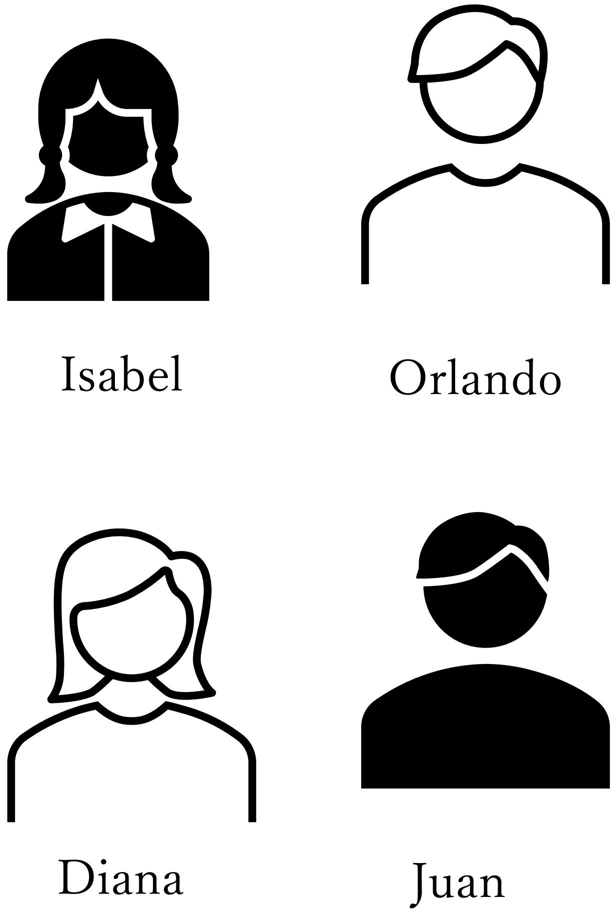
Spatial Autocorrelation
Part B

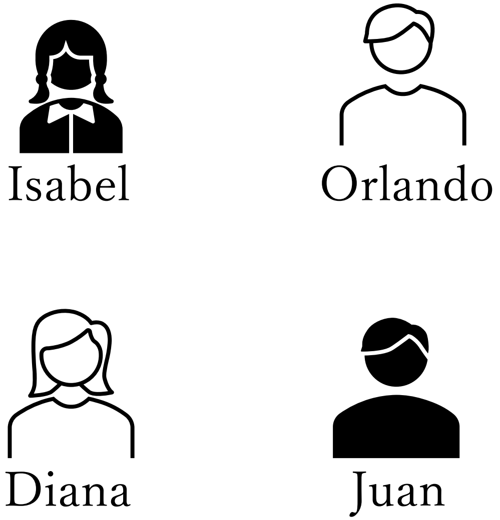
Second Session

Orlando Sabogal-Cardona
PhD researcher
University College London UCL

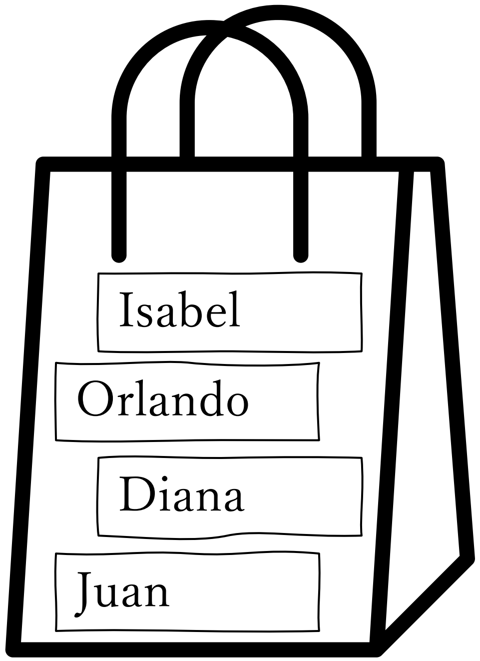
Hypothesis testing



Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY



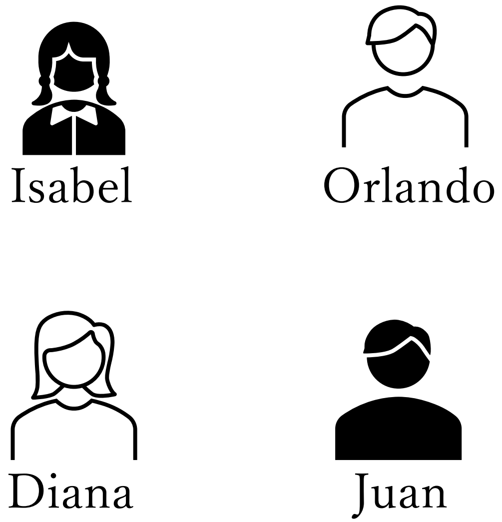
One of them should randomly do the dishes every night



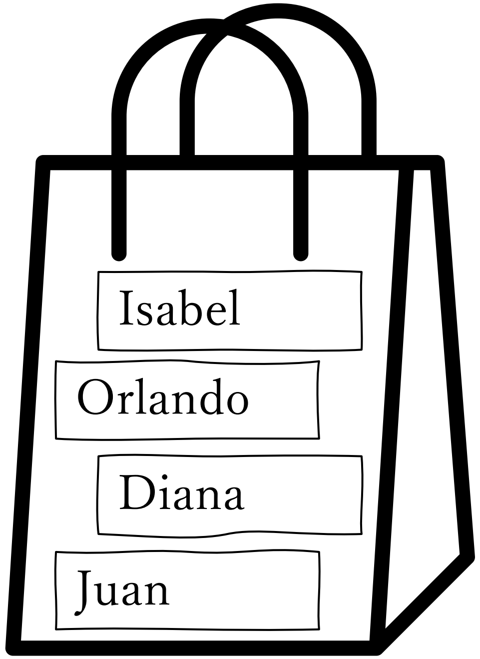
So, every night they randomly pick up one name from the bag

Probability(not doing the dishes) = 3/4 = 0.75

Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY

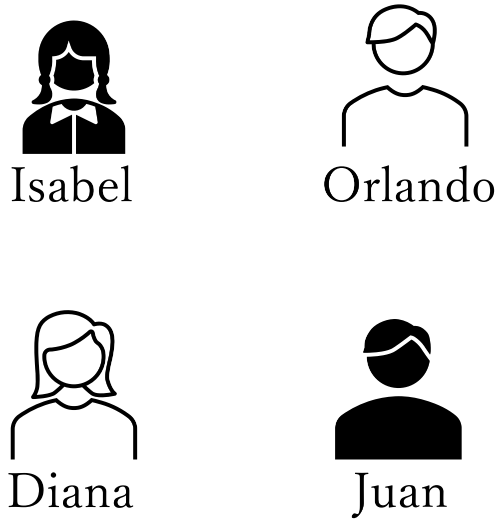


One of them should randomly do the dishes every night

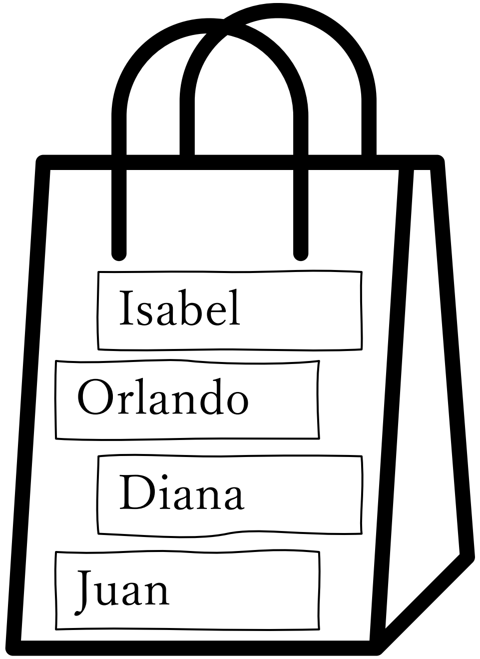


But after 4 nights, Orlando has not been selected. And he has a reputation

Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY



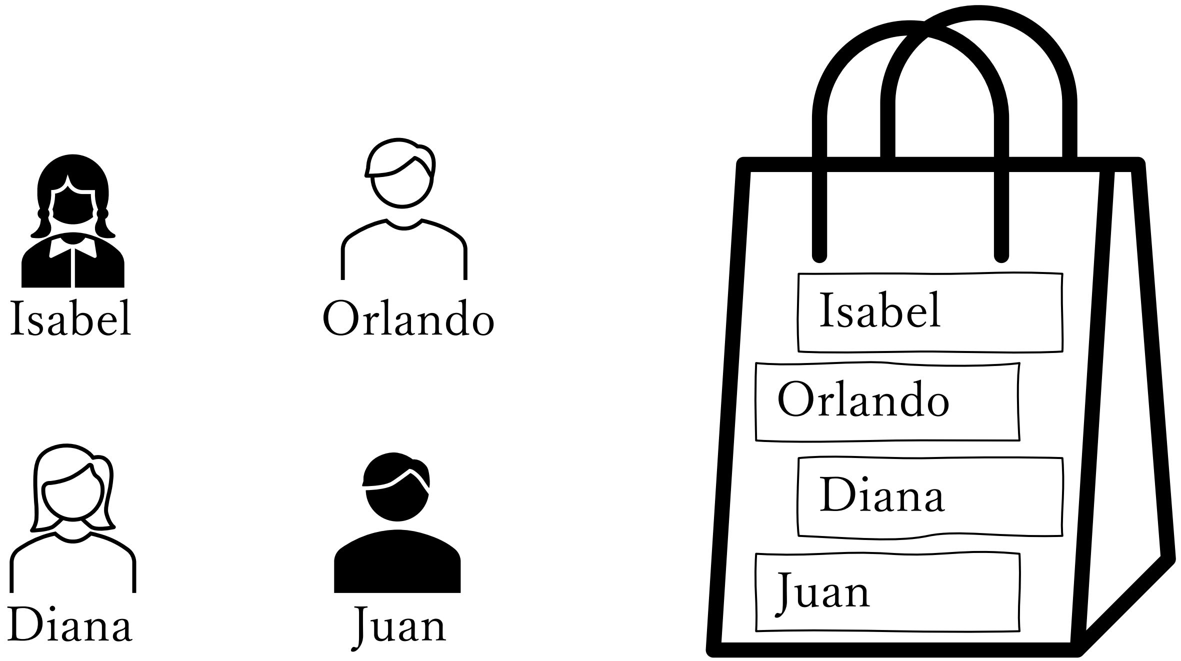
One of them should randomly do the dishes every night



But after 4 nights, Orlando has not been selected. And he has a reputation

Is Orlando cheating? Is he taking his name out of the bag?

Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY

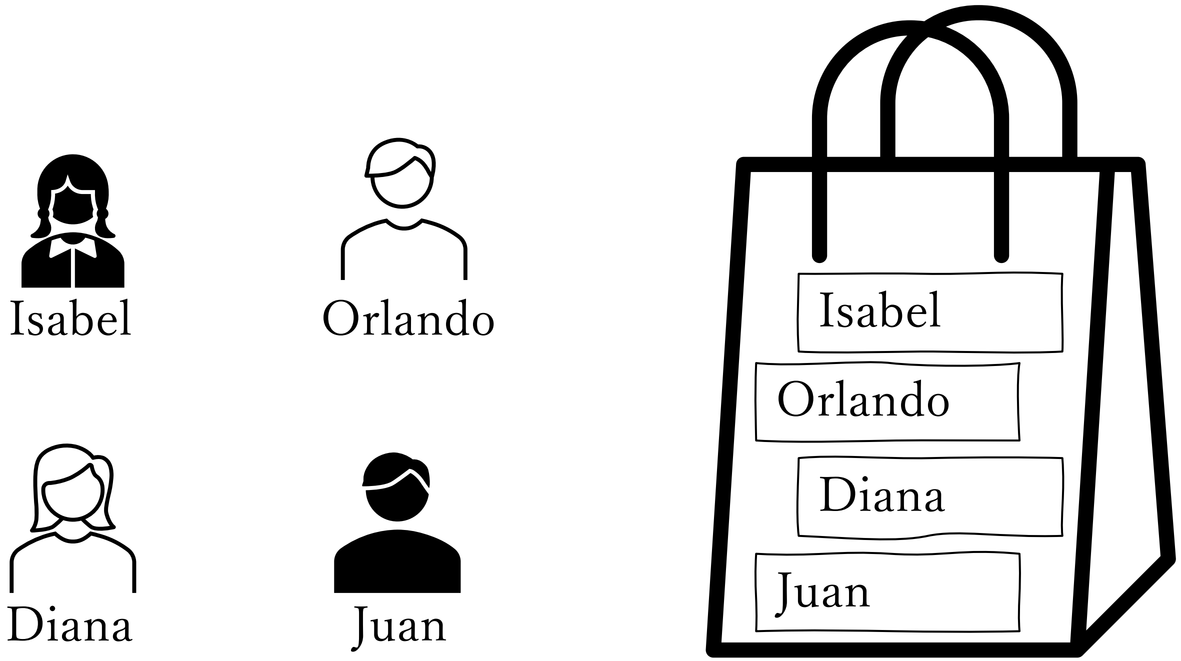


<u>IF</u> the bag has the four names, then:

Probability(of not being selected four times in a row) = Probability(not doing the dishes)^4

Probability(of not being selected four times in a row) = (0.75)^4 = 0.316

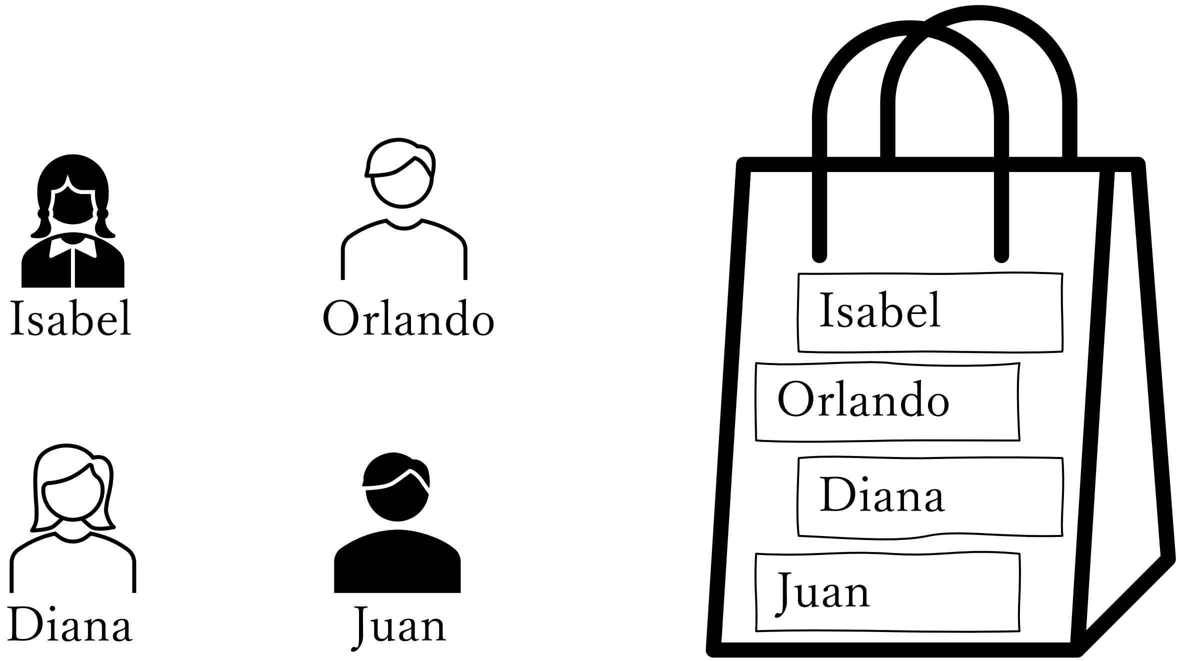
Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY



In other words: if Orlando is not cheating and the bag is correct, then the probability of observing what we are observing is 0.32.

