Tutorial 2

Statistical Computation and Analysis

Spring 2025

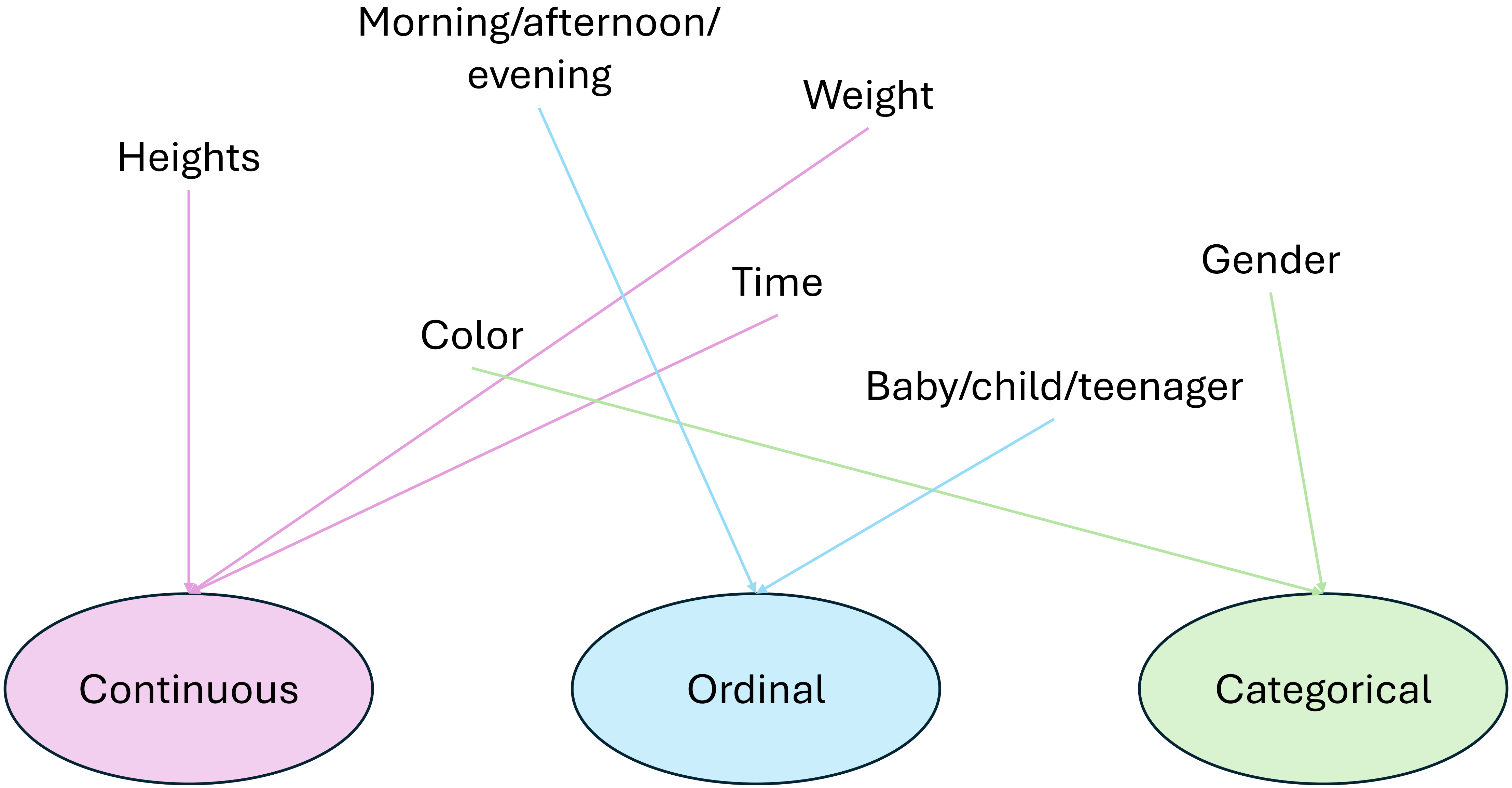
Tutorial Outline

- Data
- Experiments
- Summary statistics
  - Central tendency
  - Spread
  - Skewness
  - Kurtosis
- Probabilities and random variables
- Distributions
- Conditional probabilities
- Bayes theorem

Data

- A set of observations or measurements
- Finite
  - We can get more data, but it will still be data
- Types of data:
  - Continuous – there can be another data point between any two data values.
  - Categorical – values can’t be put on a scale.
  - Ordinal – data can be ordered but are not real numbers.

Data

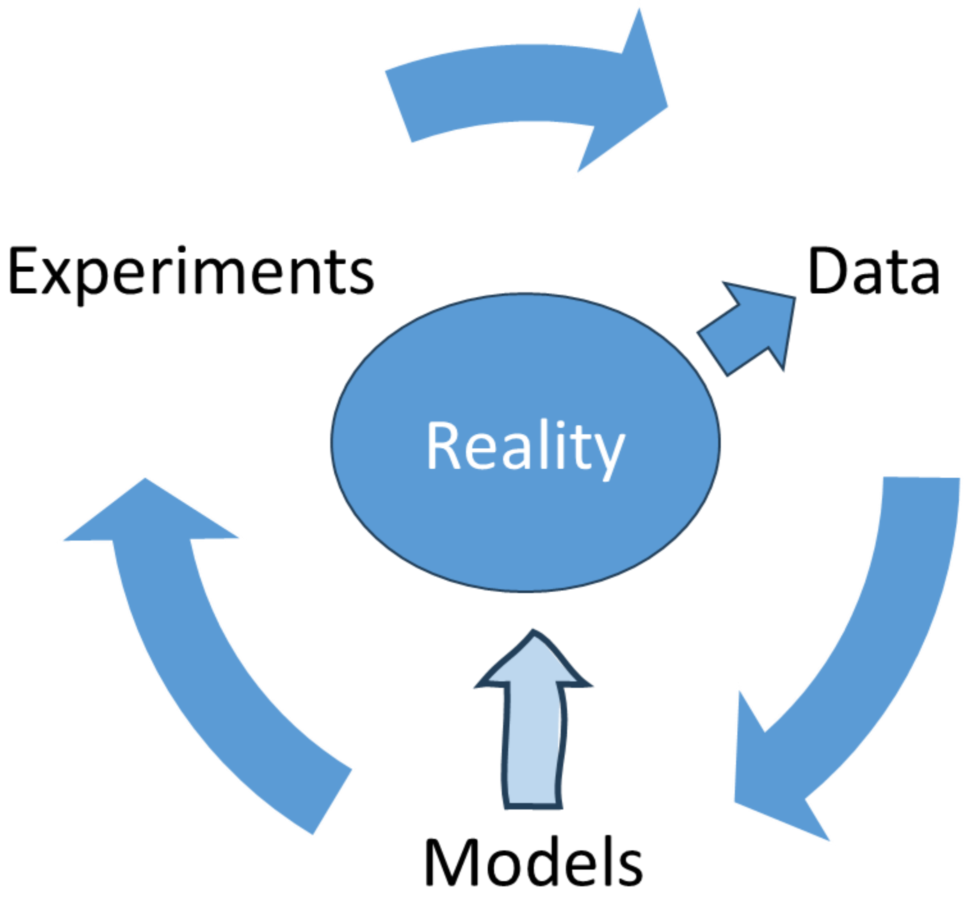

Morning/afternoon/
evening
Weight
Heights
Time
Gender
Color
Baby/child/teenager
Continuous
Ordinal
Categorical

Data

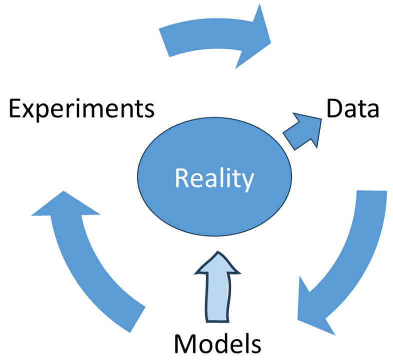
- We need to understand what we are measuring
- What is the precision?
  - How much precision do we need to draw conclusions?
    - Elephant age by weight in kg
    - Can we use this precision for measuring amounts of medicine?
- Is the data valid?
  - Does the data measure what you want?
    - Is musical ability indicative of who will be a good doctor?
    - Is a psychometric grade indicative of who would be a good doctor?
- Is the data reliable?
  - A measure that is stable – the value will stay the same:
    - If we measure again
    - If someone else measures
    - If we measure at different times (if the measurement is not meant to change over time)

Data

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



Experiment



Our experiment will be asking questions:

