L12: Hypothesis Testing or Comparing Two Samples

Eng 10200

Title: A Data Science and Statistical Approach to Programming

Instructor: Michael Grossberg

How do we know what is true?

A PRIORI REASONING (MATH BASED ON ASSUMPTIONS)

A POSTERIORI REASONING (EVIDENCE/OBSERVATIONS/EXPERIENCE)

A priori reasoning

- Socrates is human
- Humans are mortal
- Therefore: Socrates is mortal

Limited what we can prove this way (based on strong assumptions)

A Posteriori reasoning

- Make an observation
- Base a generalization on it …

See a pink submarine therefore all submarines are pink

A Posteriori reasoning

- Make an observation
- Base a generalization on it …

WRONG!

See a pink submarine therefore all submarines are pink

All this proves is

One side of one submarine appears pink

Common but wrong practice

- Step 1. Come up with a hypothesis
- Step 2. Collect evidence that will support your hypothesis
- If you can find lots of evidence that agrees with your hypothesis declare it supported

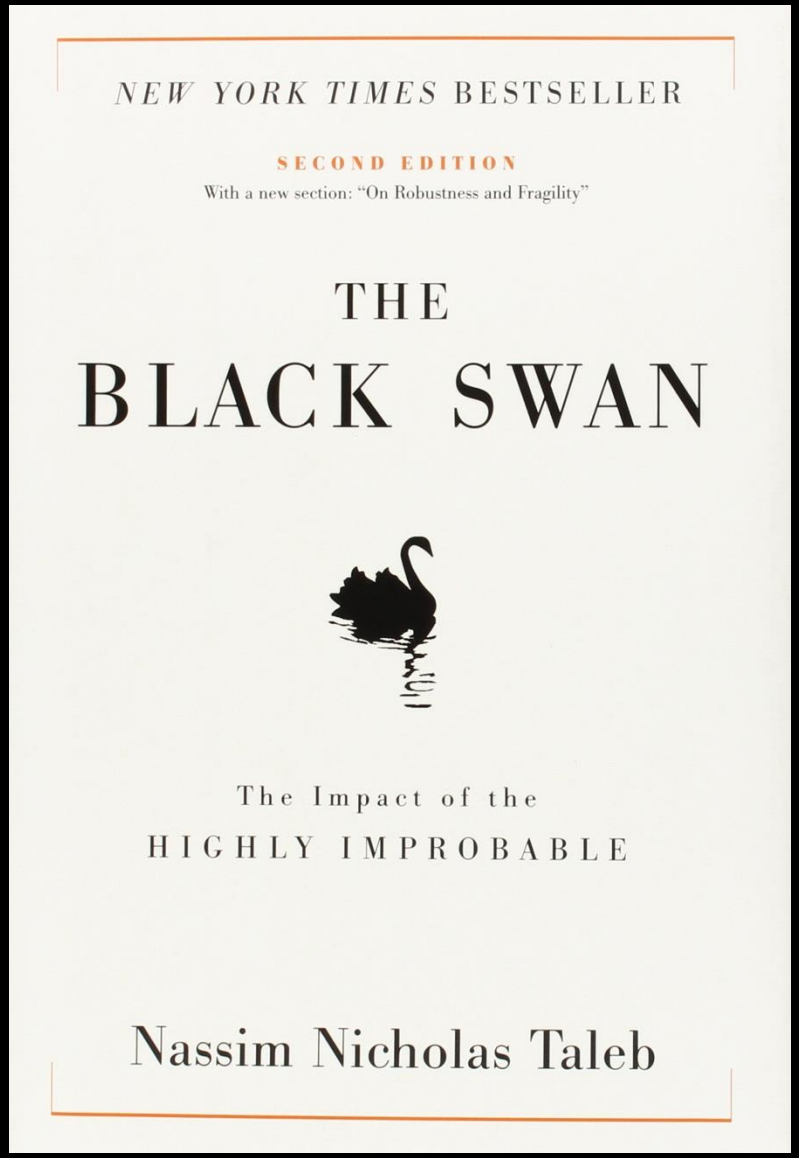
What is wrong?

Cherry picking

- If you look for evidence that supports your hypothesis you are most likely to find it
- Odd numbers are prime: 3, 5, 7, 11, 17 …..