As 0.32 is relatively high, we do not have evidence against good Orlando.

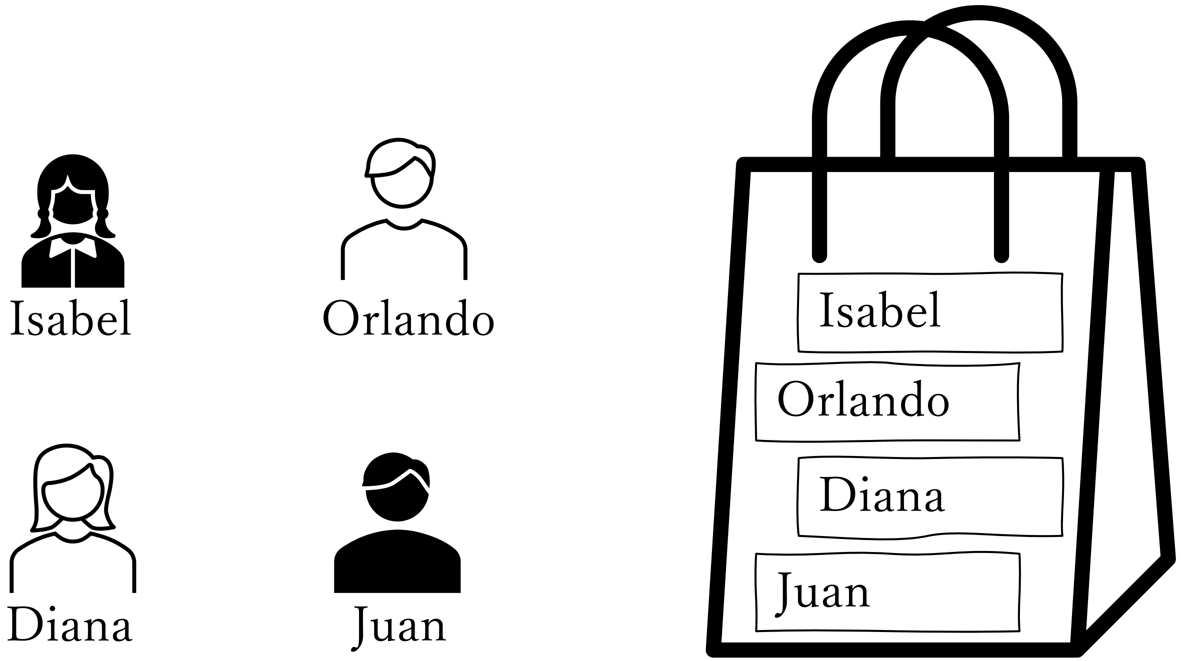
Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY



Null hypothesis: Orlando is not doing anything, nothing is happening

Alternative hypothesis: Orlando is cheating, something is happening

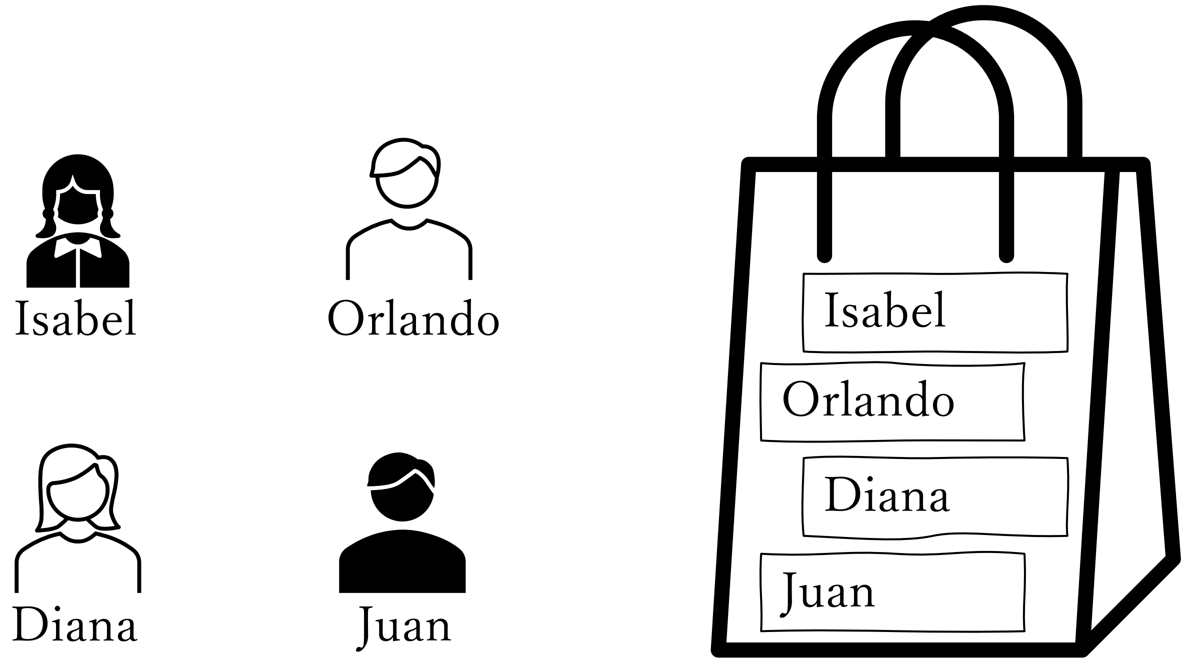
Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY



Null hypothesis: Orlando is not doing anything, nothing is happening
Alternative hypothesis: Orlando is cheating, something is happening

We assume the Null to be true

Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY



Null hypothesis: Orlando is not doing anything, nothing is happening

Alternative hypothesis: Orlando is cheating, something is happening

We assume the Null to be true

High probability ________

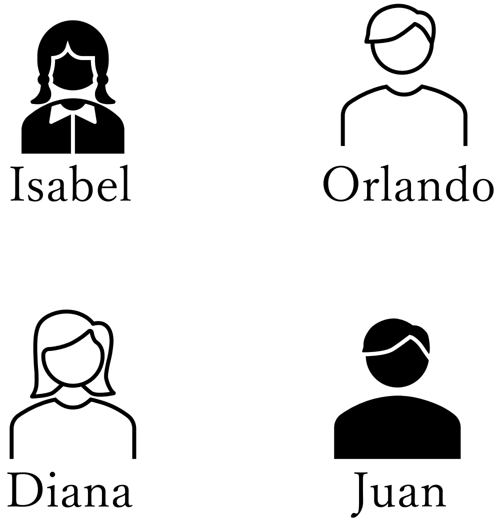
Low probability ________

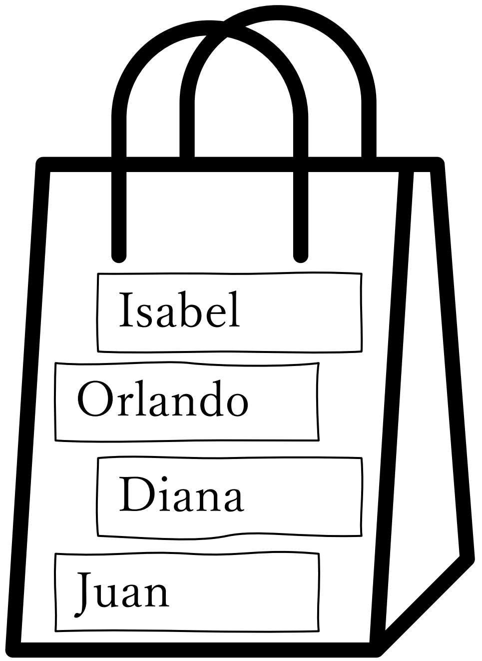
What is high? What is low?

Significance level. Alpha level

Standard practice: probability of 0.5.

Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY





Null hypothesis: Orlando is not doing anything, nothing is happening

Alternative hypothesis: Orlando is cheating, something is happening

We assume the Null to be true

High probability

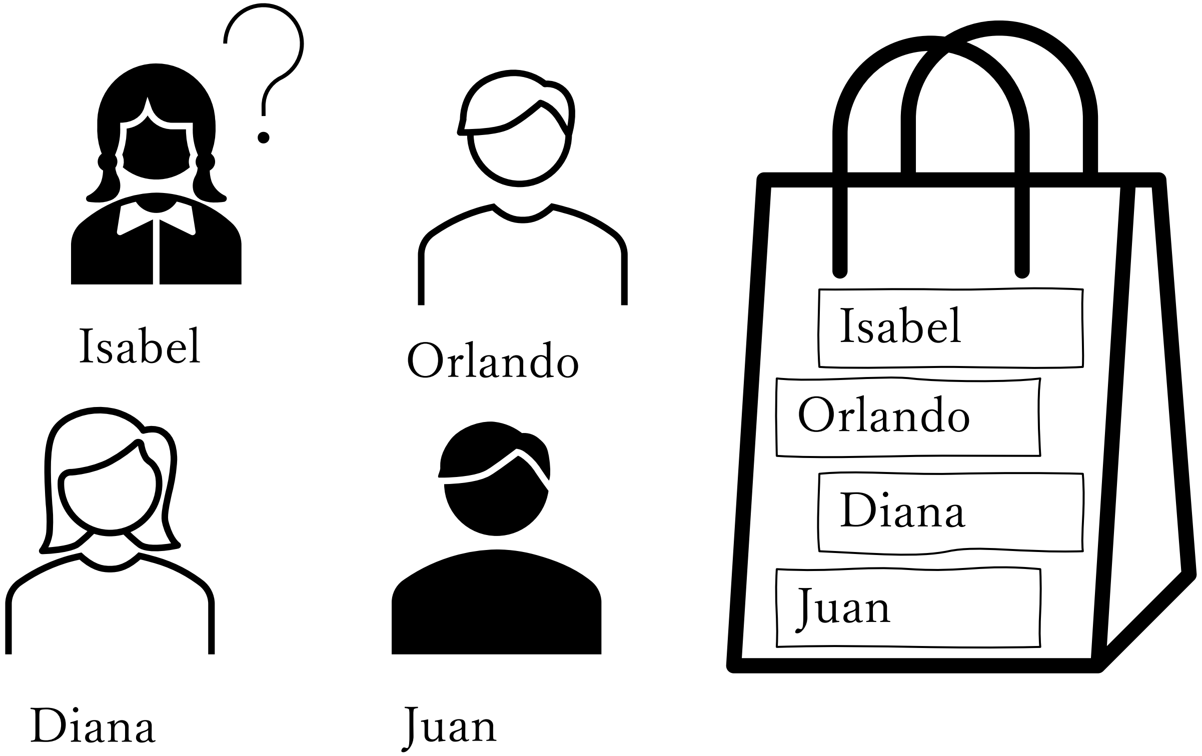
"The Null should be ok". We fail to reject the Null. No evidence to support the Alternative

Low probability

"The Null is not ok". We reject the Null. There is evidence to support the Alternative

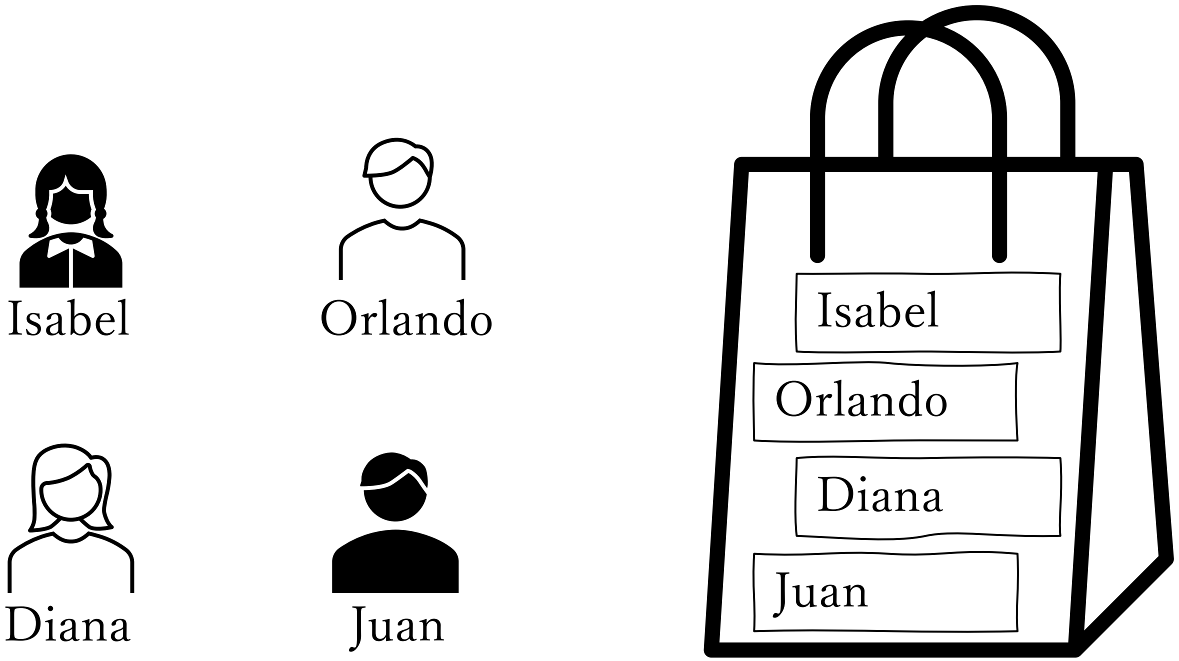
Results are "statistically significant"

Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY



What about Isabel?
After 20 nights, her name has never come out

Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY



<u>**IF**</u> the bag has the four names, then:

Probability(of not being selected 20 times in a row) = Probability(not doing the dishes)^20

Probability(of not being selected 20 times in a row) = (0.75)^20 = 0.003

Example adapted from: https://www.youtube.com/watch?v=5D1gV37bKXY

Example: flipping a coin

Now Isabel and Orlando are playing a simple game. They are flipping a coin, with Isabel winning if it lands on heads and Orlando winning if it lands on tails.