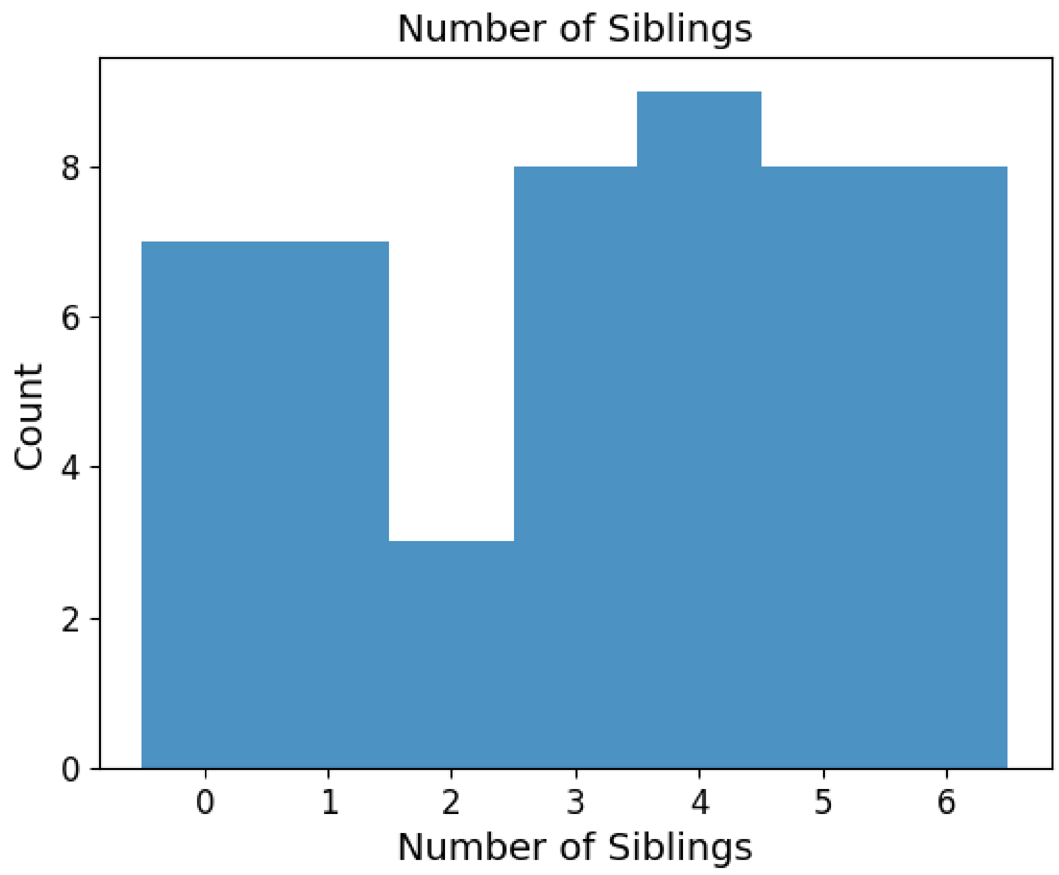
1. How many siblings do you have? (discrete)

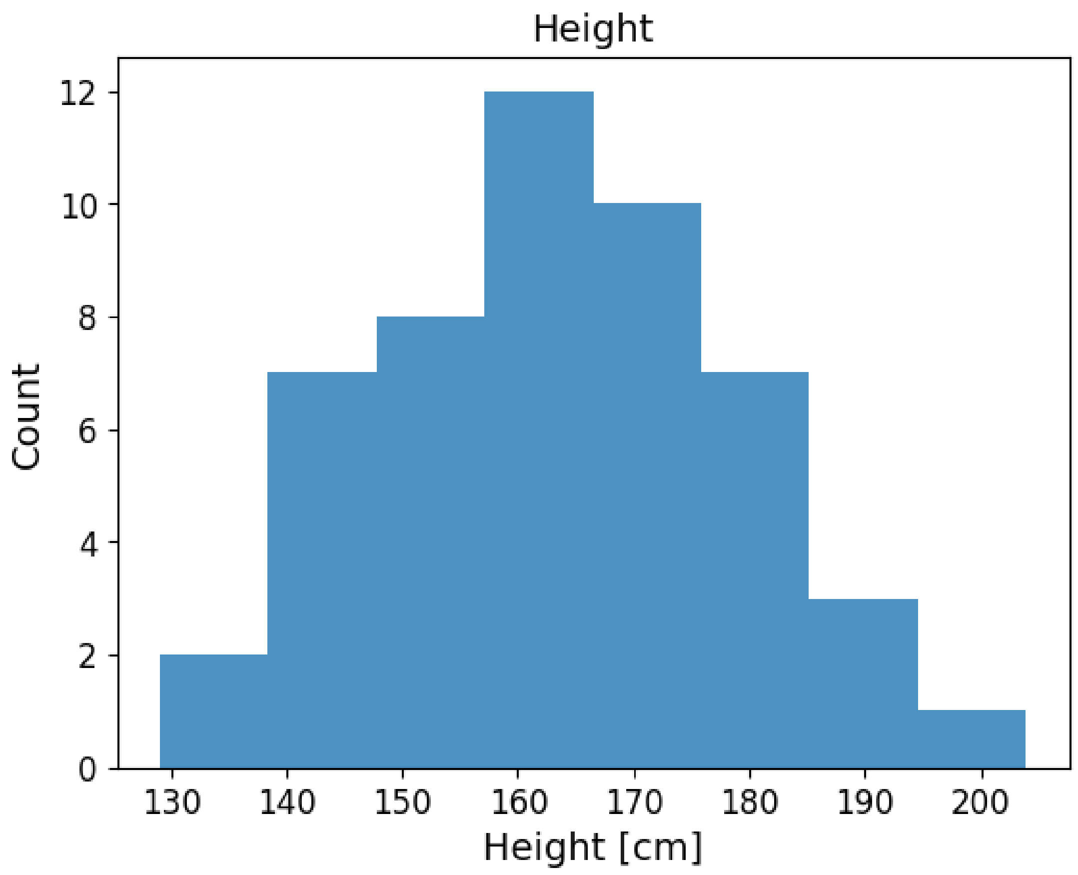
2. What is your height? (continuous)

https://docs.google.com/forms/d/e/1FAIpQLSfqCmuVLoOisvyWWSiETPb5XT-fdRqU36iP0RMp3_s4wQOI8Q/viewform?usp=sharing

Plotting Results

With random data:





Experiment Conclusions

- We can use (simple or complex) experiments to collect data.
- If we run the experiment again on different participants, we will get different measurements.
- Each measurement is a poor measurement of reality.
- At the end of the experiment, we have some measurements of heights or number of siblings.
- What do we do with these measurements?

Summary Statistics

- We can use metrics to summarize different aspects of our collected data.
  - Central tendency
  - Spread
  - Skewness
  - Kurtosis

Central Tendency

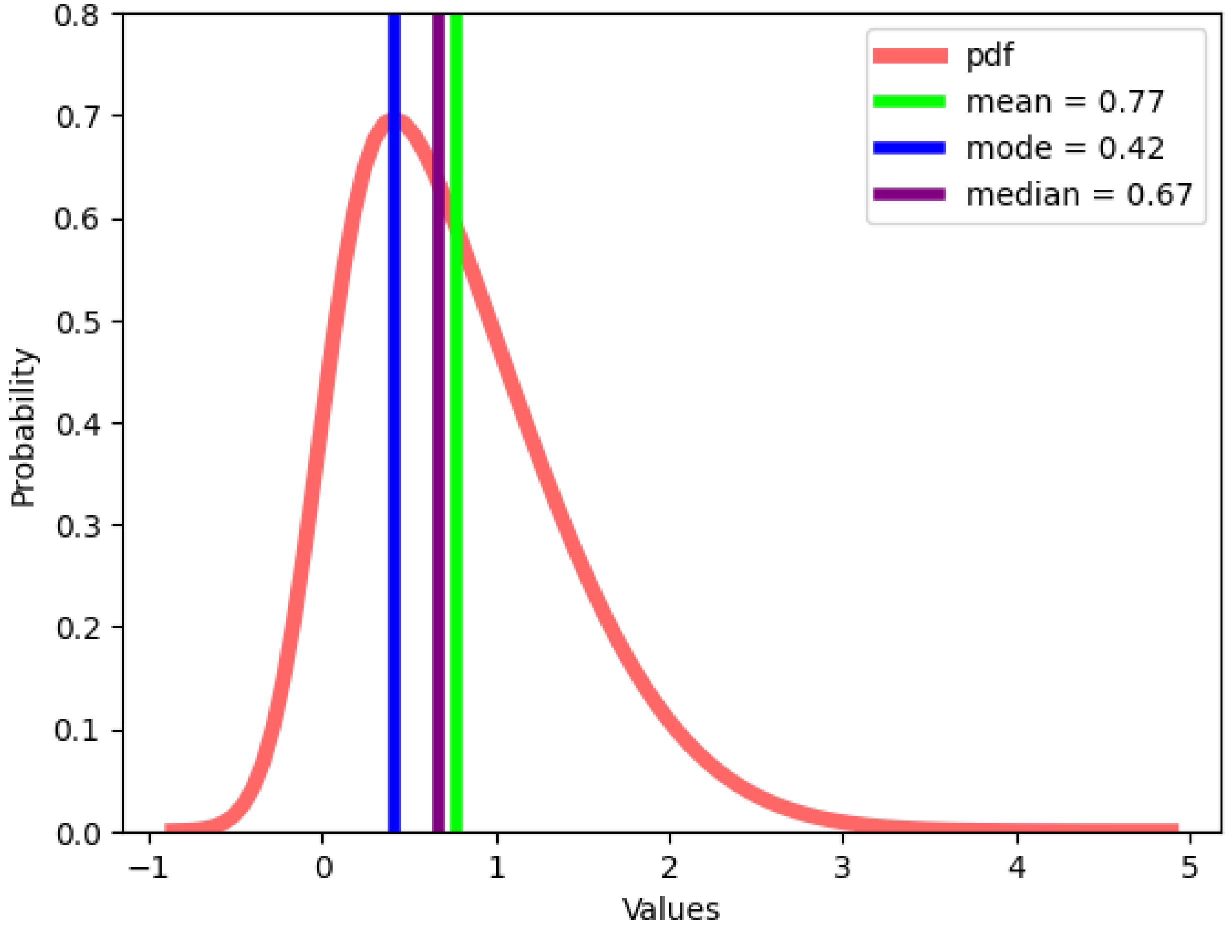
- Measure for middle of the data
  - Mean: $$\mu = \int_{-\infty}^{-\infty} xp(x)dx = E(x)$$
  - Median : $$\int_{-\infty}^{x_{med}} p(x)dx = 0.5$$
  - Mode – the most frequent value.