The Black Swan

The Black Swan



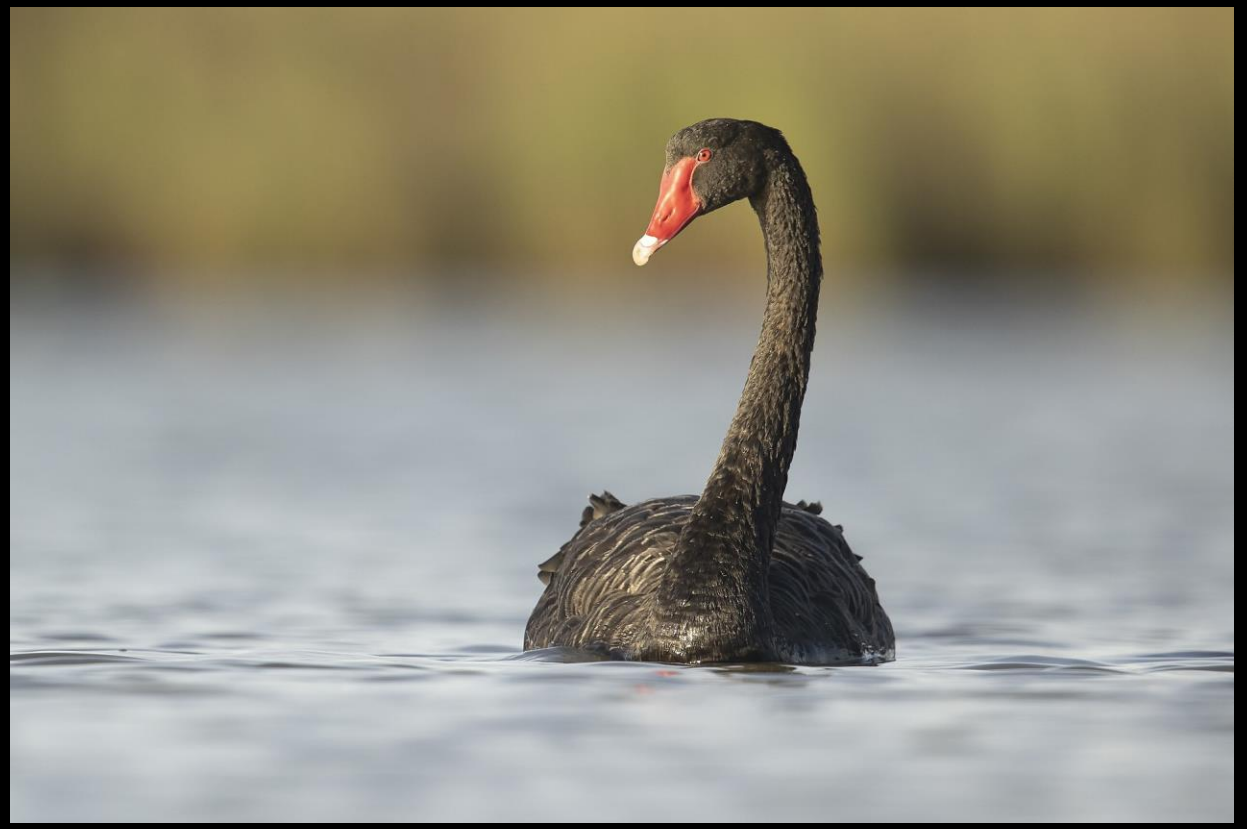
All swans are white

- Until 1790 Europeans believed that all swans were white

The term “black swan” meant something that was impossible

Europeans discover Black Swans

- In 1790 John Latham documents a black swan in Australia



Sherlock Holmes (Arthur Conan Doyle)

"How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?"

Almost

Formulate a hypothesis that can be falsified

Try very hard to falsify it

If every attempt to prove the hypothesis fails, there is supporting evidence

Hypothesis Testing in Statistics

- The null hypothesis: The deviation of a statistic from a unremarkable baseline, is coincidental (not statistically significant)

The birthday problem

- How many people do you need to have a greater than 50% probability of two people with the same birthday?