After 10 flips, Isabel won 8 times. Is she cheating?

Do not panic!

In real life you do not have to come with a way to compute probabilities for every situation.

On the contrary, there are already well-defined tests for very specific cases:
- Z tests, T tests, Chi-squared tests, F tests, etc.
- Mean, compare mean, variance, compare variances, proportions, etc.

Keep in mind that hypothesis testing is about making "statistical inference"

But why do hypothesis tests work? Why do they make sense?

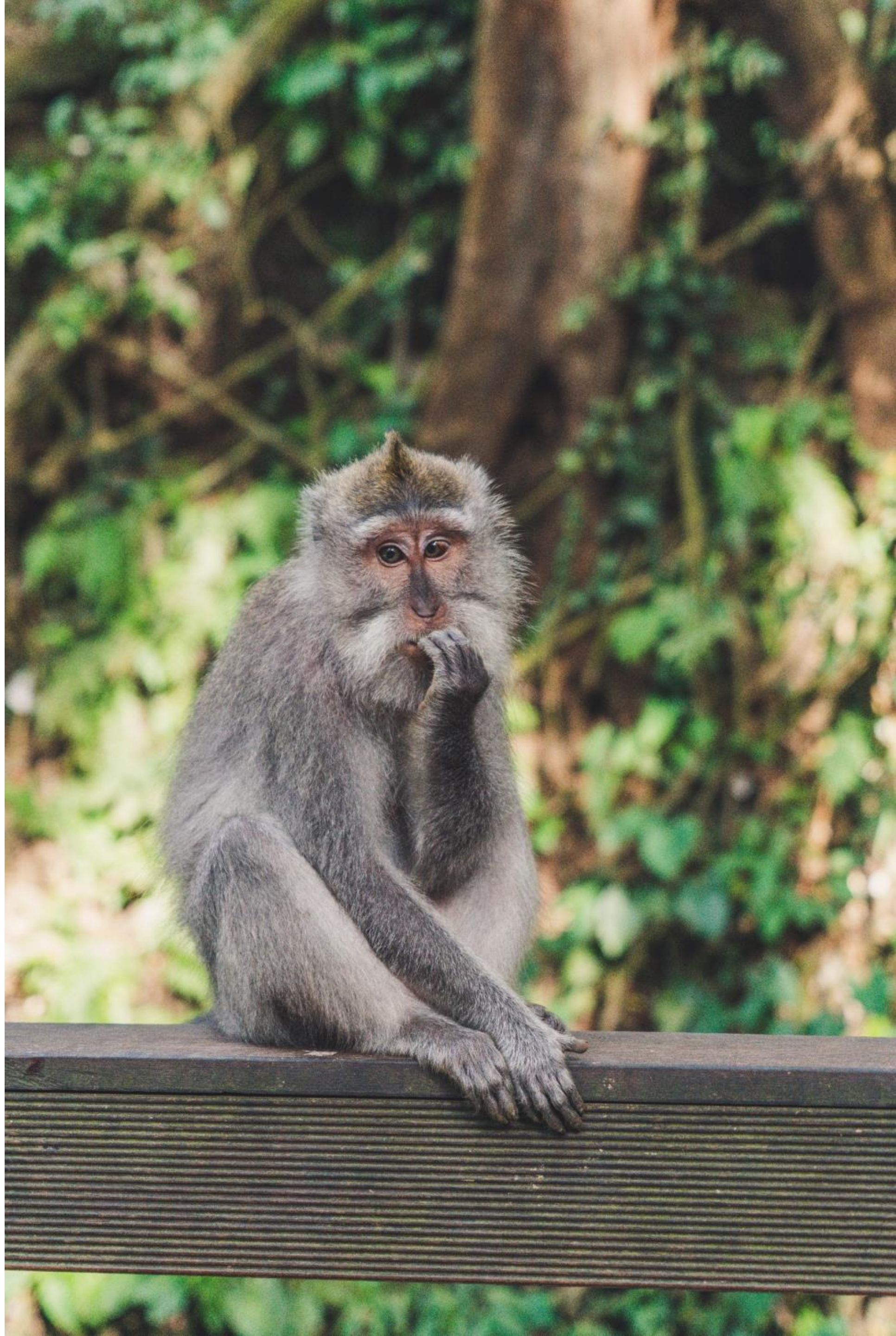


Photo by Paolo Nicolello on Unsplash

Central Limit Theorem CLT

- One of the most fundamentals and profounds concepts in statistics (and science)
- The central limit theorem provides the theoretical foundation for the use of hypothesis testing to make inferences about population parameters based on sample data.

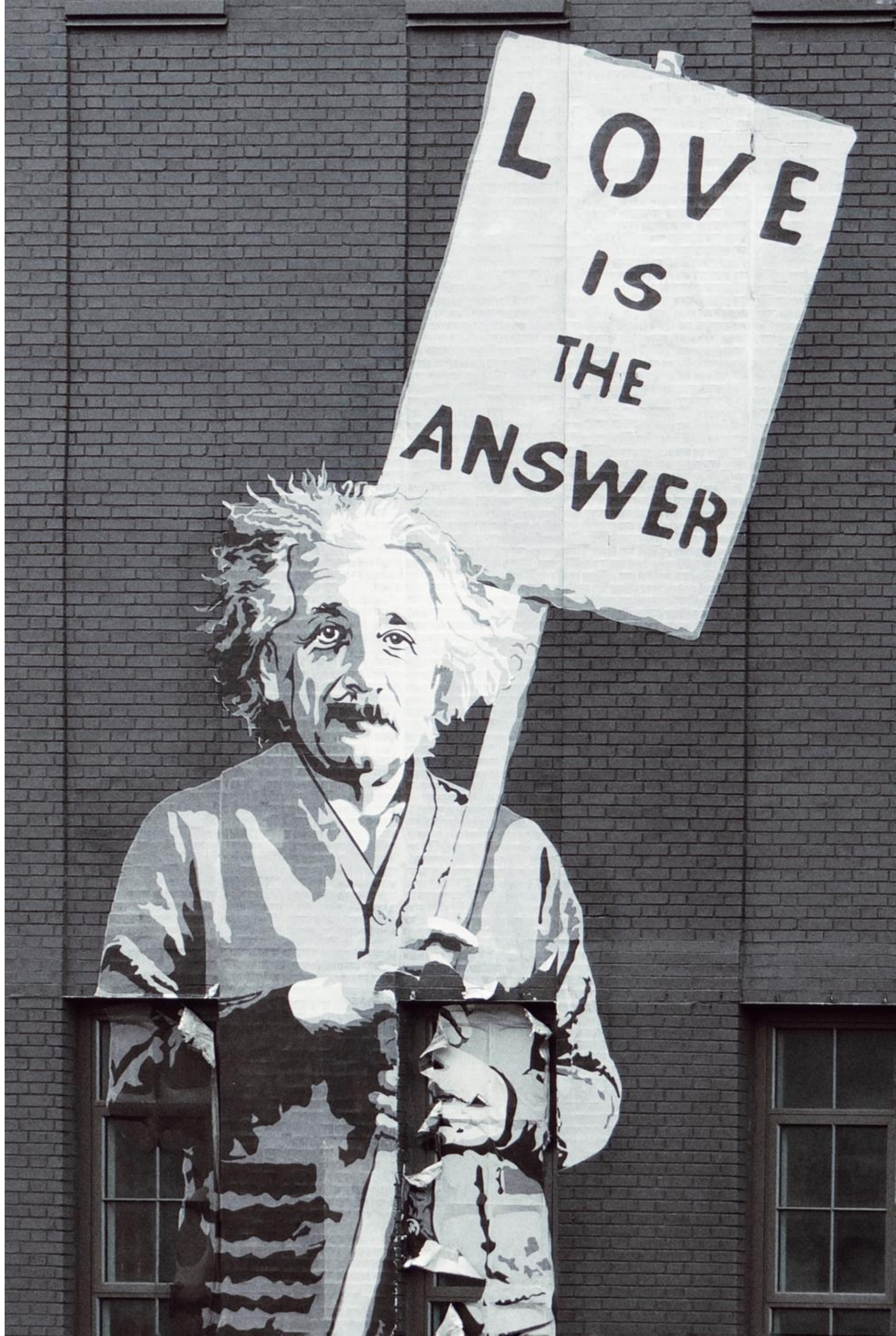
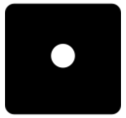


