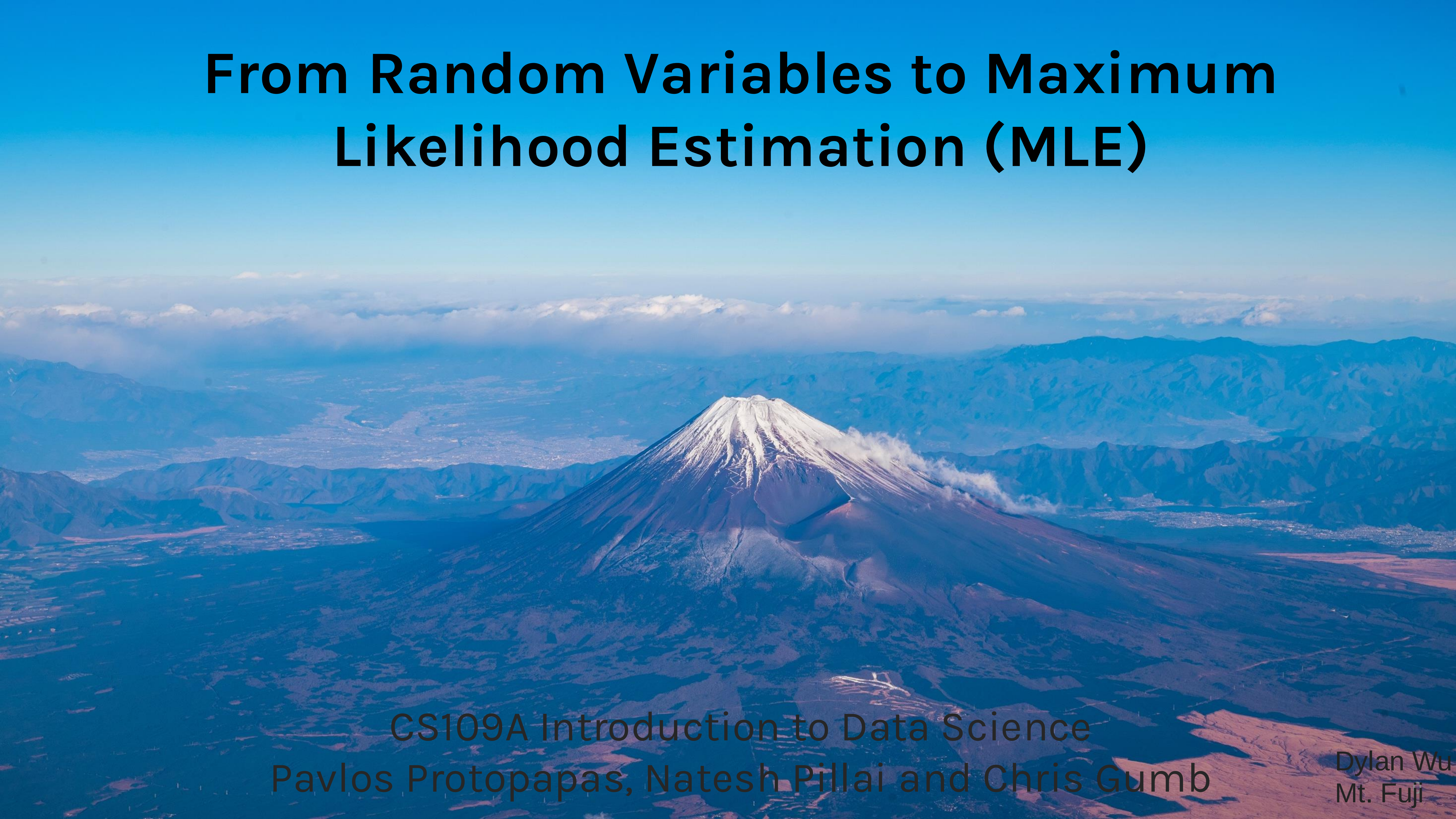
From Random Variables to Maximum Likelihood Estimation (MLE)



CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai and Chris Gumb

Dylan Wu
Mt. Fuji

Outline

- What is a random variable?
- Point estimates of random variables. Confidence Intervals, Histogram, Probability, and PDF/PMF
- Known random variables: Uniform, Binomial. Normal
- Joint Distributions
- Modeling Data with Probability Distributions
- Likelihood Theory
- Modeling Linear Regression Probabilistically

Outline

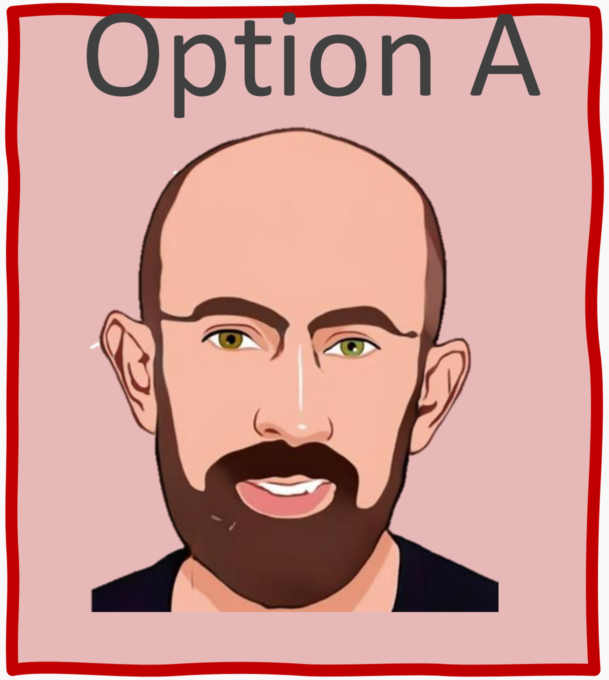
- **What is a random variable?**
- Point estimates of random variables. Confidence Intervals, Histogram, Probability, and PDF/PMF
- Known random variables: Uniform, Binomial. Normal
- Joint Distributions
- Modeling Data with Probability Distributions
- Likelihood Theory
- Modeling Linear Regression Probabilistically

CS109A Olympics

CS109A 100m dash

WHO WILL WIN THE 100M DASH?

CS109A 100m dash



THE PROFESSOR

AVERAGE Pace: 13 seconds

Consistency: High

CS109A 100m dash

Option B

THE HOT SHOT

AVERAGE Pace: 13.0 secor

Consistency: VERY LOW

Option A

CS109A 100m dash

THE BANGALORE CHAMPION

AVERAGE Pace: 13.1 seconds

Consistency: MEDIUM

Option C

Option A

Option B

CS109A 100m dash

THE Researcher

AVERAGE Pace: 14 seconds

Consistency: LOW

Option D

Option A

Option B

Option C

Option A

Option B

Option C

Option D

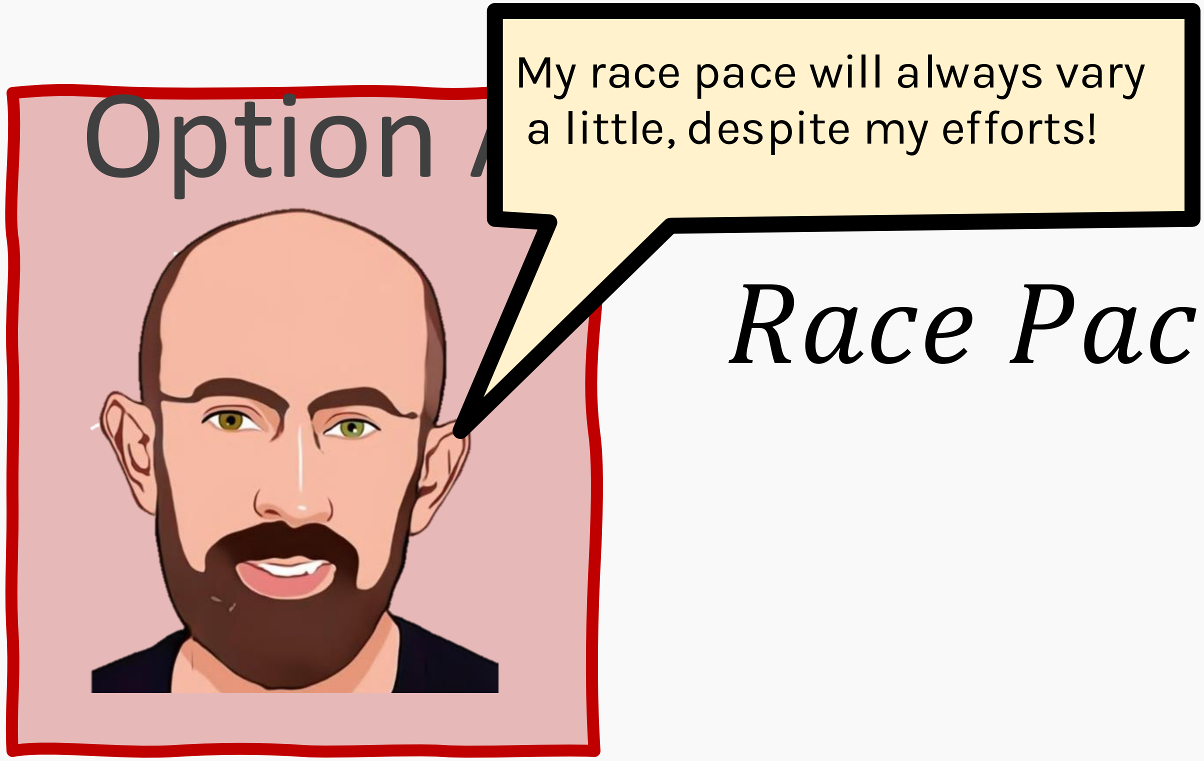
	Option A	Option B	Option C	Option D
Race #1	13.12	13.52	14.25	13.51
Race #2	[illegible]	[illegible]	[illegible]	15.01
Race #3	[illegible]	[illegible]	[illegible] 🏆	13.63
Race #4	12.87 🏆	13.52	13.12	13.91
Race #5	13.22	13.24	12.78	12.32 🏆

RACE TIME

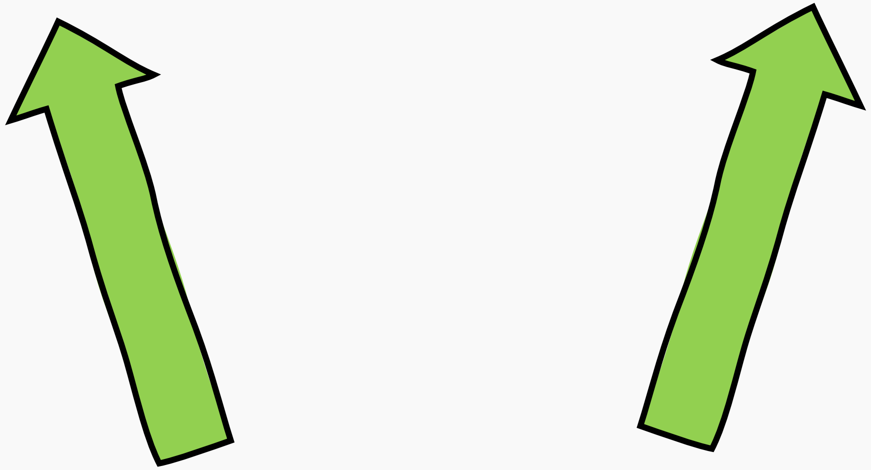
PROTOPAPAS

CS109A 100m dash

Let X be the race pace for a given 100m dash, then X is called a random variable



$$Race\ Pace = Average\ Pace + \ \epsilon$$



Constant

Varying

RECAP: Python variables

We have seen variables as something we assign a value to.

- Integer <int>
- Float <float>
- List <list>
- Dictionary <dict>

```
In [2]: a = 2
In [3]: b = 2.5
In [4]: pavloslist = [1,2,3,4,5]
In [5]: pavlosdict ={‘John':2,'Pavlos’:2,Eric':7}

In [6]: type(a)
Out[6]: int
In [7]: type(b)
Out[7]: float
In [8]: type(pavloslist)
Out[8]: list
In [9]: type(pavlosdict)
Out[9]: dict
```

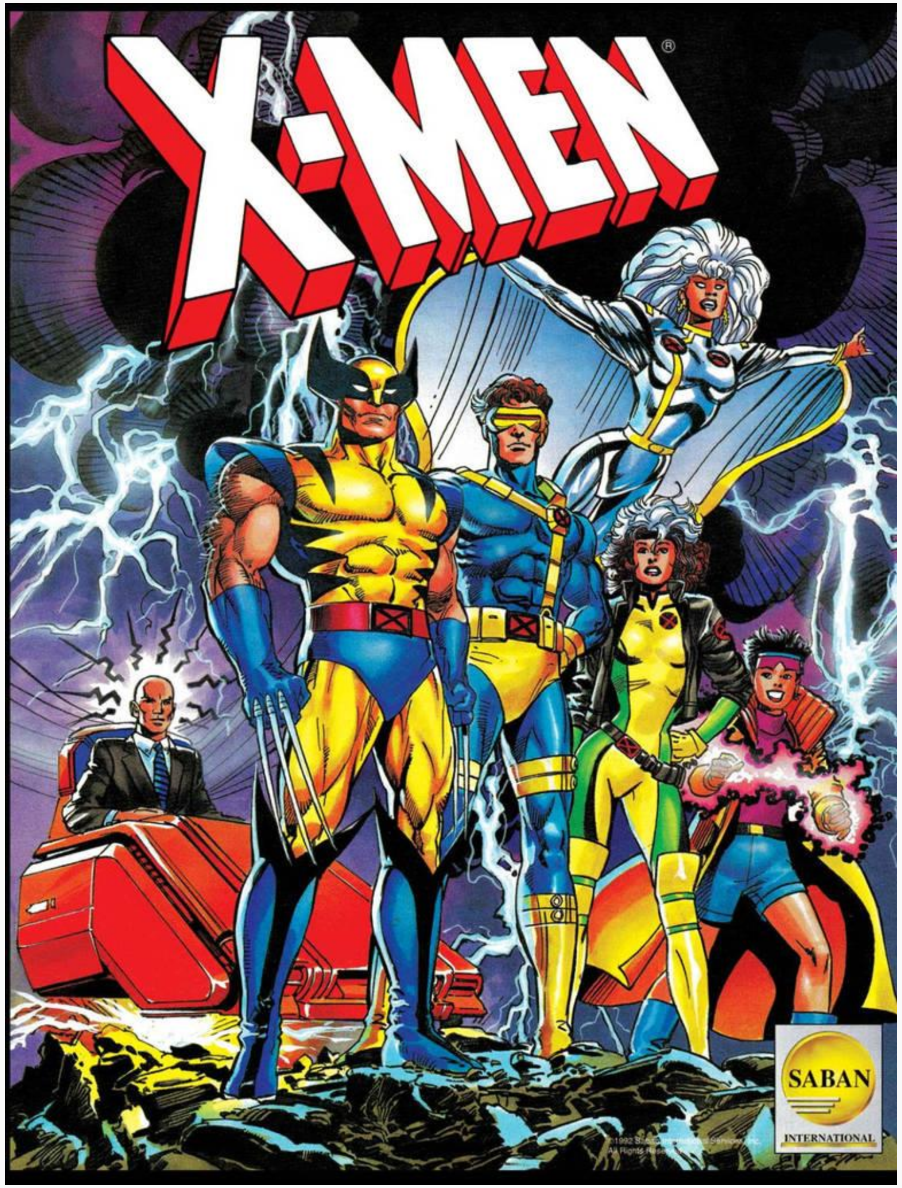
Random Variable

- A random variable can be thought of as a numeric outcome of a random experiment.

- Unlike the python variables defined before, the value of a random variable is not fixed.

- The output of a random variable could be either discrete (can only take on specific values) or continuous (can take on any value within a range).

```
In [26]: x = RandomVariable()
In [27]: x.random
Out[27]: 0.5632899481539281
In [28]: x.random
Out[28]: 0.630954141651853
```



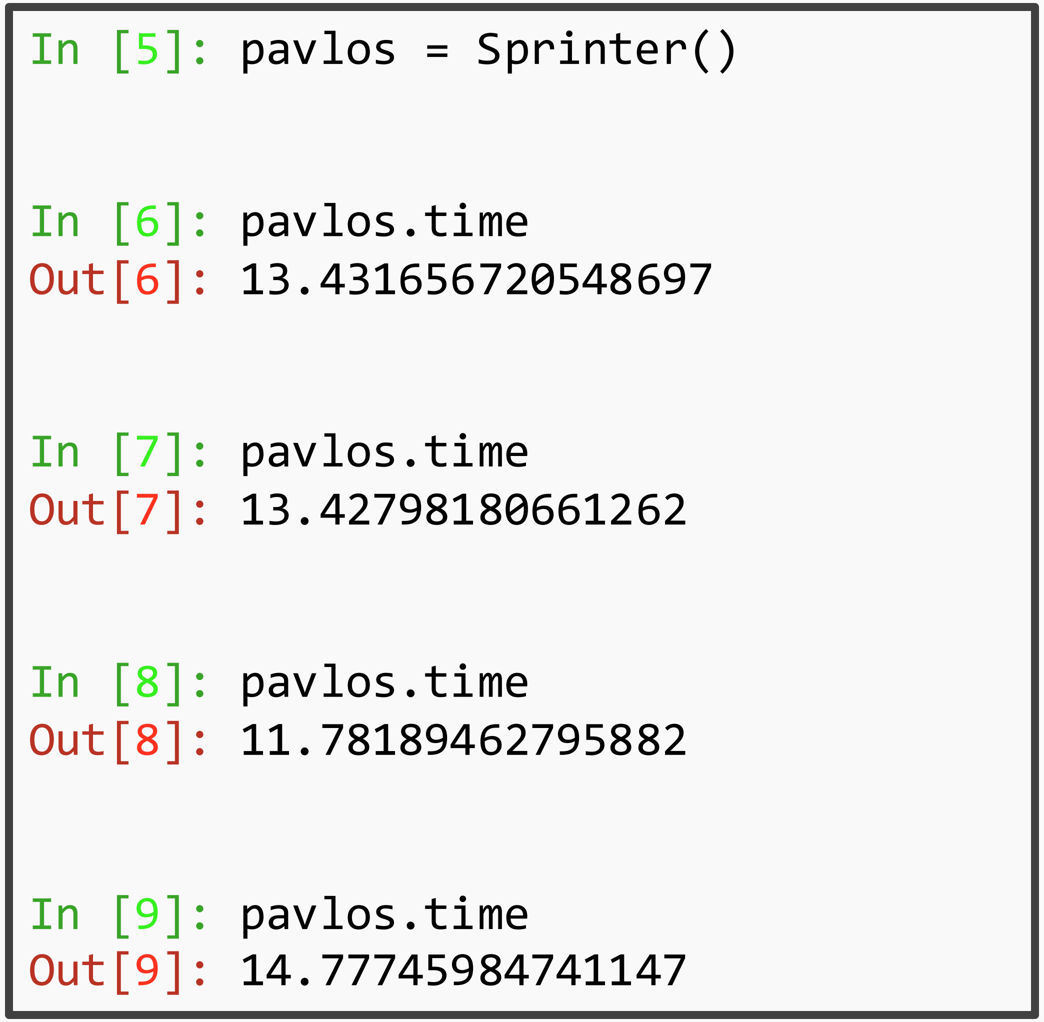
Random Variable

Questions about a random variable?

