How sharp is the peak?

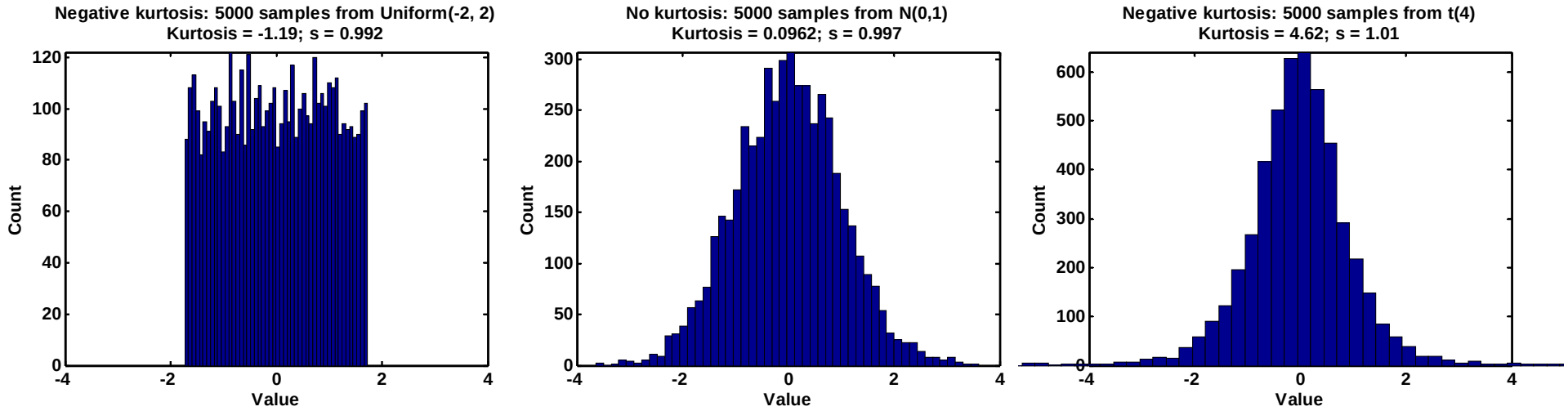
Measure of assymetry: Skewness



- For the normal distribution, we only need mean and variance
- For any other distribution, it will vary in either skewness or kurtosis
- Skewness measures the symmetry of the distribution