Central Tendency

- Measure for middle of the data
  - Mean – the sum of all values divided by the total number of values.
  - Median – the middle number in an ordered dataset.
  - Mode – the most frequent value.

Central Tendency

- Where is the middle?



Central Tendency

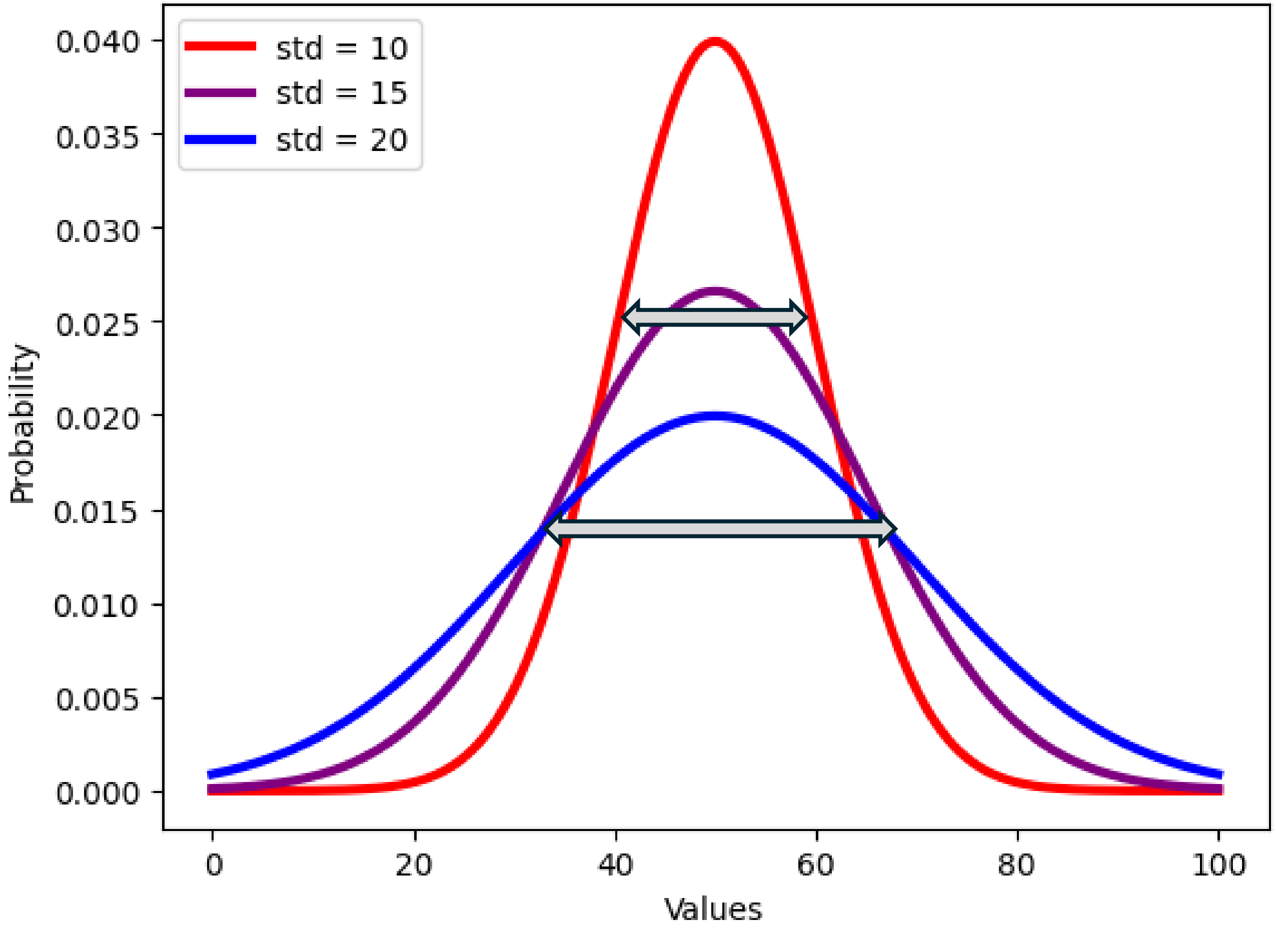
- Measure for middle of the data
  - Mean
    - Convenient to work with mathematically
    - Very effected by outliers
    - Can be used for continuous or ordinal data (for ordinal, we need to assign numbers to the categories)
  - Median
    - Less effected by outliers
    - Can be used for continuous or ordinal data
  - Mode
    - Less effected by outliers
    - The only central tendency measurement for categorical data

Spread

$$Var(x) = E[(x-\mu)^2]$$

$$std(x) = \sqrt{Var(x)}$$

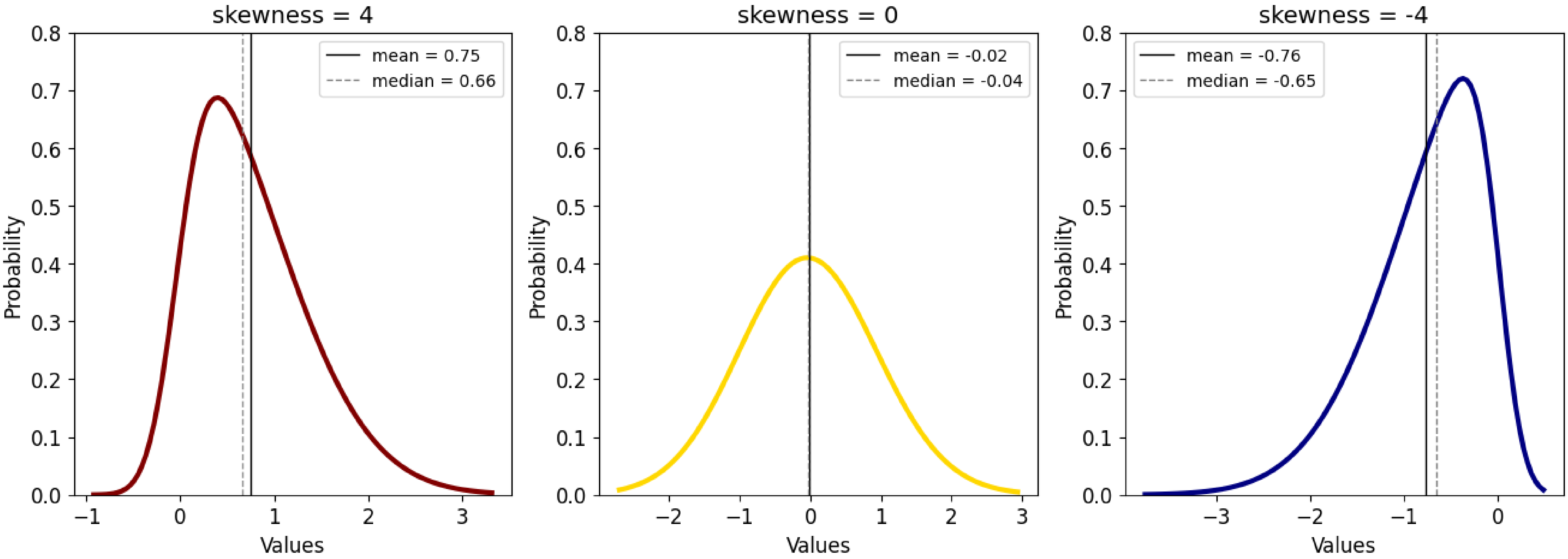
- How far are we from the middle?



Skewness

- How symmetric is the data?

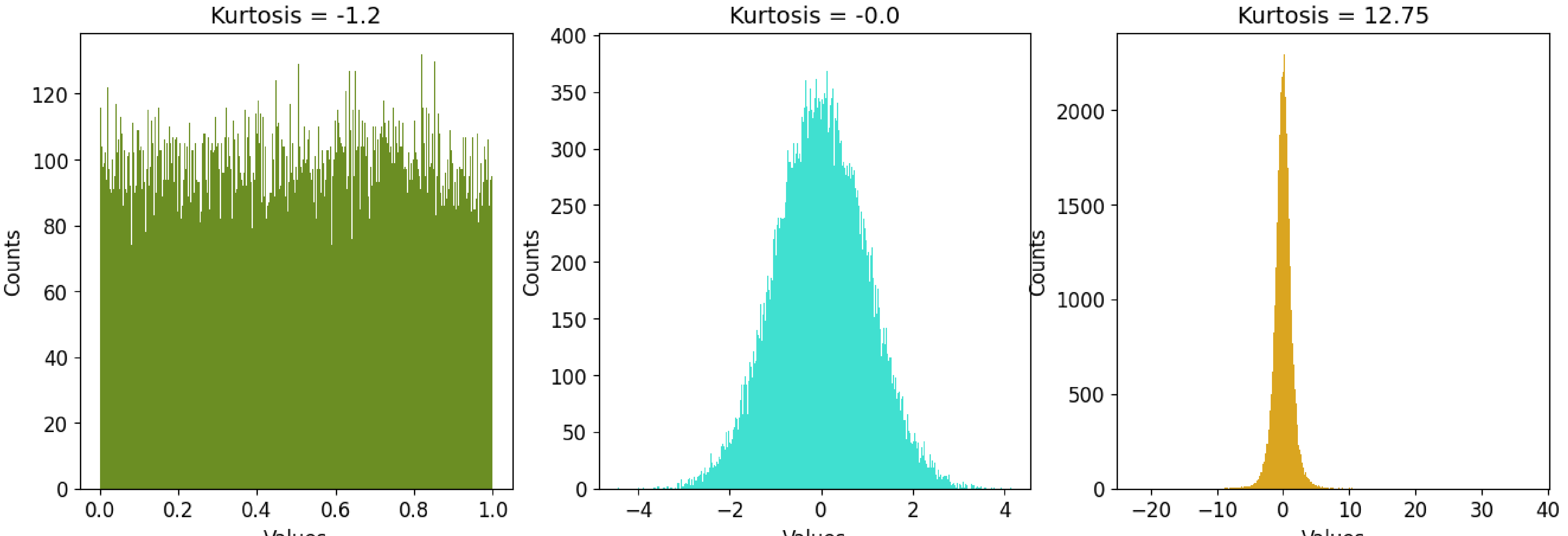
$$\text{Skewness} = \frac{1}{\sigma^3} E\left((x_i - \mu)^3\right)$$