$$X = Average\ Pace + \epsilon$$

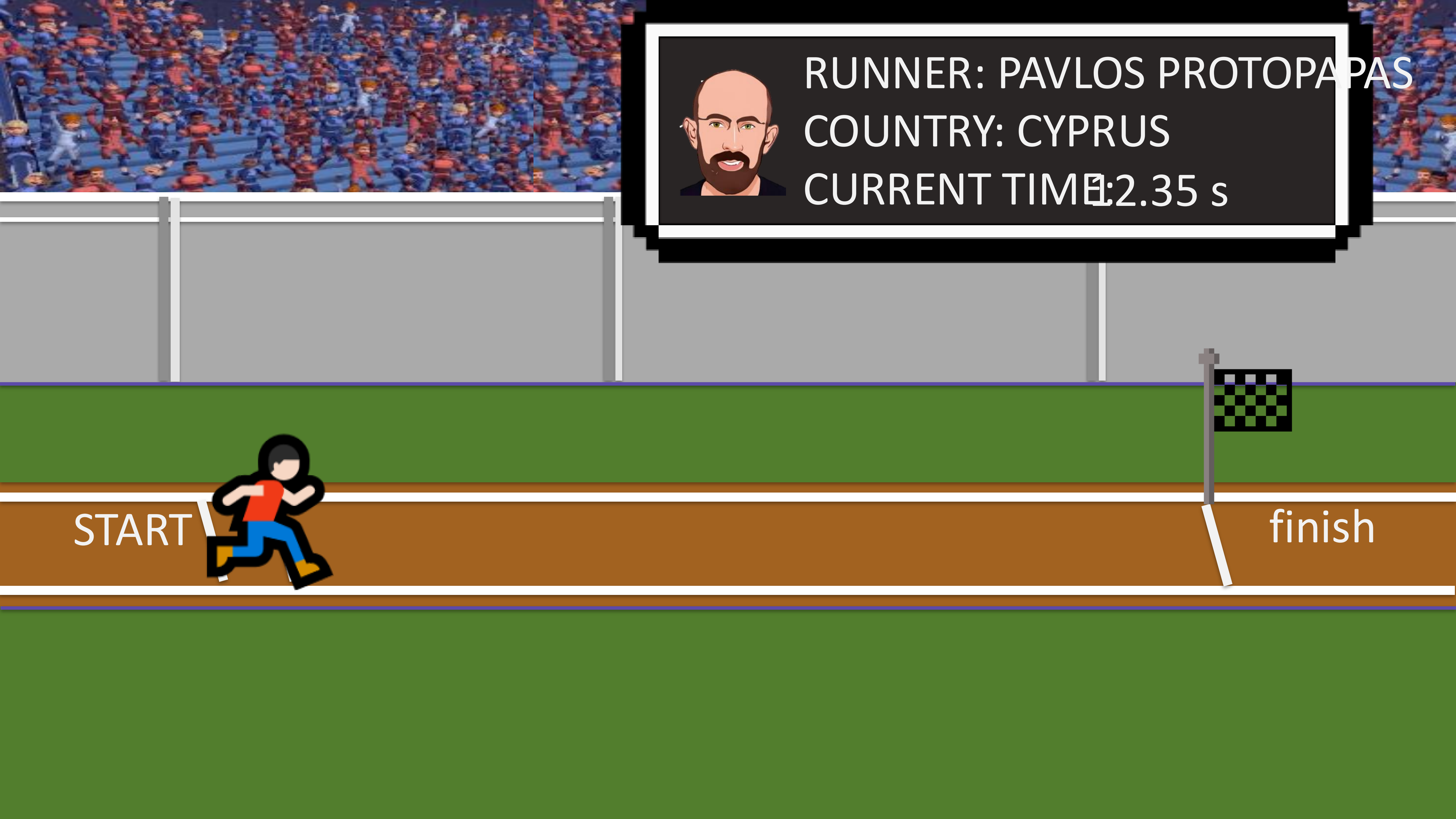
- What are the possible values of X?
- What is the maximum value of X?
- What is the minimum value of X?
- What is the expected value of X?
- Are the values of X spread out, or consistent?

AND MANY MORE QUESTIONS ...



```
In [5]: pavlos = Sprinter()


In [6]: pavlos.time
Out[6]: 13.43165720548697


In [7]: pavlos.time
Out[7]: 13.4279818066126 2


In [8]: pavlos.time
Out[8]: 11.7818946279588 2


In [9]: pavlos.time
Out[9]: 14.7774598474114 7
```

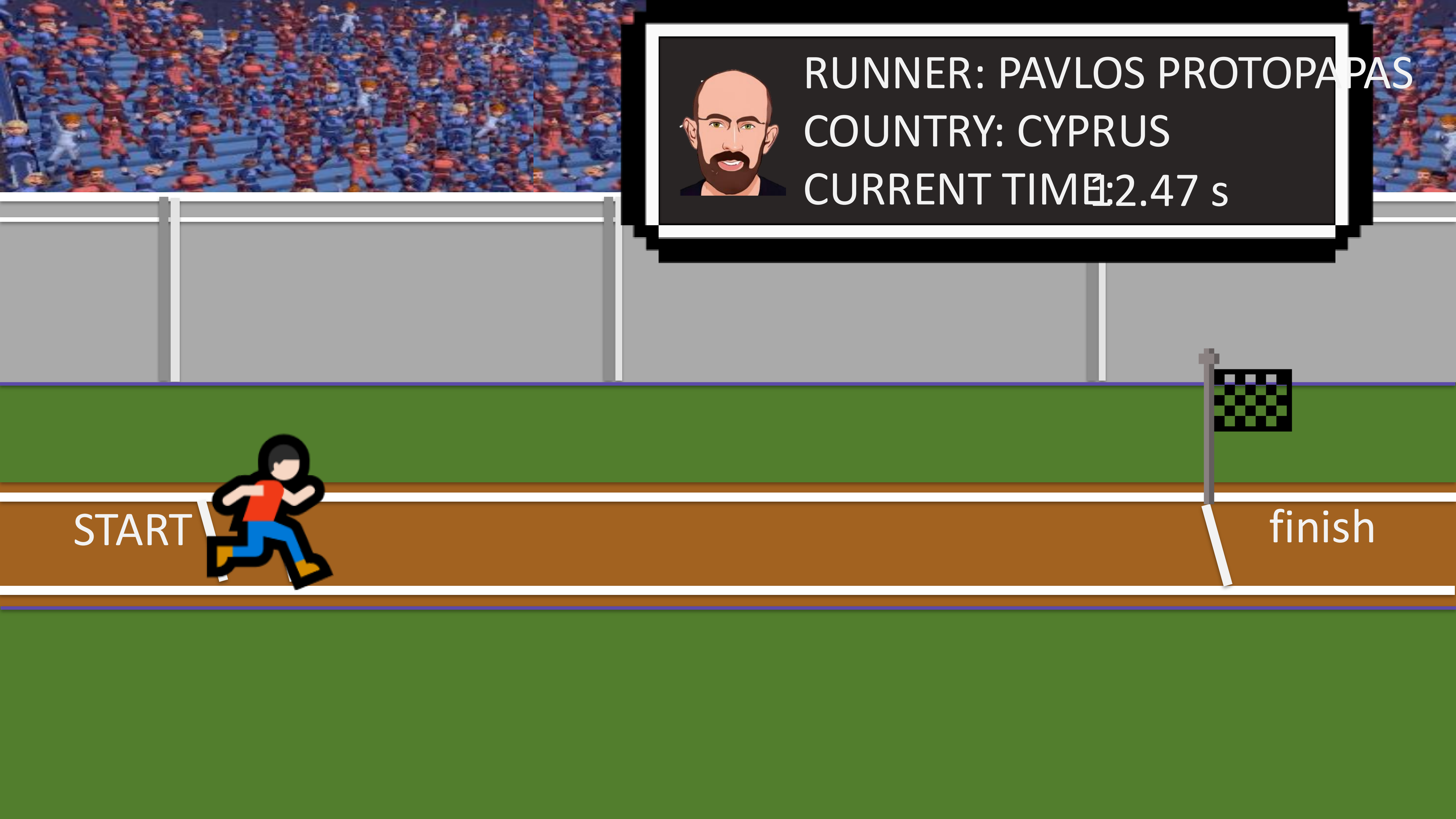
Simulations



RUNNER: PAVLOS PROTOPAPAS

COUNTRY: CYPRUS

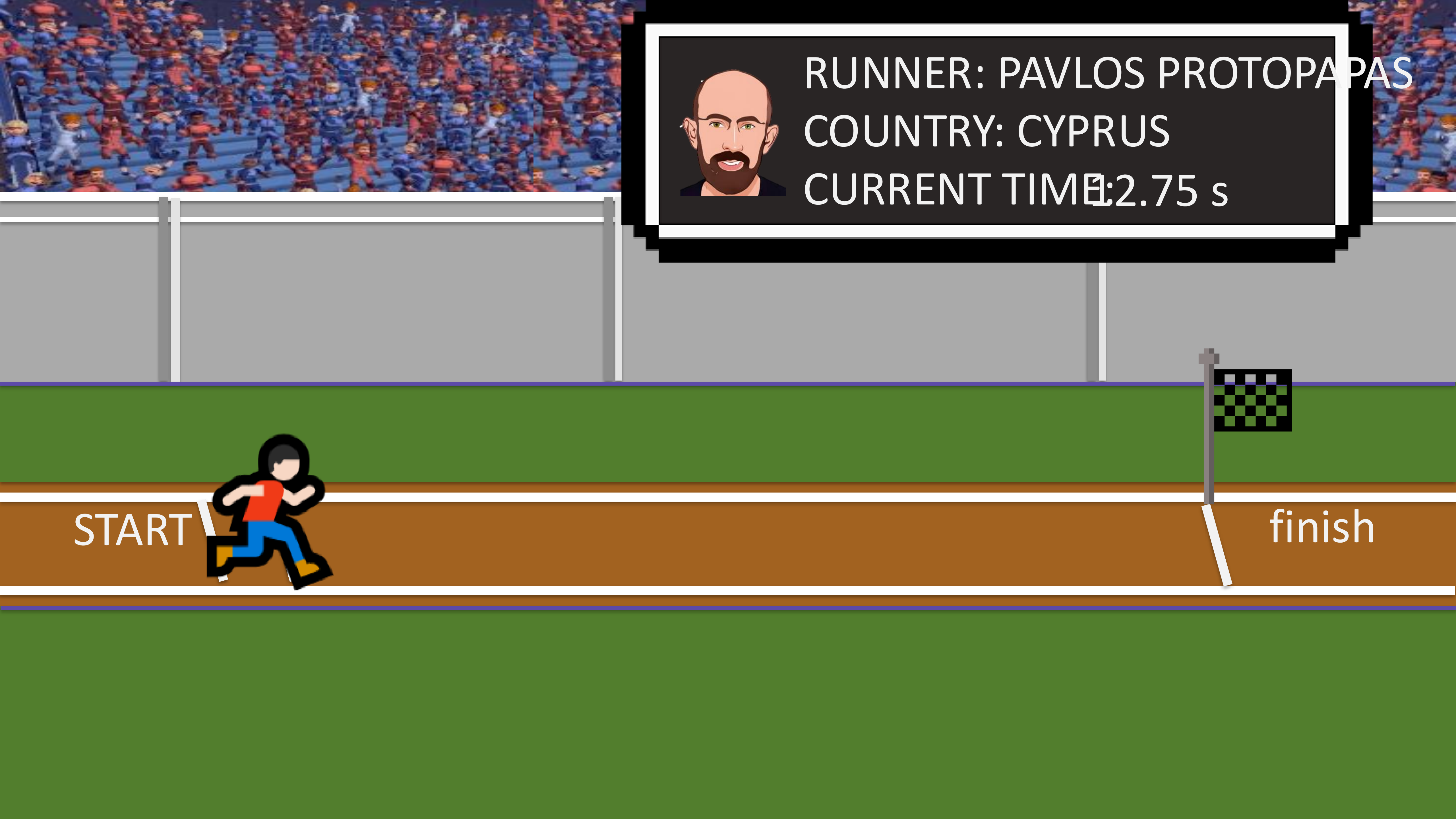
CURRENT TIME: 12.35 s



RUNNER: PAVLOS PROTOPAPAS

COUNTRY: CYPRUS

CURRENT TIME: 12.47 s

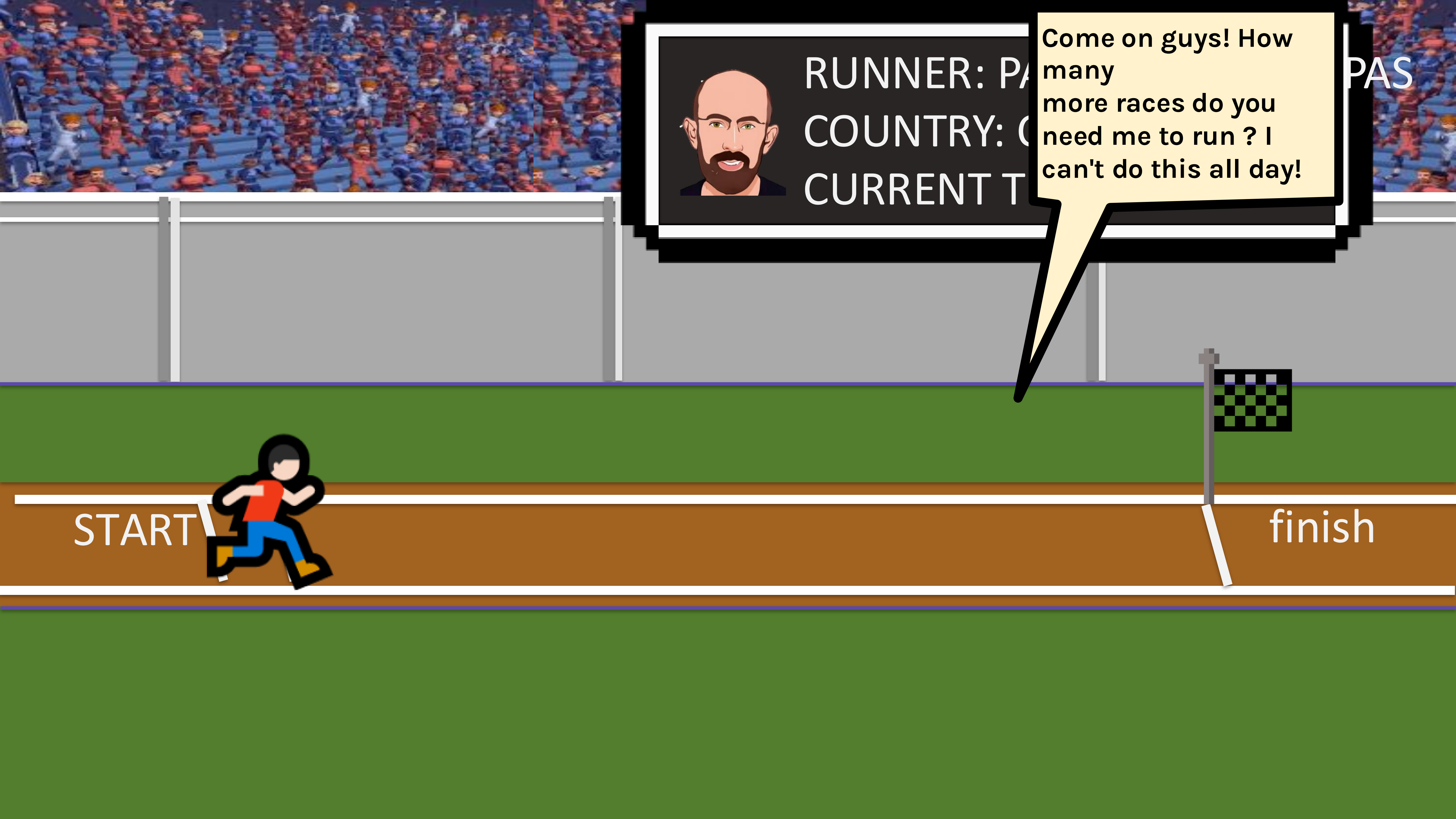

RUNNER: PAVLOS PROTOPAPAS
COUNTRY: CYPRUS
CURRENT TIME:12.75 s
START
finish

Random Variable

$$X = Average\ Pace + \ \epsilon$$

[13.75, 15.21, 13.65, 13.58, 12.93, 14.23, 12.81, 11.50, 13.09, 12.26, …]

- We could run the experiment multiple times and record the results.
- This will give us a list of possible values of the random variable X.
- An exhaustive list of all possible values is often called the **population** space of the random experiment.


RUNNER: PA
PAS
COUNTRY: C
CURRENT T
Come on guys! How many more races do you need me to run ? I can't do this all day!
START
finish

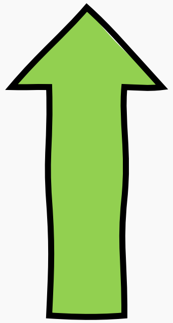
Random Variable

$$\left[13.75, 15.21, \ 13.65, 13.58, 12.93, 14.23, 12.81, \ 11.50, 13.09, 12.26, \ldots\right]$$

Random Variable

[13.75, 15.21, 13.65, 13.58, 12.93, 14.23, 12.81, 11.50,13.09,12.26, ...]

[13.75, 13.65, 12.93, 12.81, 12.26]



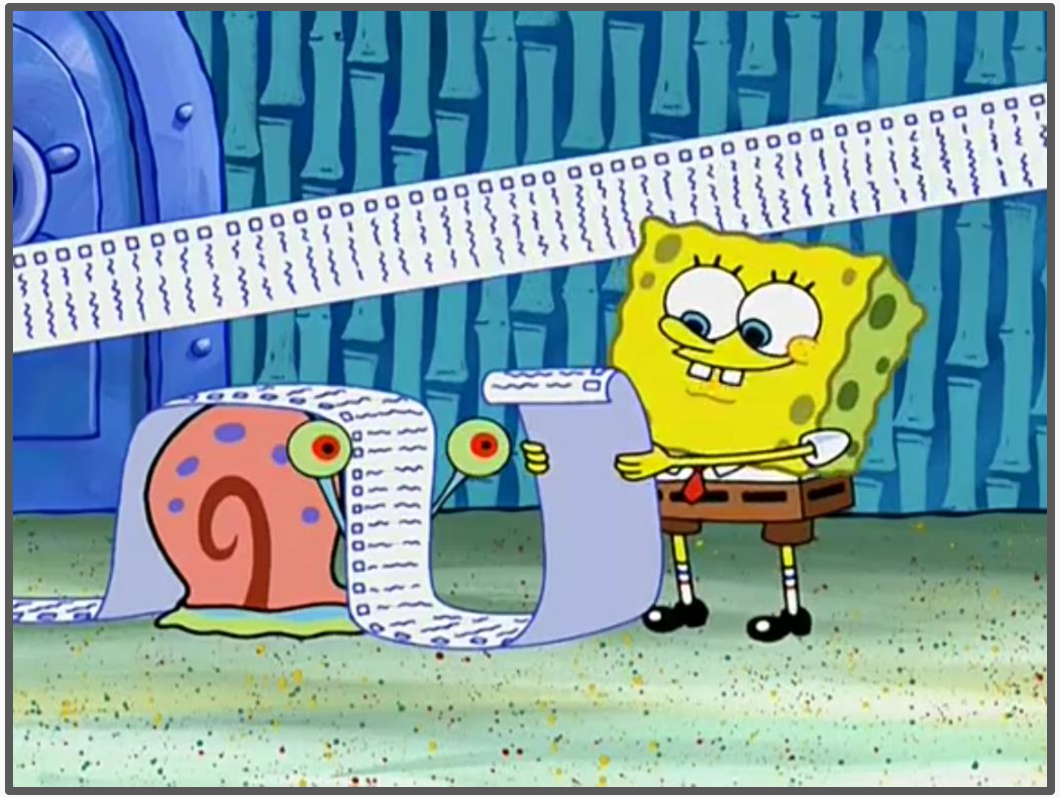
Sample

Properties of a Random Variable

ISSUES?

The outcomes of a random variable captured over multiple simulations is difficult to interpret and consequently difficult to compare to other random variables.

- **ISSUE #1**: We do not have estimates to compare with other random variables.
- **ISSUE #2**: It is difficult to visualize the spread of the outcome.