Photo by Hannes Richter on Unsplash

We can roll a dice several times
and add the results

We can roll a dice several times and add the results

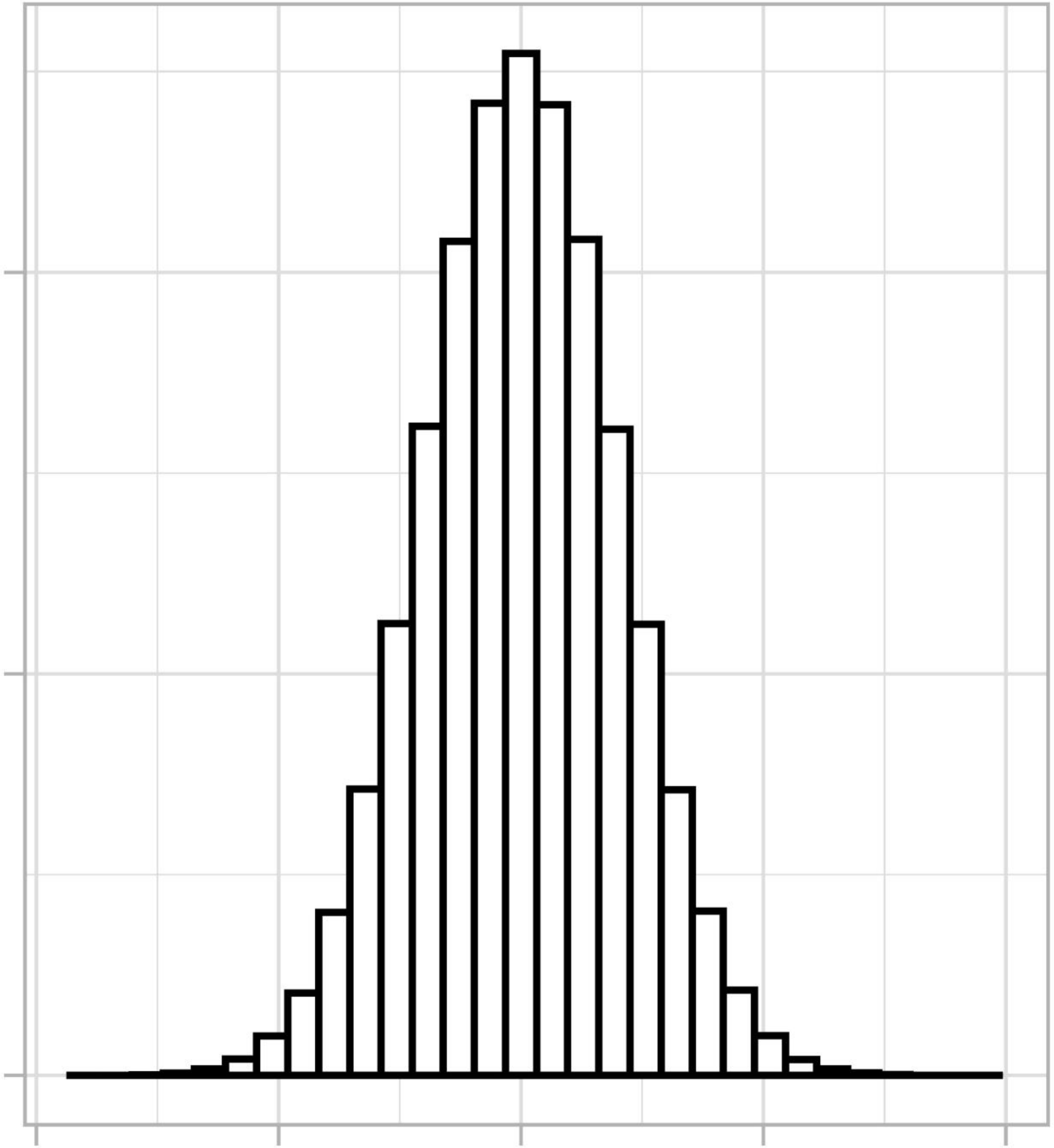
 +  +  +  +  +  = 16

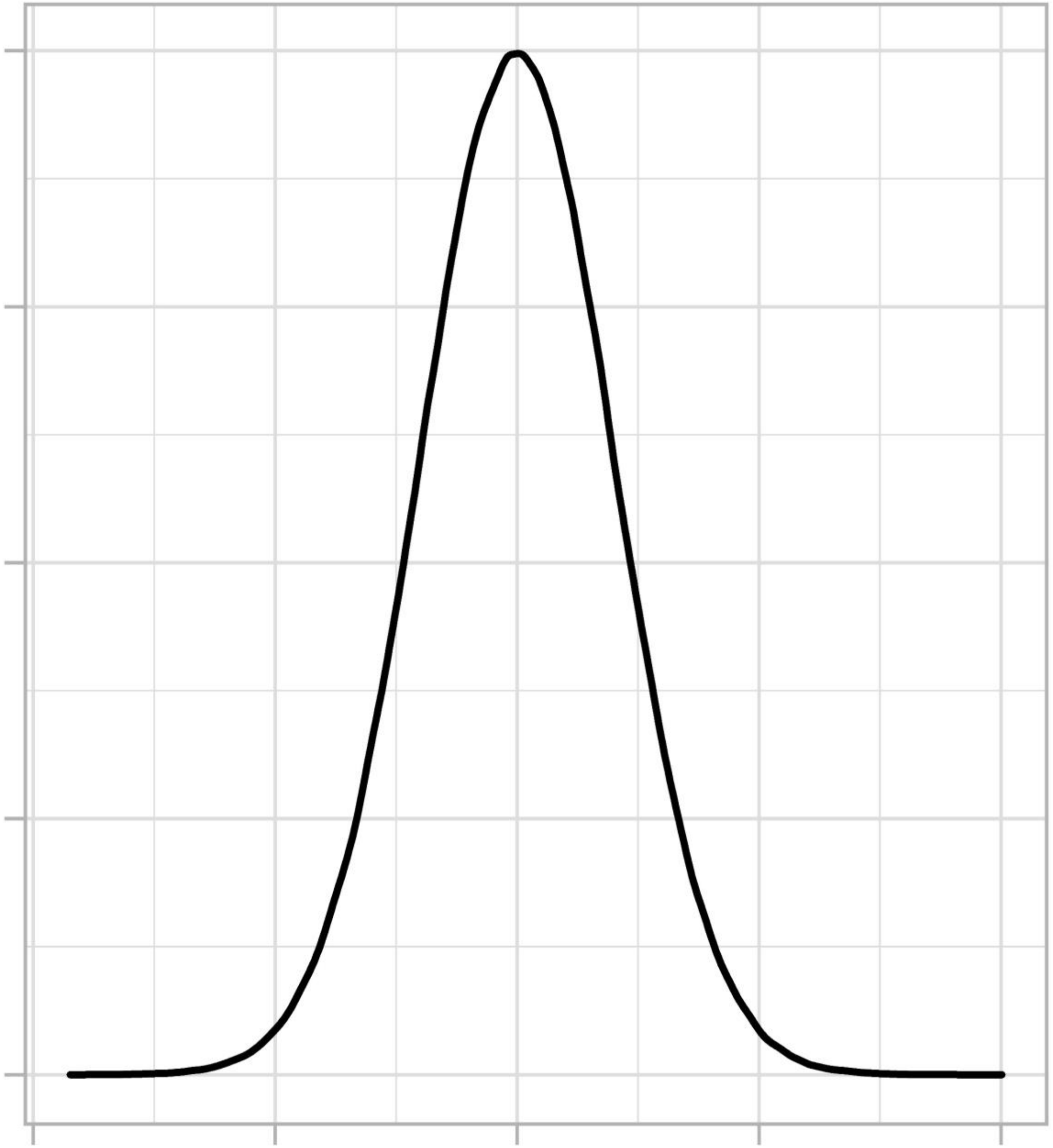
We can roll a dice several times and add the results

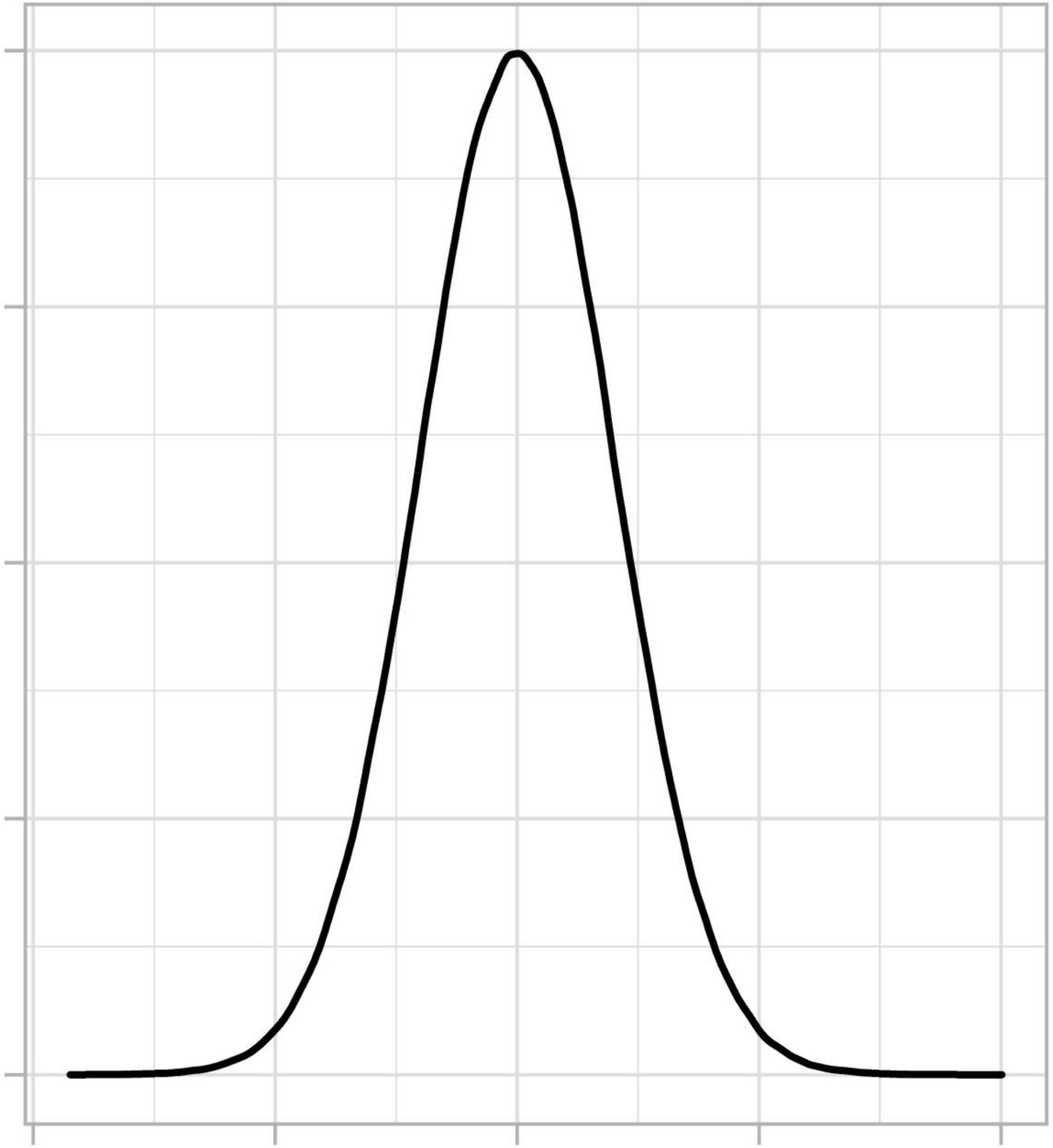
⚂ + ⚀ + ⚂ + ⚂ + ⚃ + ⚁ = 16

⚅ + ⚅ + ⚃ + ⚂ + ⚀ + ⚂ = 23

⚀ + ⚁ + ⚁ + ⚃ + ⚂ + ⚂ = 15







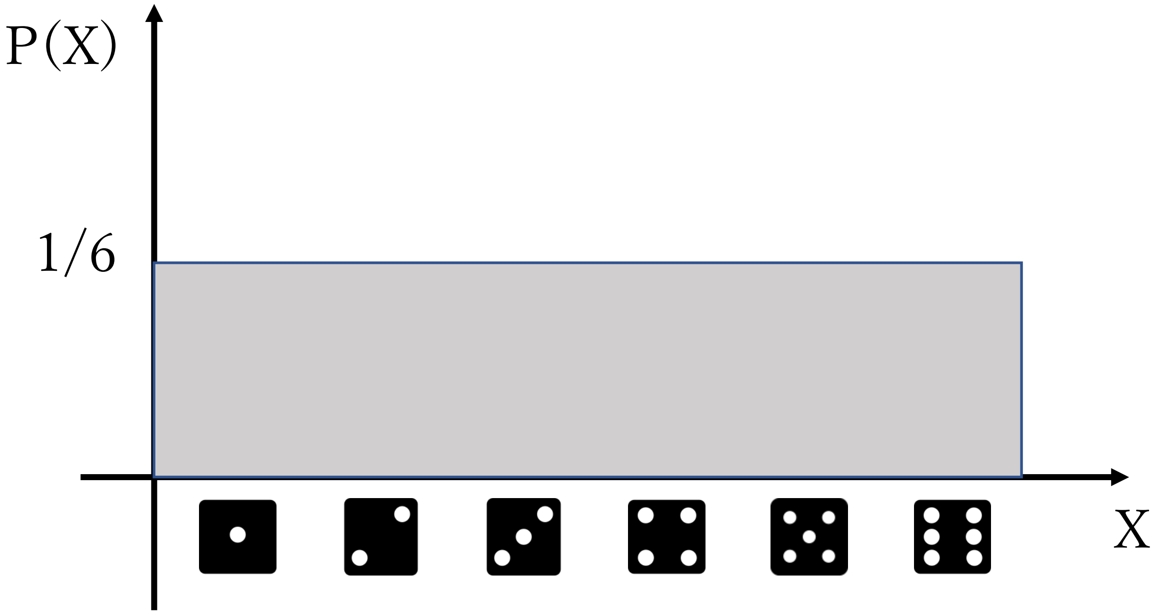
This pattern will emerge regardless of the original distribution

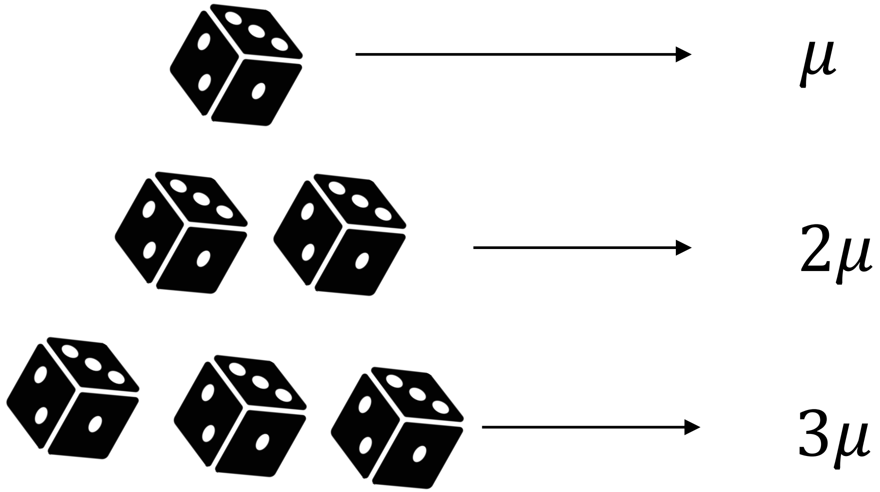
$$\mu = E[X] = \sum_x P(X = x) * x$$

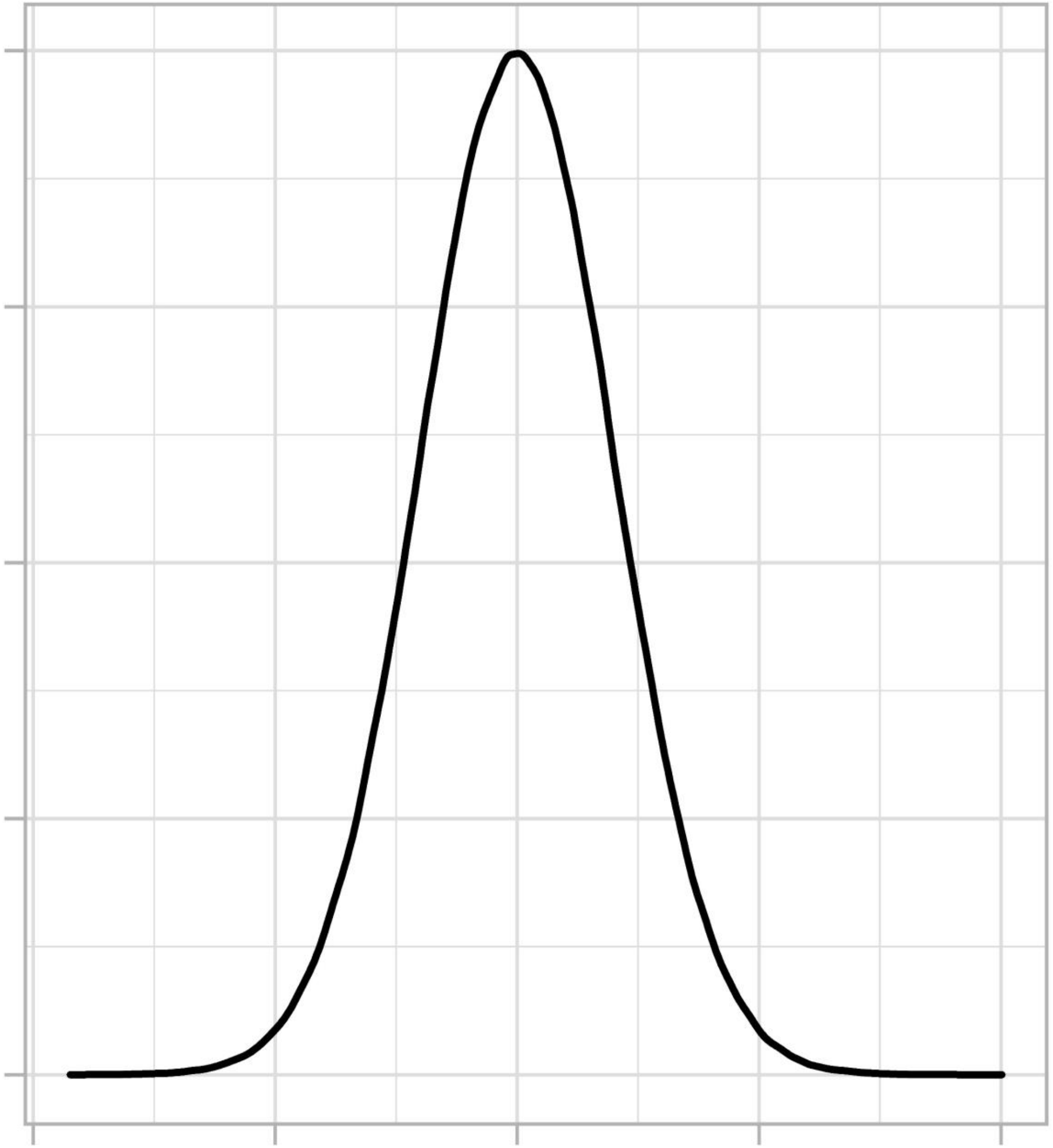
$$Var(X) = E[(X - \mu)^2]$$

$\sigma = \sqrt{Var(X)}$: Standard deviation

In the original distribution of X
(in this case, rolling the dice):