Kurtosis

- Size of tails
- Pointiness of peak

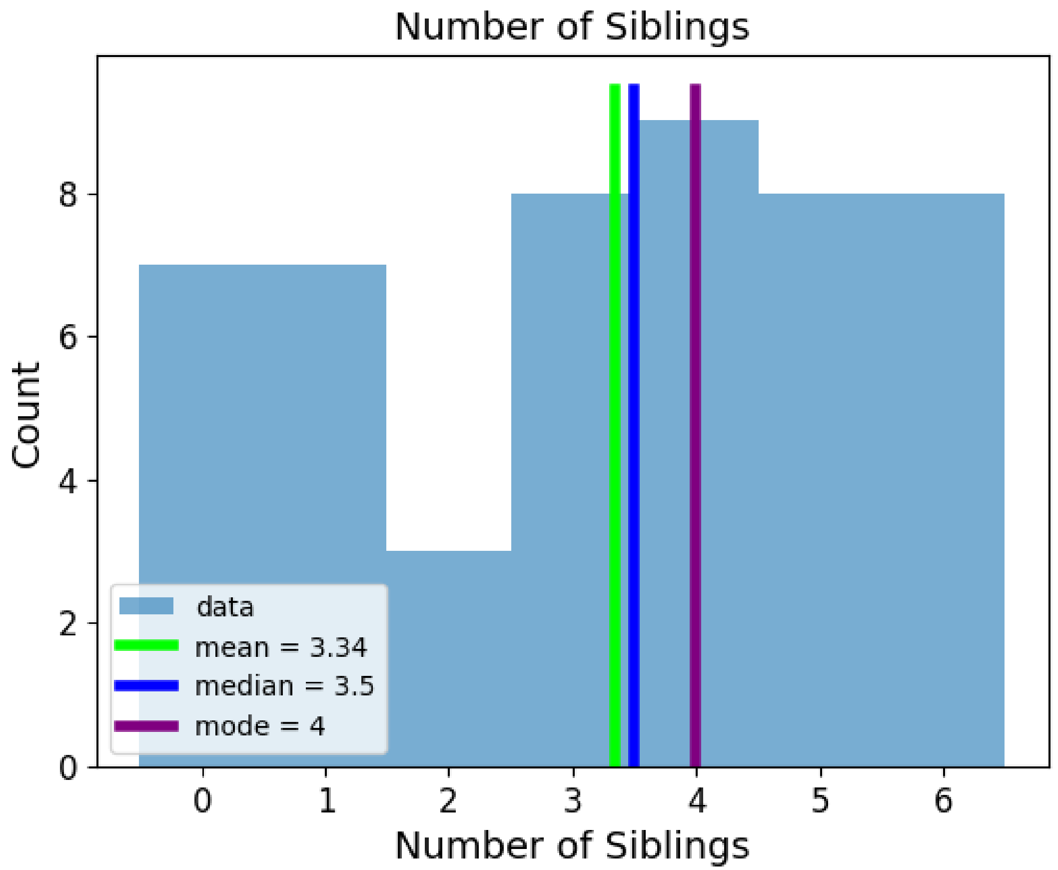
$$\text{Kurtosis} = \frac{1}{\sigma^4} E\left((x - \mu)^4\right) - 3$$

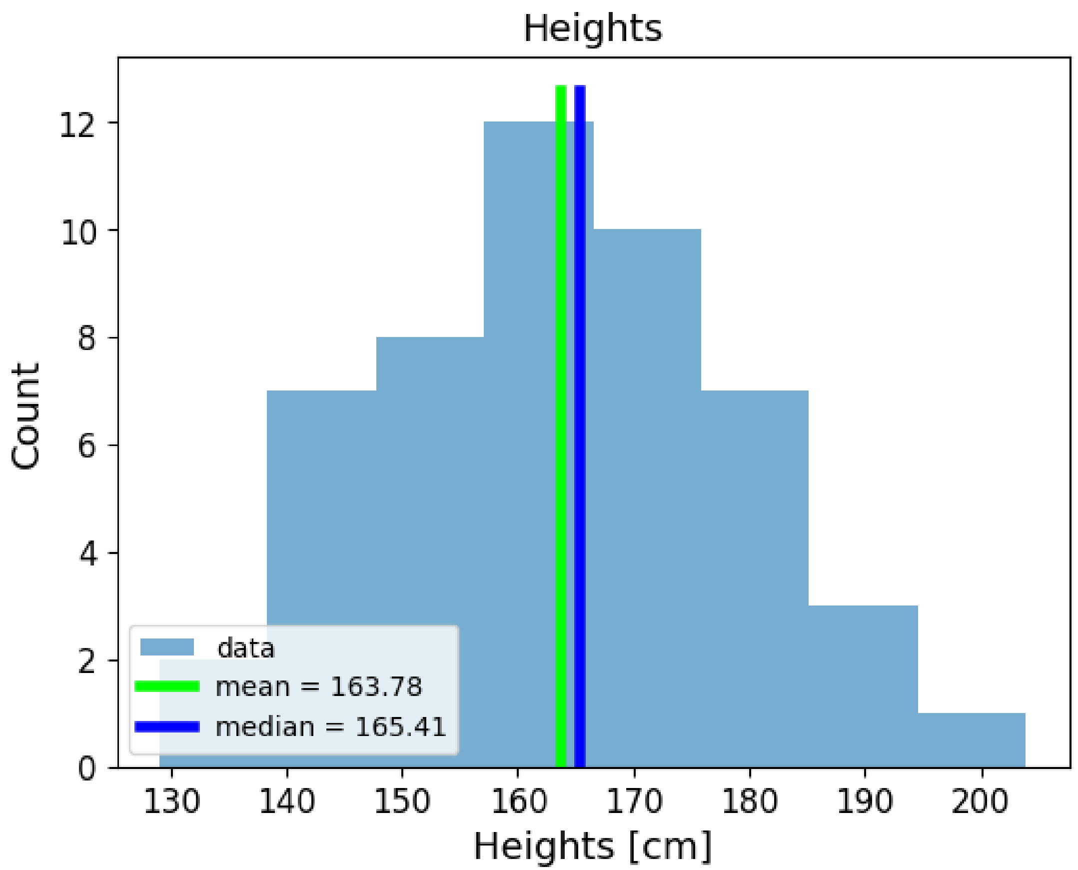


Central Dendency

With random data:

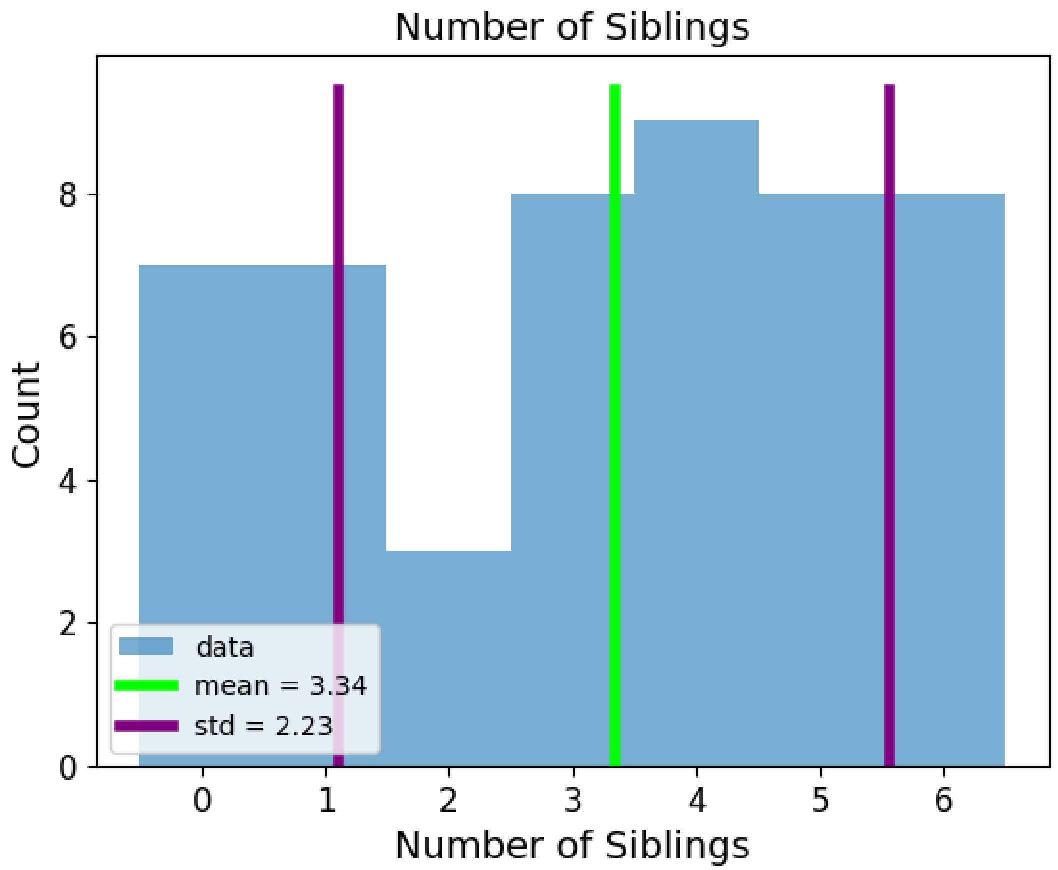
Why didn't we calculate the mode for the heights?

