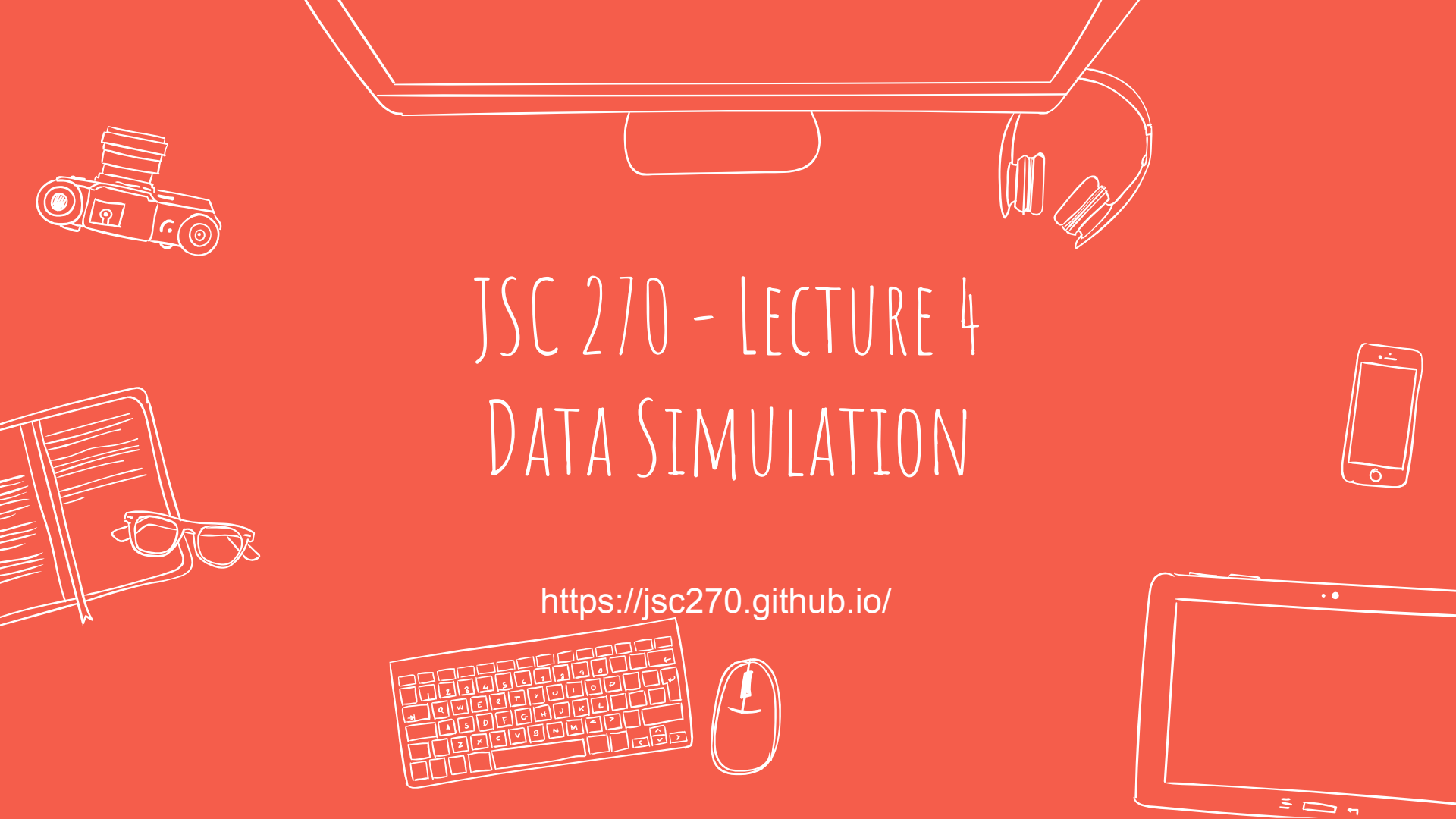


JSC 270 - Lecture 4
Data Simulation

https://jsc270.github.io/

Announcements

Today: Visualization Talk by Dr Fanny Chevalier 2-3pm

Next Monday: Talk on Reproducibility by Dr Benjamin Haibe-Kains also 2-3pm

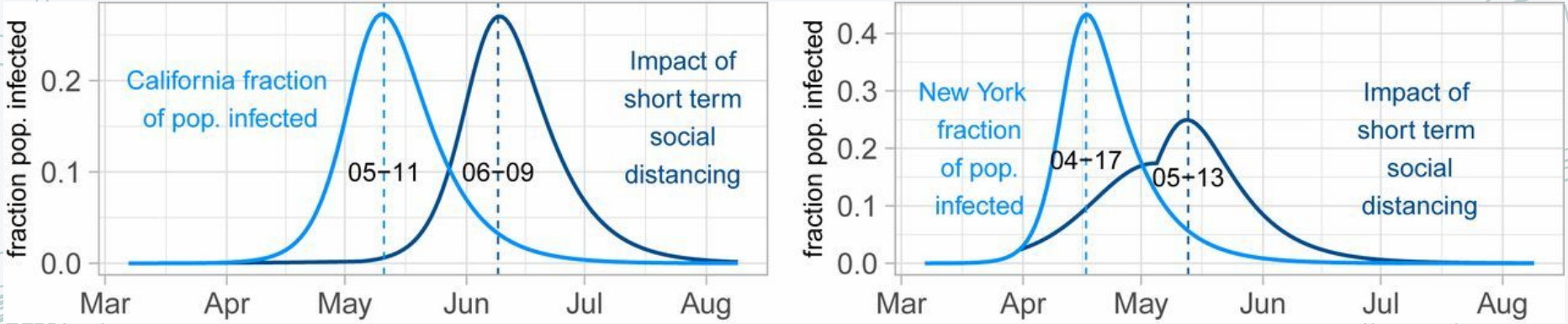
Assignment grades will be available by Lab time