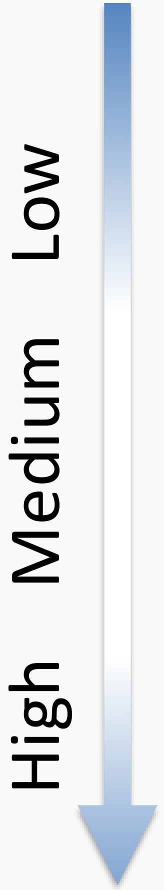
Properties of a Random Variable

ISSUES?

The outcomes of a random variable captured over multiple simulations is difficult to interpret and consequently difficult to compare to other random variables.

- **ISSUE #1**: We do not have estimates to compare with other random variables.
- **ISSUE #2**: It is difficult to visualize the spread of the outcome.

Description of Random Variables



Point Estimates

Confidence Intervals

Histogram

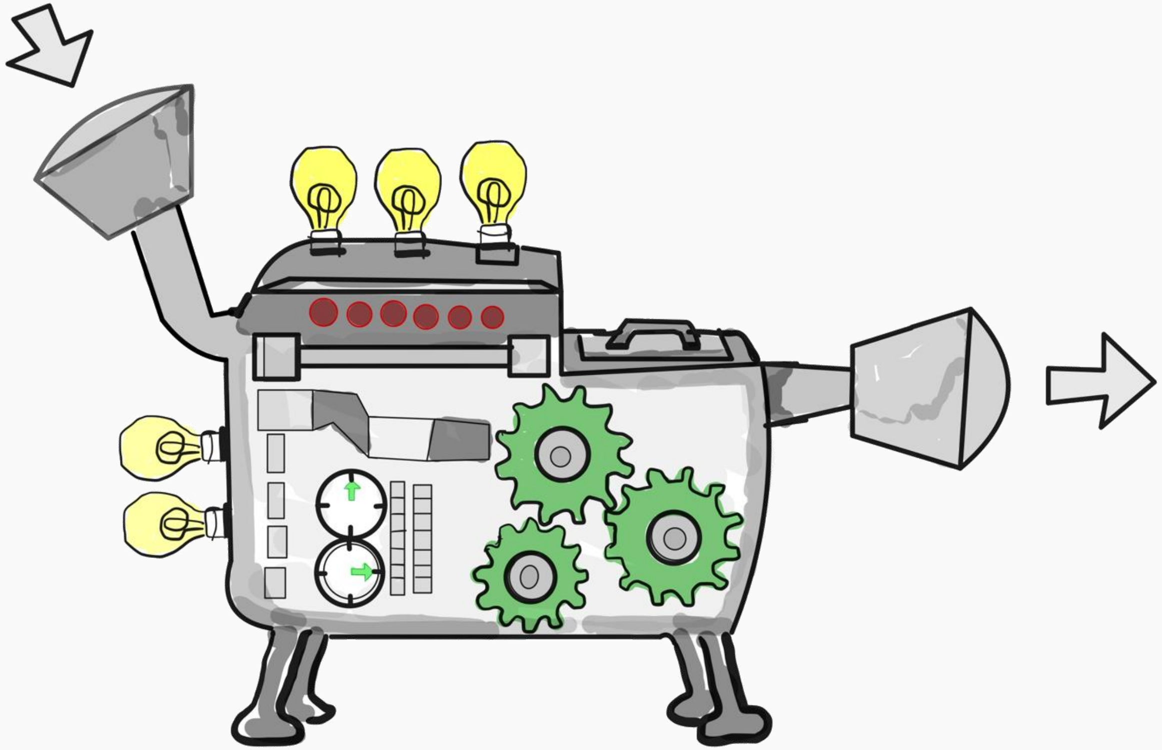
Probability Density Function or Probability Mass Functions

Point estimates

Sample

$[13.75, 13.65, 12.93, 12.81, 12.26]$

MAX



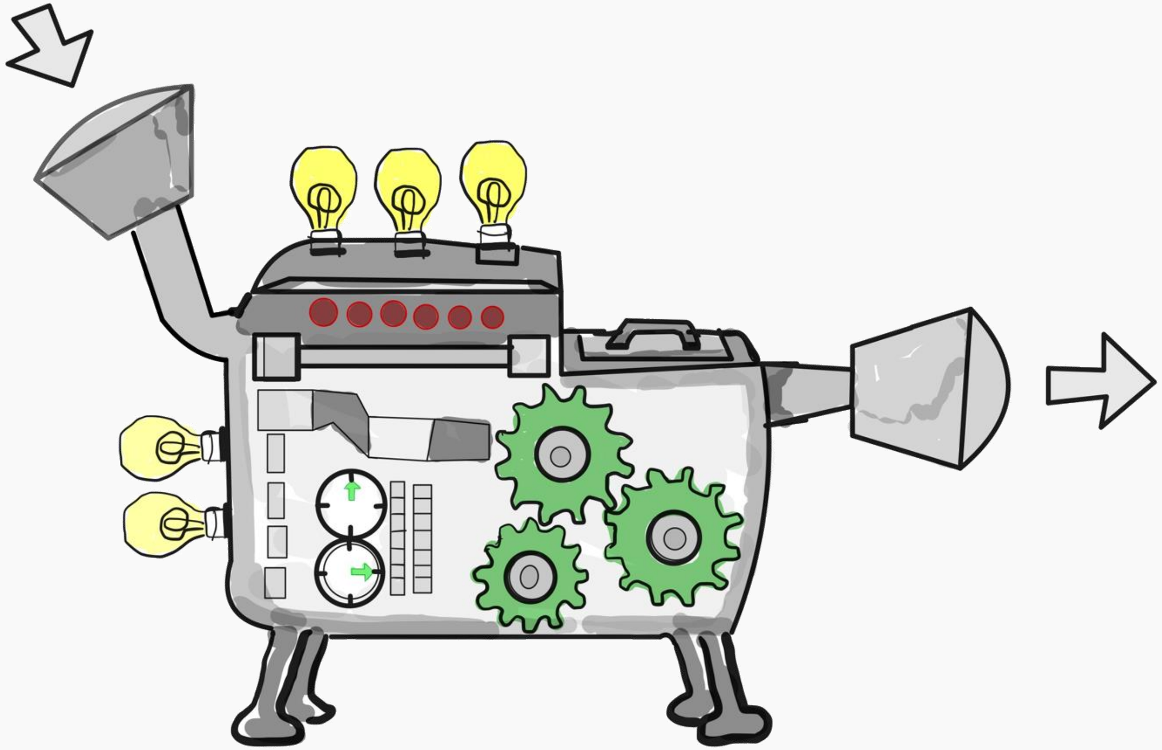
13.75

Point estimates

MIN

Sample

[13.75, 13.65, 12.93, 12.81, 12.26]



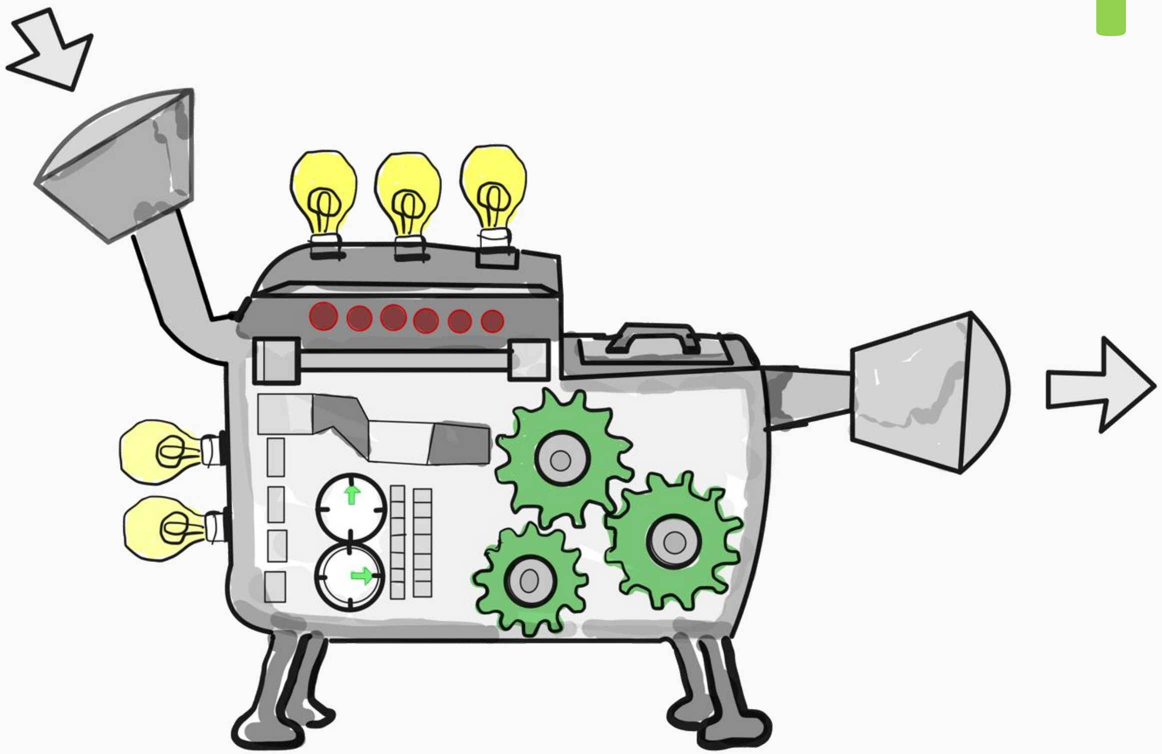
12.26

Point estimates

Sample

$[13.75, 13.65, 12.93, 12.81, 12.26]$

mean



13.08

What we want

POINT ESTIMATES

- Point estimates can be defined as numbers that give some information of the random variable.
- Commonly used point estimat include max, min, mean, medi mode, variance, interquartile range, etc.
- Two major categories describe the central tendency & the spread.

Population Parameters		Sample Statistics
Mean	μ	$\bar{X}$
		s^2
Standard Deviation	σ	

Remember! Population estimates are in Greek and sample estimates are written in roman style!

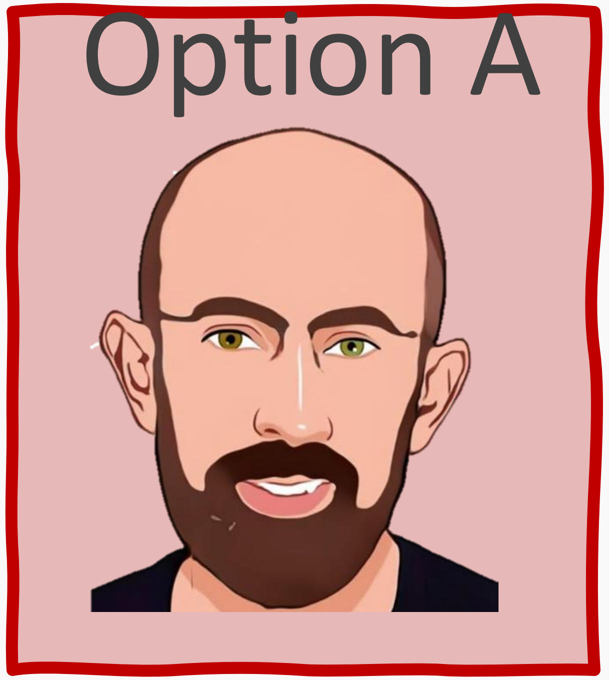
Random Variables - Point estimates

Option A

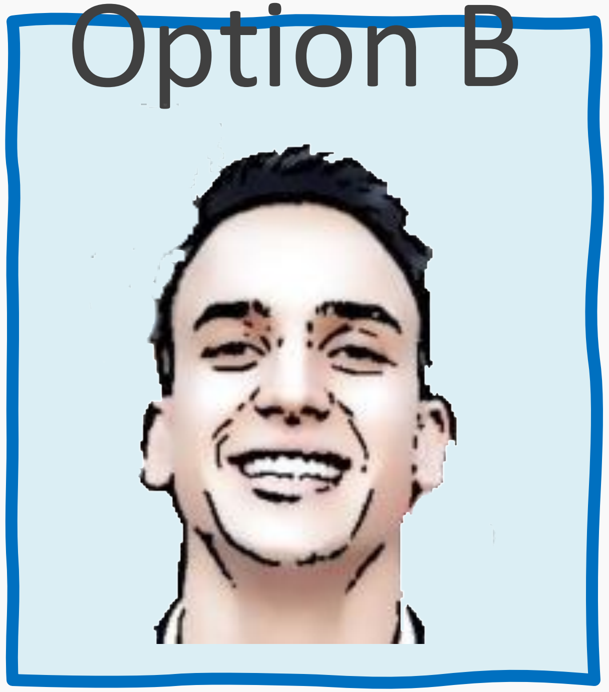
V/S

Option B

Random Variables Comparison - Point estimates



13.31	MAX	14.78
12.67	MIN	11.31
13.00	MEAN	13.00
13.01	MEDIAN	13.00
12.67	MODE	12.25



Measure of Central Tendency

Central Tendency

- Mean is the same as the ‘average’ that we are used to. If we know all the outcomes of the population:

$$\text{Population mean}, \mu = \frac{\sum_i x_i}{n}$$

- Sample statistics are calculated in a manner which best approximates the population parameters.
- Sample mean is calculated like population mean:

$$\text{Sample mean}, \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measure of Spread

Spread

- Standard deviation is a measure of how spread out the data is from the mean. Assuming all of population is known:

$$\text{Population std}(\sigma): \sqrt{\frac{\Sigma(x_i - \mu)^2}{n}}$$

- Sample std has a slight correction te to population std:

Sample std is used as an estimate for the population std, using n-1 gives more accurate results.

$$\text{Sample std}(s): \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}$$

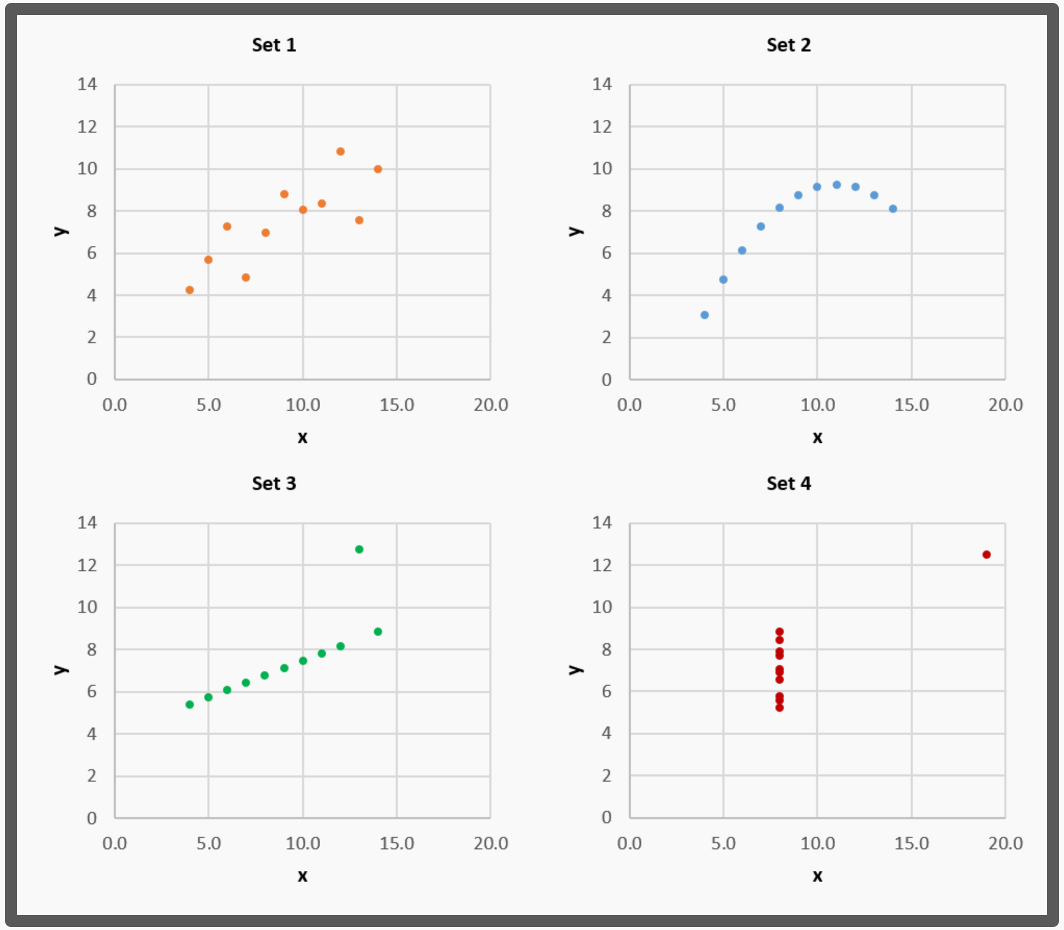
More on this at this link

Confidence Interval

Confidence Intervals

- Point estimates can often be misleading and lead to imprecise understanding of the random variable.

Anscombe's Quartet

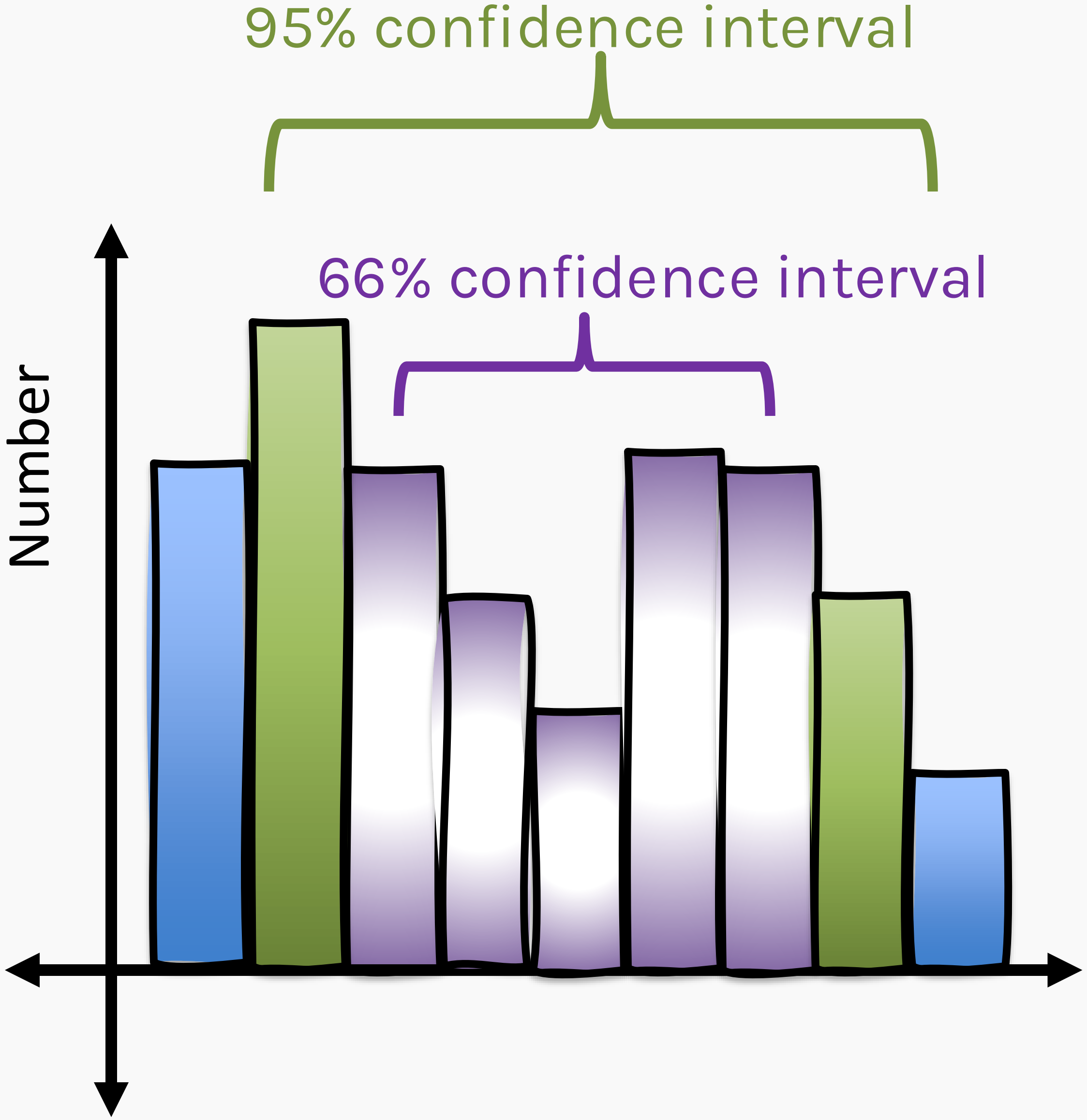


Confidence Intervals

- Point estimates can often be misleading and lead to imprecise understanding of the random variable.