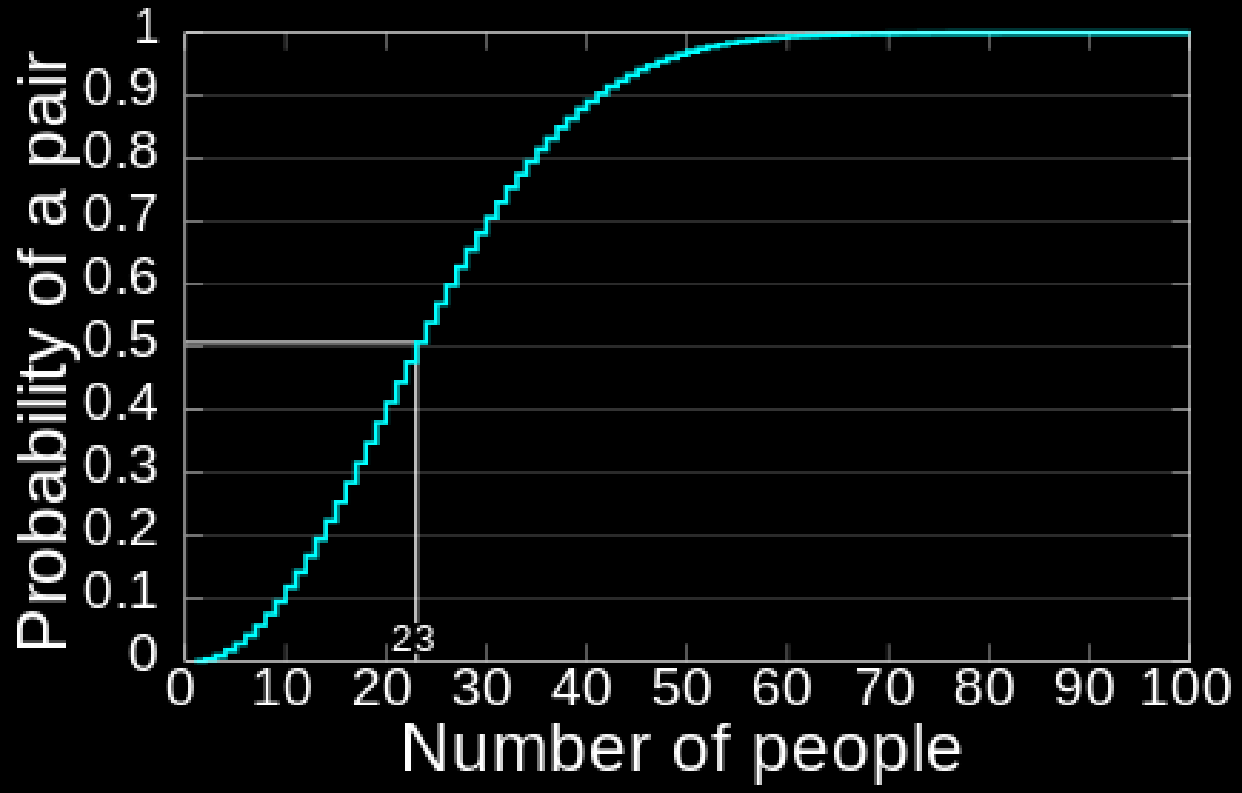
The Birthday Problem

$$V_{nr} = \frac{n!}{(n-k)!} = \frac{365!}{(365-23)!}$$

$$V_t = n^k = 365^{23}$$

$$P(A) = \frac{V_{nr}}{V_t} \approx 0.492703$$

$$P(B) = 1 - P(A) \approx 1 - 0.492703 \approx 0.507297(50.7297\%)$$



23

(Statistical) Model

We assume our data comes from some probability distribution

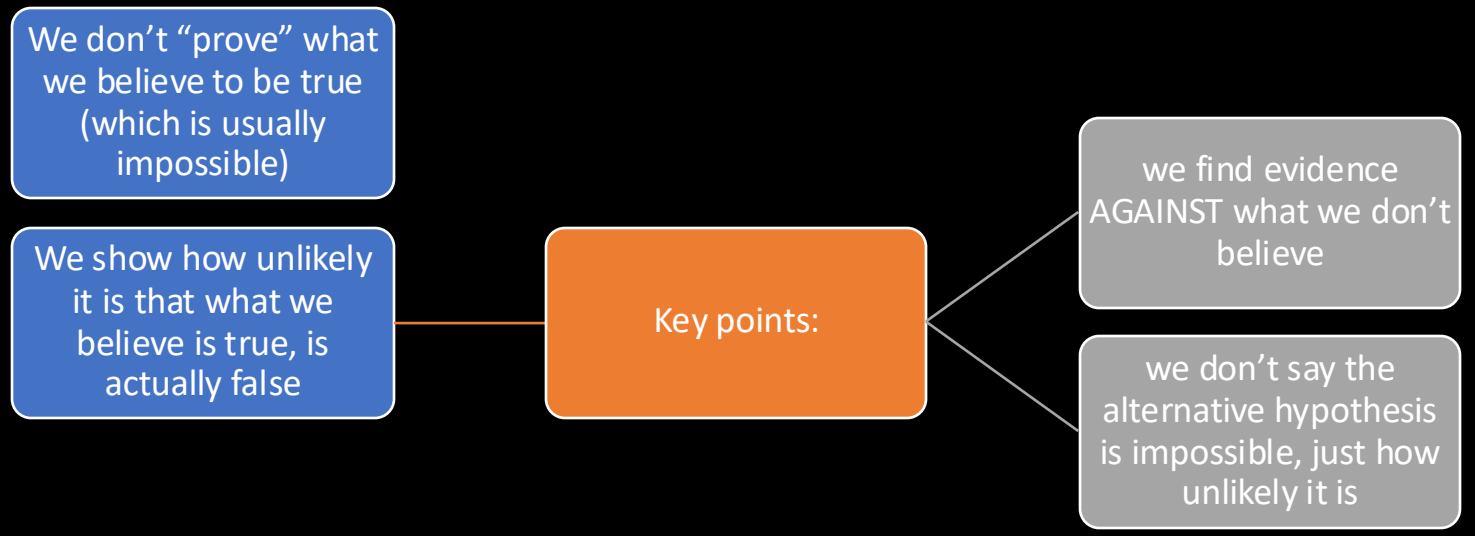
Prediction: If we look at more data, it will be distributed according to a hypothesized distribution

Inference: Estimating some or all the parameters of the distribution that the data is a sample of, from the data

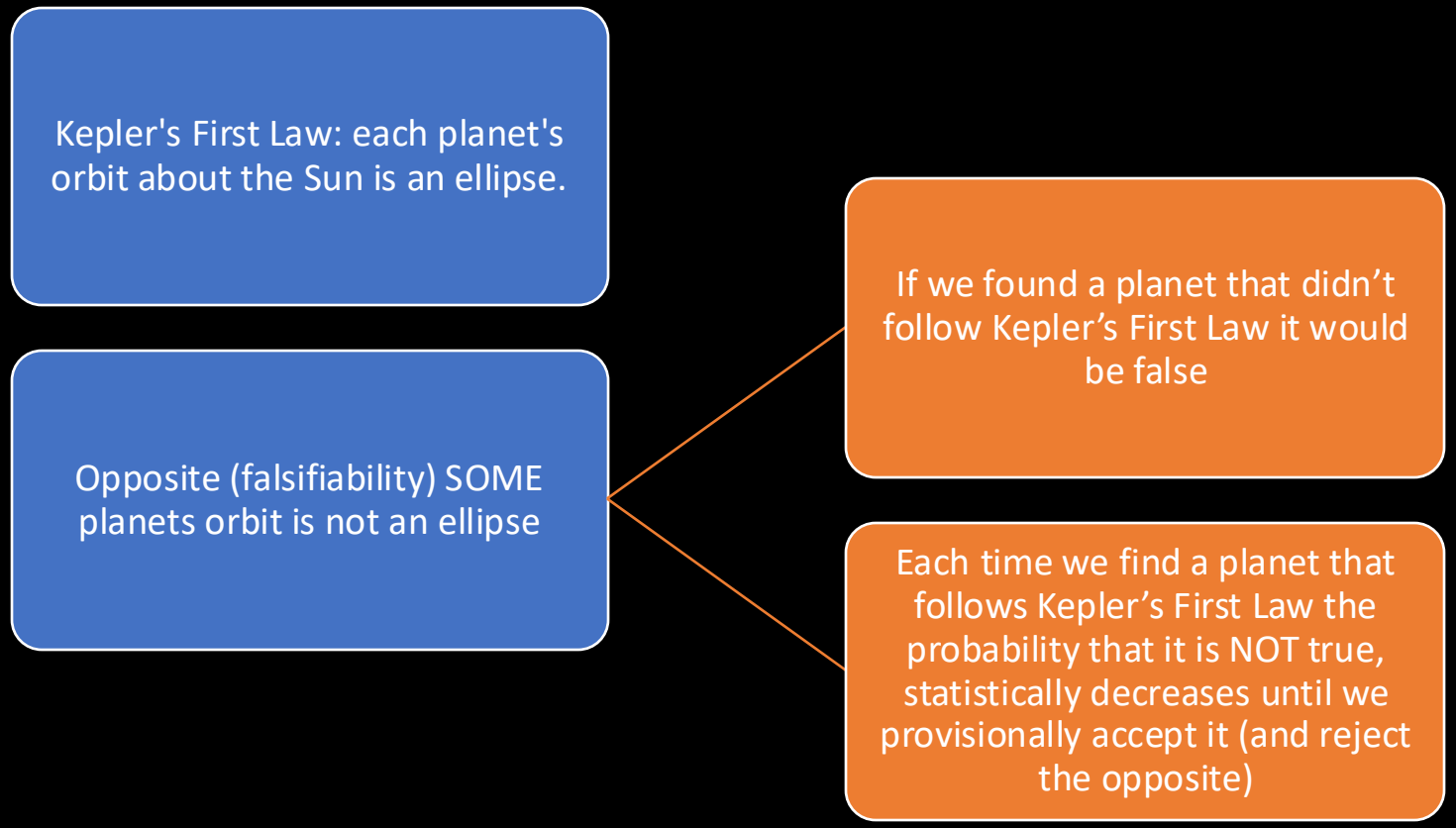
(statistical) Hypothesis

- A predicate on the probability distribution that underlies the data
  - Example: The pattern of heads and tails I see, come from a fair coin with a 50%/50% probability distribution
  - Example: The probability of being hired for a particular job is statistically independent of the gender of the applicant
  - Example: The probability distribution of a stock price today only depends on the price yesterday (random walk)