$$\text{Skewness} = \frac{1}{\sigma^3} E\left(\left(x-\mu\right)^3\right)$$

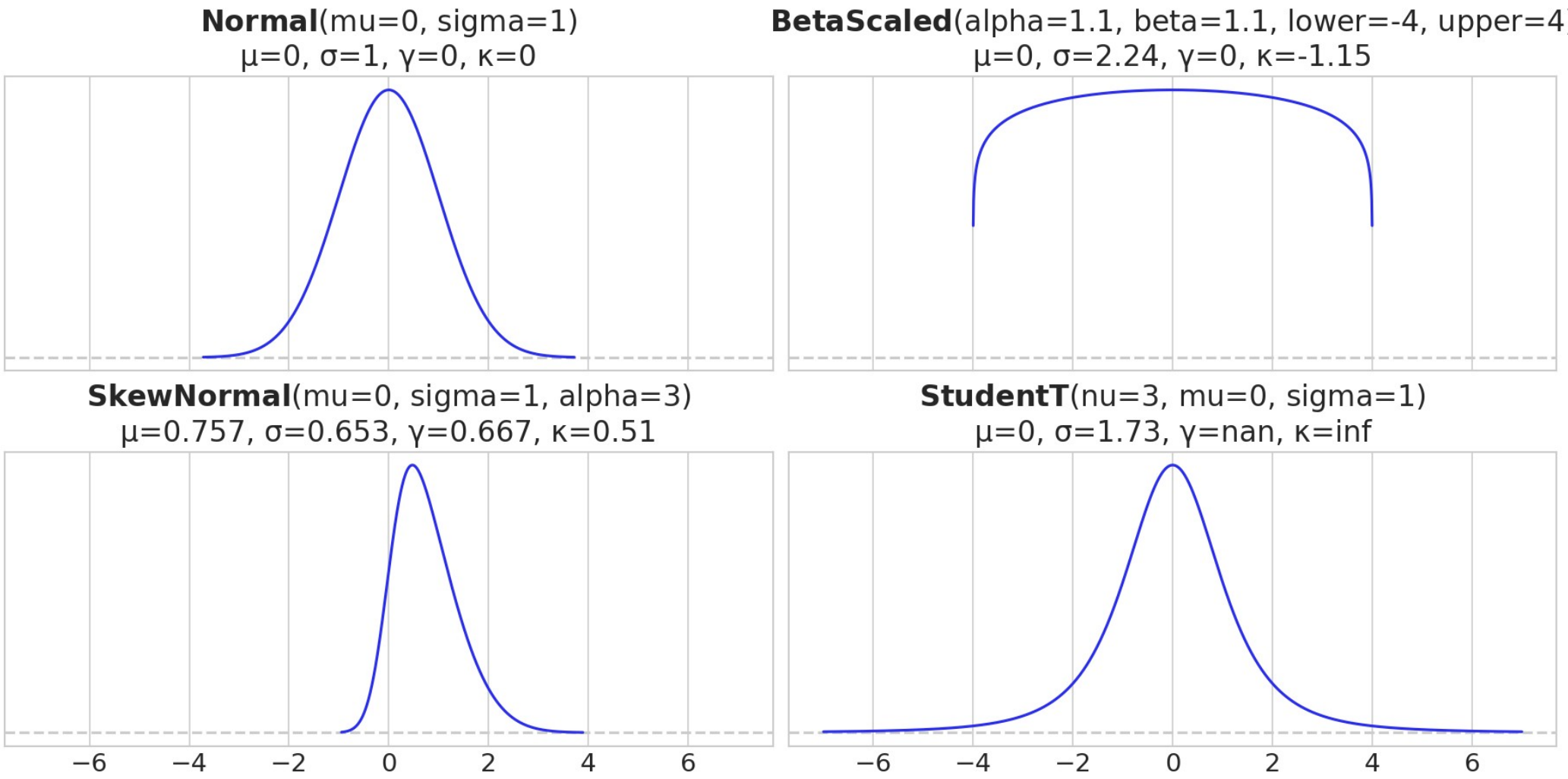
Sharpness of peak: Kurtosis



- Kurtosis measures the size of the tails
- This also captures the pointiness of the peak

$$\text{Kurtosis} = \frac{1}{\sigma^4} E\left(\left(x-\mu\right)^4\right) - 3$$

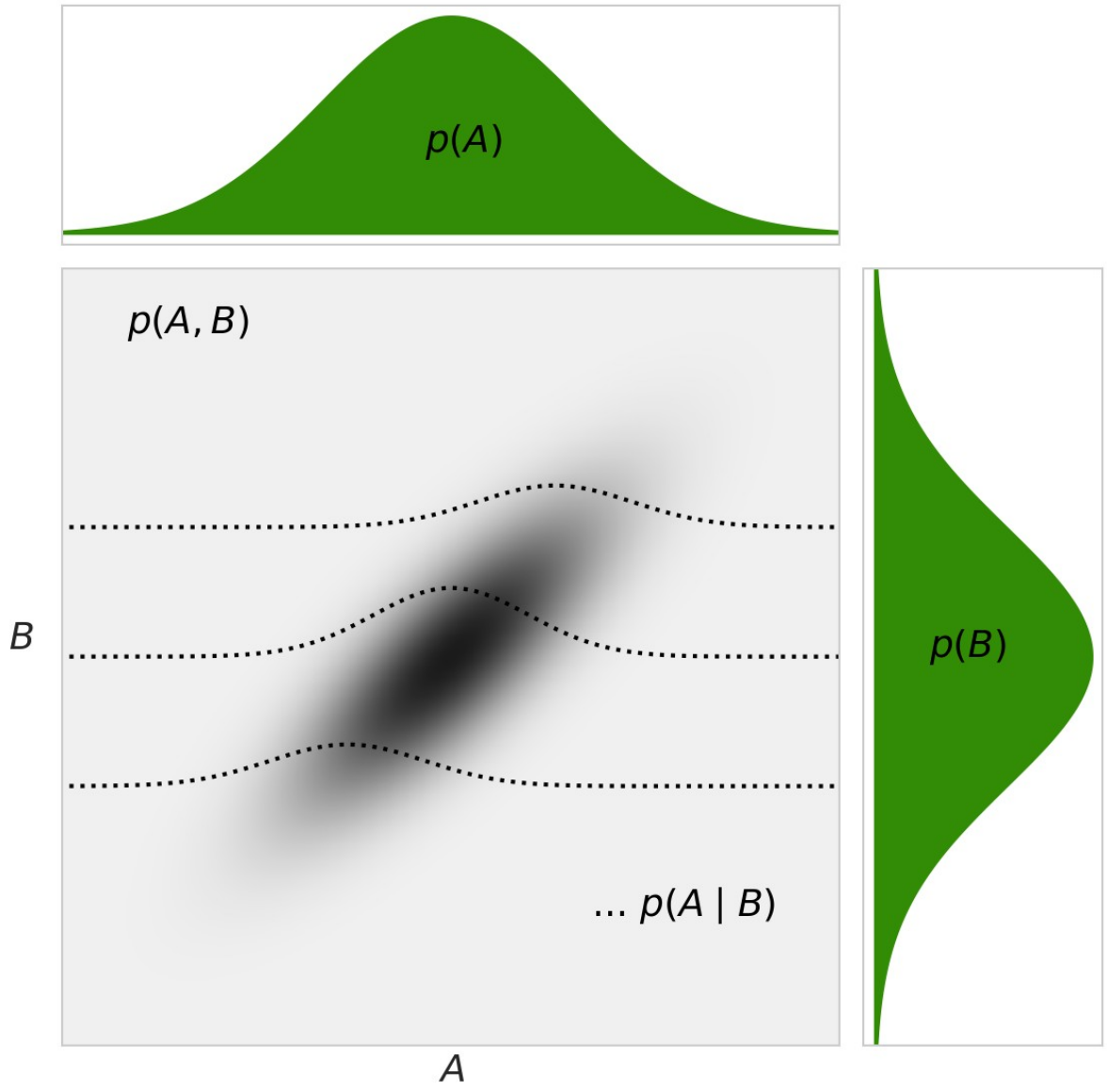
Mean, standard deviation, skewness and kurtosis of 4 distributions