- Unlike point estimates, a confidence interval is a range that represents the likely output of a random experiment.

- We often set the confidence level before examining the data and it is expressed as **%**, e.g., 95% confidence



Confidence Intervals

[13.75, 15.21, 13.65, 13.58, 12.93, 14.23, 12.81, 11.50,13.09,12.26, …]

Step #1: Sort the original data from lowest to highest

Confidence Intervals

[13.75, 15.21, 13.65, 13.58, 12.93, 14.23, 12.81, 11.50,13.09,12.26, …]

Step #1: Sort the original data from lowest to highest

[11.50, 12.26 12.81, 12.93, 13.09, 13.58,13.65,13.75, 14.23, 15.21,, …]

Step #2: Find the lower confidence range using np.percentile

Confidence Intervals

[13.75, 15.21, 13.65, 13.58, 12.93, 14.23, 12.81, 11.50,13.09,12.26, ...]

Step #1: Sort the original data from lowest to highest

[11.50, 12.26 12.81, 12.93, 13.09, 13.58,13.65,13.75, 14.23, 15.21,, ...]

Step #2: Find the lower confidence range using np.percentile

```
np.percentile( [11.50, 12.26 12.81,12.93,13.09,13.58, 13.65, 13.75,14.23, 15.21, , ...  ],2.5)  = 12.80
```

[11.50, 12.26 12.81, 12.93, 13.09, 13.58,13.65,13.75, 14.23, 15.21,, ...]

2.5% of data are on the left of this value

Confidence Intervals

[11.50, 12.26 12.81, 12.93, 13.09, 13.58,13.65,13.75, 14.23, 15.21,, …]

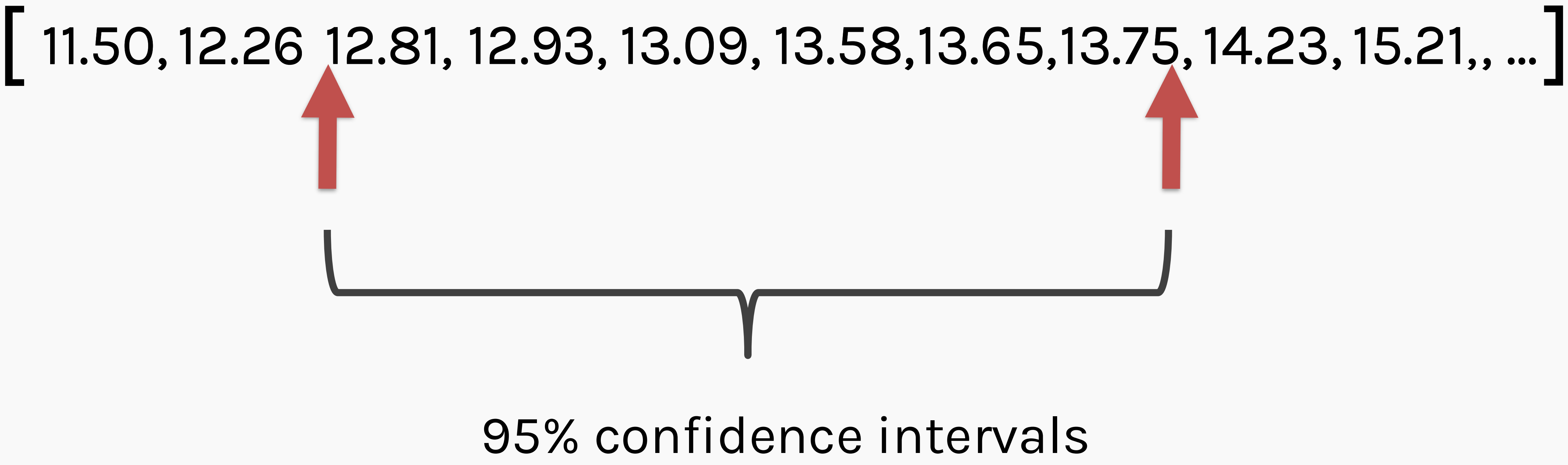
Step #3: Find the upper confidence range again using np.percentile

`np.percentile(` [11.50, 12.26 12.81,12.93,13.09,13.58, 13.65, 13.75,14.23, 15.21, , …]`,97.5)` = 13.71

[11.50, 12.26 12.81, 12.93, 13.09, 13.58,13.65,13.75, 14.23, 15.21,, …]

```
2.5% of data are on the right of this value
```

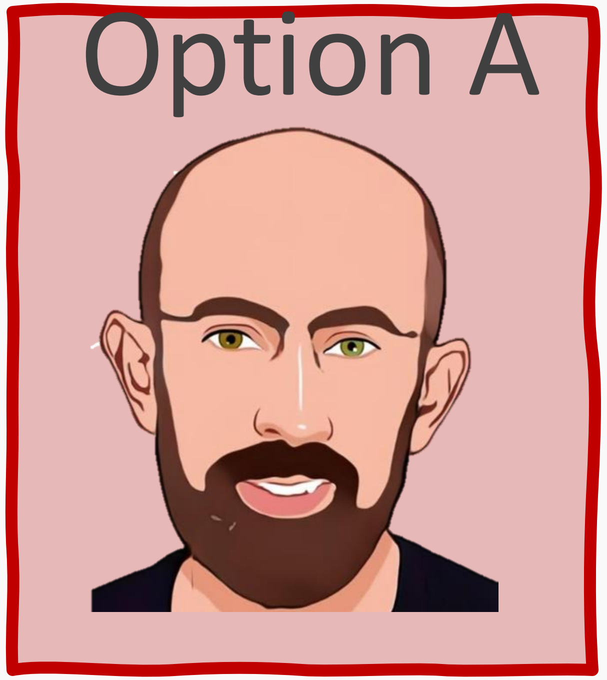
Confidence Intervals



[11.50, 12.26 12.81, 12.93, 13.09, 13.58,13.65,13.75, 14.23, 15.21,, …]

95% confidence intervals

Random Variables - Point estimates

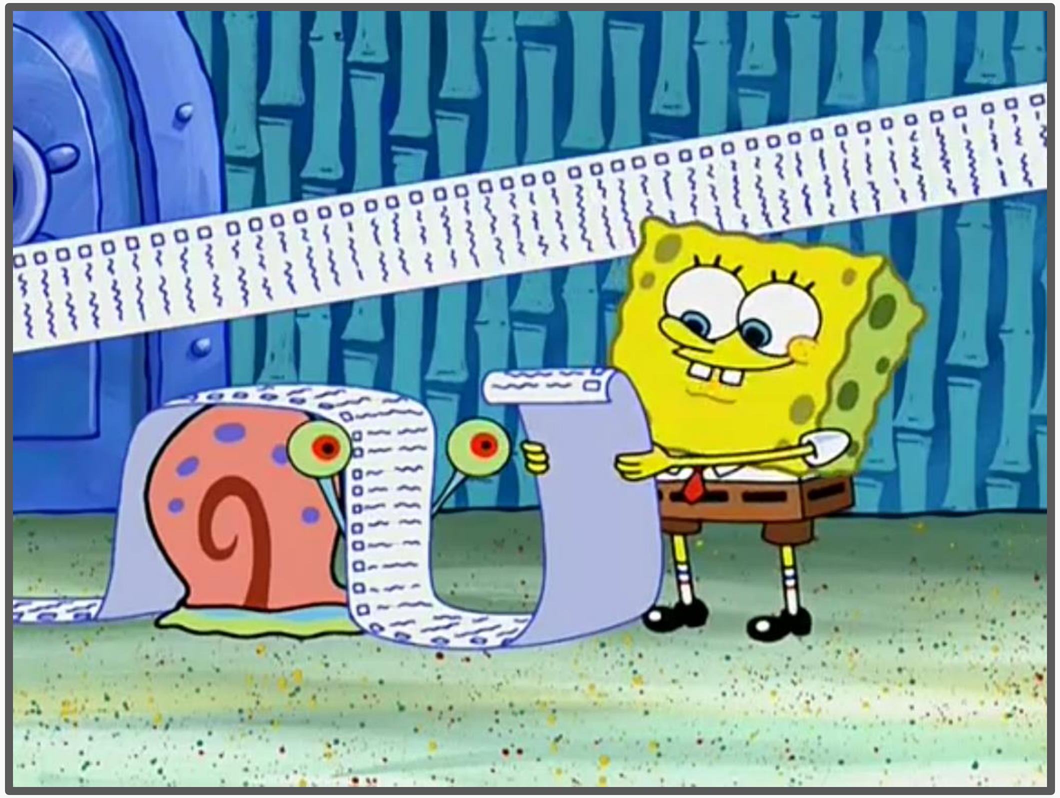


13.31	MAX	14.78
12.67	MIN	11.31
13.00	MEAN	13.00
13.01	MEDIAN	13.00
12.67	MODE	12.25
12.80, 13.20	CI	12.80, 13.20

Properties of a Random Variable

ESTIMATE ISSUES

- Point or interval estimates of random variables do not guarantee a unique description of the output.
- Due to its approximate nature, it may lead to confounding of different processes.
- A popular example of this is the Anscombe's Quartet; a set of four datasets with same estimates but different distributions.