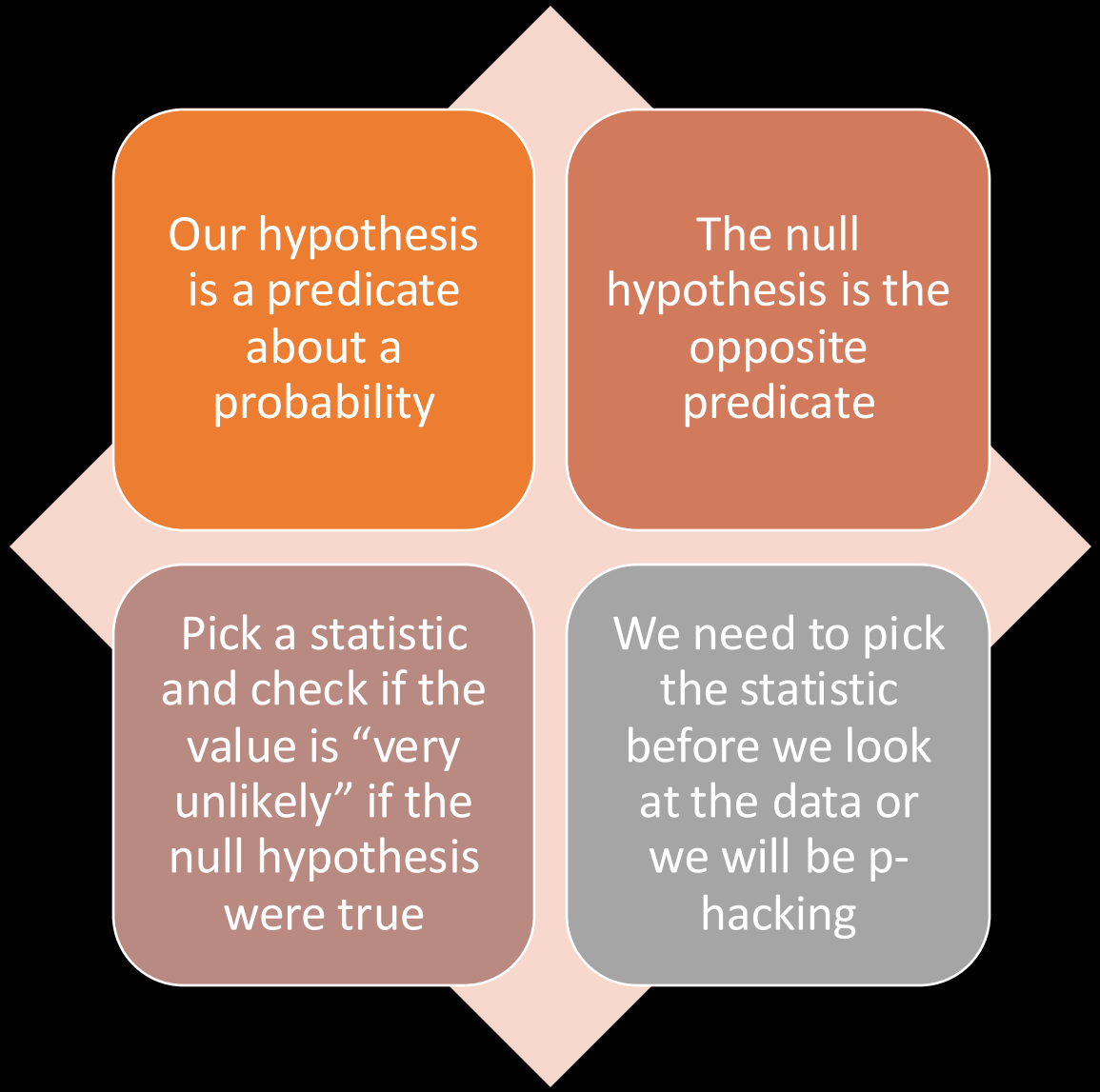
Hypothesis Test



A scientific hypothesis test

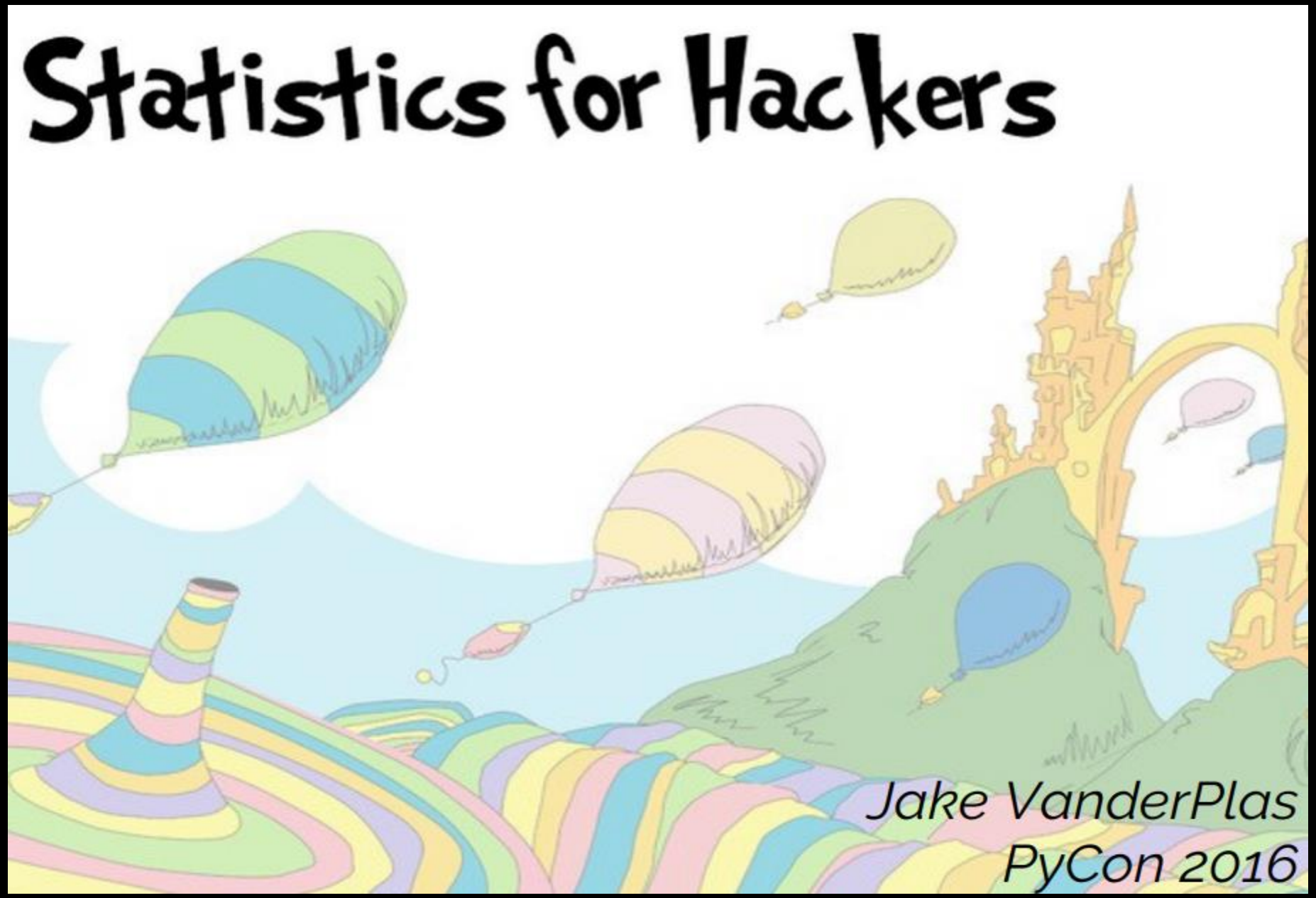


Statistical Hypothesis Test



Better Explanation

Statistics: Make the computer do the work



https://files.speakerdeck.com/presentations/518cae54da12460e895163d809e25933/15h15_-_Jake_Vanderplas_-_Statistics_for_Hackers-1bsngy0cnlmve.pdf