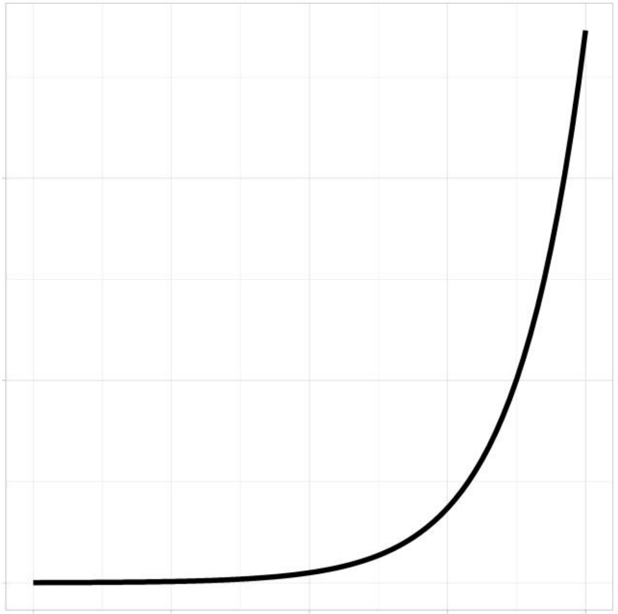




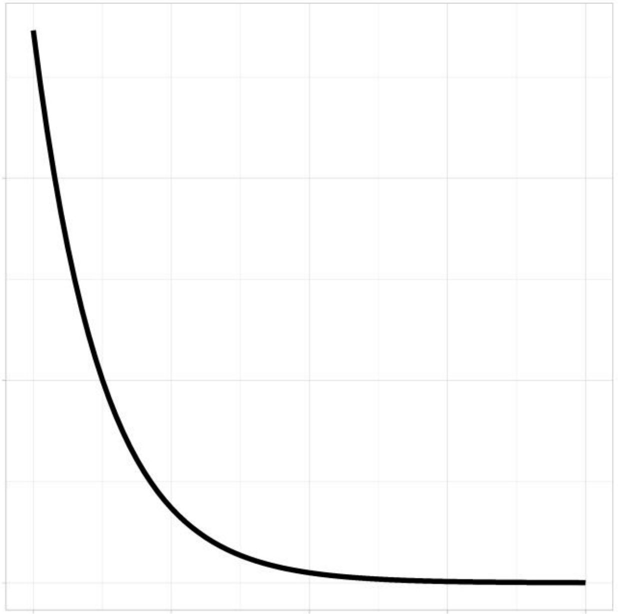


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

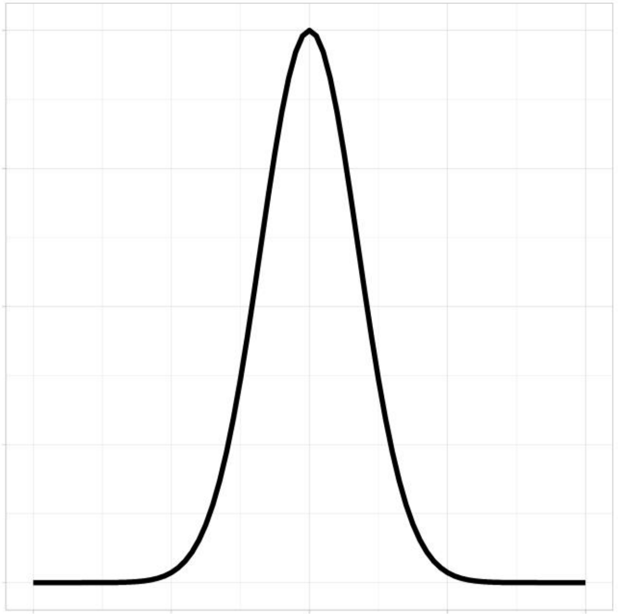
e^{x}: *Exponential growth*



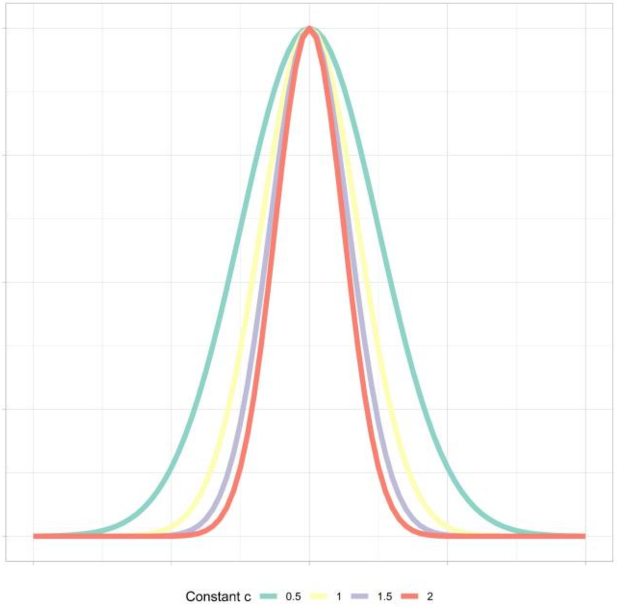
e^{-x}: *Exponential decay*



e^{-x^2}: *Bell shape*



e^{-cx^2}



$$e^{-x^2} \longrightarrow \frac{1}{\sqrt{\pi}} e^{-x^2}$$

Given that

$Area = \sqrt{\pi}$: area under curve

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

Given that

$$\frac{1}{\sqrt{\pi}} e^{-x^2} = \frac{1}{\sqrt{\pi}} e^{-x^2} * \frac{\sigma\sqrt{2}}{\sigma\sqrt{2}}$$

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

Given that

$$e^{-x^2} = e^{-\frac{x^2\sigma^2}{\sigma^2}} \qquad e^{ab} = e^{a^b}$$

This is a valid probability distribution

Central Limit Theorem CLT

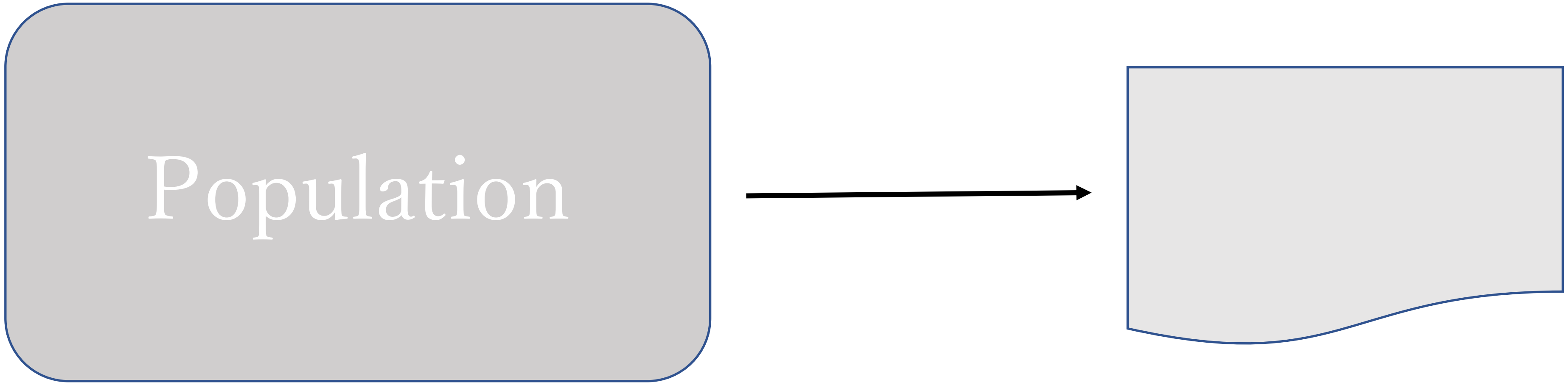
The distribution of sample means from any population approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution

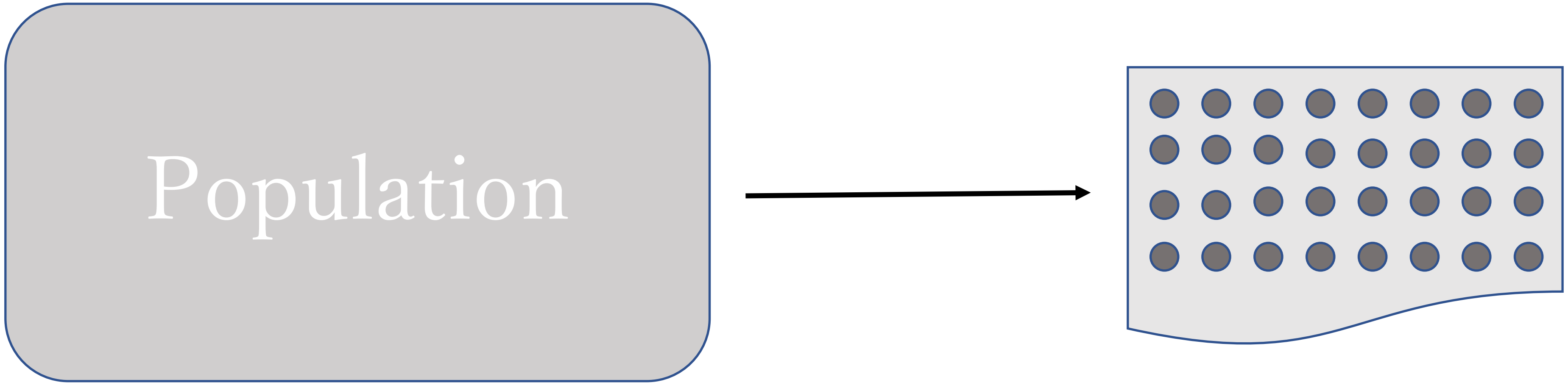
Three assumptions

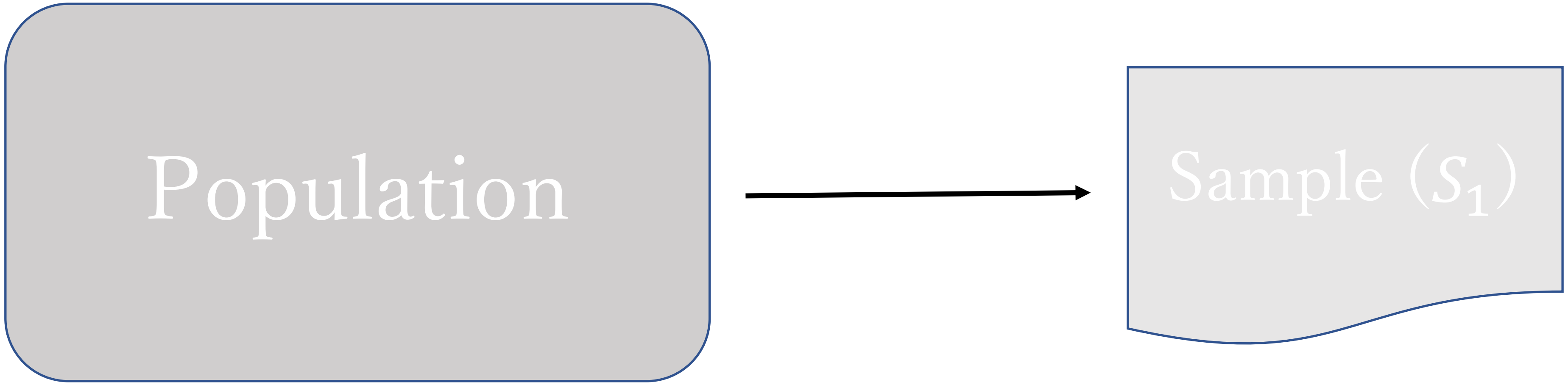
- All Xi are independent from each other
- Each Xi is drawn from the same distribution
- Variance is between 0 and infinite