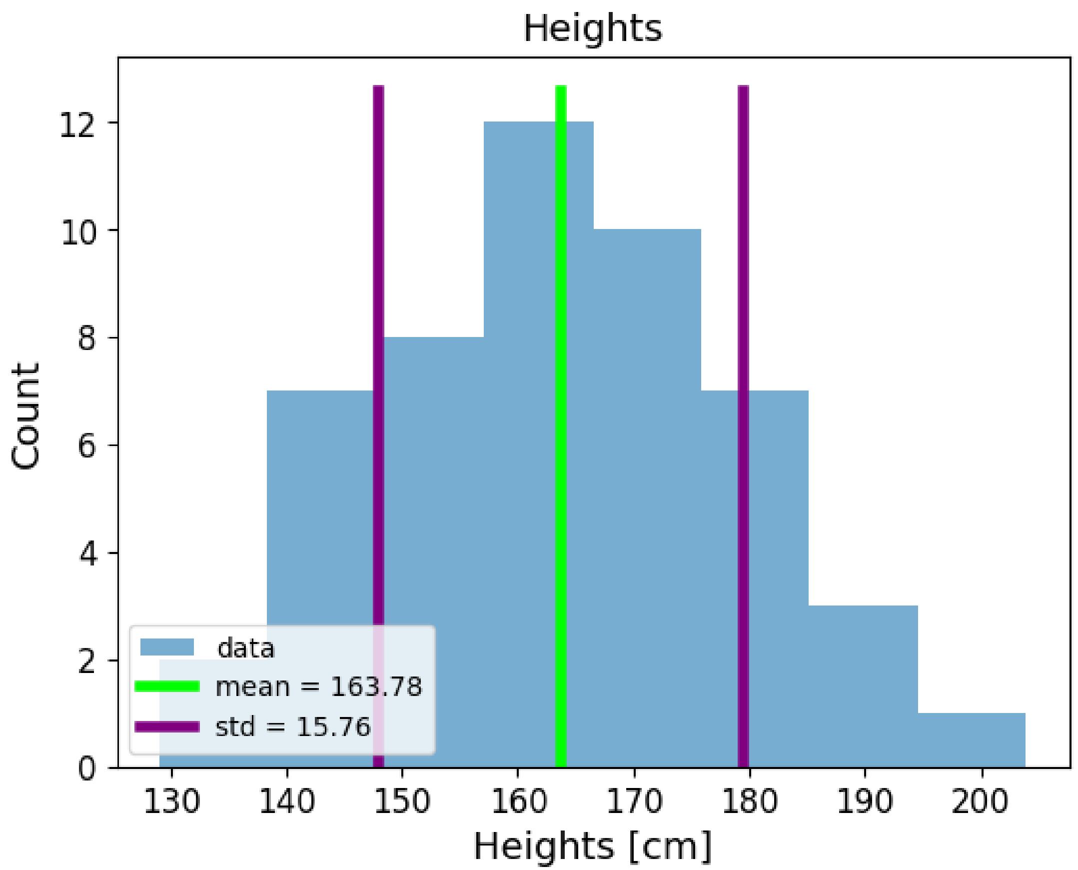




Spread

With random data:





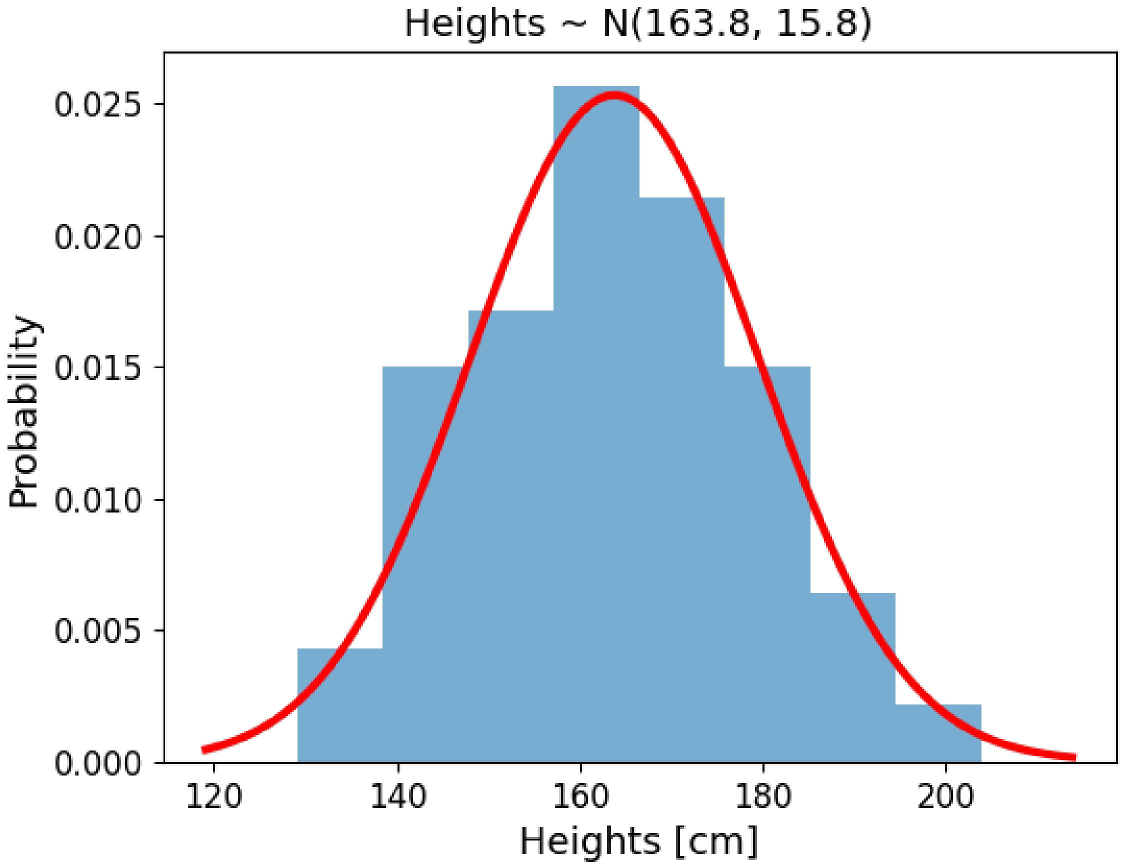
Skewness and Kurtosis

```
#skewness
num_sibs_skew = stats.skew(num_sibs)
heights_skew = stats.skew(heights)

#kurtosis
num_sibs_kurtosis = stats.kurtosis(num_sibs)
heights_kurtosis = stats.kurtosis(heights)

print(f'Skewness of Number of Siblings = {round(num_sibs_skew, 2)}')
print(f'Skewness of Heights = {round(heights_skew, 2)}')

print(f'Kurtosis of Number of Siblings = {round(num_sibs_kurtosis, 2)}')
print(f'Kurtosis of Heights = {round(heights_kurtosis, 2)}')
```

```
Skewness of Number of Siblings = 0.02
Skewness of Heights = 0.05
Kurtosis of Number of Siblings = -1.03
Kurtosis of Heights = -0.28
```

Distribution

We can also model the distribution of the data.



Probability

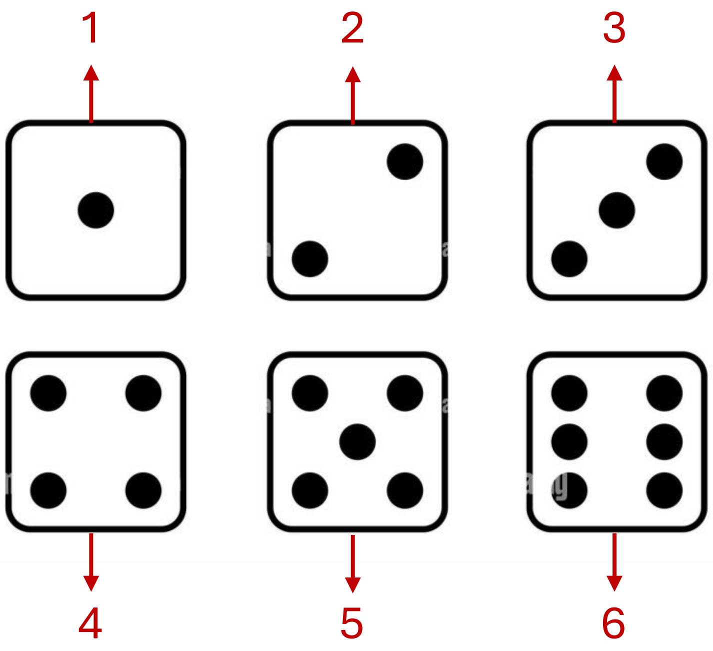
- **Question:** how do people feel about the sunny weather where they live?
- **Experiment:** ask 3 people if they like the weather or not (yes/no).
- **Sample space:** all possible outcomes of an experiment
  - $S = \{(y, y, y), (y, y, n), (y, n, y), (n, y, y), (y, n, n), (n, y, n), (n, n, y), (n, n, n)\}$
- **Events:** subset of the sample space
  - Example: $event\ A = (y, y, y)$
  - $event\ B = \{(y, y, n), (y, n, y), (n, y, y), (y, n, n), (n, y, n), (n, n, y), (n, n, n)\}$
- **Probability of an event:** if all events are equally likely:
  - $P(A) = \frac{size(A)}{size(S)}$

Probability

- Considering all events equally likely makes calculating probability easier, but it’s not necessarily true.
  - It is NOT true that all yes-no questions have a 50-50 chance (do you like going to the doctor?)
  - What is the probability of seeing a purple horse?
    - Natural color
    - Cartoon
- We’ll assign a distribution.