1F Conditional probability

The joint distribution

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$



Baye's rule

$$p(c|r) = \frac{p(c,r)}{p(r)} \qquad p(r|c) = \frac{p(c,r)}{p(c)}$$

$$p(c|r)\,p(r) = p(c,r) \qquad p(r|c)\,p(c) = p(c,r)$$

$$p(c|r)\,p(r) = p(r|c)\,p(c)$$

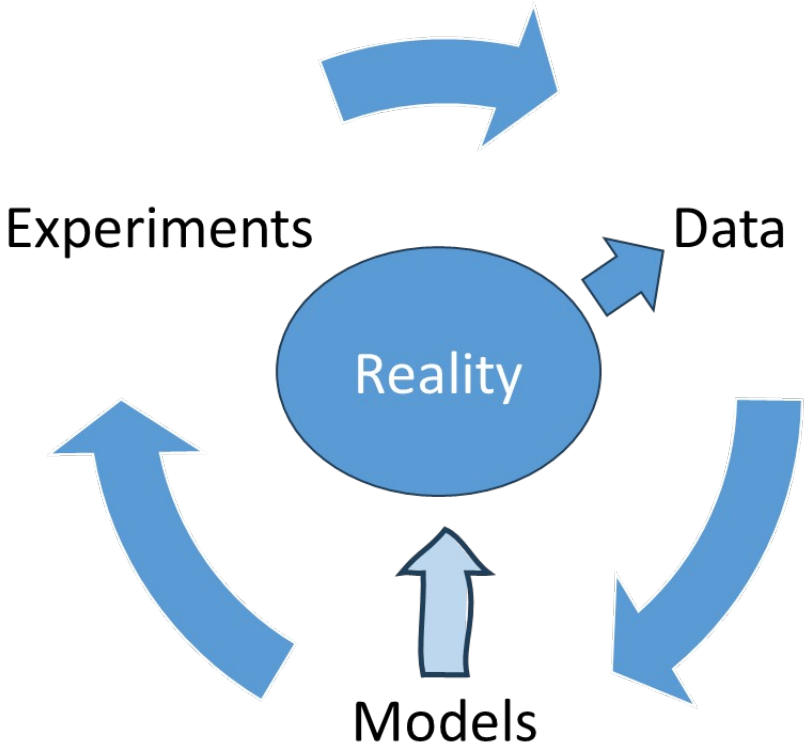
$$p(c|r) = \frac{p(r|c)\,p(c)}{p(r)} = \frac{p(r|c)\,p(c)}{\sum_{c^*} p(r|c^*)\,p(c^*)}$$

Applied to data and models

$$p(c|r) = \frac{p(r|c)\,p(c)}{\sum_{c^*} p(r|c^*)\,p(c^*)}$$

$$p(\text{model}|\text{data}) = \frac{p(\text{data}|\text{model})\,p(\text{model})}{\sum_{\text{models}} p(\text{data}|\text{model})\,p(\text{model})}$$

We will use Baye's rule to update our models



$$p\left(\text{model}\middle|\text{data}\right)=\frac{p\left(\text{data}\middle|\text{model}\right)p\left(\text{model}\right)}{\sum_{\text{models}}p\left(\text{data}\middle|\text{model}\right)p\left(\text{model}\right)}$$