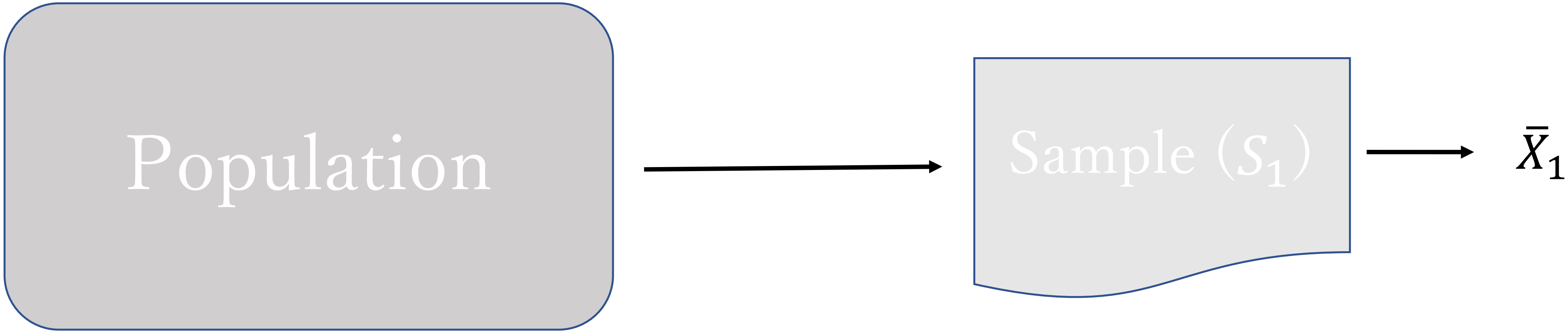
Population

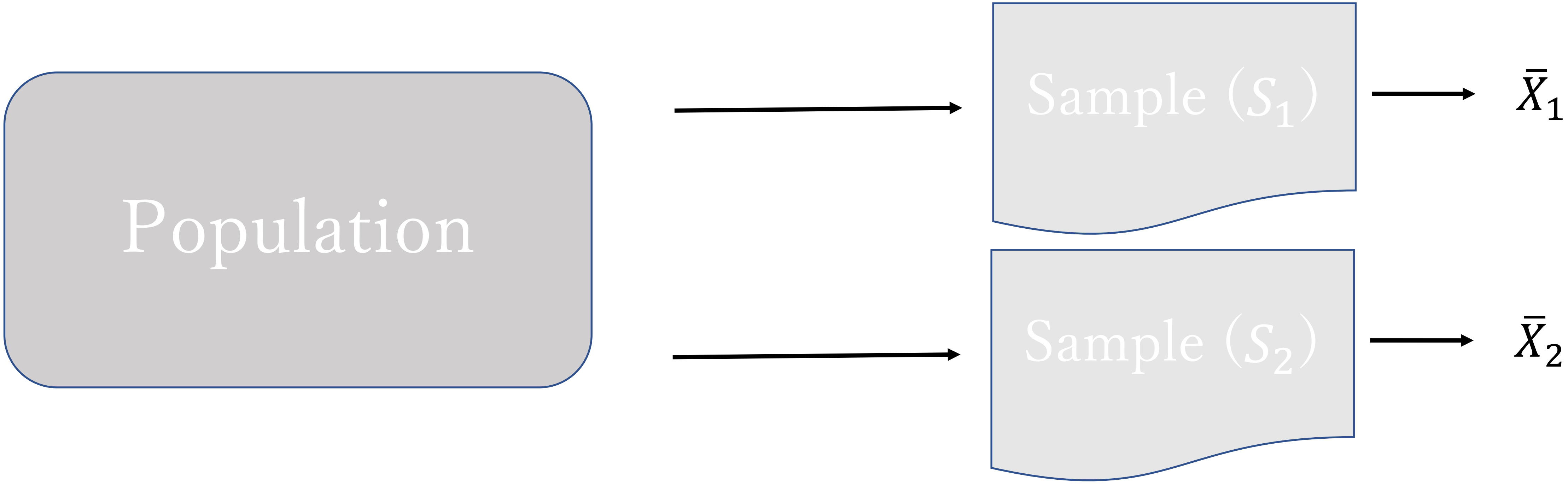
Select a sample (batch, lot, group) with sample size N

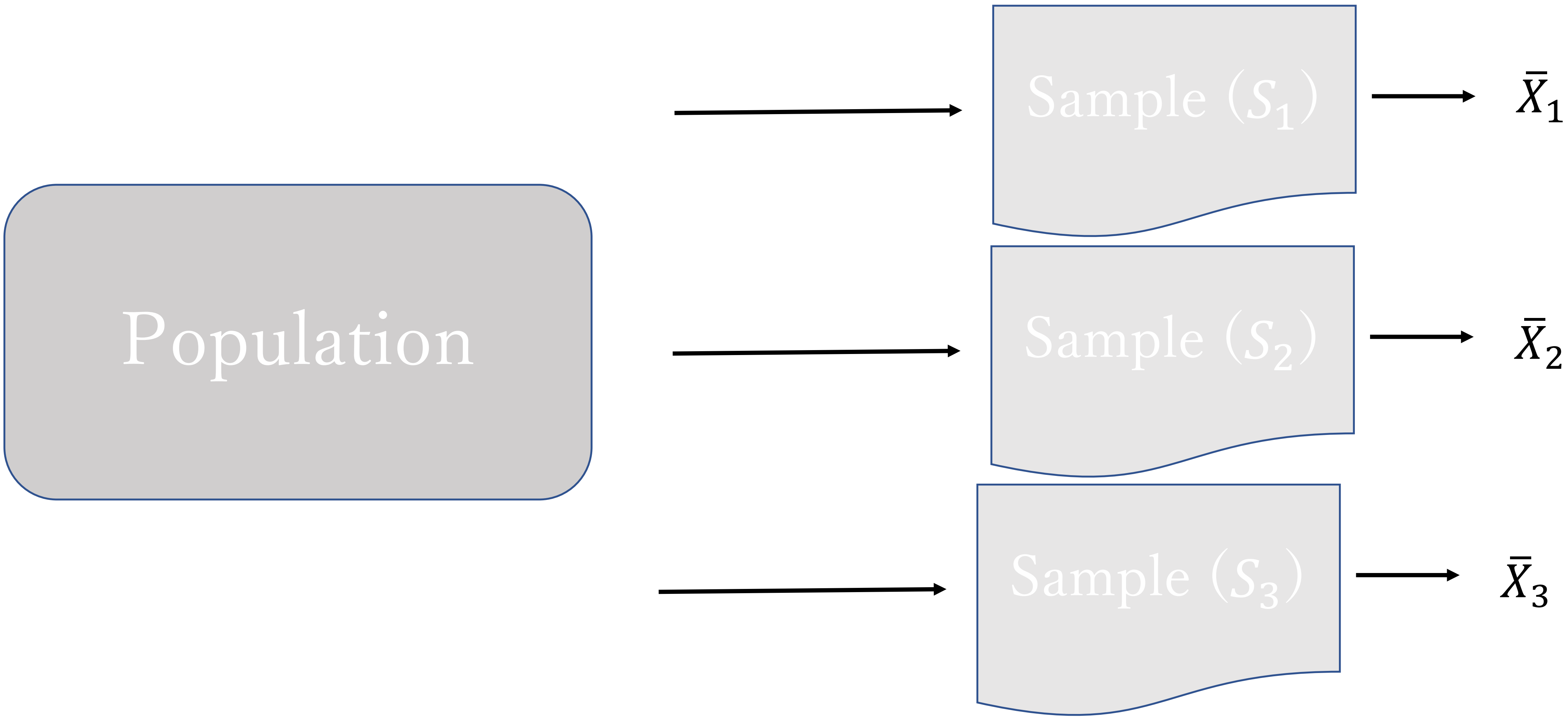


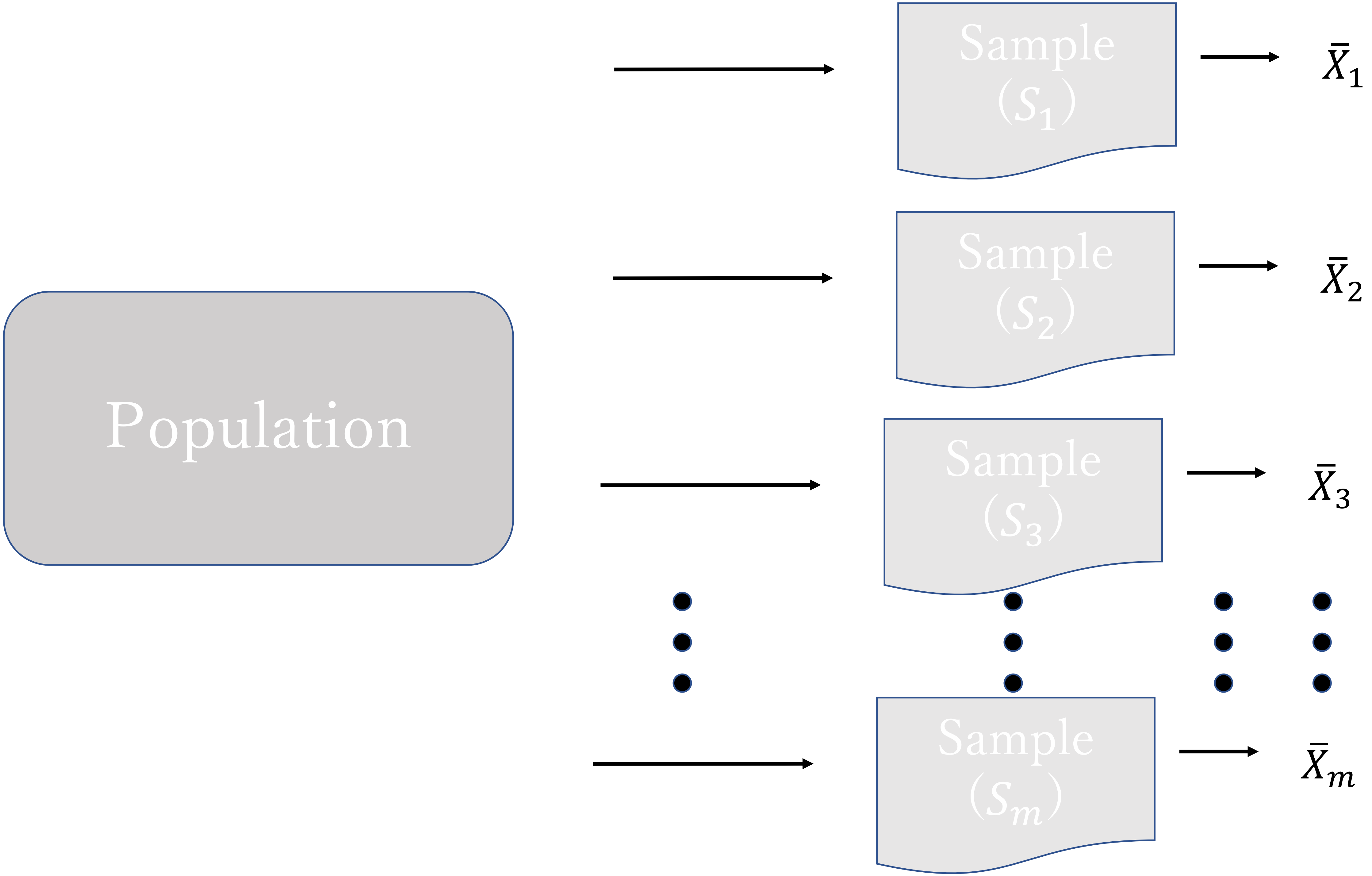

Population


Population
Sample (S_1)


Population
Sample (S_1)
$\bar{X}_1$


Population
Sample (S_1)
$\bar{X}_1$
Sample (S_2)
$\bar{X}_2$


Population
Sample (S_1)
Sample (S_2)
Sample (S_3)
$\bar{X}_1$
$\bar{X}_2$
$\bar{X}_3$


Population
Sample
(S_1)
$\bar{X}_1$
Sample
(S_2)
$\bar{X}_2$
Sample
(S_3)
$\bar{X}_3$
Sample
(S_m)
$\bar{X}_m$

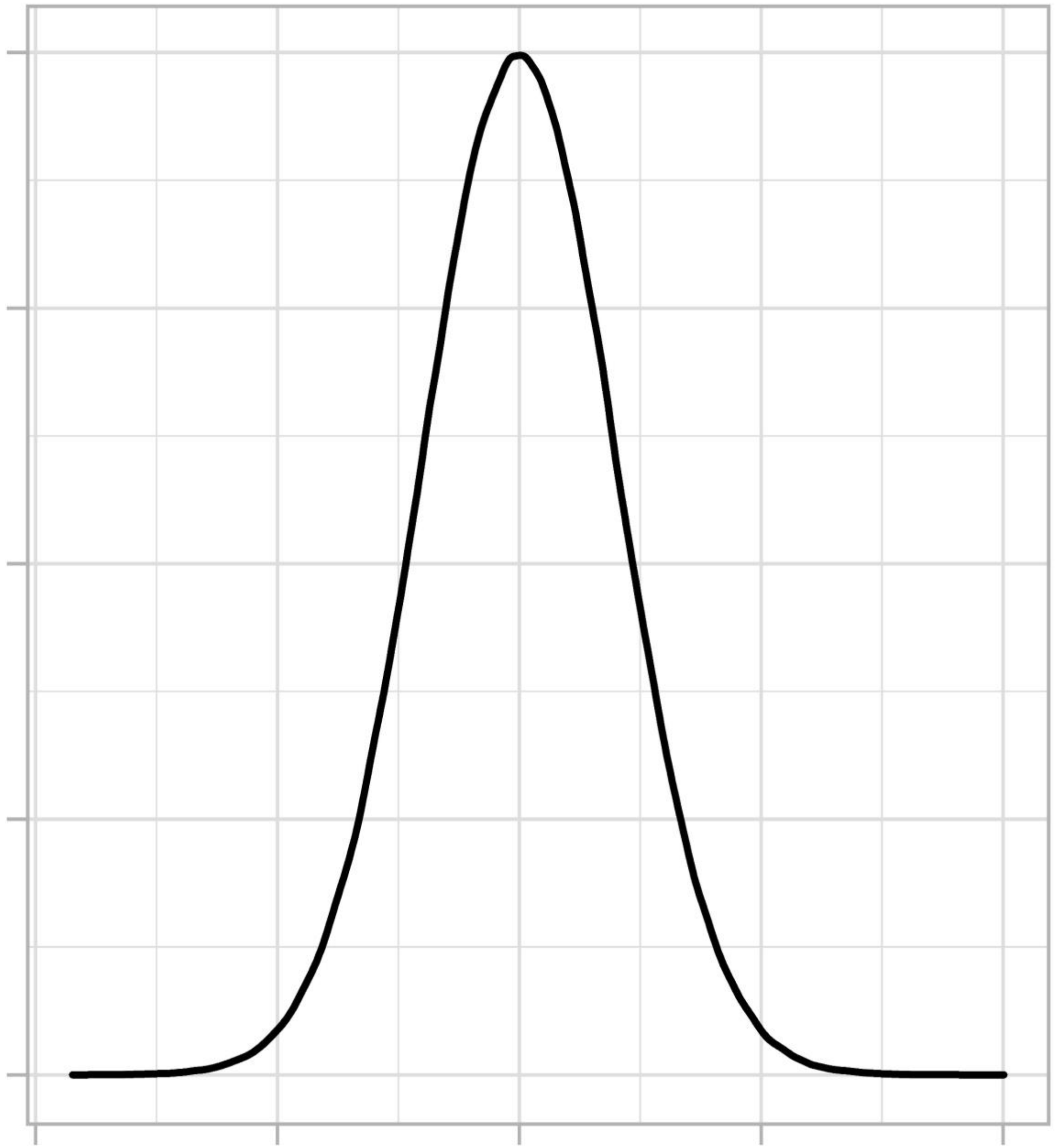
$$\{\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots, \bar{X}_m\}$$

We draw “m” samples from the population and compute “m” mean values

$$\{\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots, \bar{X}_m\}$$

We draw "m" samples from the population and compute "m" mean values

What do you think a histogram might look like?