Fantastic job on reflection quizzes and perusall!
Grades for these will be available after class today

Simulations

Can you think of any examples yourselves?

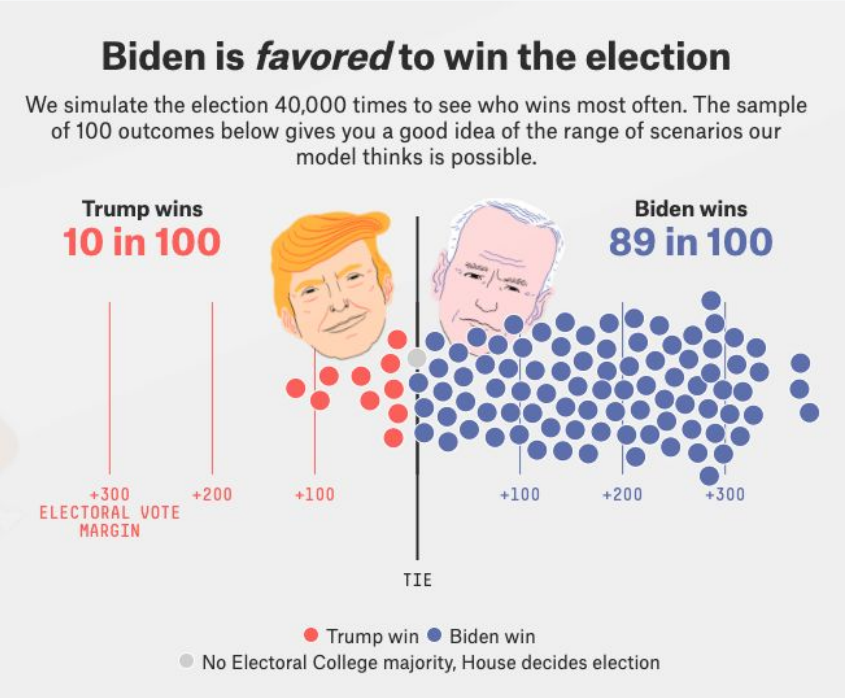
Example 1. COVID19 spread

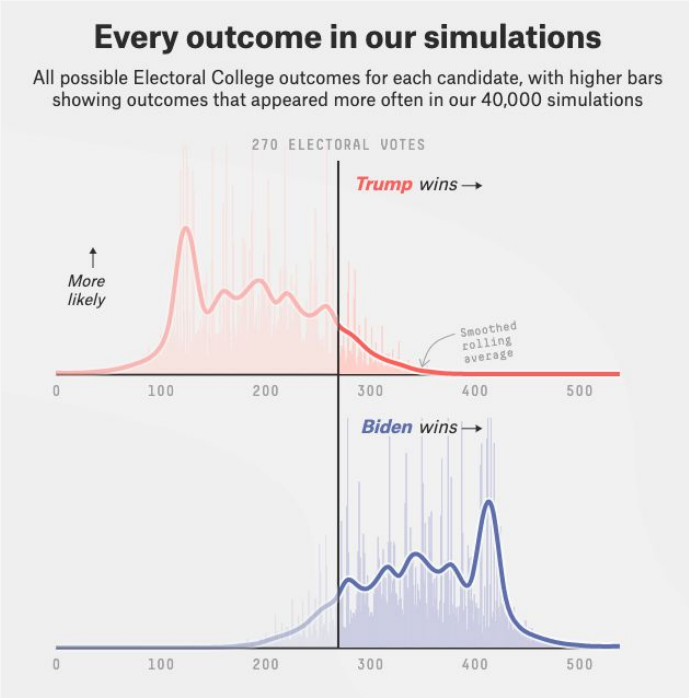


SIR Model. (Left) short term distancing (Right) longer term distancing

https://www.pnas.org/content/117/29/16732

Example 2. Elections





https://projects.fivethirtyeight.com/2020-election-forecast/?fbclid=IwAR0pCClrbTV4bqL2DTrDf2q32mZcq-G37wHBa9TM2u3ZX7G82hhtMxbB9OI

EXAMPLE 3. DALL-E (OPEN AI, JAN 5, 2021)