Random Variable

- Function that maps the sample space into real numbers (R)



- P(X=3), P(X=x), P(x≤4)
- Discrete / Continuous
- Instead of calculating probabilities of events, we may be more interested in finding out the list of probabilities for all possible events = **probability distribution**.

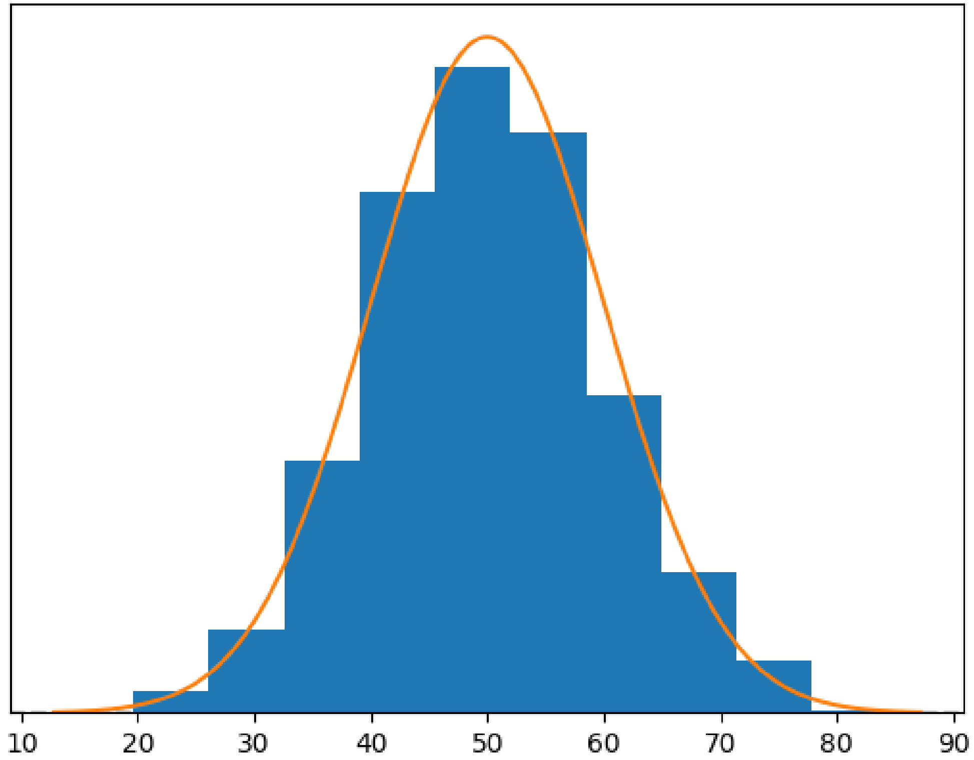
Preliz

- We will use this library a lot.
- https://github.com/arviz-devs/preliz
- Use installation from the tutorial notebook.

Normal Distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

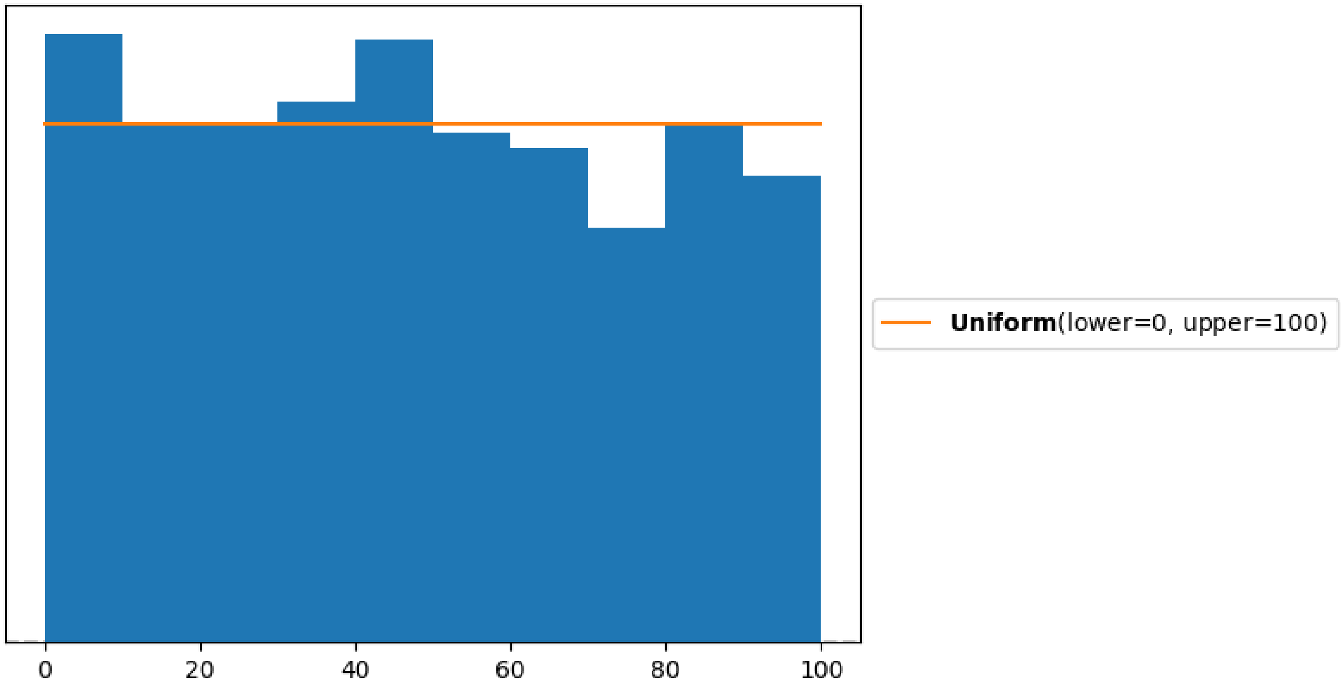
- μ – mean
- σ – standard deviation
- f(x) – probability **density** function



Continuous Uniform

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b \\ 0, & else \end{cases}$$

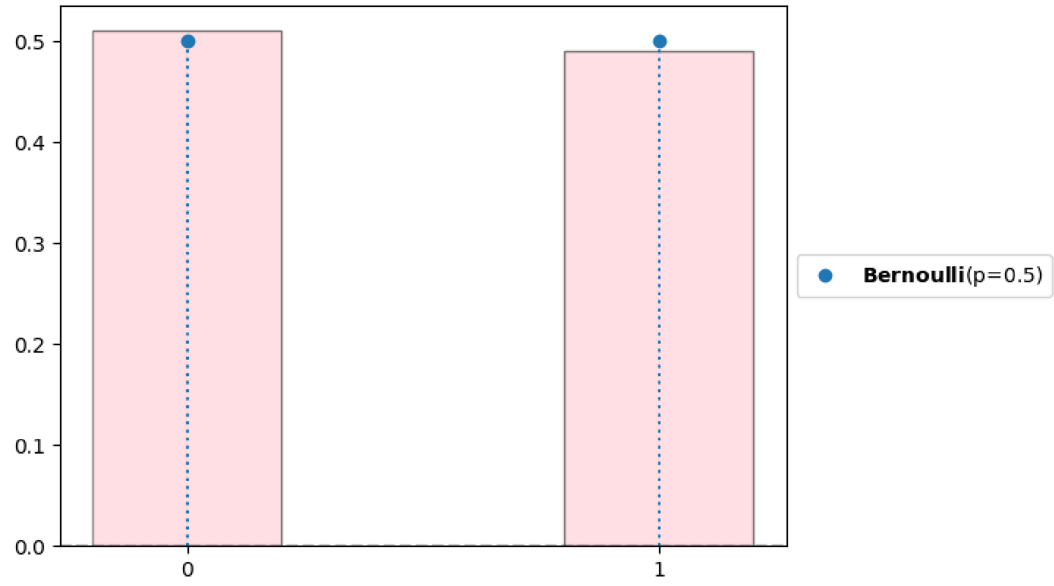
- f(x) – probability **density** function