One last example

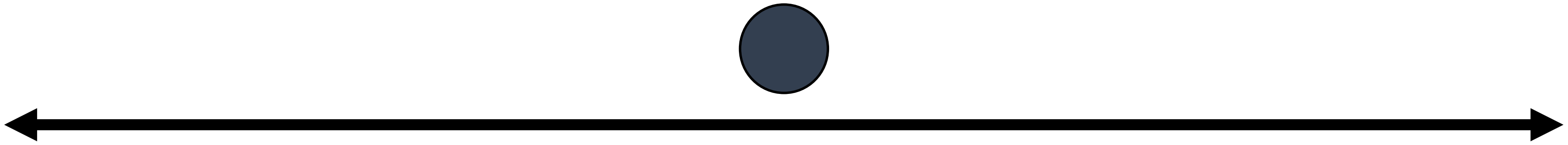
Average weight in the city?

Average weight in the city?

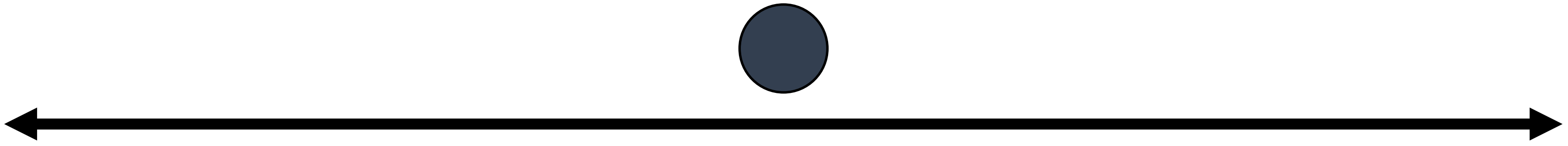
Sample mean is 80 Kg

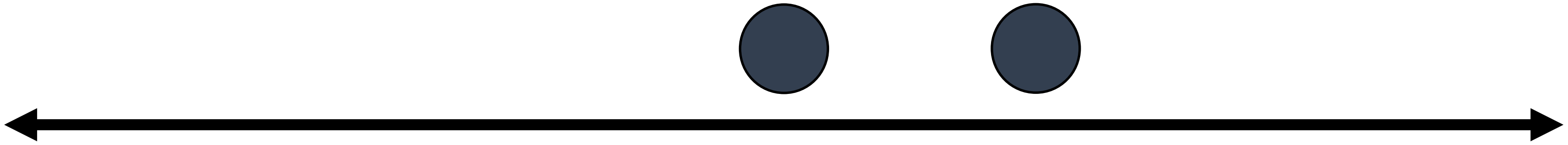
Sample mean is 80 Kg

Sample mean is 80 Kg

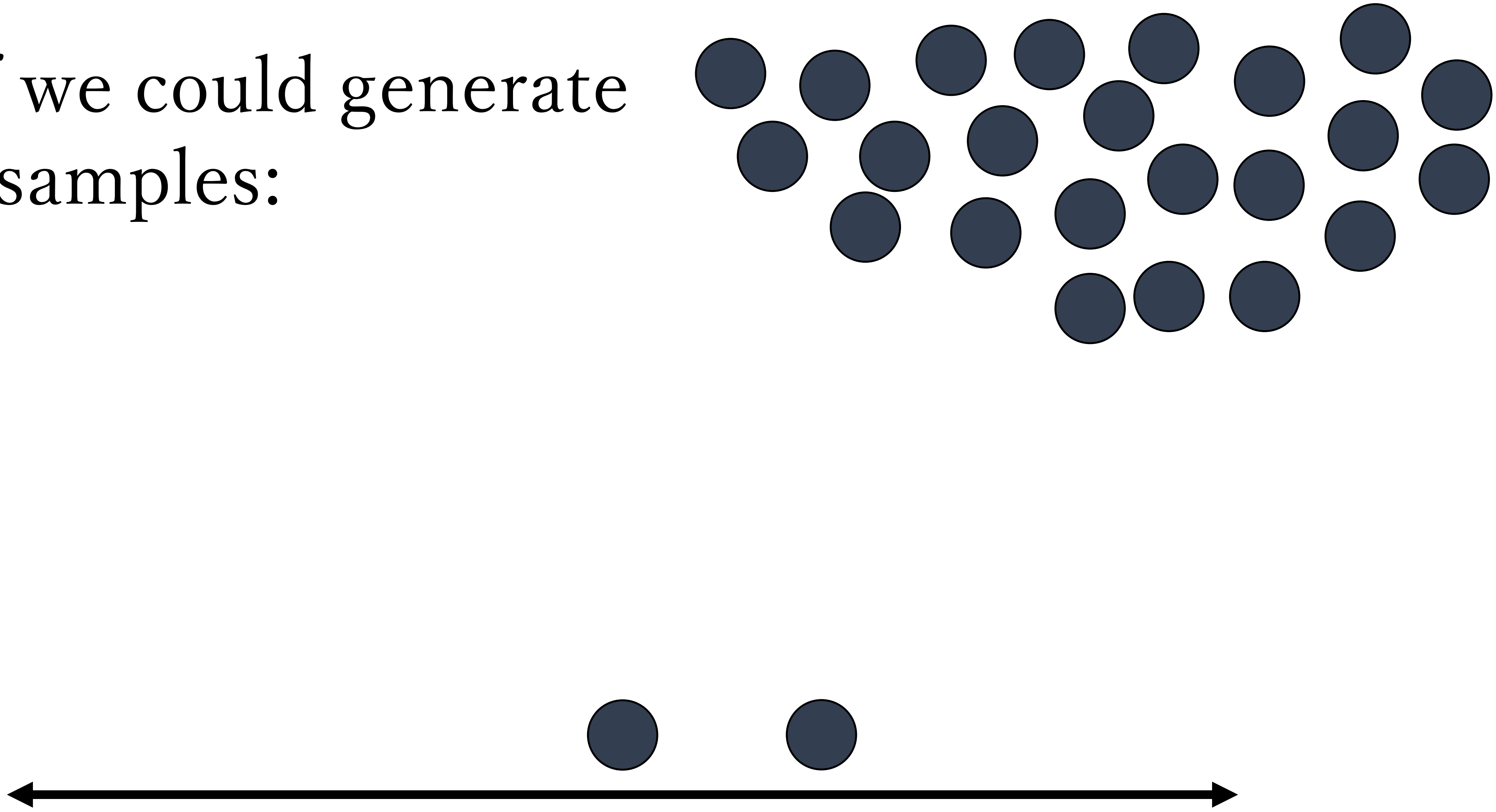


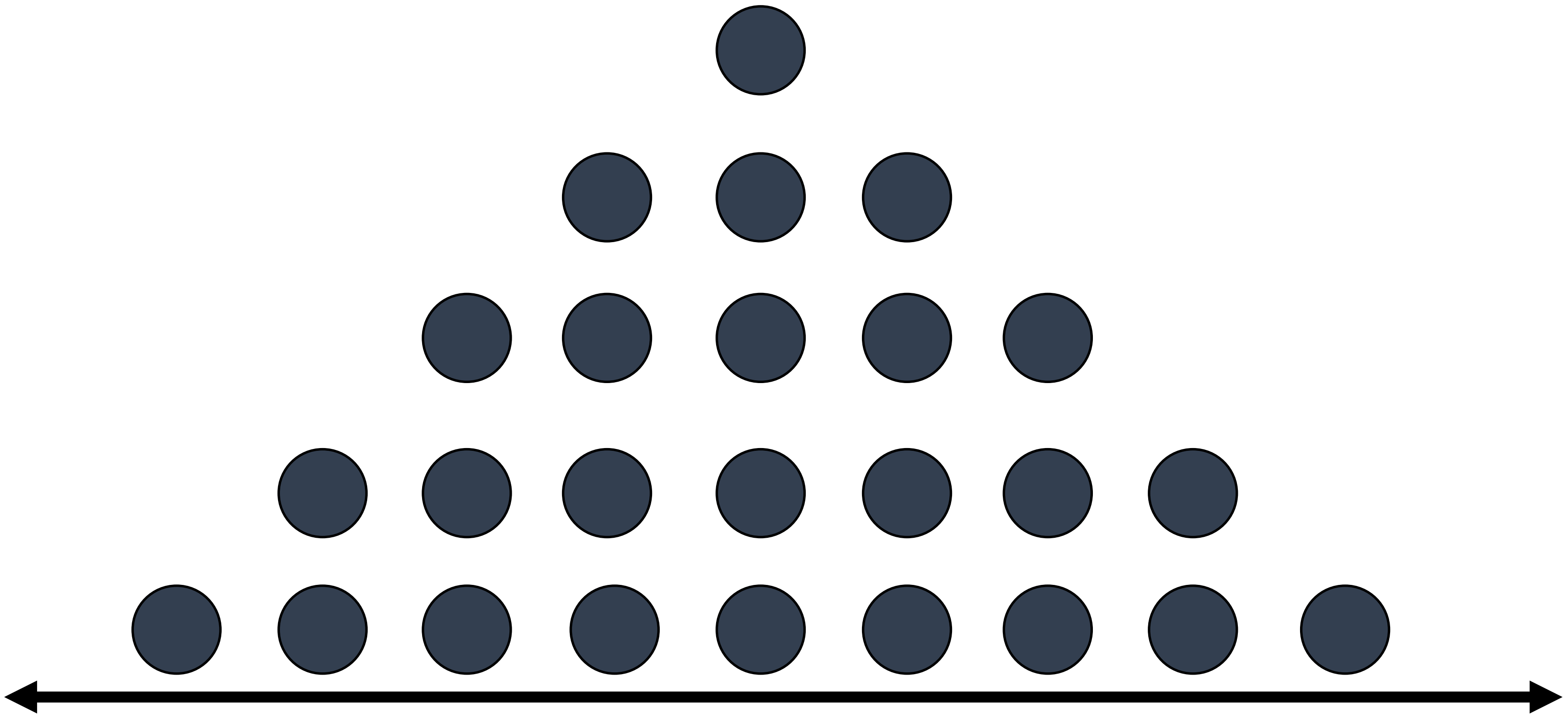
Hypothetically, if you could generate another sample:





And if we could generate many samples:





We need a way to test if two categorical variables are associated

We need a way to test if two categorical variables are associated

But first we need to revisit two concepts:

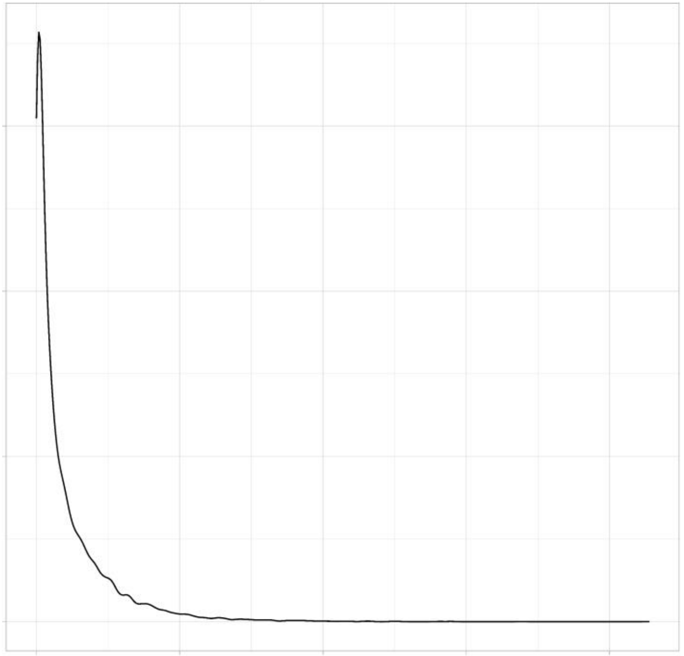
- Hypothesis testing
- Chi-squared distribution