GPT-3 based model

Generates images from English text from text-image pairs

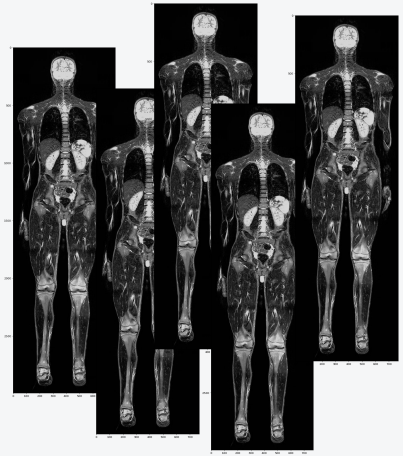
TEXT PROMPT

an armchair in the shape of an avocado […]

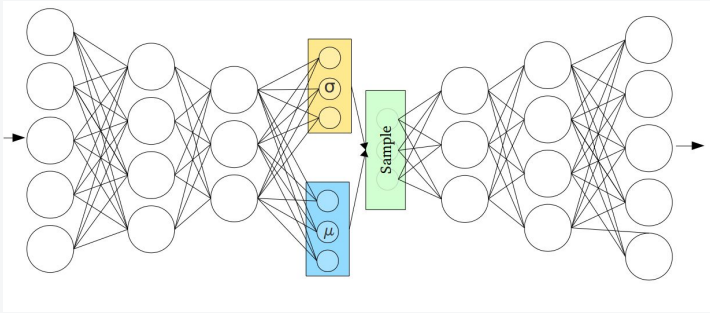
AI-GENERATED IMAGES

https://openai.com/blog/dall-e/

Example 4. Medical Imaging



Healthy wbMRI



https://arxiv.org/abs/2006.00727

Example 5. Social Networks

https://arxiv.org/pdf/0912.5410.pdf
http://reports-archive.adm.cs.cmu.edu/anon/ml/CMU-ML-06-107.pdf

Why we do simulations?

Why we do simulations

- To think through the design of a study
- To test/verify a hypothesis
- To study the system (complexity and dynamics) in a controlled environment
- To augment data
- To generate new knowledge/discovery
- Educational/training purposes

When are simulations not appropriate?

When are simulations not appropriate?

- ❌ When the system is deterministic
- ❌ When not enough is known about the system such that the assumptions you would have to make are unlikely to be accurate

Simulating Coin Flips

How many heads are we likely to see if we throw 5 slightly unfair (0.45) coins?

Simulating Coin Flips

How many heads are we likely to see if we throw 5 slightly unfair (0.45) coins?

What do you expect?

Simulating Coin Flips

How many heads are we likely to see if we throw 5 slightly unfair (0.45) coins?

How do we simulate this to study the outcome?

Simulating Coin Flips

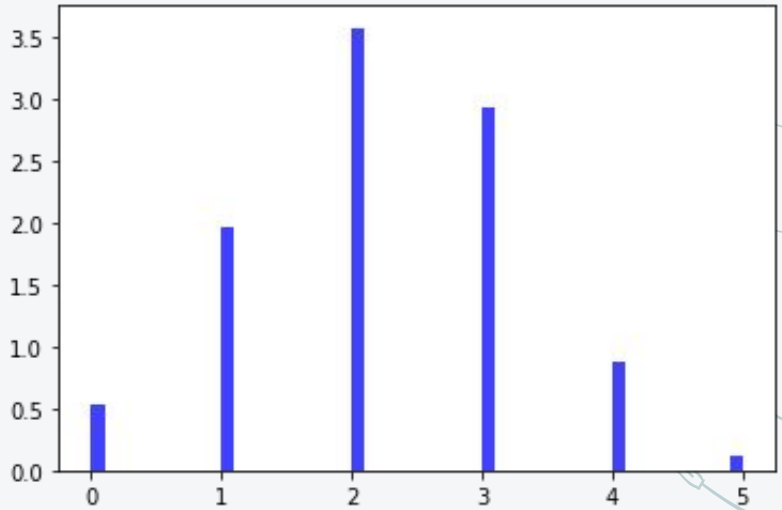
How many heads are we likely to see if we throw 5 slightly unfair (0.45) coins?

$X \sim Bin(5, 0.45)$

Simulating Coin Flips

How many heads are we likely to see if we throw 5 slightly unfair (0.45) coins?

$X \sim Bin(5, 0.45)$

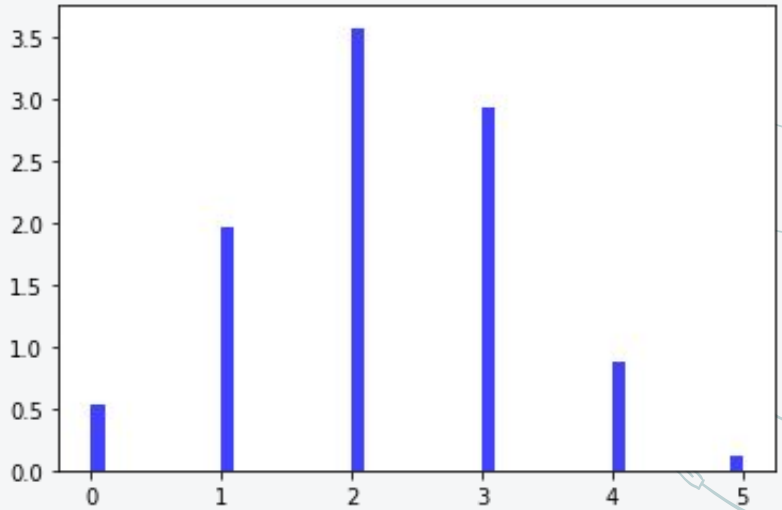


Simulating Coin Flips

How many heads are we likely to see if we throw 5 slightly unfair (0.45) coins?