https://www.youtube.com/watch?v=Iq9DzN6mvYA

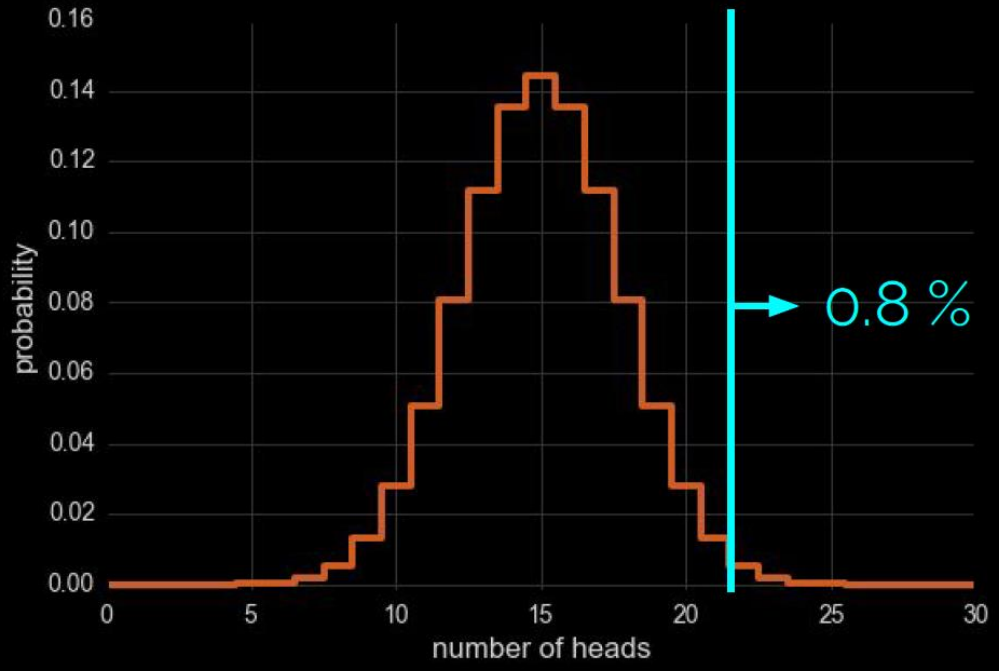
Warm-up:

Coin Toss

You toss a coin **30** times and see **22** heads. Is it a fair coin?

Null Hypothesis H_0

- It is a fair coin! Any deviation from baseline is a coincidence!
- How likely is our observation if H_0 is true?

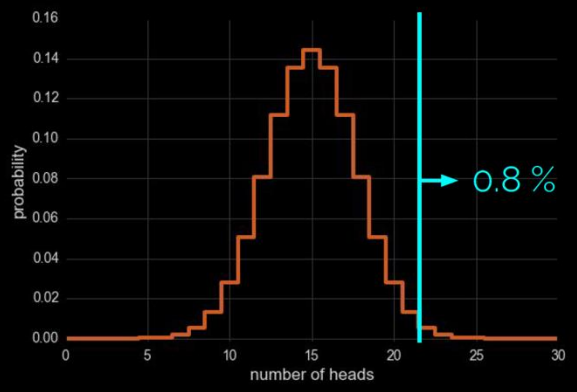


Unlikely to be fair
Reject Null Hypothesis
Observation statistically significant

How did we compute?

Math

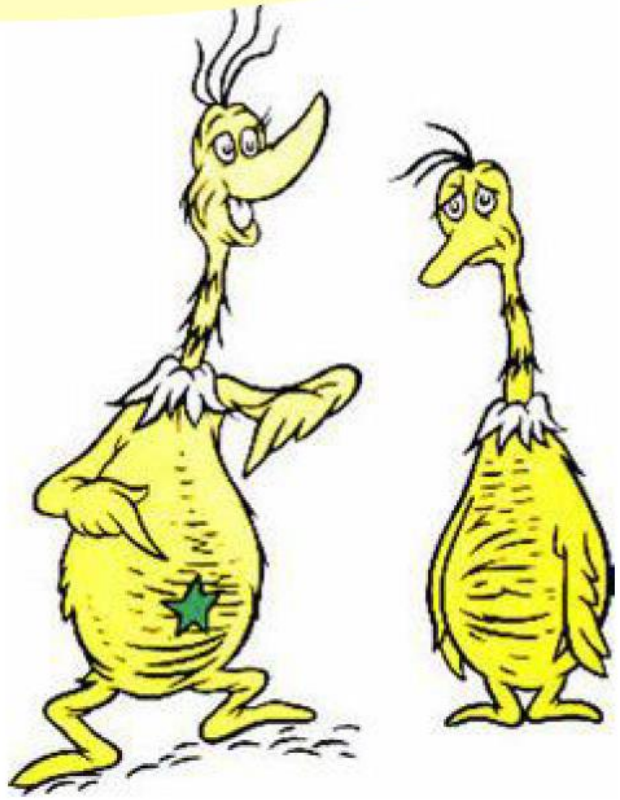
$$P(N_H, N_T) = \binom{N}{N_H}\left(\frac{1}{2}\right)^{N_H}\left(1-\frac{1}{2}\right)^{N_T}$$



Programming Simulation

```
M = 0
for i in range(10000):
    trials = randint(2, size=30)
    if (trials.sum() >= 22):
        M += 1
p = M / 10000 # 0.008149
```