Chi-squared distribution

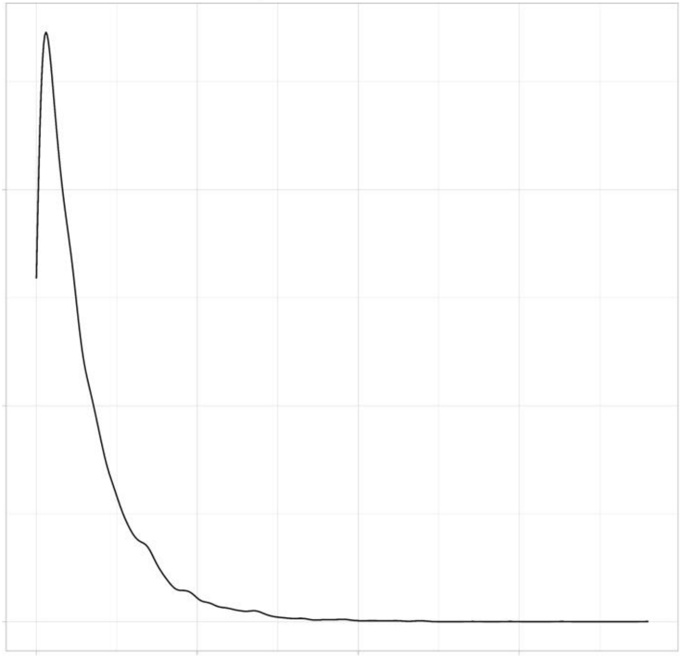
Chi-squared distribution

- Describes the distribution of the sum of squared standard normal random variables.
- The Chi-squared distribution is defined by a single parameter, which is called the degrees of freedom (df). The degrees of freedom represent the number of independent standard normal random variables that are squared and summed to obtain the Chi-squared random variable.
- As the degrees of freedom increase, the Chi-squared distribution becomes more and more similar to a normal distribution

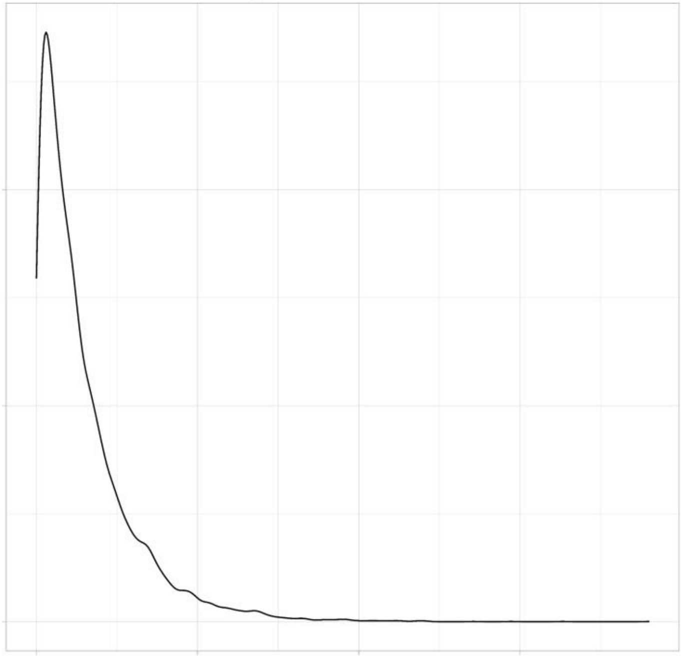
Chi-squared distribution with 1 df



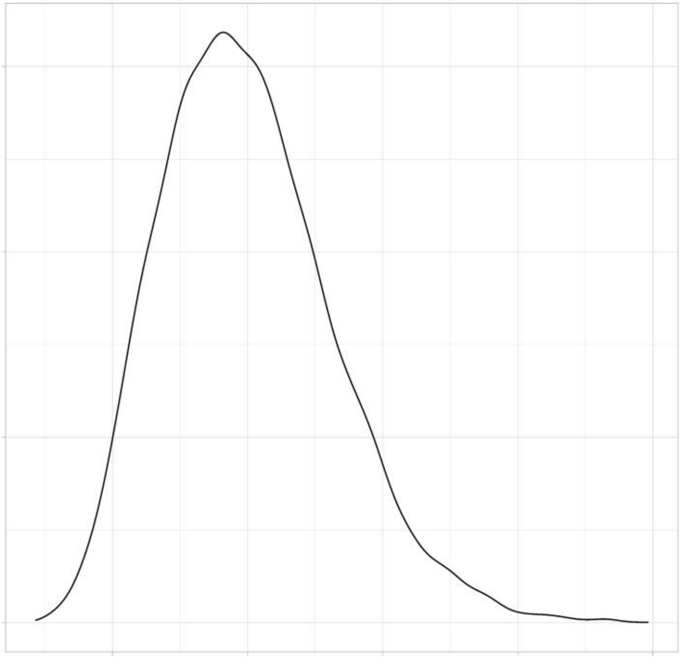
Chi-squared distribution with 3 df



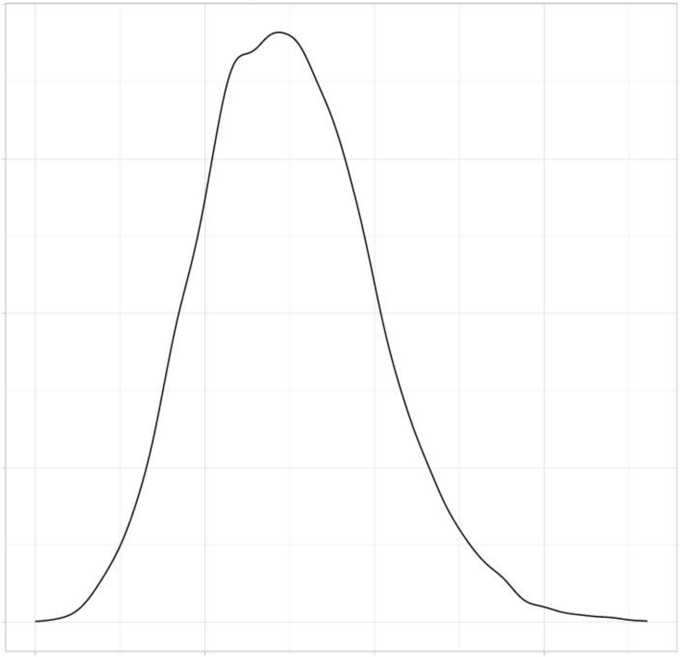
Chi-squared distribution with 10 df



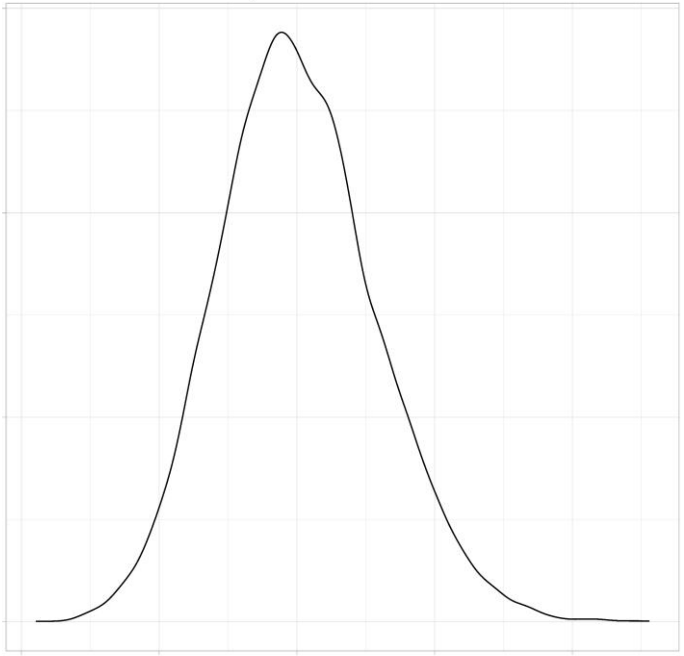
Chi-squared distribution with 20 df



Chi-squared distribution with 50 df



Chi-squared distribution with 100 df



Back to the association test of categorical variables…

You already imagine that we solve this by using a Chi-squared hypothesis test

Contingency table (an example)

Education : {Non-university, University}
Income : {Low, Medium, High}

Contingency table

Education : {Non-university, University}
Income : {Low, Medium, High}

Education	Income
Non-university	Low
University	High
University	Low
Non-university	High
⋮	⋮
Non-university	Low

Contingency table

Education : {Non-university, University}
Income : {Low, Medium, High}

Education	Income
Non-university	Low
University	High
University	Low
Non-university	High
⋮	⋮
Non-university	Low

	Low	**Mediun**	**High**
Non-university	300	200	100
University	100	100	200

Contingency table

	Low	Mediun	High	
Non-university	300	200	100	600
University	100	100	200	400
	400	300	300	1000

Contingency table

	Low	Mediun	High	
Non-university	300	200	100	600
University	100	100	200	400
	400	300	300	1000

Null: Education and income are not related
Alternative: Education and income are related

f_c: *Frequency of the colum*

f_r: *Frequency of the row*

n: *Sample size*

f_c: *Expected frequency*

f_c: *Observed frequency*

df : *Degrees of freedom*

χ^2: *test statistic*

$$f_e = \frac{f_c * f_r}{n}$$

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$df = (rows - 1) * (columns - 1)$$

Contingency table

	Low	Mediun	High	
Non-university	300(240)	200(180)	100(180)	600
University	100(160)	100(120)	200(120)	400
	400	300	300	1000

Null: Education and income are not related
Alternative: Education and income are related

$$f_e = \frac{f_c * f_r}{n}$$

Non-university – Low : 240
Non-university – Medium : 180
Non-university – High : 180

University – Low : 160
University – Medium : 120
University – High : 120

Contingency table

	Low	Mediun	High	
Non-university	300(240)	200(180)	100(180)	600
University	100(160)	100(120)	200(120)	400
	400	300	300	1000

Null: Education and income are not related
Alternative: Education and income are related

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2 = \frac{(300-240)^2}{240} + \frac{(280-180)^2}{180} + \frac{(100-180)^2}{180} + \frac{(100-160)^2}{160} + \frac{(100-120)^2}{120} + \frac{(200-120)^2}{120}$$

$$\chi^2 = 185.2778$$

$$df = 2$$

$$the\ p\ value\ is\ very\ close\ to\ zero\ (p < 0.0001)$$

$$the\ critical\ value\ for\ an\ alpha\ of\ 5\%\ is\ 5.991$$

$$qchisq(0.95, 1)$$

As I see it:

Every cell is considered as a squared normal distribution. The squared is calculated considering the differences between the observed and expected value

Adding all cells we get a Chi-squared distribution

So, we assume not association (the null), meaning that values in the cells are independent and random. And, if the probability of observing the data (of producing the data) is high, then the null holds. But, if the probability is low (p value < 0.05), then the null does not hold (we reject it), and we lean towards the alternative (there should be some association)

A (personal) note on the formula

Pearson originally argued that the degrees of freedom were “rc-1”.
Fisher noted that it was (r-1)(c-1).

Because of the degrees of freedom

What really are “degrees of freedom”?

Photo by Aarón Blanco Tejedor on Unsplash

Pearson response:

““I hold that such a view (Fisher’s) is entirely erroneous, and that the writer has done no service to the science of statistics by giving it broad-cast circulation in the pages of the Journal of the Royal Statistical Society. … I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us.”

From “An introduction to categorical data analysis” by Alan Agresti

Pearson response:

"“I hold that such a view (Fisher’s) is entirely erroneous, and that the writer has done no service to the science of statistics by giving it broad-cast circulation in the pages of the Journal of the Royal Statistical Society. … I trust my critic will pardon me for comparing him with Don Quixote tilting at the windmill; he must either destroy himself, or the whole theory of probable errors, for they are invariably based on using sample values for those of the sampled population unknown to us.”

From “An introduction to categorical data analysis” by Alan Agresti

By the way, Fisher was right.

Take away message: Do not worry if you do not completely understand everything. Even Pearson struggled to understand what is supposed to be a straightforward idea

Limitations

It only test for association. It does not answer all questions
It ignores the order of the categories

Wrap up (Chi-squared test of association):

You want to know if two categorical variables are associated

↓

Null Hypothesis: there is not an association
Alternative Hypothesis: there is an association

↓

You run a Chi-squared test of association on a contingency table
Function chisq.test() in R

↓

If p value => 0.05, you conclude that there is not an association
If p value < 0.05, you conclude that there is an association

Back to spatial autocorelation and the Moran's I

Spatial autocorrelation, so…

- We need to find a way to "measure" spatial autocorrelation
- And we need to find a way to assess if that measurement is correct (do we believe in it?)

We do this by using spatial autocorrelation indexes. The most famous is Moran's I

Unpacking Moran's I:

- Neighbours
- Mathematical form
- Hypothesis testing (we will take some time here)

Morans's I

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Null: spatial randomness, spatial pattern in the data have occurred by chance (we want to reject it)

The test involves generating a distribution of Moran's I values under the null hypothesis by randomly permuting the attribute values among the spatial locations while preserving the spatial structure. This process is repeated multiple times (e.g., 999 or 9,999 permutations) to create a reference distribution of Moran's I values

Wrap up - Morans's I

You want to know if there is spatial autocorrelation

↓

Null Hypothesis: spatial randomness

↓

You calculate the neighborhood structure and the Moran's I.
Functions poly2nb(), nb2listw, moran.test(), and moran.test()

↓

If p value => 0.05, you conclude that there is not an association
If p value < 0.05, you conclude that there is an association

Alternatives to Moran's I

Two other popular alternatives are Geary's C and Getis-Ord (Gi*)

Geary's: is calculated as the sum of squared differences between each pair of neighboring locations divided by the sum of squared differences for all locations. It ranges from 0 to 2, where values closer to 0 indicate positive spatial autocorrelation and values closer to 2 indicate negative spatial autocorrelation.

Getis-Ord: It evaluates whether individual locations have attribute values that are significantly clustered or dispersed compared to neighboring locations

Moran's I and Geary's C can handle binary data.
Moran's I is considered more powerful, except for binary data.

For categorical data: join count

Thank you

Orlando Sabogal-Cardona
PhD researcher
University College London UCL