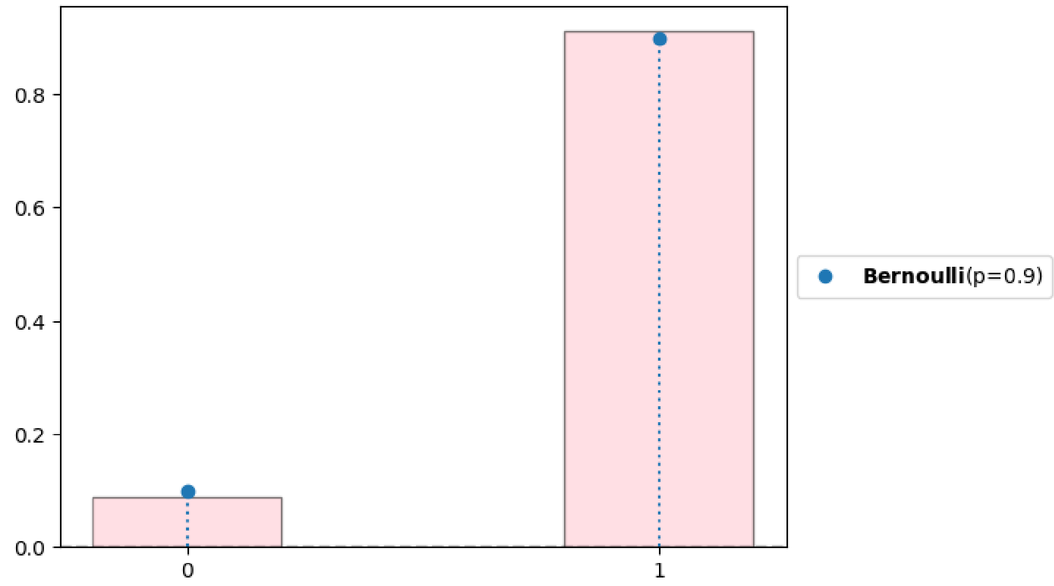


Bernoulli

$$f(x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

- p = the probability of success
- 1-p = the probability of failure
  - Coin toss – fair coin p=0.5
  - sucess / failure in a single trial
- f(x) – probability **mass** function

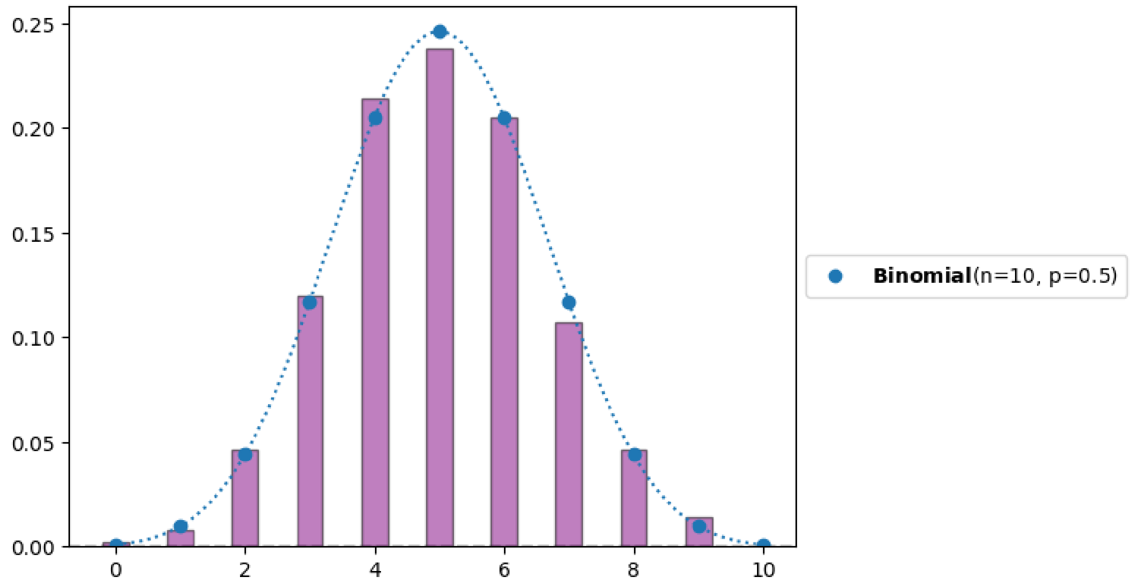


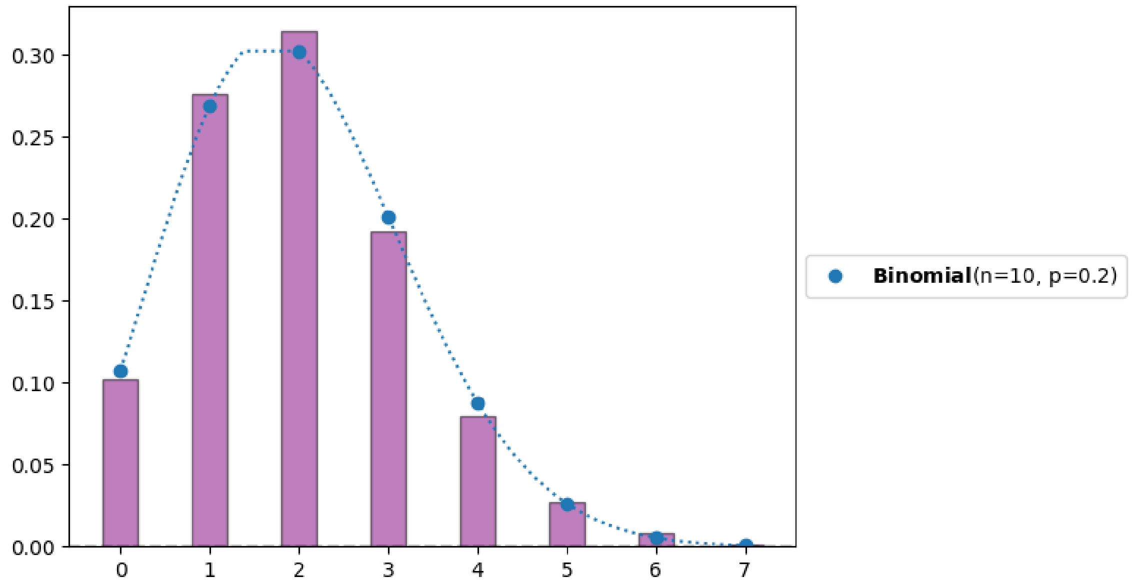


Binomial

$$f(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- p = the probability of success
- n = the number of experiments
- k = the number of successes
- f – probability **mass** function