Histogram

Histogram

Sample

$$\left[13.75, 13.65, 12.93, 12.81, 12.26\right]$$

Histogram

13.75, 13.75,

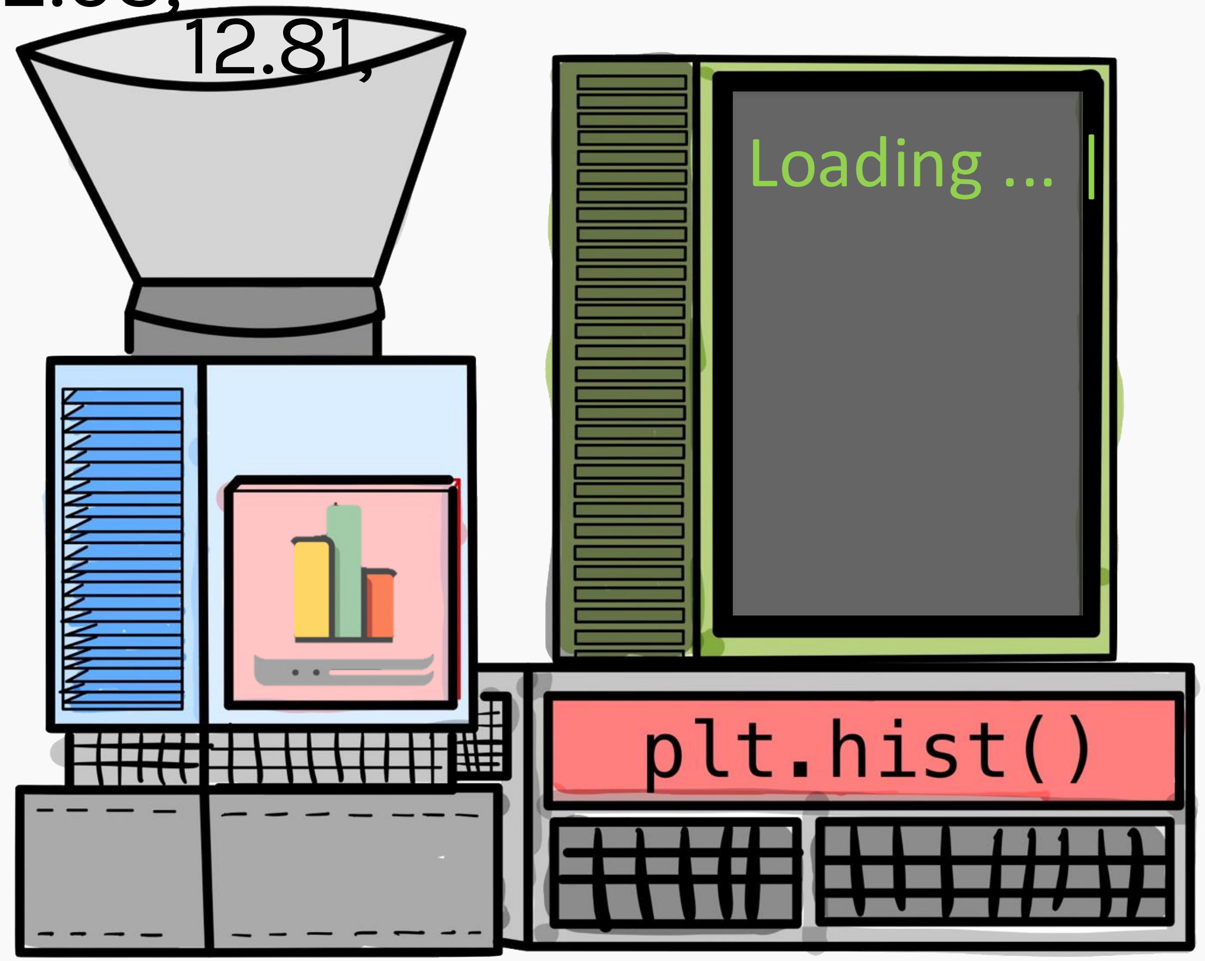
12.81,

13.75, 13.65,

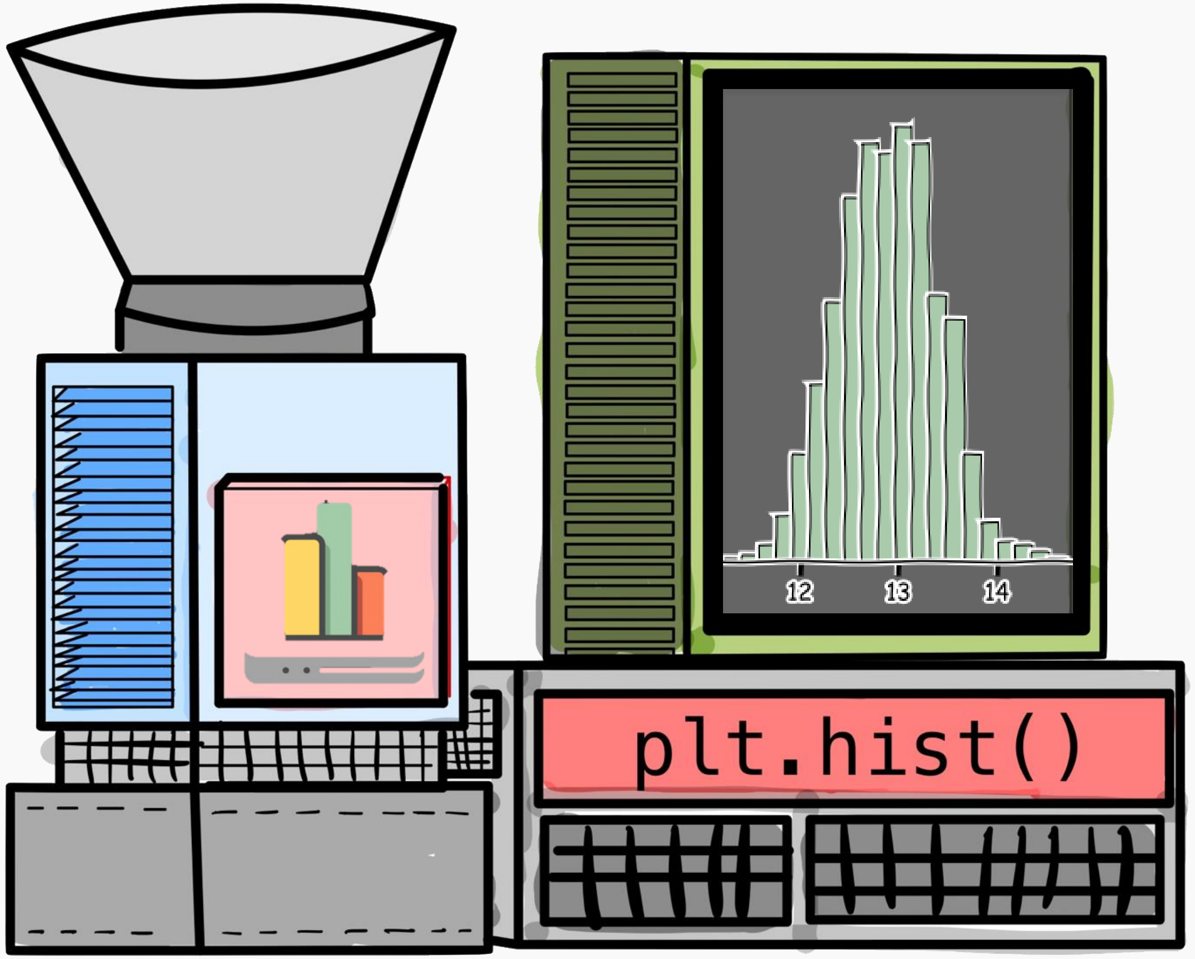
12.26

12.93,

12.81,



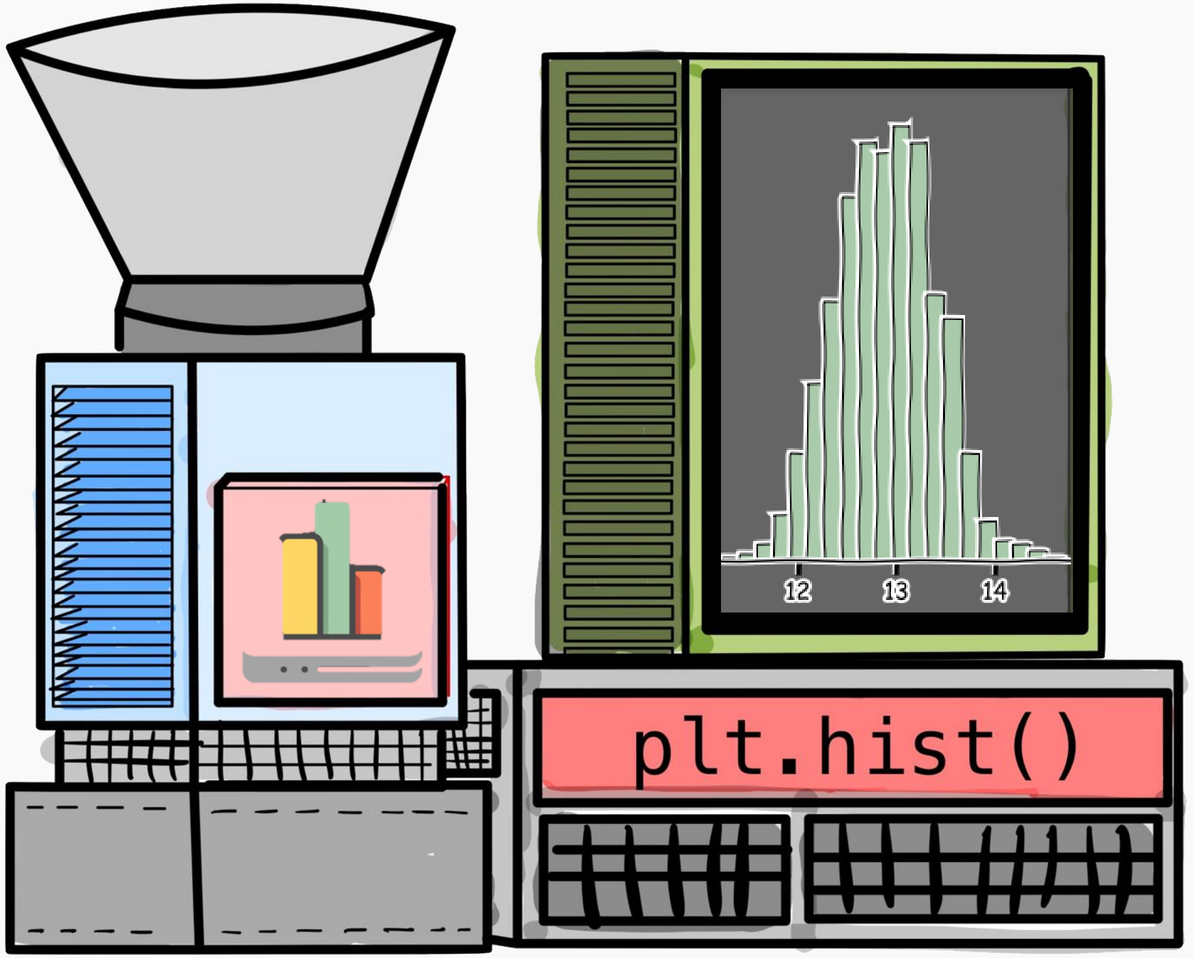
Histogram



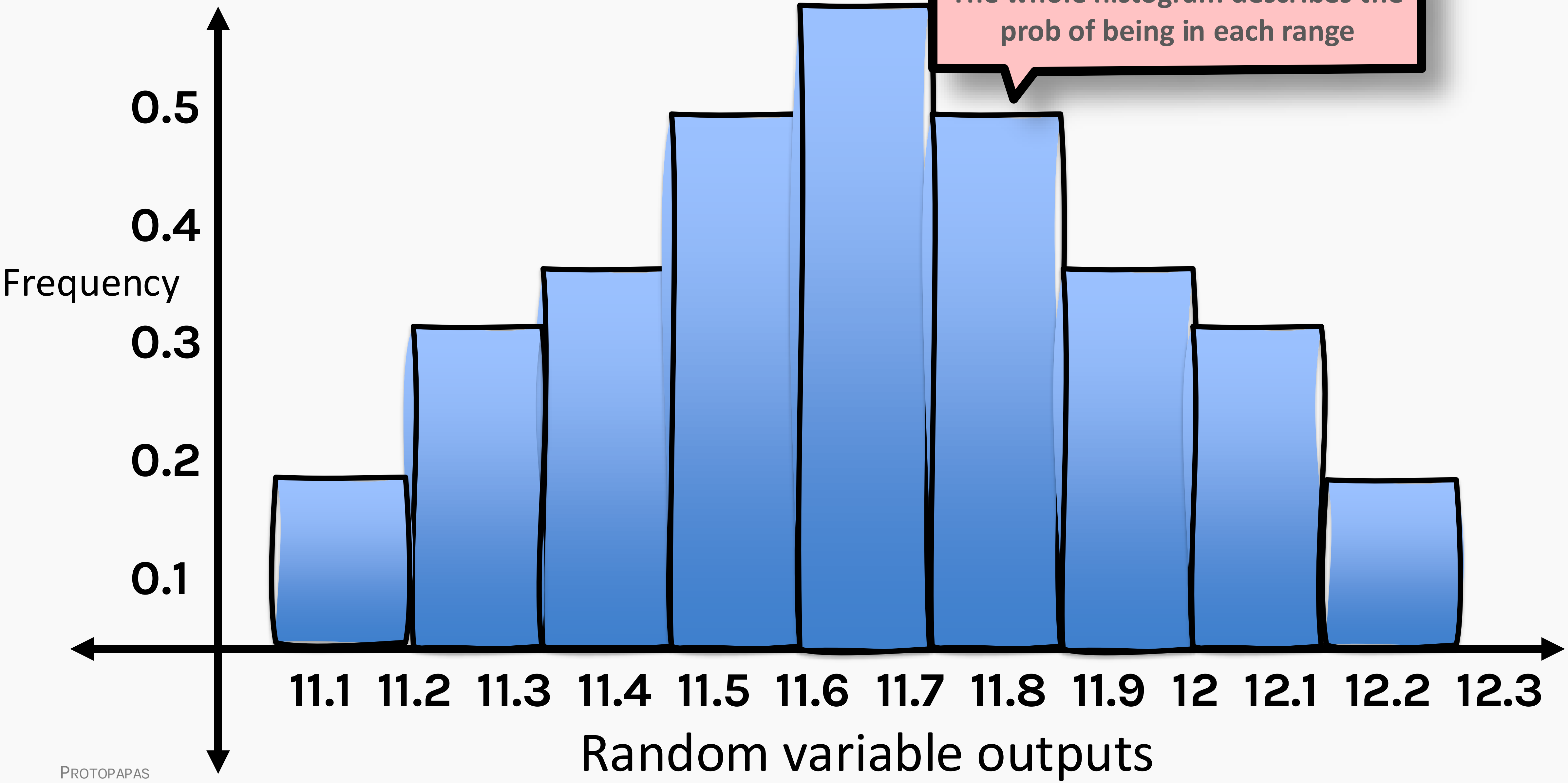
Histogram

DISTRIBUTIONS

- Histogram (from the Greek word *histos* meaning pole & *gram* meaning chart) is a visual representation of the sample.
- It is defined by the relative frequency on the y-axis and the outcomes of the random variable on the x-axis.
- It can be decorated with point estimates for better description.

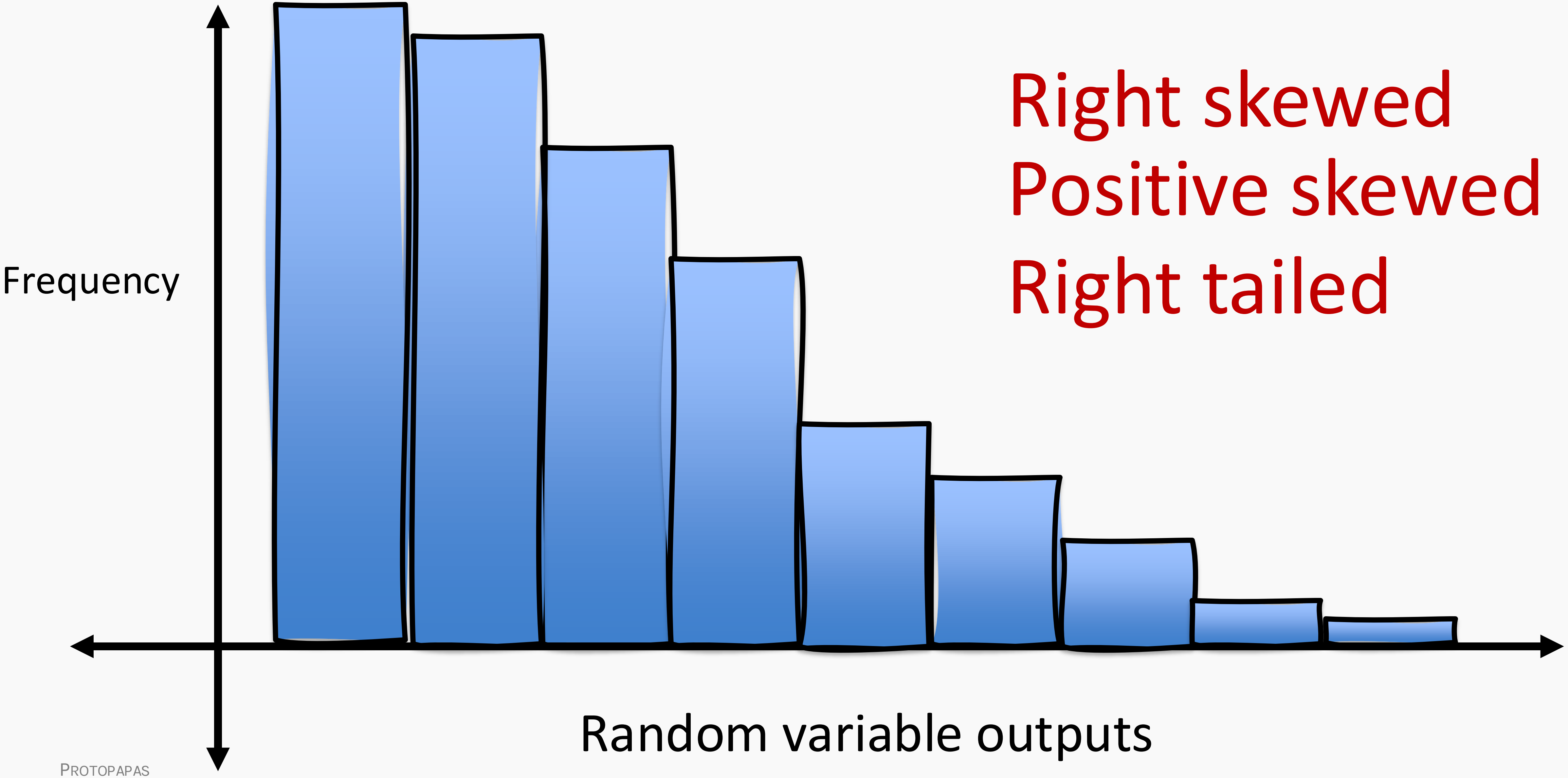


Anatomy of a histogram



This bin describes the probability of the variable to be in this range. The whole histogram describes the prob of being in each range

Anatomy of a histogram

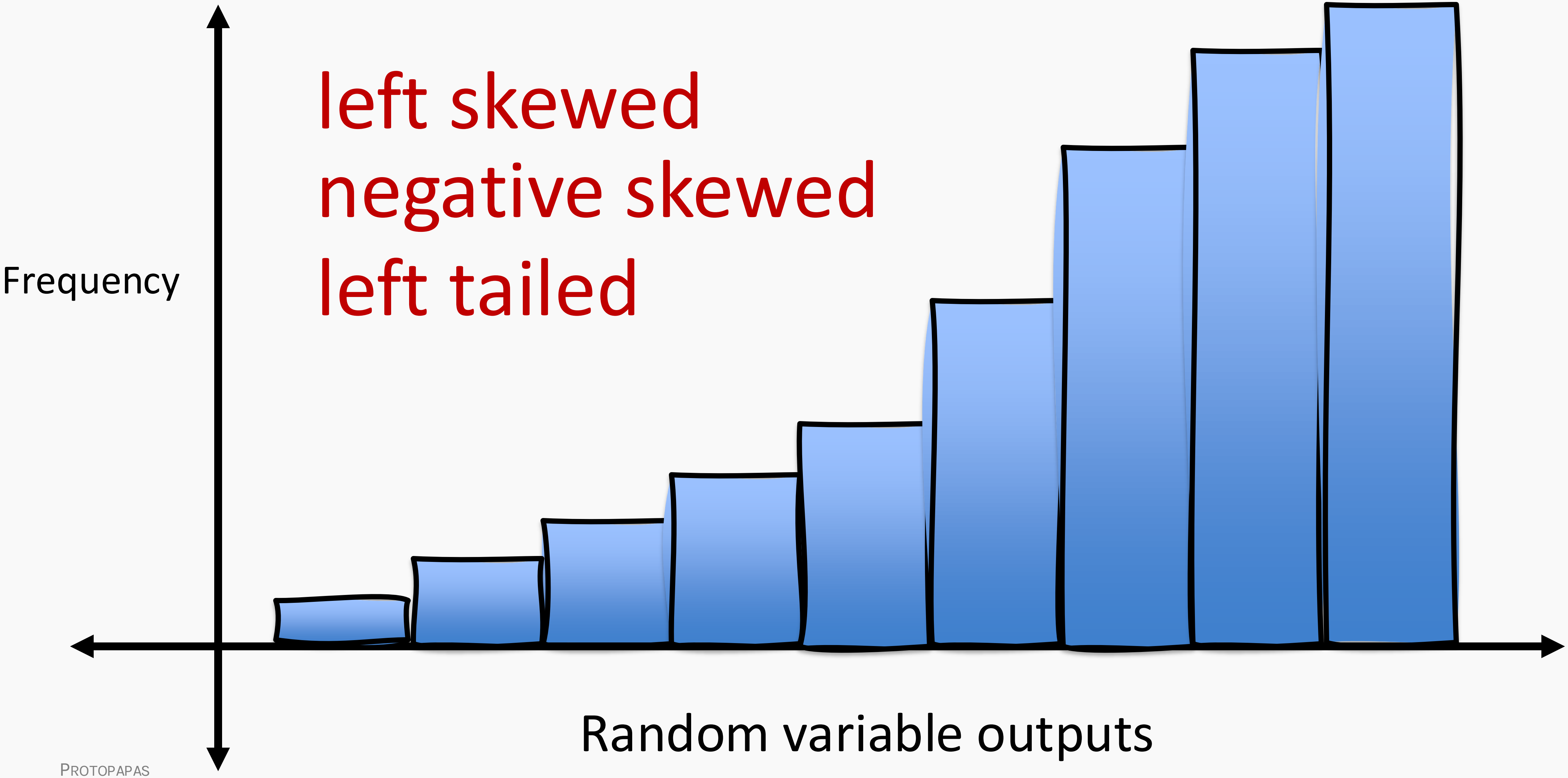


Right skewed

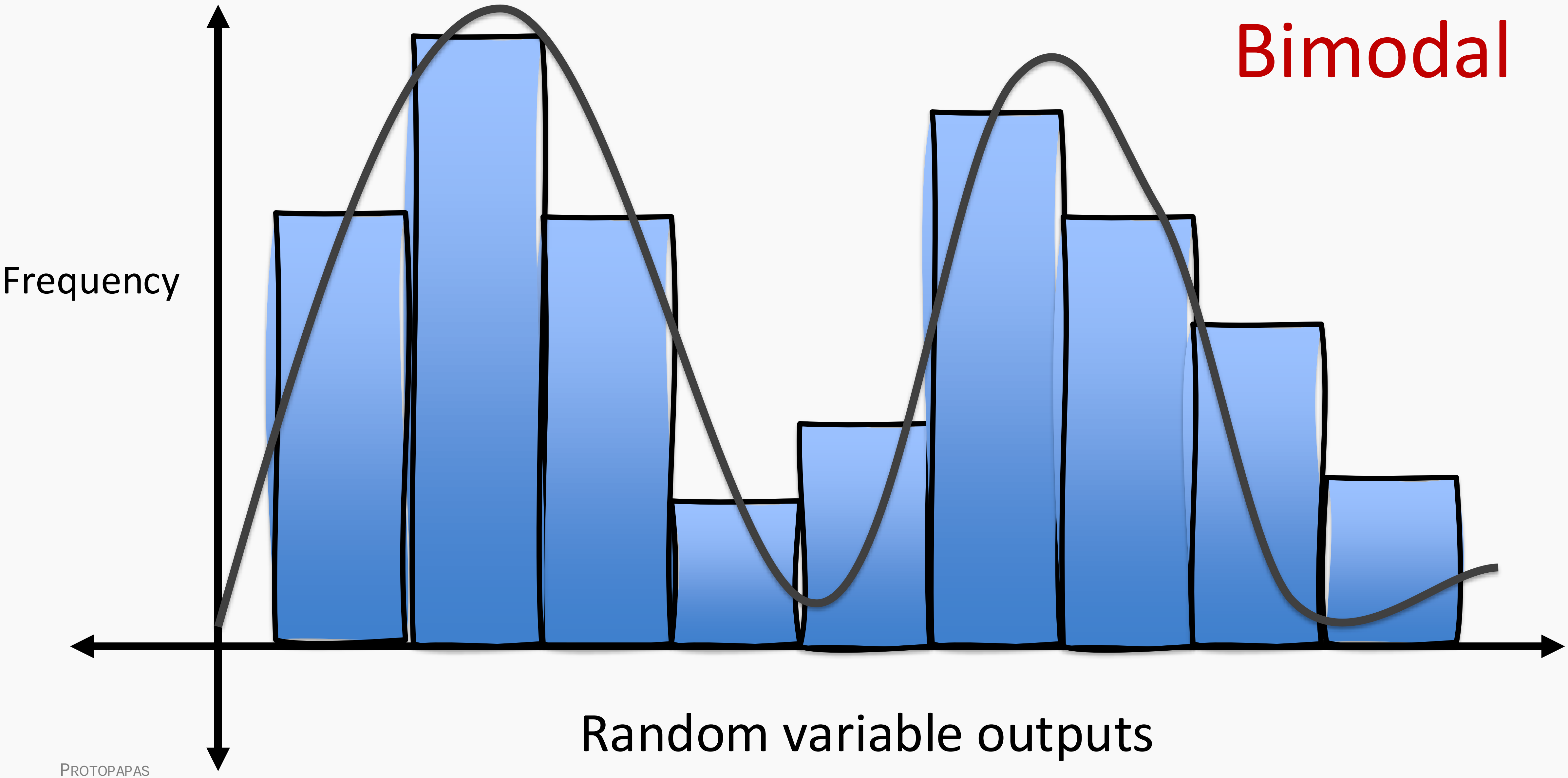
Positive skewed

Right tailed

Anatomy of a histogram



Anatomy of a histogram

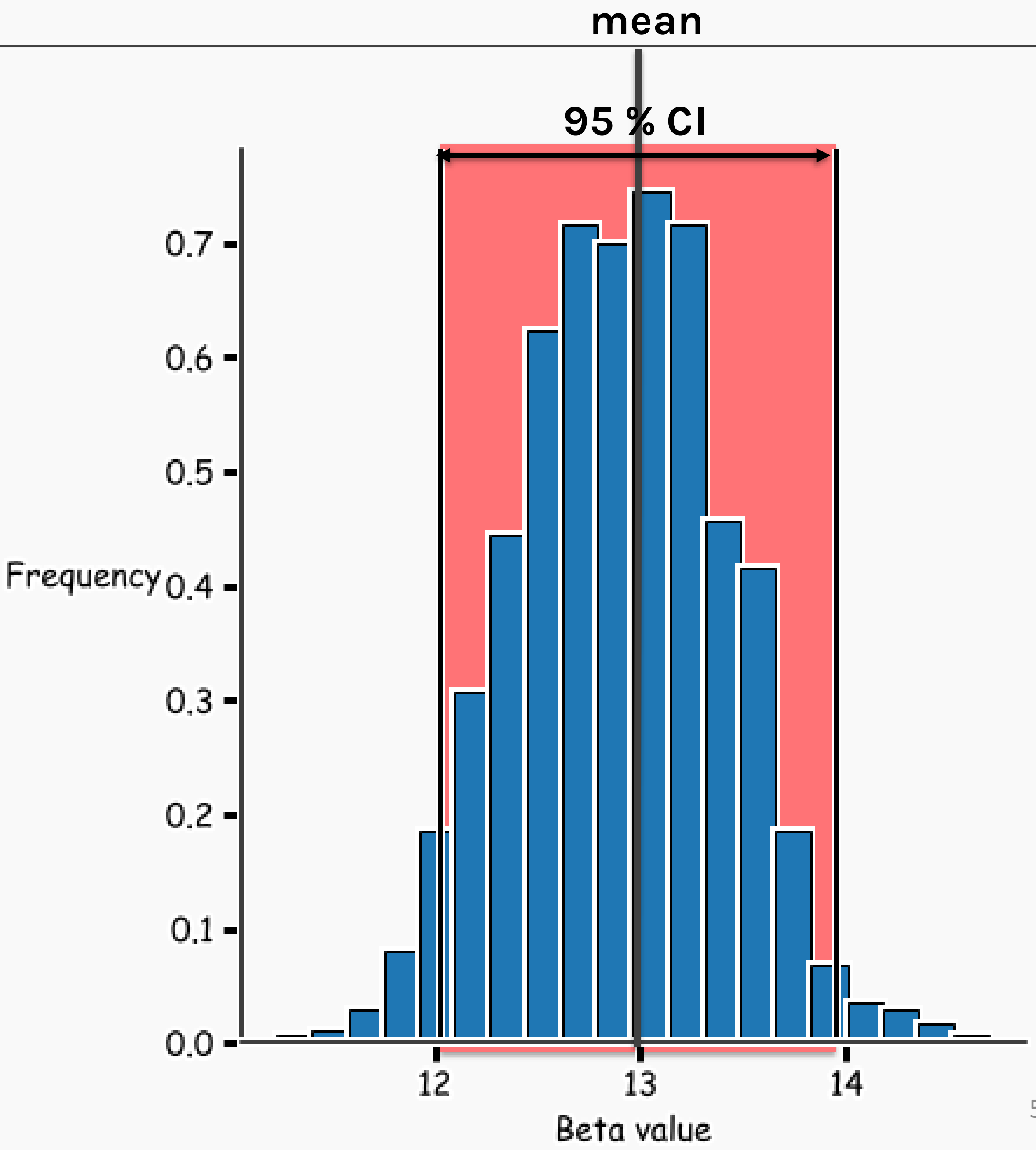


Histogram

ISSUES?