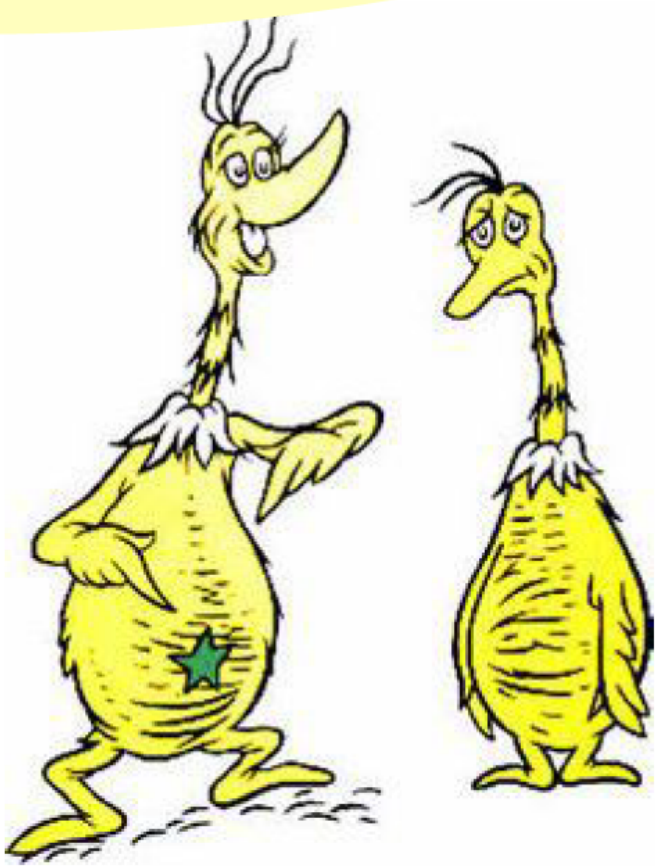
Sneeches:
Stars and Intelligence



Now, the Star-Belly Sneetches
had bellies with stars.
The Plain-Belly Sneetches
had none upon thars . . .

*inspired by John Rauser's
Statistics Without All The Agonizing Pain

Sneeches:
Stars and Intelligence



Test Scores

★		×	
84	72	81	69
57	46	74	61
63	76	56	87
99	91	69	65
		66	44
		62	69

★ mean: 73.5
× mean: 66.9
difference: 6.6

Is this difference of 6.6 statistically significant?

★ mean: 73.5
× mean: 66.9
difference: 6.6

We can’t simulate sneeches

We don’t have a formula for sneech intelligence either

Classic Method

(Welch's t-test)

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Classic Method

(Welch's t-test)

$$t = \frac{73.5 - 66.9}{\sqrt{\frac{316.3}{8} + \frac{124.8}{12}}} = 0.932$$

Classic Method

(Student's t distribution)

$$p(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Classic Method

(Student's t distribution)

$$p(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Degree of Freedom: "The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it."

-Wikipedia

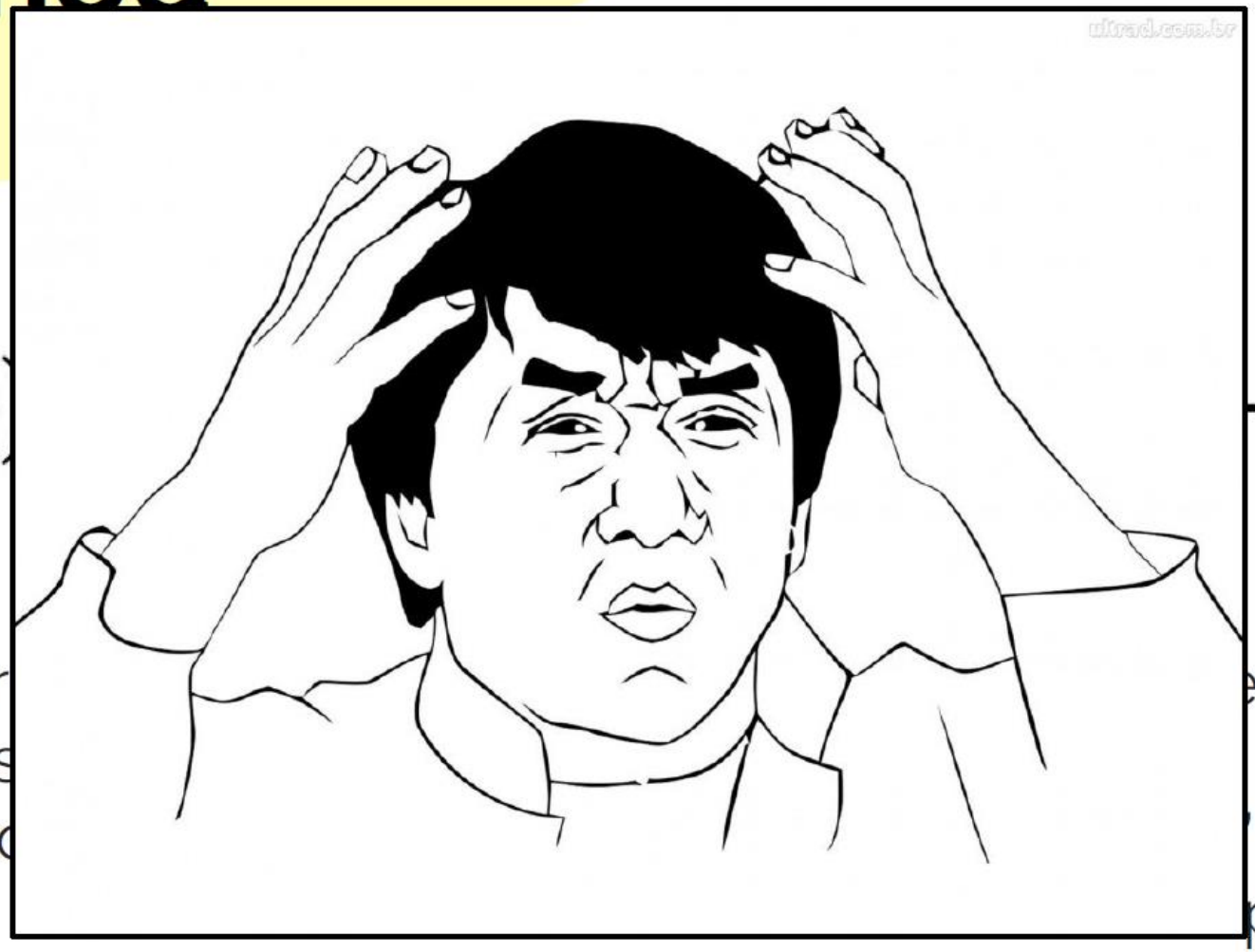
Classic Method

(Student's t distribution)

$$p(t;\nu) = \left(- \frac{\nu+1}{2} \right)$$

Degr ent
ways
witho

pedia



Classic Method

(Welch–Satterthwaite equation)

$$\nu \approx \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} \right)^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}}$$