X ~ Bin(5,0.45)

How did we do this?



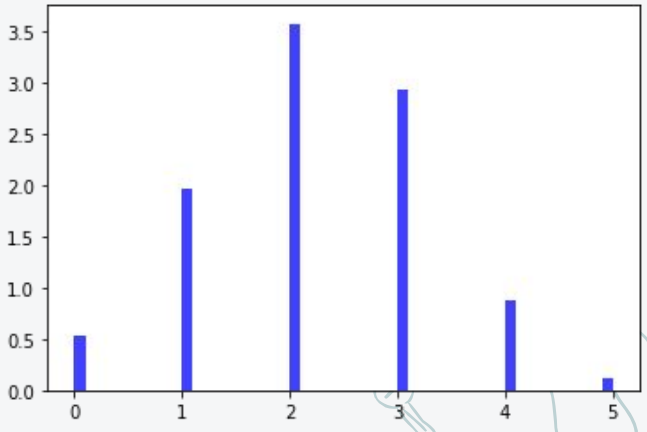
Simulating Coin Flips

How many heads are we likely to see if we throw 5 slightly unfair (0.45) coins?

$X \sim Bin(5, 0.45)$

How did we do this?

Sample from the binomial!



How Exactly do we sample from a given binomial distribution?

Assumption - there exists a uniform random number generator

Step 1. Generate 5 numbers from 0 to 1

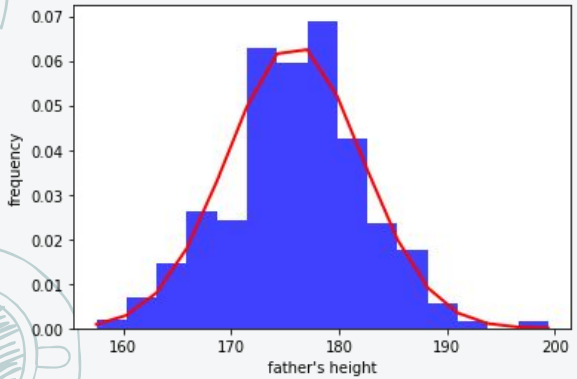
Step 2. Record how many numbers fall below 0.45

Step 3. Repeat (100/1000/... times)

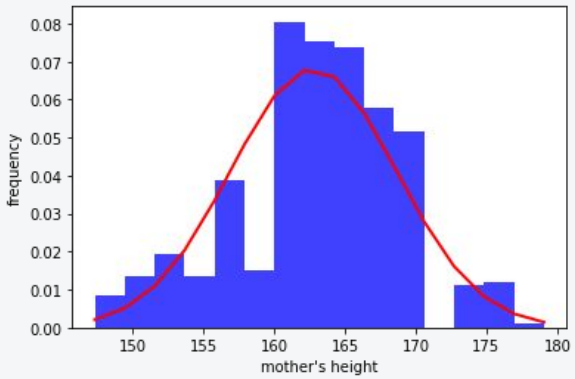
Galton's data - an example of Sampling from a Gaussian

Generating new height data based on Galton's data

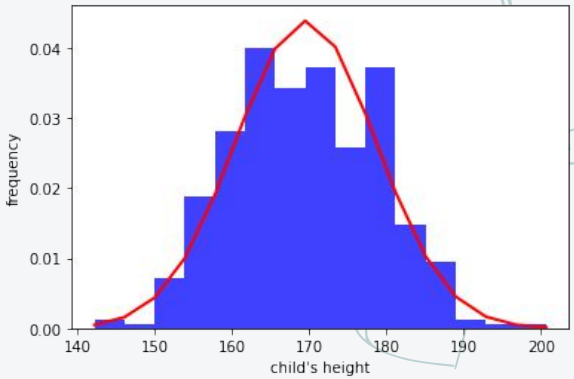
We have a set of heights from the population



N(175.85,6.3)



N(162.8,5.9)