The outcomes of a random variable captured over multiple simulations is difficult to interpret and consequently difficult to compare to other random variables.

- ~~**ISSUE #1**: We do not have estimates to compare with other random variables.~~
- ~~**ISSUE #2**: It is difficult to visualize the spread of the outcome.~~



What is probability?

What is probability?

Q: What is probability?

A: A common definition: the long-run, relative frequency* of a random phenomenon/experiment/event.

Q: What values can probabilities take on?

A: Any value between 0 and 1 (including the endpoints).

Q: Why do we care?

A: Because data can be thought of as random realizations of a *data generating process* (whether though sampling or a theoretical construct).

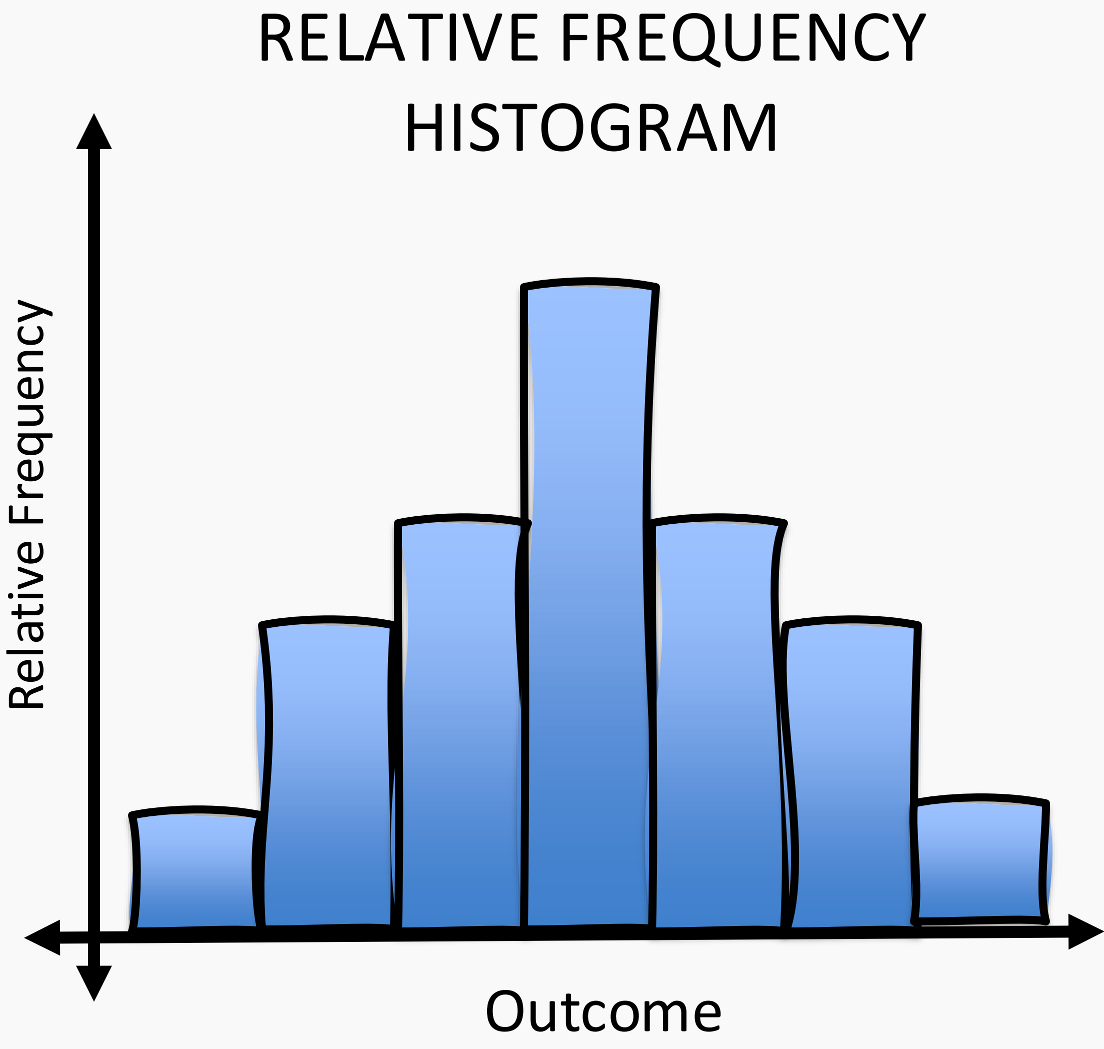
*Note: this ignores the Bayesian definition of probability: a measure of belief.

Known Distributions

Histogram as a Probability Density Function (PDF)

- Recall that a histogram describes the probability of being in a given range (relative frequency of the "bins").
- We can describe the probability of being a range with a function.
- This function could be defined for either discrete or continuous variables:
  - In case of continuous, this is the **probability density function (PDF)** for values in a range. This is usually written as $f(x)$.
  - In case of discrete, the range is the discrete value itself. The function is the **probability mass function (PMF)**, denoted as $P(X = x)$ or $p(x)$.

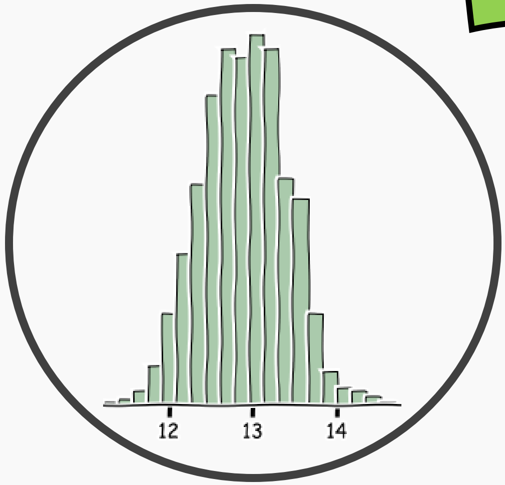
*Note: probabilities for a continuous random variable can be represented as areas under the curve, and thus $P(X = x) = 0$ since there is no width.



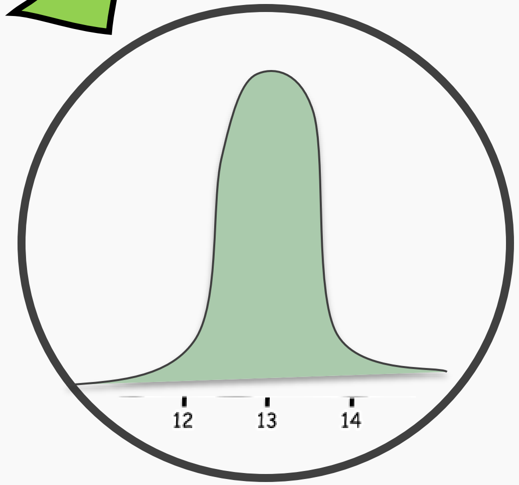
PDF vs PMF

Random Variable

X



Discrete Random Variable

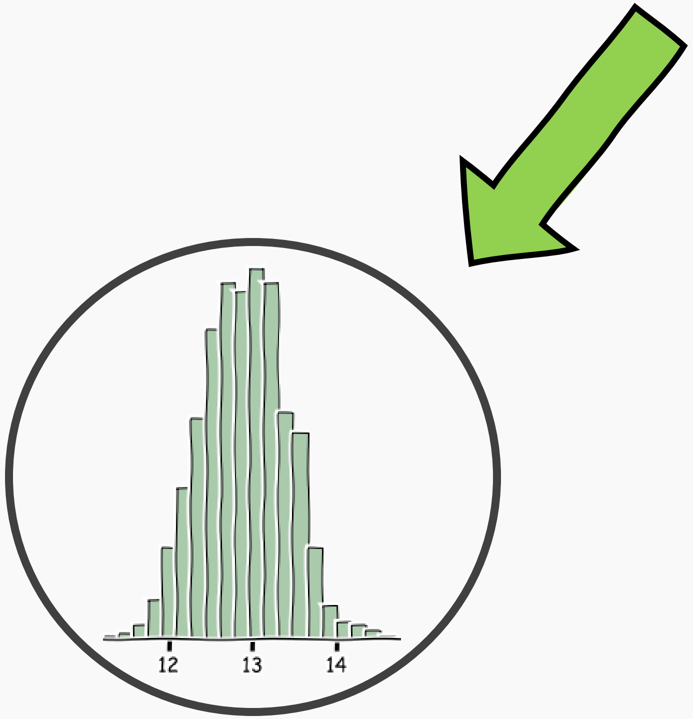


Continuous Random Variable

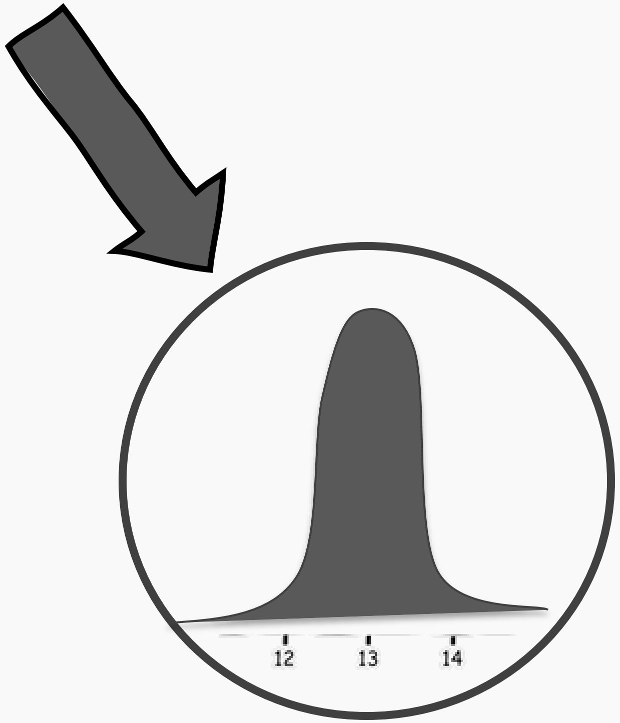
PDF vs PMF

Random Variable

X



Discrete Random Variable



Continuous Random Variable

Discrete Uniform Distribution

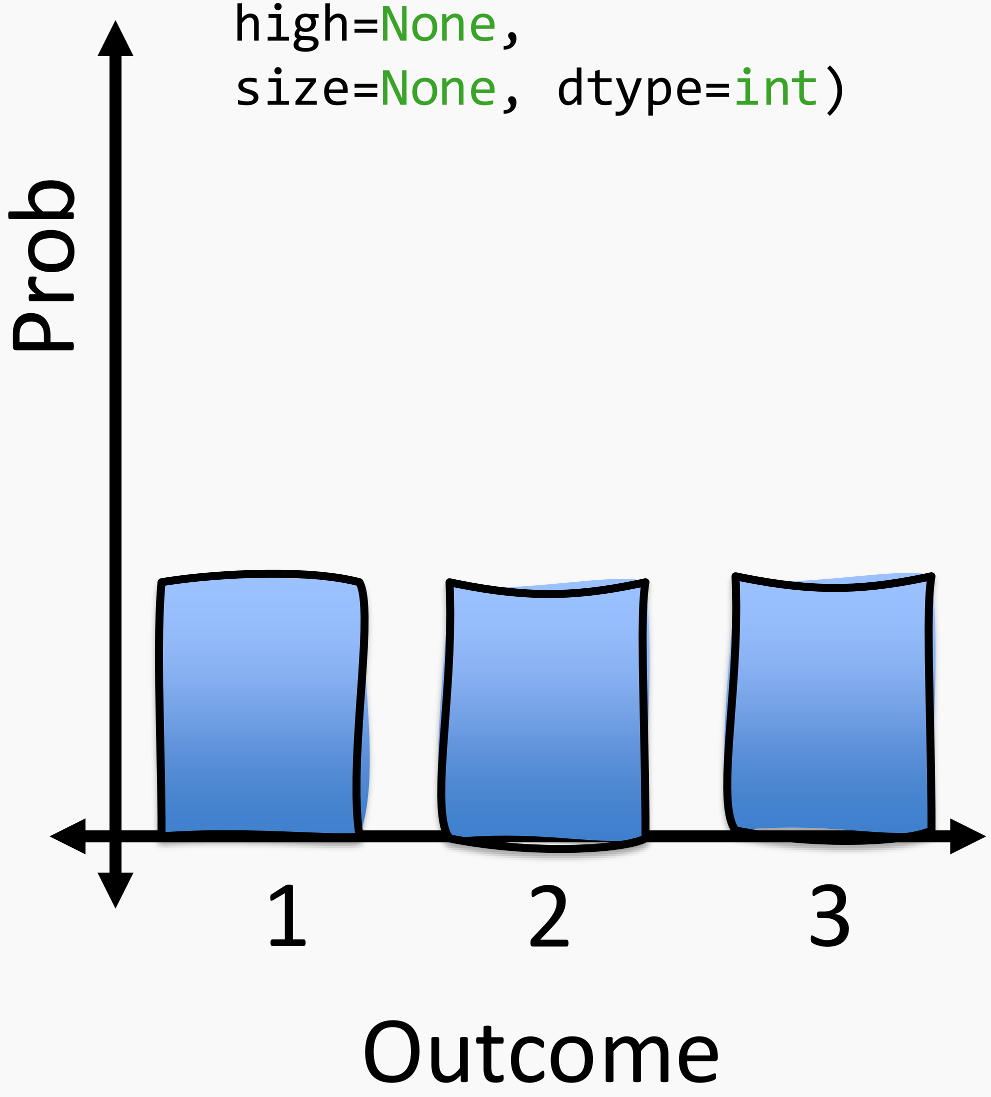
- This distribution occurs when there are a finite number of equally likely outcomes possible.

PMF: $P(X = 1) = \frac{1}{N}$

mean $\mu = \frac{a+b}{2}$, where a and b should be the first and last outcomes in the range

Variance $\sigma^2 = \frac{(b-a+1)^2-1}{12}$

```
np.random.randint(low,
high=None,
size=None, dtype=int)
```



Bernoulli Distribution

- This distribution can be thought of as a model of possible outcomes of an experiment that asks a yes-no question.
- E.g., If you toss a coin, will you get a *head* or a *tail*?

PMF: $P(X = x) = p^x(1-p)^{1-x}$

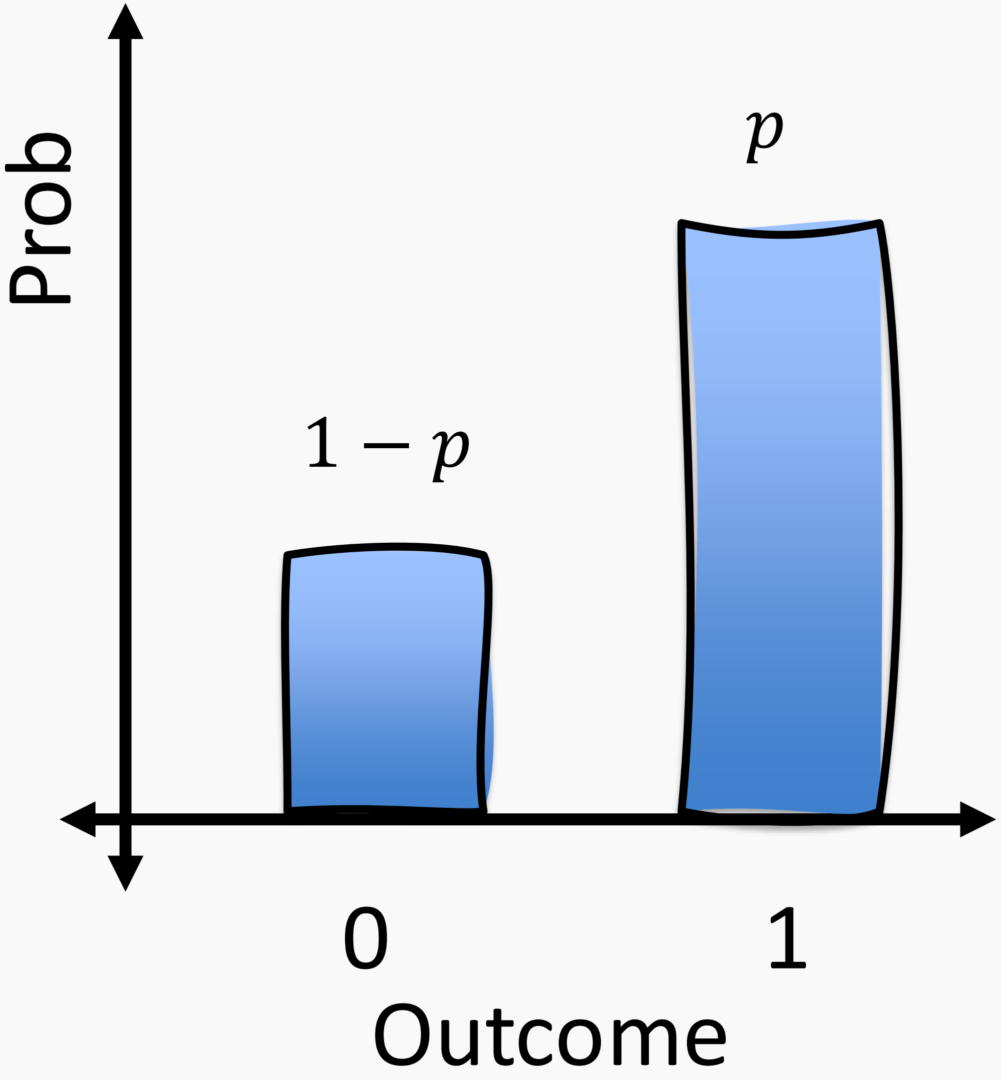
where p is the probability of success and

$q = 1 - p$ is the probability of failure.

mean $\mu = p$

Variance $\sigma^2 = pq$

```
np.random.binomial(1, p, size=None)
```



Binomial Distribution

- A binomial distribution with parameters n and p is the distribution of the number of k successes in a sequence of n independent experiments, each asking a yes-no questions. This is often written as: $X \sim Binom(n, p)$.