Classic Method

(Welch–Satterthwaite equation)

$$\nu \approx \frac{\left(\frac{316.3}{8} + \frac{124.8}{12}\right)^2}{\frac{316.3^2}{8^2(8-1)} + \frac{124.8^2}{12^2(12-1)}} = 10.7$$

Classic Method

α (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
df							
1	6.3138	12.7065	31.8193	63.6551	127.3447	318.4930	636.0450
2	2.9200	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.7470	4.6041	5.5976	7.1732	8.6103
5	2.0150	2.5706	3.3650	4.0322	4.7734	5.8934	6.8688
6	1.9432	2.4469	3.1426	3.7074	4.3168	5.2076	5.9589
7	1.8946	2.3646	2.9980	3.4995	4.0294	4.7852	5.4079
8	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	5.0414
9	1.8331	2.2621	2.8214	3.2498	3.6896	4.2969	4.7809
10	1.8124	2.2282	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247	4.4369
12	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296	4.3178
13	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520	4.2208
14	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	4.1404
15	1.7530	2.1314	2.6025	2.9467	3.2860	3.7328	4.0728
16	1.7459	2.1199	2.5835	2.9208	3.2520	3.6861	4.0150
17	1.7396	2.1098	2.5669	2.8983	3.2224	3.6458	3.9651
18	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105	3.9216
19	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.7247	2.0860	2.5280	2.8454	3.1534	3.5518	3.8495

Classic Method

α (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
df							
1	6.3138	12.7065	31.8193	63.6551	127.3447	318.4930	636.0450
2	2.9200	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.7470	4.6041	5.5976	7.1732	8.6103
5	2.0150	2.5706	3.3650	4.0322	4.7734	5.8934	6.8688
6	1.9432	2.4469	3.1426	3.7074	4.3168	5.2076	5.9589
7	1.8946	2.3646	2.9980	3.4995	4.0294	4.7852	5.4079
8	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	5.0414
9	1.8331	2.2621	2.8214	3.2498	3.6896	4.2969	4.7809
10	1.8124	2.2282	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.7959	2.2010	2.7181	3.1058	3.4966	4.0247	4.4369
12	1.7823	2.1788	2.6810	3.0545	3.4284	3.9296	4.3178
13	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520	4.2208
14	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	4.1404
15	1.7530	2.1314	2.6025	2.9467	3.2860	3.7328	4.0728
16	1.7459	2.1199	2.5835	2.9208	3.2520	3.6861	4.0150
17	1.7396	2.1098	2.5669	2.8983	3.2224	3.6458	3.9651
18	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105	3.9216
19	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.7247	2.0860	2.5280	2.8454	3.1534	3.5518	3.8495

Classic Method

α (1 tail)	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
α (2 tail)	0.1	0.05	0.02	0.01	0.005	0.002	0.001
df							
1	6.3138	12.7065	31.8193	63.6551	127.3447	318.4930	636.0450
2	2.9200	4.3026	6.9646	9.9247	14.0887	22.3276	31.5989
3	2.3534	3.1824	4.5407	5.8408	7.4534	10.2145	12.9242
4	2.1319	2.7764	3.7470	4.6041	5.5976	7.1732	8.6103
5	2.0150	2.5706	3.3650	4.0322	4.7734	5.8934	6.8688
6	1.9432	2.4469	3.1426	3.7074	4.3168	5.2076	5.9589
7	1.8946	2.3646	2.9980	3.4995	4.0294	4.7852	5.4079
8	1.8595	2.3060	2.8965	3.3554	3.8325	4.5008	5.0414
9	1.8331	2.2621	2.8214	3.2498	3.6896	4.2969	4.7809
10	[illegible]	[illegible]82	2.7638	3.1693	3.5814	4.1437	4.5869
11	1.7959	[illegible]10	2.7181	3.1058	3.4966	4.0247	4.4369
12	[illegible]	[illegible]88	2.6810	3.0545	3.4284	3.9296	4.3178
13	1.7709	2.1604	2.6503	3.0123	3.3725	3.8520	4.2208
14	1.7613	2.1448	2.6245	2.9768	3.3257	3.7874	4.1404
15	1.7530	2.1314	2.6025	2.9467	3.2860	3.7328	4.0728
16	1.7459	2.1199	2.5835	2.9208	3.2520	3.6861	4.0150
17	1.7396	2.1098	2.5669	2.8983	3.2224	3.6458	3.9651
18	1.7341	2.1009	2.5524	2.8784	3.1966	3.6105	3.9216
19	1.7291	2.0930	2.5395	2.8609	3.1737	3.5794	3.8834
20	1.7247	2.0860	2.5280	2.8454	3.1534	3.5518	3.8495