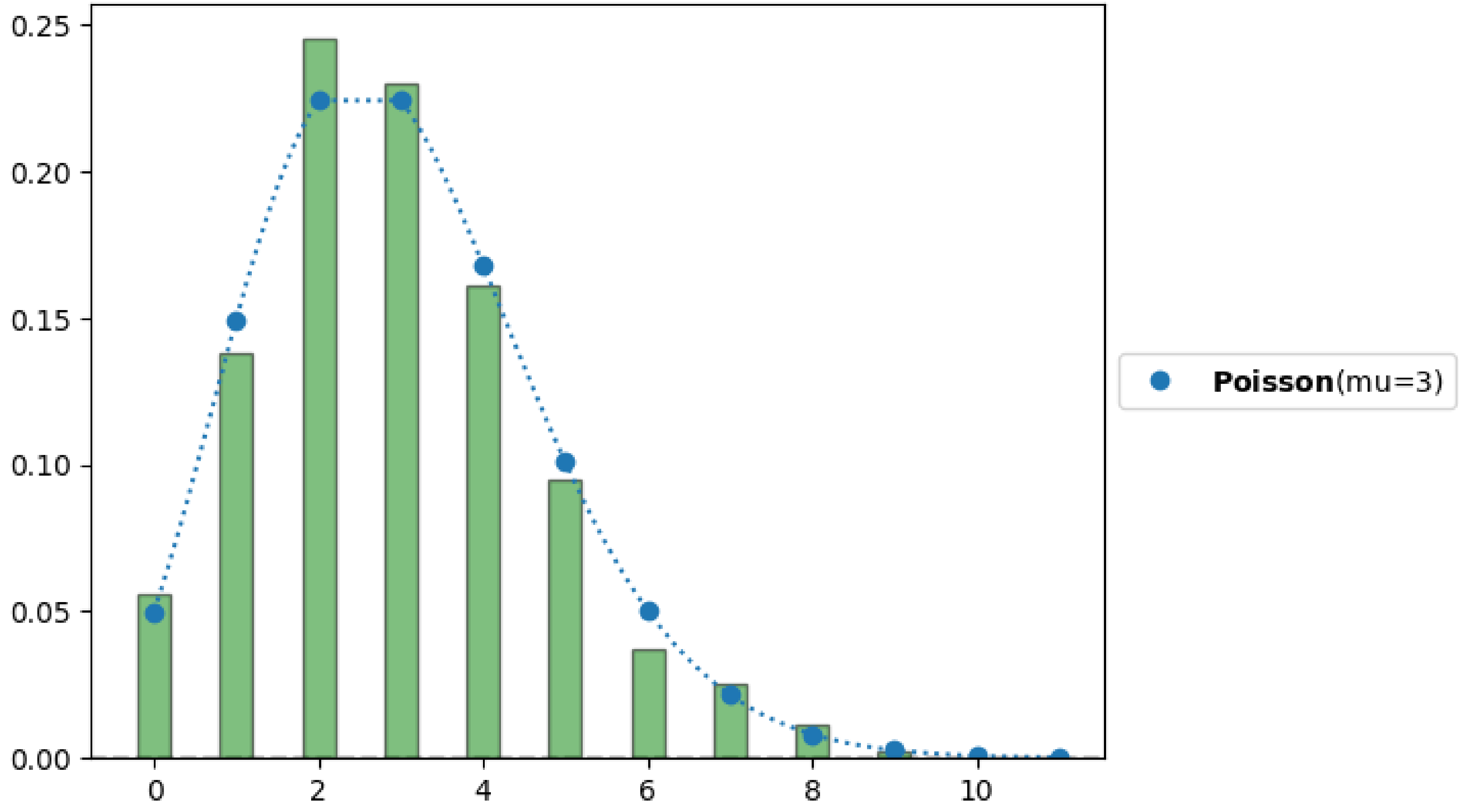


Poisson

$$f(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

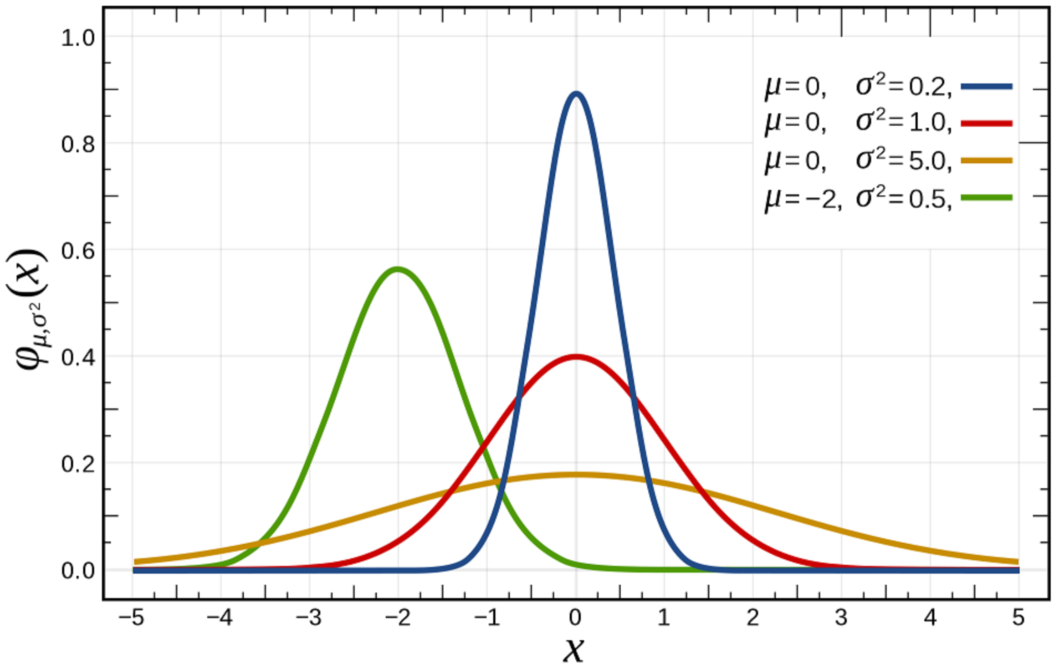
- The probability of a given number of events occurring in a fixed interval
  - Events occur at a known constant mean rate - λ
  - Events occur independently of the time since the last event
- k = the number of occurrences
- f – probability **mass** function

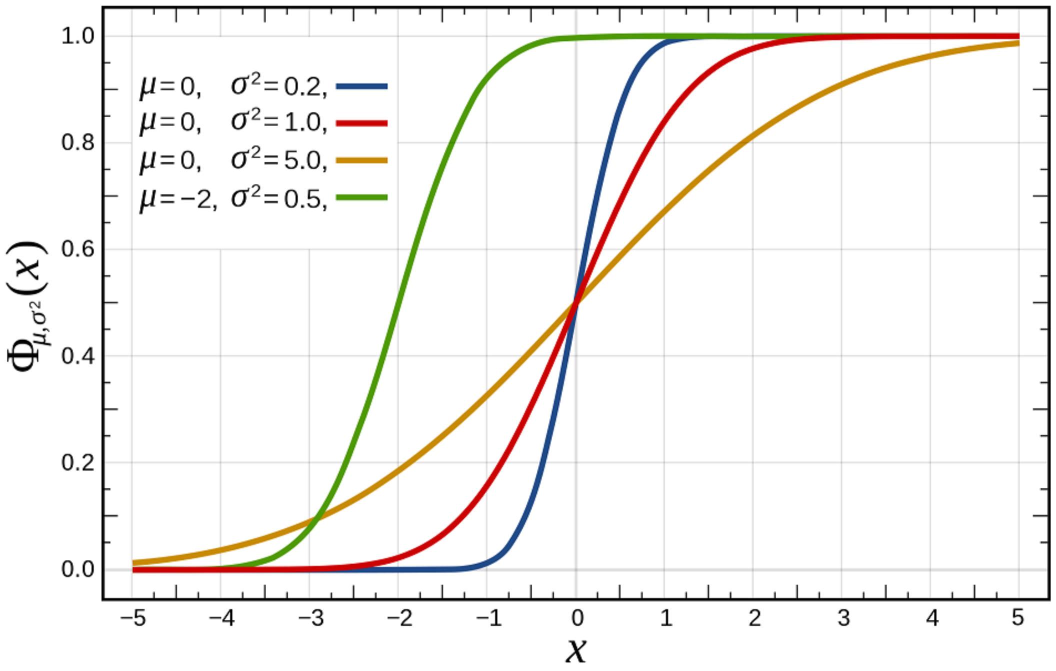
Poisson



Cumulative Distribution Functions

- $P(X \leq x)$





$$cdf_{N(\mu,\sigma^2)}(x) = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

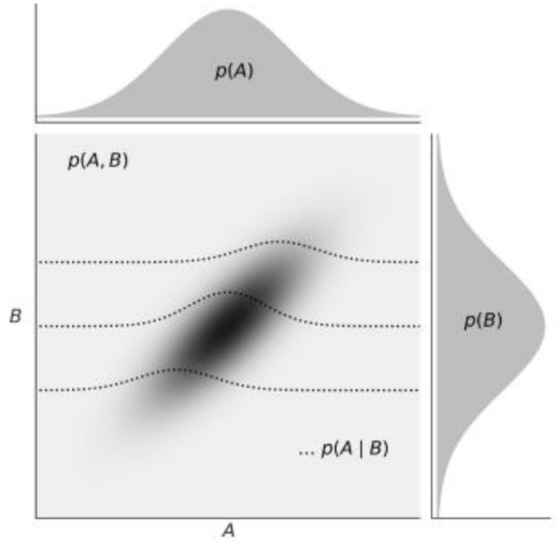
$$\operatorname{erf} z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, \mathrm{d}t.$$

Conditional Probability

- What is the chance of event A if we know that event B has occurred?
  - What is the chance of needing an umbrella (A) if we know it is raining (B)?
  - $P(A|B) = \frac{P(A,B)}{P(B)}, P(B) > 0$
- If A and B are independent:
  - $P(A|B) = \frac{P(A,B)}{P(B)} = \frac{\mathrm{P(A)P(B)}}{P(B)} = P(A)$

Conditional Probability

- Dark colors = higher probability
- Higher A ↔higher B
- Dashed lines show P(A|B) for 3 different values of B.
- Marginal Distributions: P(A), P(B)



Conditional Probability

- Fair dice: what is the probability of rolling 3?
  - $P(3) = \frac{1}{6}$
- What is the probability of rolling 3 if we know we rolled an odd number?
  - $P(3|\{1,3,5\}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$

Conditional Probability

- You have a standard deck of 52 cards.
- What is the probability that the card you drew is a King?
  - $P(King) = \frac{4}{52} = \frac{1}{13}$
- Suppose we know that the card you drew is a **face card** (Jack, Queen, or King).

Conditional Probability

- What is the probability that the card you drew is a King, given that it is a face card?
- $P(King|Face\ card) = \frac{P(King\ and\ Face\ Card)}{P(Face\ card)}$
  - $P(King\ and\ Face\ Card)$ = {every king card is also a face card} = $\frac{4}{52}$
  - $P(Face\ Card) = \frac{12}{52}$
  - $P(King|Face\ card) = \frac{\frac{4}{52}}{\frac{12}{52}} = \frac{1}{3}$

Conditional Probability

- In a class of 100 students, 80 students passed the exam, and 50 students studied hard. Of those who studied hard, 45 passed the exam.
- What is the probability that a student passed the exam, given that they studied hard?

  - $P(Passed\ |Studied) = \frac{P(Passed\ and\ Studied)}{P(Studied)}$
  - $P(Passed\ and\ Studied) = \frac{45}{100}$
  - $P(Studied) = \frac{50}{100}$
  - $P(Passed\ |Studied) = \frac{\frac{45}{100}}{\frac{50}{100}} = 0.9$
- $P(Passed) = 0.8$

Bayes Theorem

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

- $P(c|r) \neq P(r|c)$
- The probability of a person being the Pope given that this person is Argentinian is not the same as the probability of being Argentinian given that this person is the Pope.
  - There are around 47,000,000 Argentinians alive and a single one of them is the current Pope - $P(pope|Argentinian) = \frac{1}{47{,}000{,}000}$
  - $P(Argentinian|\text{pope}) = 1$

Bayes Theorem

- $P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model})P(\text{model})}{\sum_{\text{models}} P(\text{data}|\text{model})P(\text{model})}$

- Prior distribution: what do we know about the values of the model before seeing the data.
- Likelihood: how we will introduce our data.
- Posterior distribution: the result of Bayesian analysis -> reflects all we know about our question given our data and model.
  - Probability distribution for the parameters in our model
  - Not a single value
- Marginal likelihood: probability of observing the data averaged over all the possible values the parameter can take
  - Normalization factor