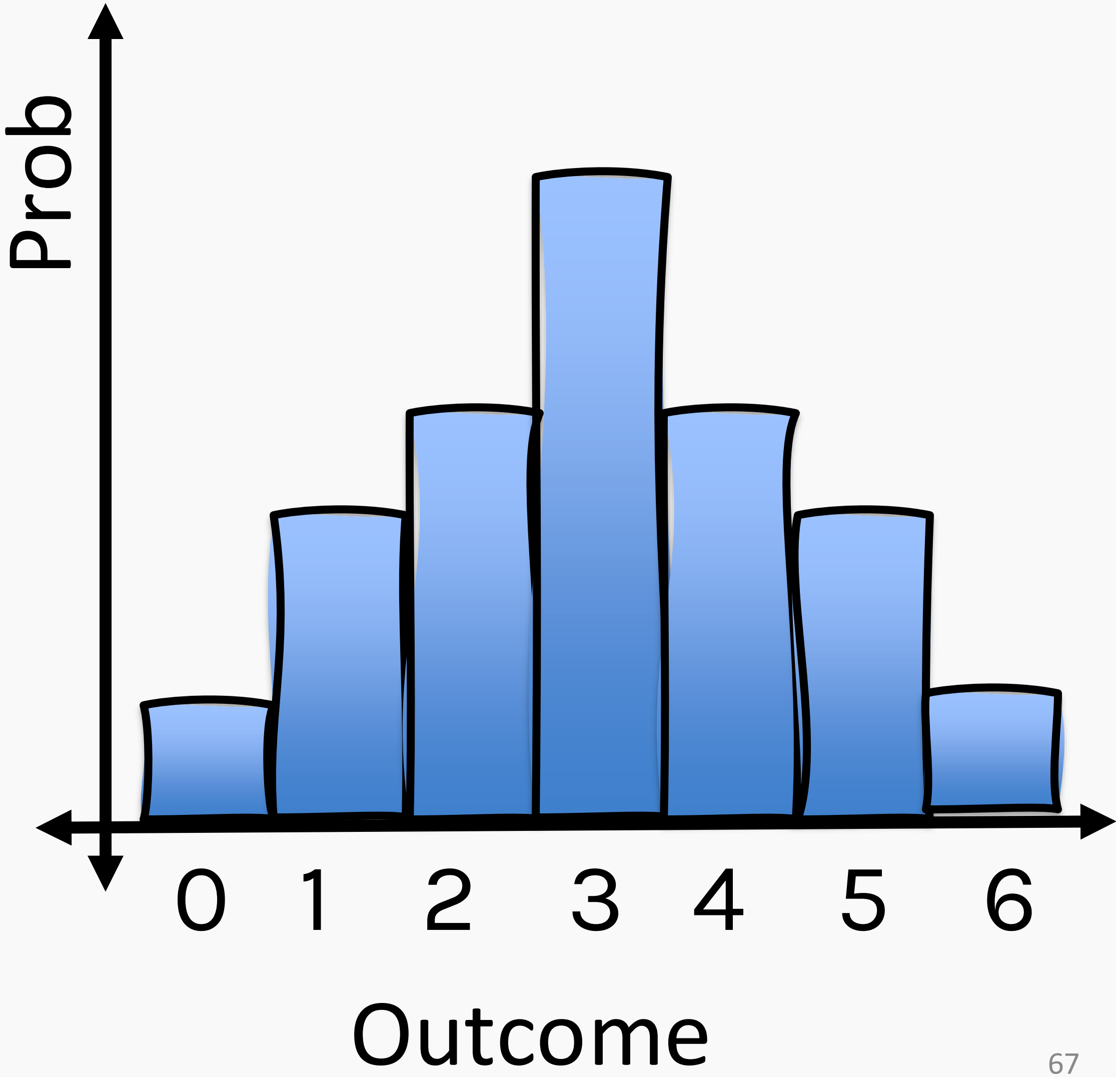
PMF: $P(X = x) = \binom{n}{k} p^k q^{n-k}$

where p is the probability of success and $q = 1 - p$ is the probability of failure.

mean $\mu = np$

Variance $\sigma^2 = npq$

```
np.random.binomial(n, p, size=None)
```



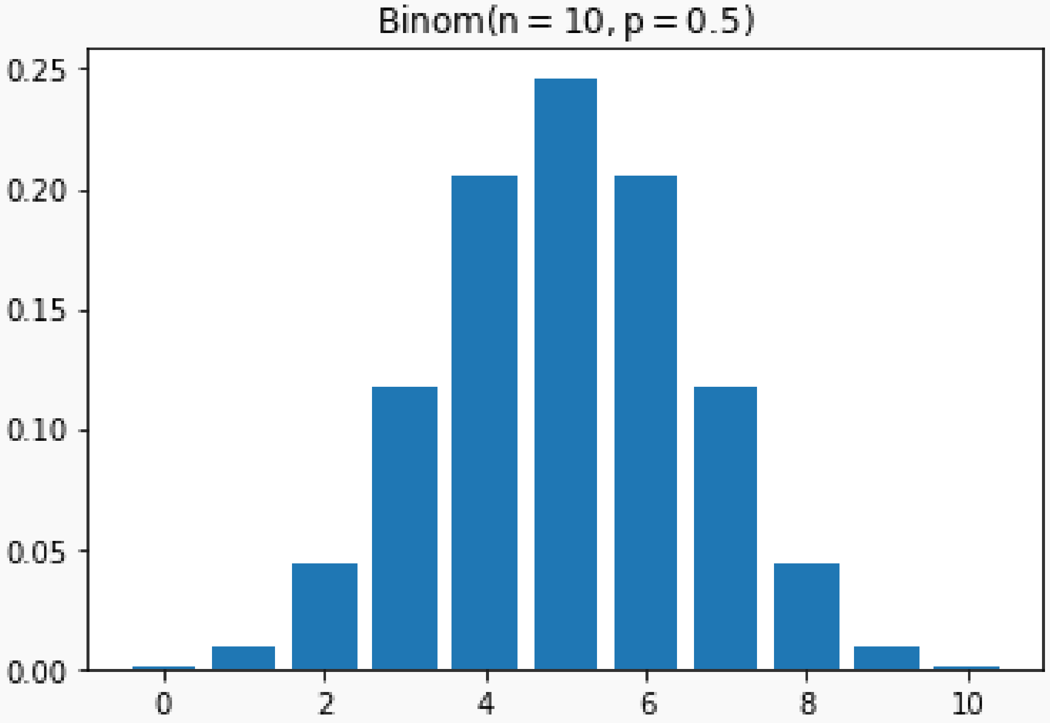
Binomial Distribution

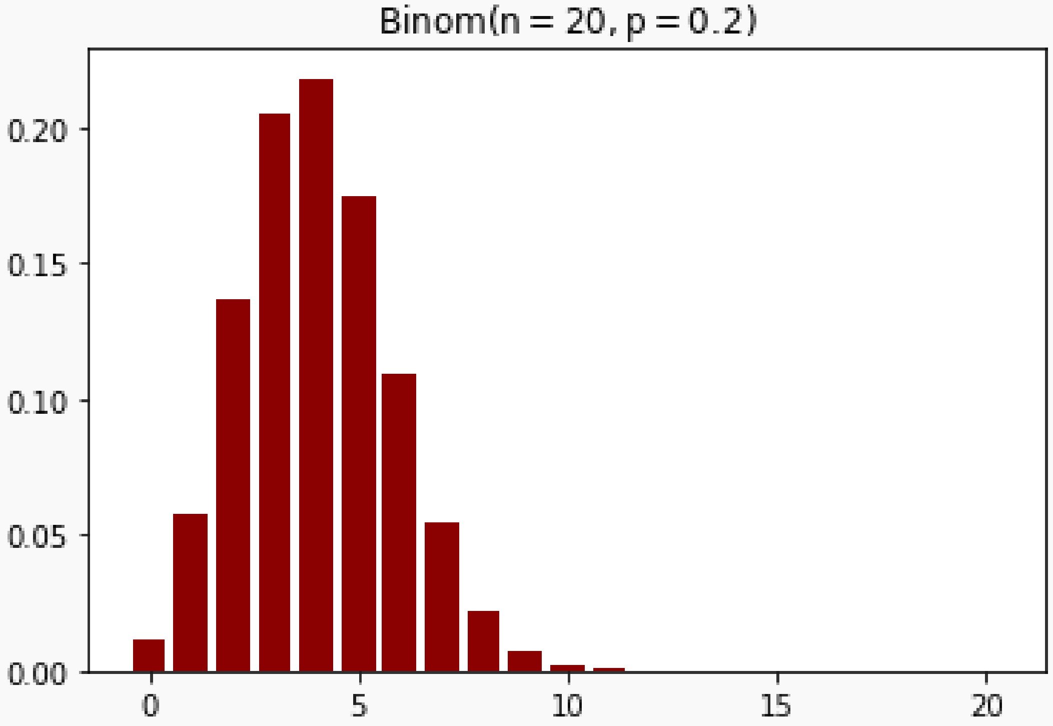
Think counting the number of heads when flipping a biased coin n times.

The binomial distribution is useful to describe polling data (proportion of people who will vote for Biden), survey data (will you take CS1090B next year?), or any data that are binary!

The **Bernoulli distribution** is a special case when n = 1.

Binomial Distribution Examples



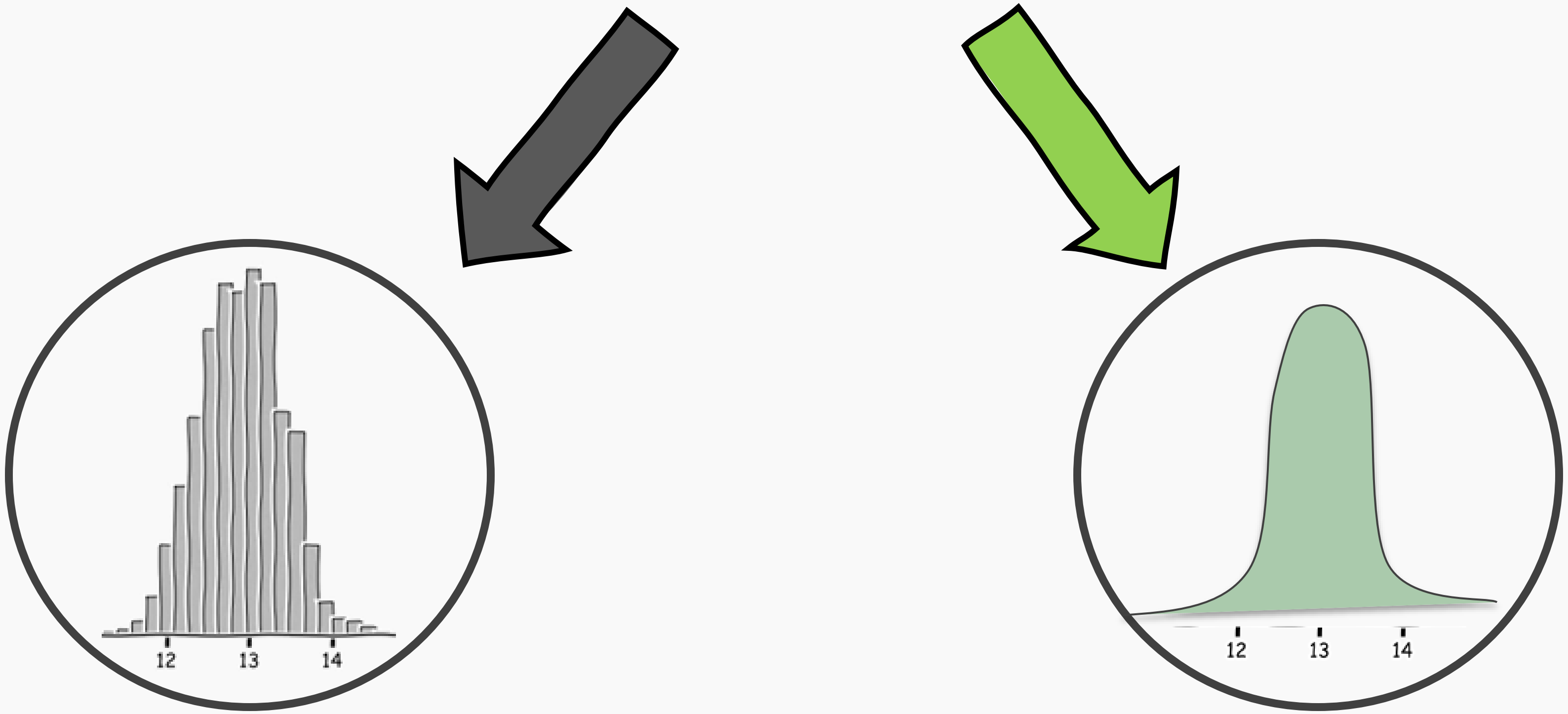


A binomial distribution has mean np and standard deviation $\sqrt{np(1-p)}$.

PDF vs PMF

Random Variable

X



~~Discrete Random Variable~~

Continuous Random Variable

Uniform continuous distribution

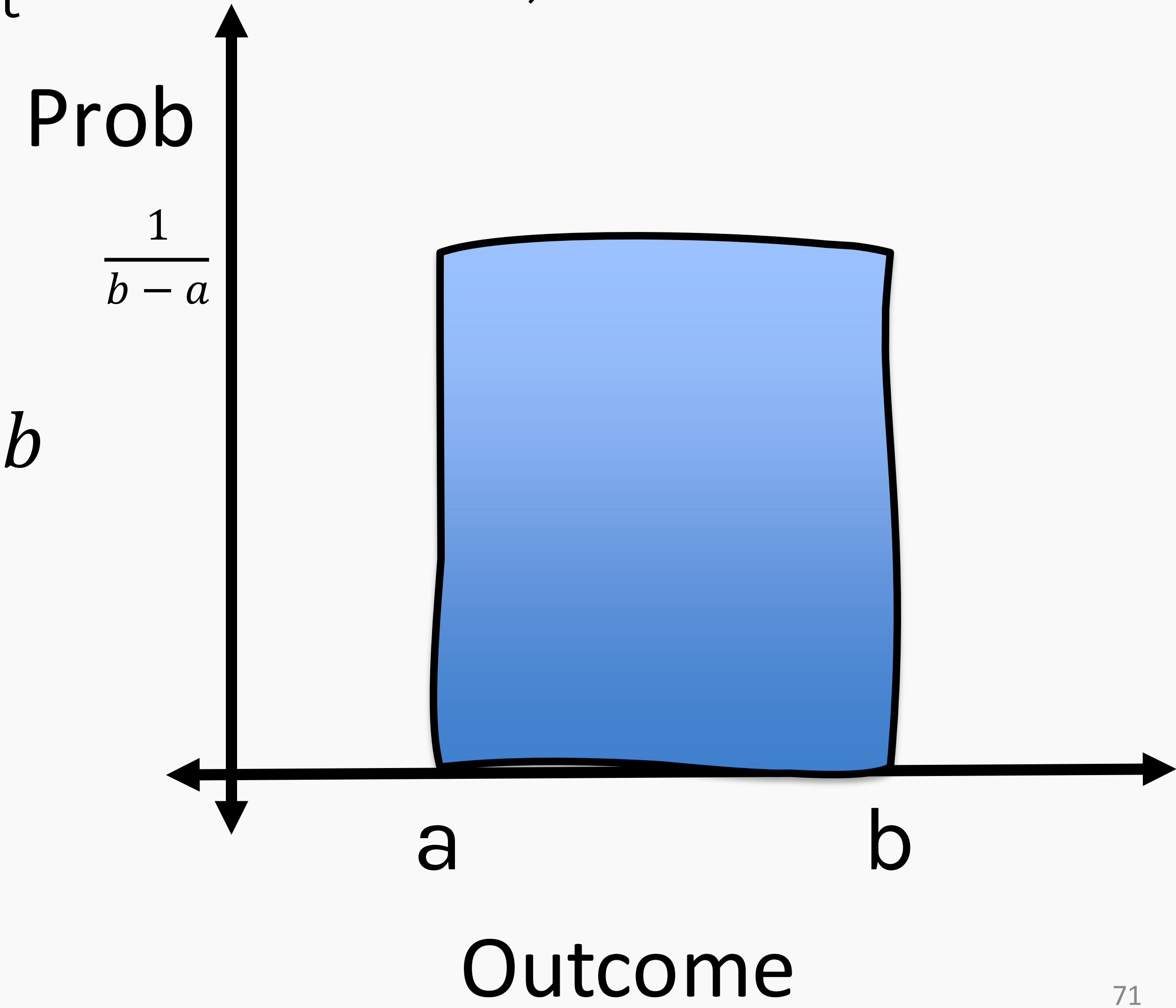
- This distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds, defined by parameters a and b.

```
np.random.uniform(low=0.0, high=1.0,
 size=None)
```

PDF: $P(X = x) = \begin{cases} \frac{1}{b-a} & for\ a \leq x \leq b \\ 0 & \end{cases}$

mean $\mu = \frac{a+b}{2}$

Variance $\sigma^2 = \frac{(b-a)^2}{12}$



Normal distribution

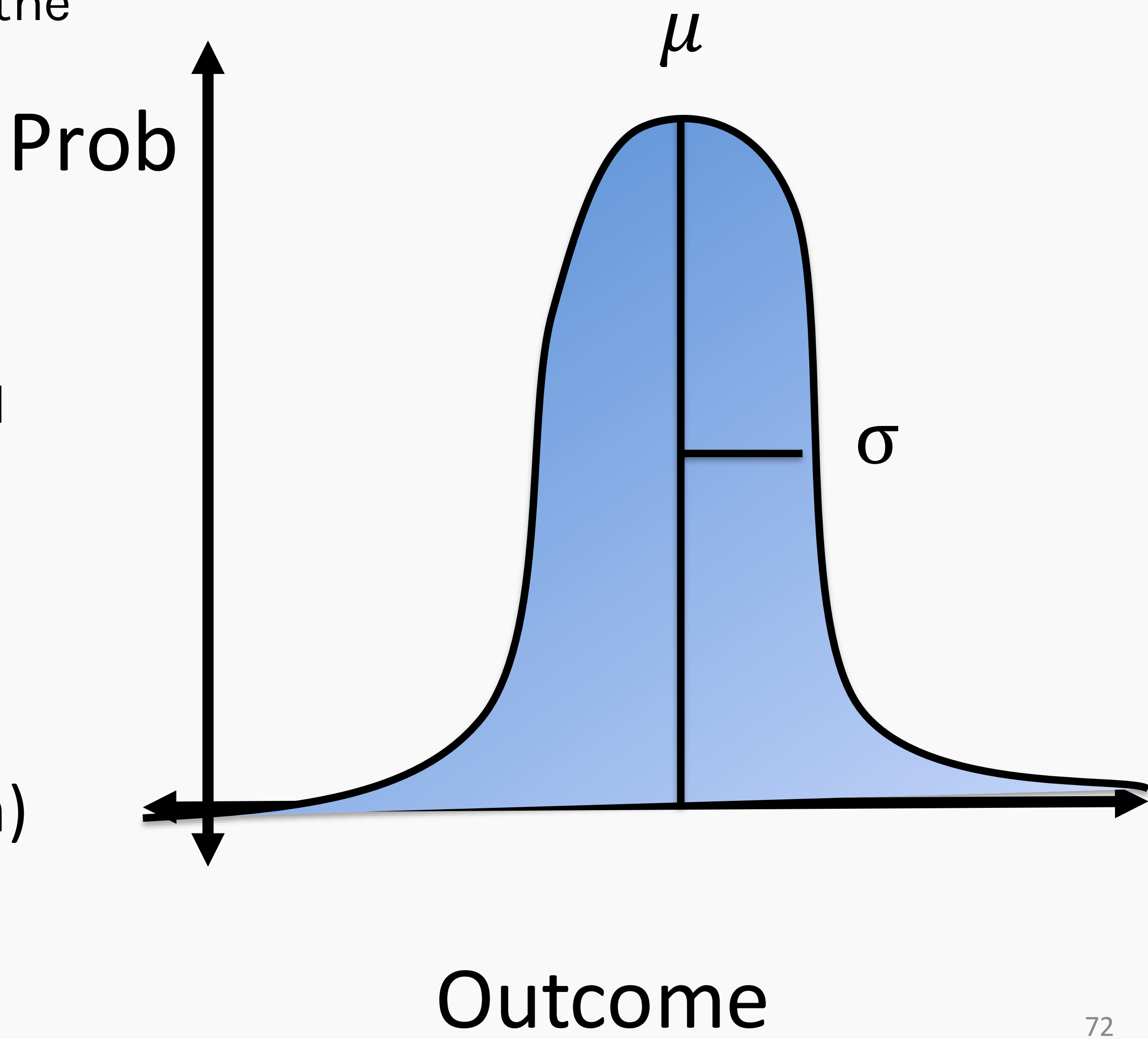
```
np.random.normal(loc=0.0, scale=1.0,
 size=None)
np.random.randn()
```

- A normal (or Gaussian) distribution is one of the most used continuous random variables.
- As a result of the Central Limit Theorem, the distribution of sample means from a sufficiently large sample size approximates a normal distribution, regardless of the original population distribution.