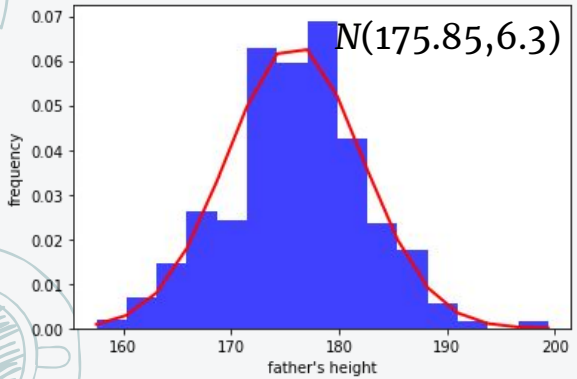


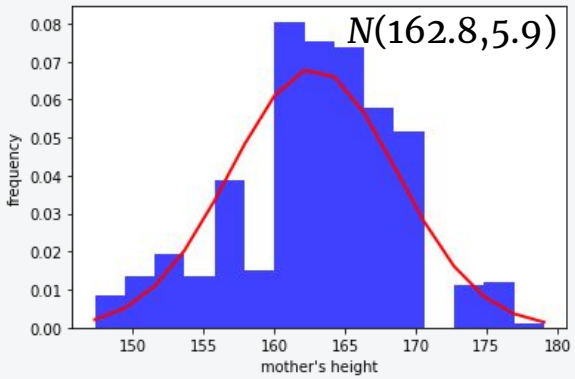
N(169.6,9)

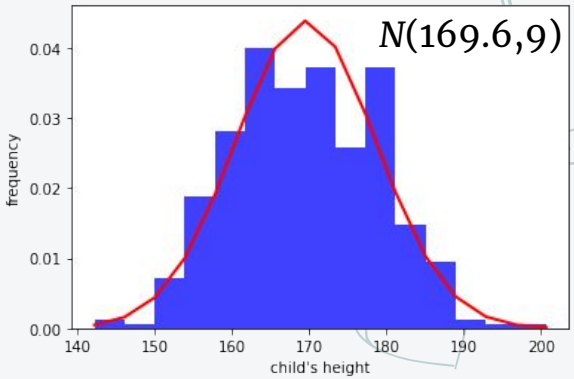
We want to generate another sample from this population (a new family). Can we just sample from these three distributions?

Generating new height data based on Galton's data

We want to generate another sample from this population (a new family). Can we just sample from these three distributions? Let's try....







Correlation matrix

	father	mother	child
father	1	0	0
mother	0	1	0
child	0	0	1

Galton's data - an example of Sampling from a Gaussian

Child's height

$$Y = 56.67 + 0.38 \cdot FH + 0.28 \cdot MH$$

What is the right algorithm for generating info about a set of new families?

Step 1.

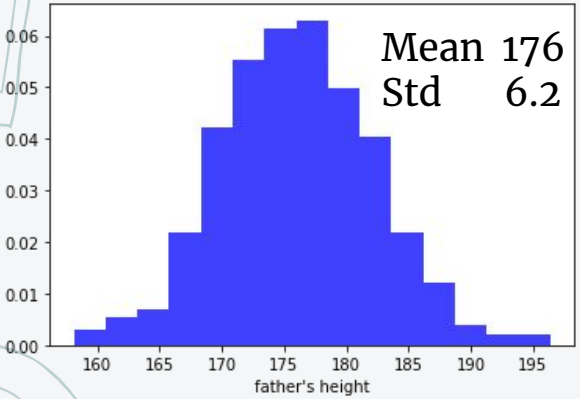
Step 2.

Step 3.

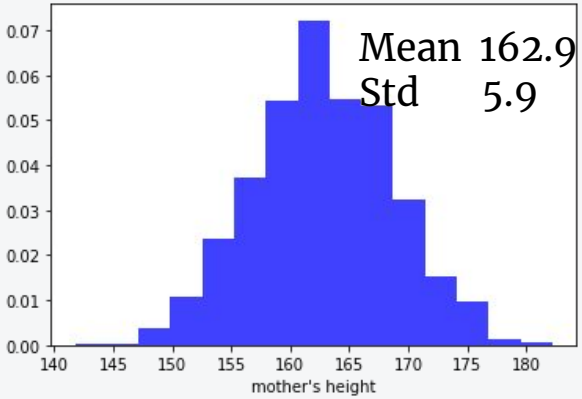
Step 4.

Let's simulate data according to this algorithm

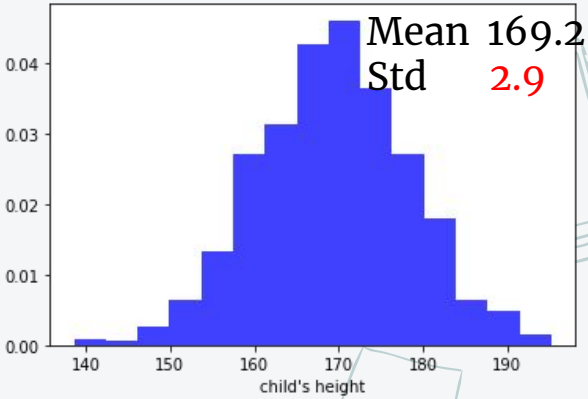
$N(175.85,6.3)$



$N(162.8,5.9)$



$N(169.6,9)$



What could be the reason(s) for the distribution of children heights to deviate?