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

mean $E[X] = \mu$ (location)

Variance $E[(X-\mu)^2] = \sigma^2$ (scale)



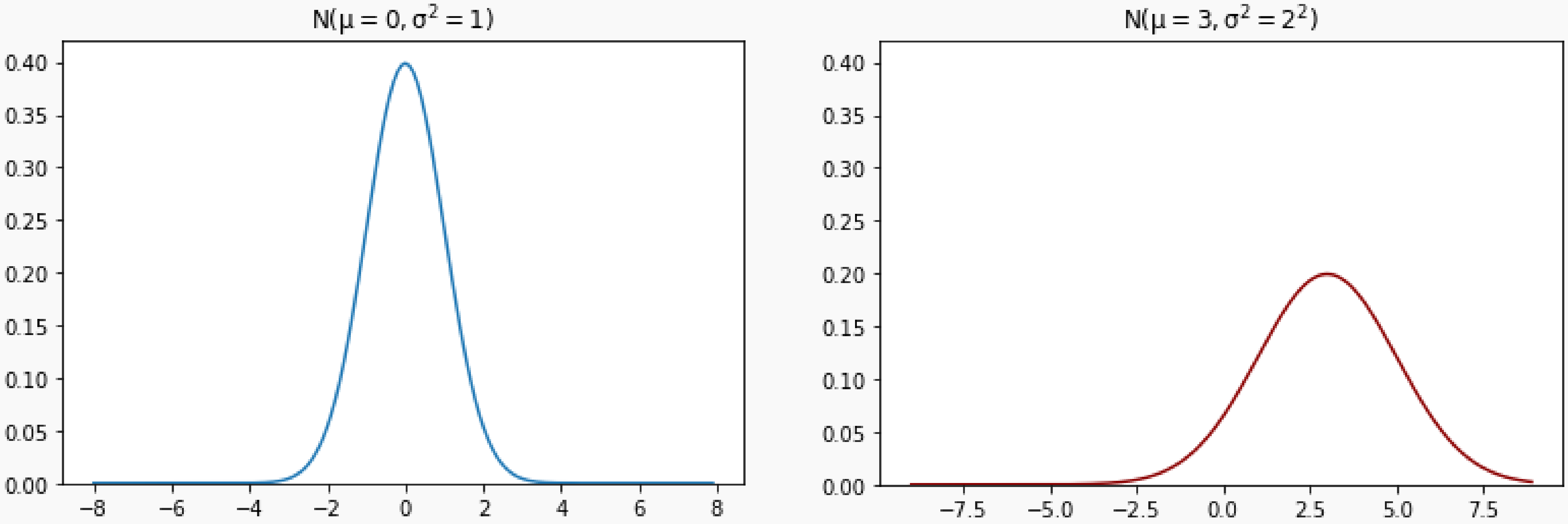
Normal Distribution

The normal distribution (sometimes called the Gaussian) is often referred to as the bell-shaped curve. But the normal distribution isn’t the only one that is bell-shaped: t distributions are also bell-shaped, for example.

The standard normal distribution is a special case: $Z \sim N(0,1)$.

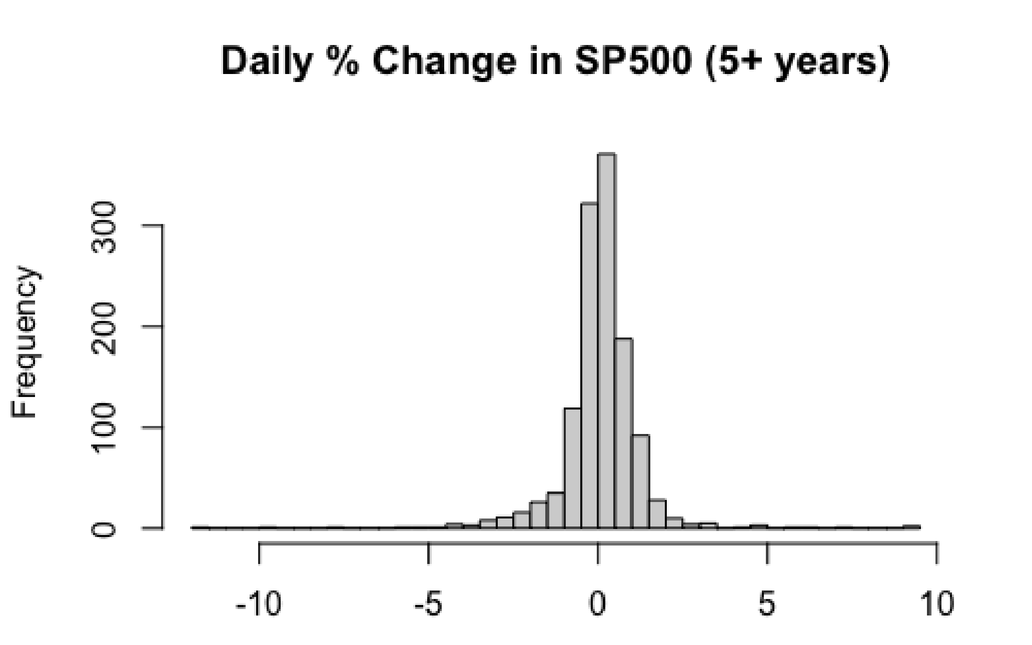
Any normal random variable can be standardized using the formula $Z = \frac{X-\mu}{\sigma}$.

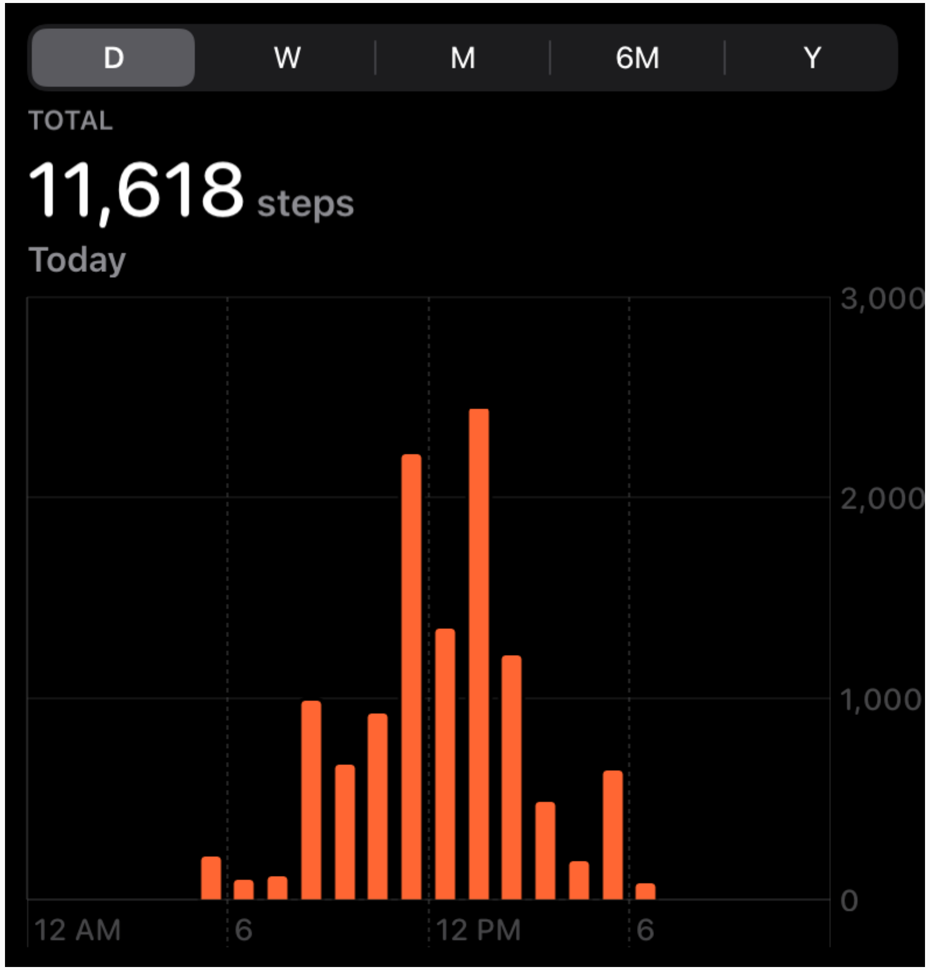
Normal Distribution Examples



A normal distribution has mean μ and standard deviation σ.

I See Normal Distributions





TOTAL

11,618 steps

Today

Central Limit Theorem

Why is the normal distribution used so often?

The **Central Limit Theorem**: random variables that are averages or sums of many other random variables will be approximately normally distributed.

More specifically: if $X_1, X_2, \dots, X_n$ are independent random variables (representing individual observations of data) with mean μ and standard deviation σ (not necessarily normal themselves), then the sample mean

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

will have approximate distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

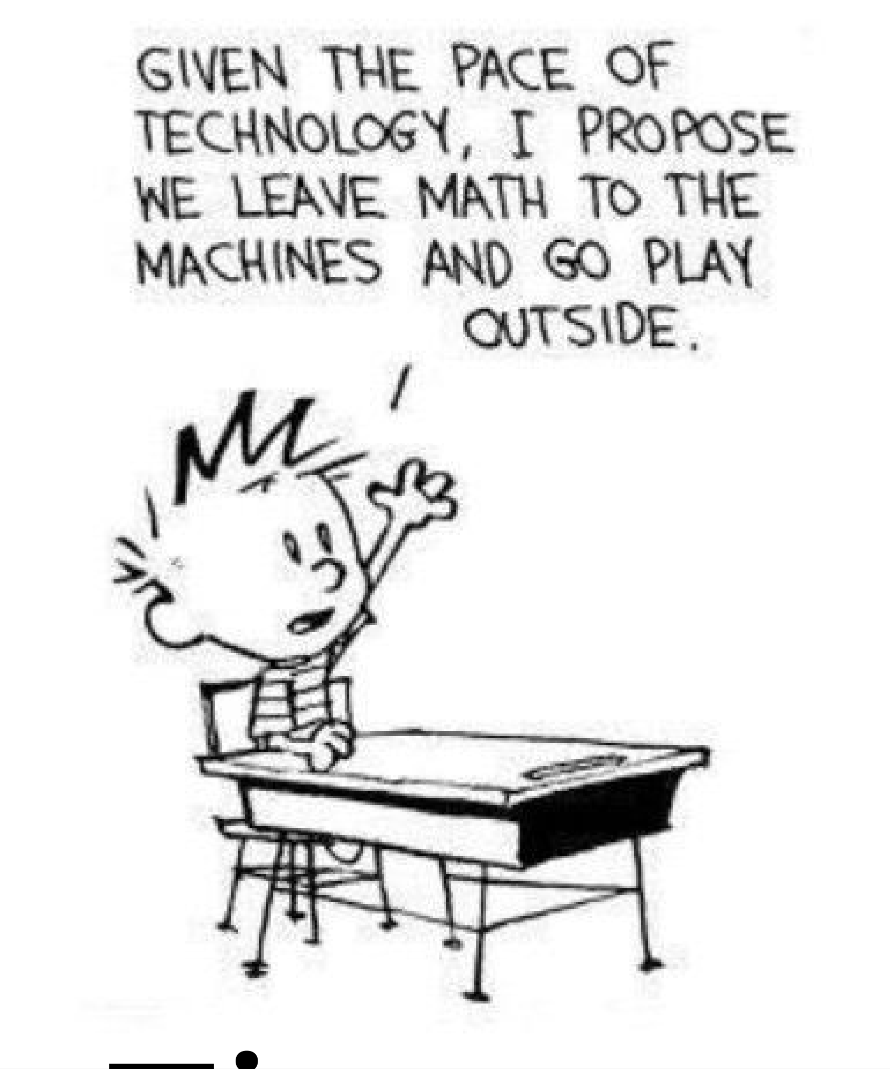
Joint Distributions

Joint Distributions

What happens to these probability distributions (PMFs and PDFs) when there are multiple random variables involved (aka, multiple observations in a data set)?

Let $f(x_1, x_2, \ldots, x_n)$ be the **joint distribution** of n separate random variables. If they all come from the same generative marginal distribution, $f(x_i)$, and are independent, what is the resulting distribution?

$$f(x_1, x_2, \ldots, x_n) = f(x_1) \cdot f(x_2) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i)$$


GIVEN THE PACE OF TECHNOLOGY, I PROPOSE WE LEAVE MATH TO THE MACHINES AND GO PLAY OUTSIDE.

⌛ Digestion Time

slido

Please download and install the Slido app on all computers you use

CS109A: Because even Harvard needs more ways to count things

ⓘ Start presenting to display the poll results on this slide.

slido

Please download and install the Slido app on all computers you use

CS109A: Where nerdy meets trendy! 🤓✨ #DataScienceRocks

ⓘ Start presenting to display the poll results on this slide.

slido

Please download and install the Slido app on all computers you use

CS109A: Where we turn data into gold... and occasionally into memes!

ⓘ Start presenting to display the poll results on this slide.

Modeling Data with Probability Distributions

Likelihood Theory

The idea of likelihood

The **likelihood** tells us:

Given the model, what is the likelihood of observing this data?

Instead of writing this function with the data (X) as the unknown, we use the same function but uses the parameter(s) as the unknown(s).

Given observed data, what values of the model's parameters are likely?

Modeling Linear Regression Probabilistically

The Simple Linear Regression Model

We've defined the linear regression model to predict the *i*-th observation's response, Y_i, from a predictor, X_i, to be:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

For any random variable, ϵ, that has zero mean

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

The error term, ϵ_i, represents the distance the observation lies from the line in the vertical distance (direction of *Y*).

The Probabilistic Regression Model

If we assume that $\epsilon_i \sim N(0, \sigma^2)$

This regression model can be rewritten as:

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

The likelihood of a measurement having value Y_i given X_i for a model β_0, β_1

$$L(\beta_0, \beta_1, \sigma^2 | Y_i, X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{2\sigma^2}}$$

The Probabilistic Regression Model

The likelihood of a measurement having value Y_i given X_i for a model β_0, β_1

$$L(\beta_0, \beta_1, \sigma^2 | Y_i, X_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{2\sigma^2}}$$

This formulation allows us to write out the joint likelihood function for this probability model.

The joint likelihood function for this probability model becomes:

$$L(\beta_0, \beta_1, \sigma^2 | \, \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{2\sigma^2}}$$

Assumptions of Linear Regression

Normality of Residuals

Linearity

$$L(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{2\sigma^2}}$$

Homoscedasticity

Independence

The Likelihood of Linear Regression

The joint likelihood function for this probability model becomes:

$$L(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{2\sigma^2}}$$

Remember, we stated that the likelihood function tells us:

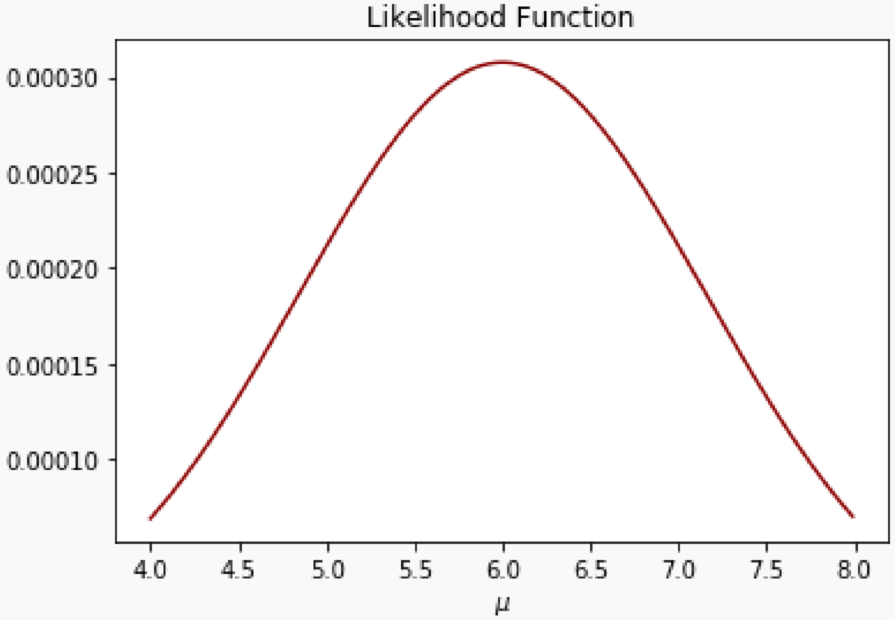
Given the observed data, what values of the model's parameters are likely?

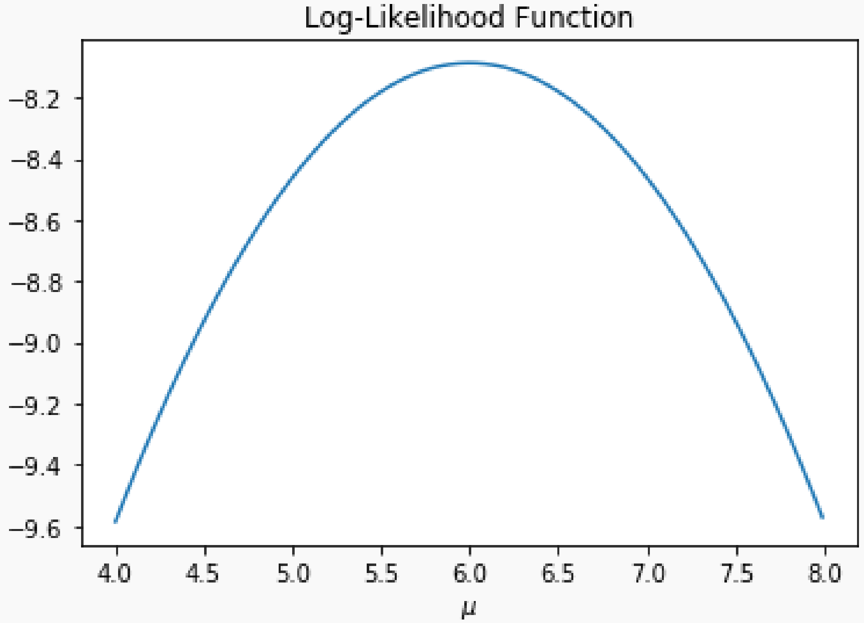
To find the model that is most likely, we simply maximize the likelihood function.

The Likelihood of Linear Regression

We observe that the maximum of a function and the maximum of the logarithm of the function yield the same values.

Let’s plot the likelihood and log-likelihood functions:





The Likelihood of Linear Regression

So, we can maximize the log-likelihood instead of the likelihood function

$$l(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{2\sigma^2}}$$

$$l(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} \ln \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2}{2\sigma^2}}$$

A little bit of algebra

$$l(\beta_0, \beta_1, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = -\sum_{i=1}^{n} \ln\left(\sqrt{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$$

The Likelihood of Linear Regression

Just two more steps before we are done:

1. We don’t need to find the σ, so we can replace any term involving σ with constants, which I call C_1, C_2

$$l(\beta_0, \beta_1, \,|\, \boldsymbol{Y}, \boldsymbol{X}) = C_1 - C_2 \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$$

2. Maximizing the log-likelihood is equivalent to minimizing the negative log-likelihood. By dropping all constants, we can define this as our LOSS function, L

$$\mathrm{L} = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2$$

$$\boxed{\mathrm{L} = \sum_{i=1}^{n} \big(Y_i - (\beta_0 + \beta_1 X_i)\big)^2}$$

Which brings us to our best friend, the Mean Squared Error (MSE)!

Thank you