Galton's data - an example of Sampling from a Gaussian

Child's height

$$Y = 56.67 + 0.38 \cdot FH + 0.28 \cdot MH + \epsilon$$

$$\epsilon \sim N(0, \sigma^2_{(y-\hat{y})})$$

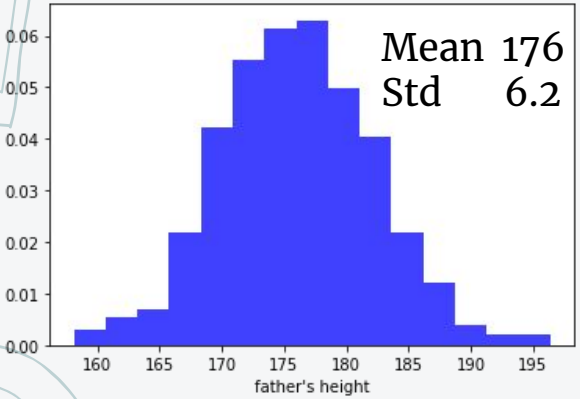
- Variance of the residuals of the model!
- 8.8

Don't forget that our model was not perfect!

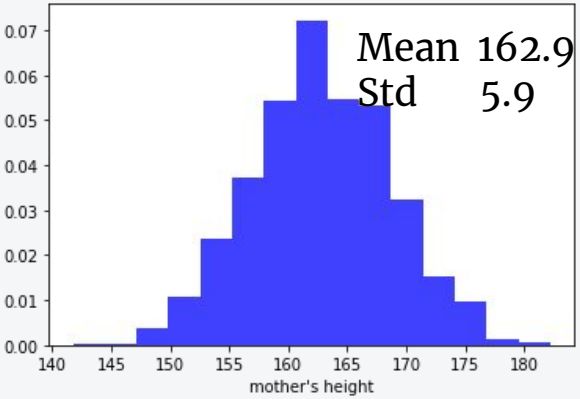
As models almost never are...

A new sample from Galton's Population

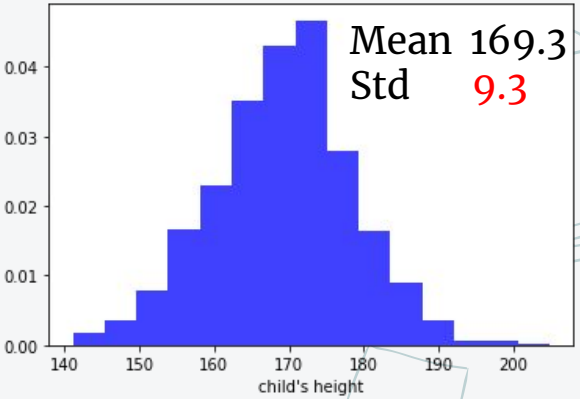
$N(175.85, 6.3)$



$N(162.8, 5.9)$



$N(169.6, 9)$



Don't forget to simulate realistic data (i.e. including the noise) using regression!

COVID19 Example

Suppose we know that people are at different levels of immunity, in fact we have four subpopulations with immunities in the 0 to 1 range

How would you simulate data from this population?

COVID19 Example

Suppose we know that people are at different levels of immunity, in fact we have four subpopulations with immunities are in the 0 to 1 range

How would you simulate data from this population?

1. Average Immunity per population ~ Uniform(0,1)
2. For simplicity, let's fix the standard dev to be 1
3. Sample each subpopulation (1,000 people) from a Beta(α,β) with the params indicated above

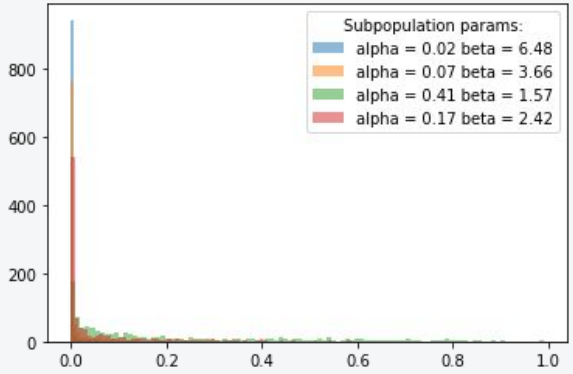
COVID19 Immunity sampled data

$\mu \sim \text{Uni}(0,1,4)$: [0.15 0.27 0.64 0.41], $\sigma = 1$

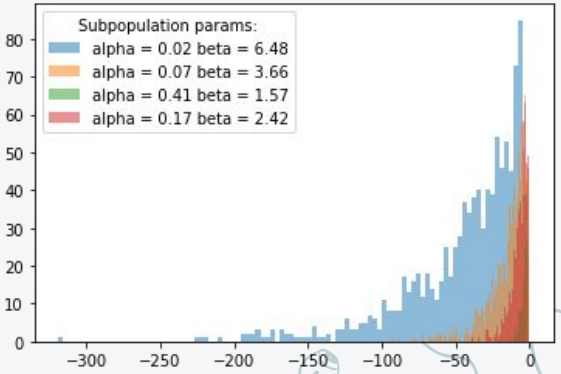
Sample each subpopulation *j* from Beta:

$$\alpha_j = \mu_j^2/\sigma_j^2$$

$$\beta_j = \sigma_j^2/\mu$$



Immunity in various populations



Log(immunity)

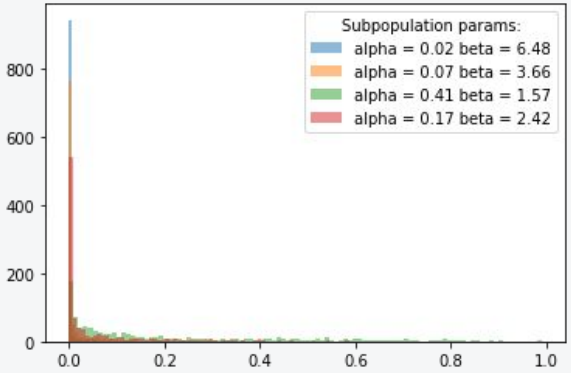
COVID19 Immunity sampled data - Hierarchical model

$\mu \sim \text{Uni}(0,1,4)$: [0.15 0.27 0.64 0.41], $\sigma = 1$

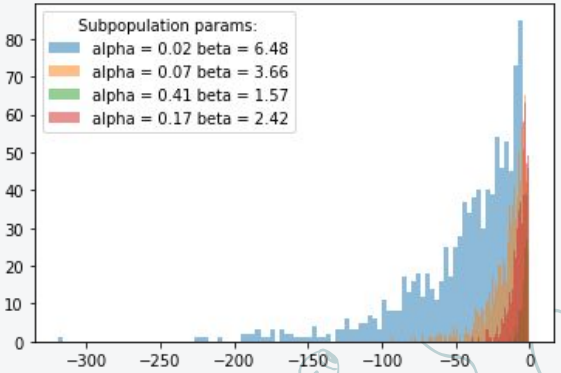
Sample each subpopulation *j* from Beta:

$$\alpha_j = \mu_j^2/\sigma_j^2$$

$$\beta_j = \sigma_j^2/\mu$$



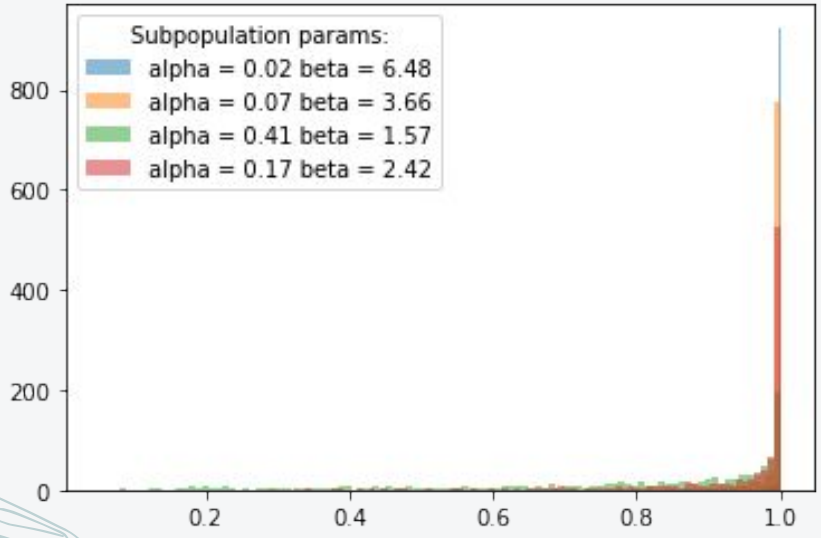
Immunity in various populations



Log(immunity)

COVID19 The probability of a person being infected

The probability of a person getting covid in various populations is 1-immunity



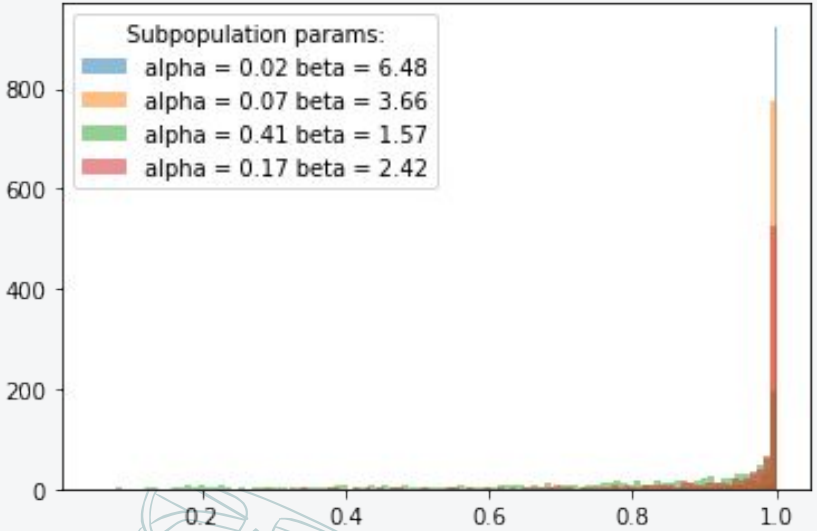
How many people would get sick if exposed in each subpopulation?

Suppose 5% of people are exposed

- Randomly generate indices (exposed people)
- If probability of getting sick is higher than randomly generated number (remember coin flips?) then this person will get sick

COVID19 NUMBER OF PEOPLE IN EACH SUBPOPULATION IF 5% EXPOSED (TOTAL 1000 PER SUBPOPULATION)

The probability of a person getting covid in various populations is 1-immunity



Subpopulation 1: 50/50

Subpopulation 2: 48/50

Subpopulation 3: 42/50

Subpopulation 4: 47/50

Summary

- Saw a few examples of data simulations
- Should be able to articulate advantages and disadvantages of data simulations
- Should be able to simulate data from
  - Binomial (e.g. coin flips)
  - Gaussian
  - Regression
  - Hierarchical model

Reading

Original paper introducing Transformer architecture - foundation for AlphaFold2 and GPT-3/DALL-E

https://arxiv.org/pdf/1706.03762.pdf

Nice example of Poisson modeled bus waiting times:
https://jakevdp.github.io/blog/2018/09/13/waiting-time-paradox/

Extra Slides

Suppose you didn't know the mean

Example: We need to simulate a dataset of heights y across various places on earth, but we only know world average height and world standard deviation

What would you do?

Suppose you didn't know the mean

What would you do?

1. Make it up based on what you are trying to model
2. Be a little more Bayesian...

Being Bayesian - Hierarchical Model

l - location
w - world

$$h_l \sim N(\mu_l, \sigma_l^2)$$
$$\mu_l \sim N(\mu_w, \sigma_w^2) \quad - \quad \text{prior}$$

How to generate data?

1. Sample mean for a location
2. Given the mean, sample subpopulation

$$h_l \sim N(\mu_l, \sigma_l^2)$$
$$\mu_l \sim N(\mu_w, \sigma_w^2)$$

Being Bayesian about heights

World height average for women is 165cm
World height std for women is 8.9cm

$$h_l \sim N(\mu_l, \sigma_l^2)$$
$$\mu_l \sim N(\mu_w, \sigma_w^2)$$

Being Bayesian about heights

World height average for women is 165cm
World height std for women is 8.9cm

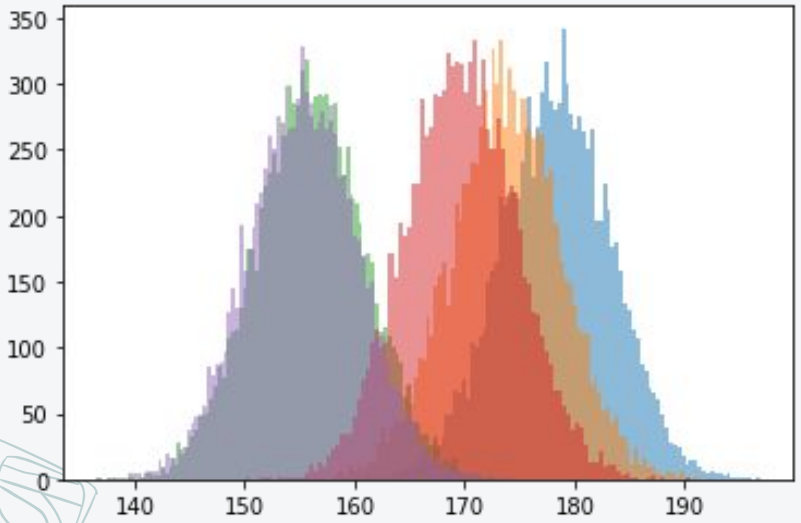
Height averages: [178.37 173.76 156 169.69 155.38]

$$h_l \sim N(\mu_l, \sigma_l^2)$$
$$\mu_l \sim N(\mu_w, \sigma_w^2)$$

Being Bayesian about heights

World height average for women is 165cm
World height std for women is 8.9cm

Height averages: [178.37 173.76 156 169.69 155.38]

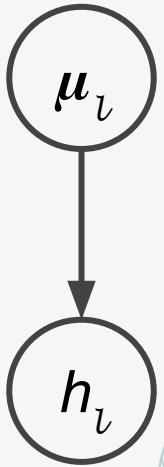


Being Bayesian

l - location
w - world

$$h_l \sim N(\mu_l, \sigma_l^2)$$
$$\mu_l \sim N(\mu_w, \sigma_w^2) \quad - \quad \text{prior}$$

Graphical Model notation



How to generate data?

1. Sample mean for a location
2. Given the mean, sample subpopulation

Should be able to sample from many models now!

Example: Latent Dirichlet Allocation (LDA) – inferring topics in a set of publications

We generate the set of publications as follows:

1. Generate the topic distribution for document i $\theta_i \sim Dir(\alpha)$
2. For each topic k, sample word distribution $\phi_k \sim Dir(\beta)$
3. For each of the positions for the word j in document i
   a. Choose topic $z_{ij} \sim Multinomial(\theta_i)$
   b. Choose a word $w_{ij} \sim Multinomial(\phi_i)$

https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation#:~:text=In%20natural%20language%20processing%2C%20the,of%20the%20data%20are%20similar.