Classic Method

$$t > t_{crit}$$

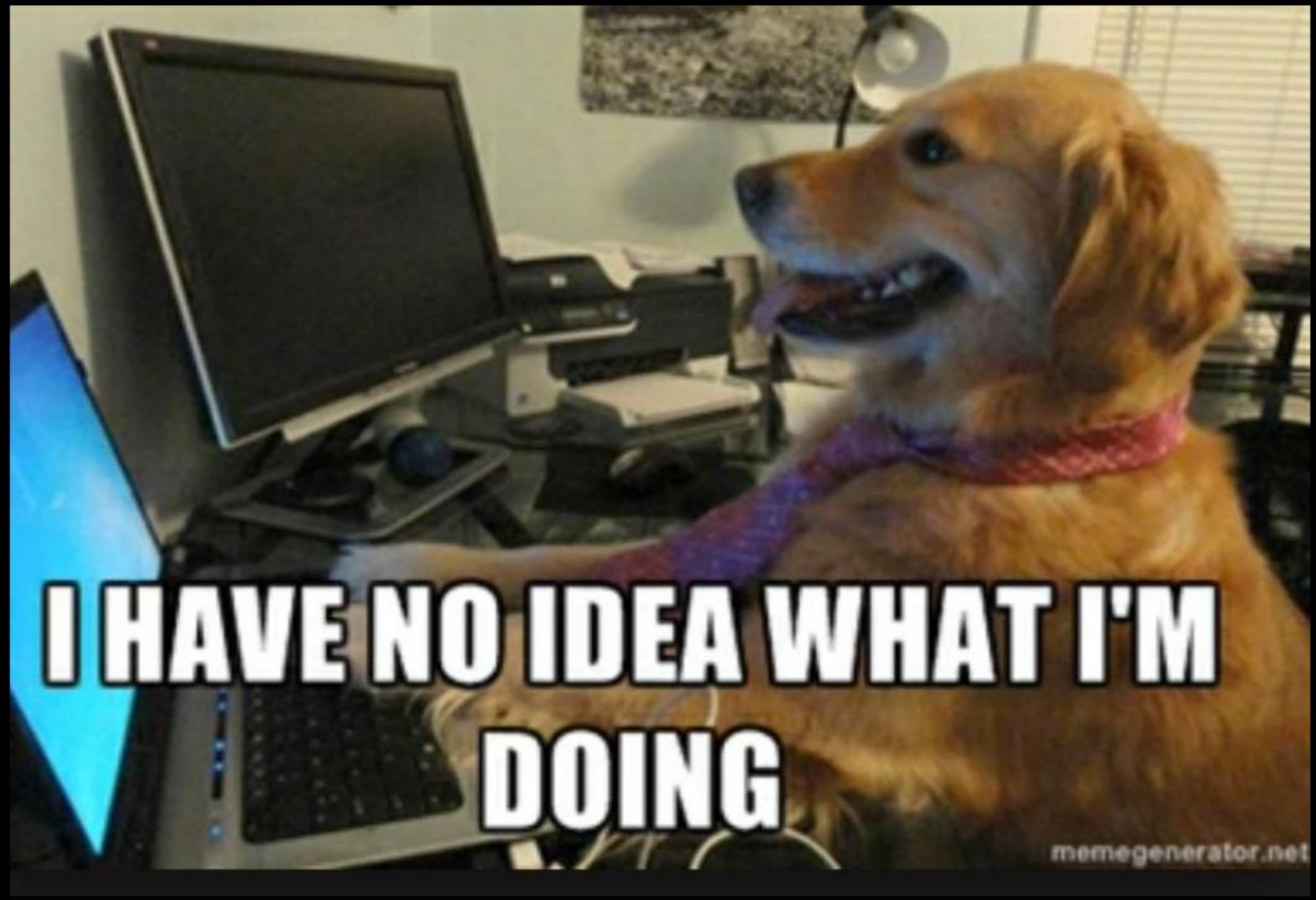
Classic Method

$$0.932 > 1.796$$

Classic Method

$$0.932 > 1.796$$

“The difference of 6.6 is not significant at the p=0.05 level”


I HAVE NO IDEA WHAT I'M DOING
memegenerator.net

The Problem:

Unlike coin flipping, we *don't* have a generative model . . .

Solution:

Shuffling

★		×	
84	72	81	69
57	46	74	61
63	76	56	87
99	91	69	65
		66	44
		62	69

Idea:
Simulate the distribution by *shuffling* the labels repeatedly and computing the desired statistic.

Motivation:
if the labels really don't matter, then switching them shouldn't change the result!

★		×	
84	72	81	69
57	46	74	61
63	76	56	87
99	91	69	65
		66	44
		62	69

1. Shuffle Labels
2. Rearrange
3. Compute means

★		×	
84	72	81	69
57	46	74	61
63	76	56	87
99	91	69	65
		66	44
		62	69

1. **Shuffle Labels**
2. Rearrange
3. Compute means

★		×	
84	81	72	69
61	69	74	57
65	76	56	87
99	44	46	63
		66	91
		62	69

1. Shuffle Labels
2. **Rearrange**
3. Compute means

★		×	
84	81	72	69
61	69	74	57
65	76	56	87
99	44	46	63
		66	91
		62	69

1. Shuffle Labels
2. Rearrange
3. **Compute means**

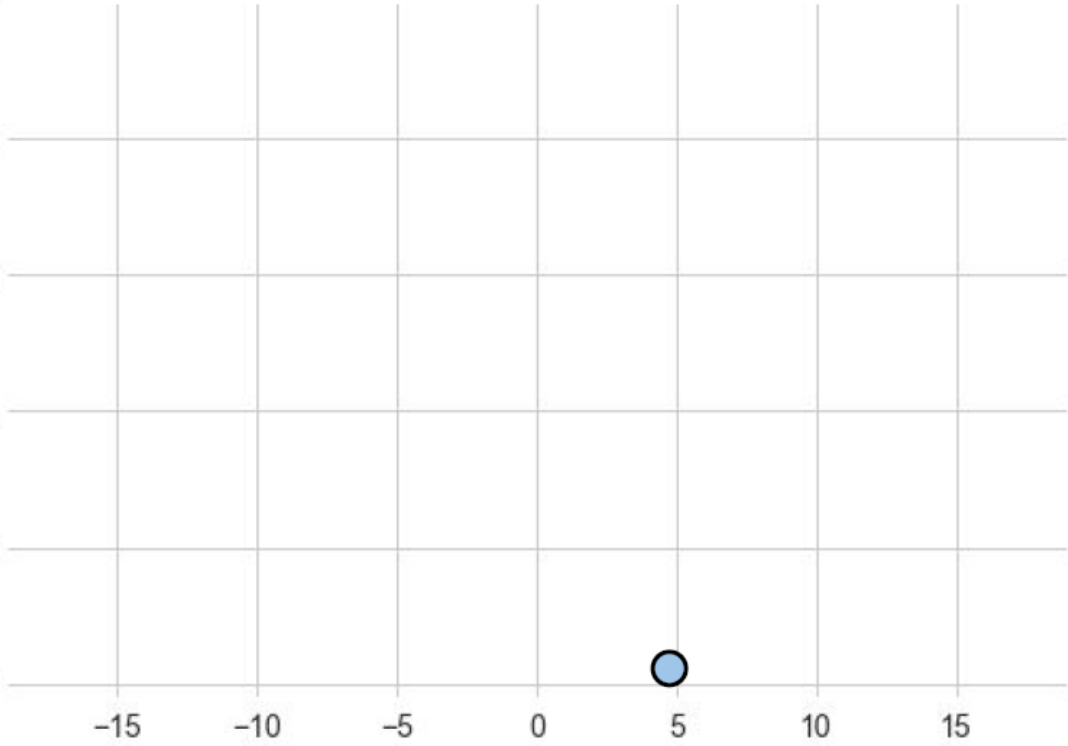
★ mean: 72.4

× mean: 67.6

difference: 4.8

★		×	
84	81	72	69
61	69	74	57
65	76	56	87
99	44	46	63
		66	91
		62	69

1. Shuffle Labels
2. Rearrange
3. **Compute means**



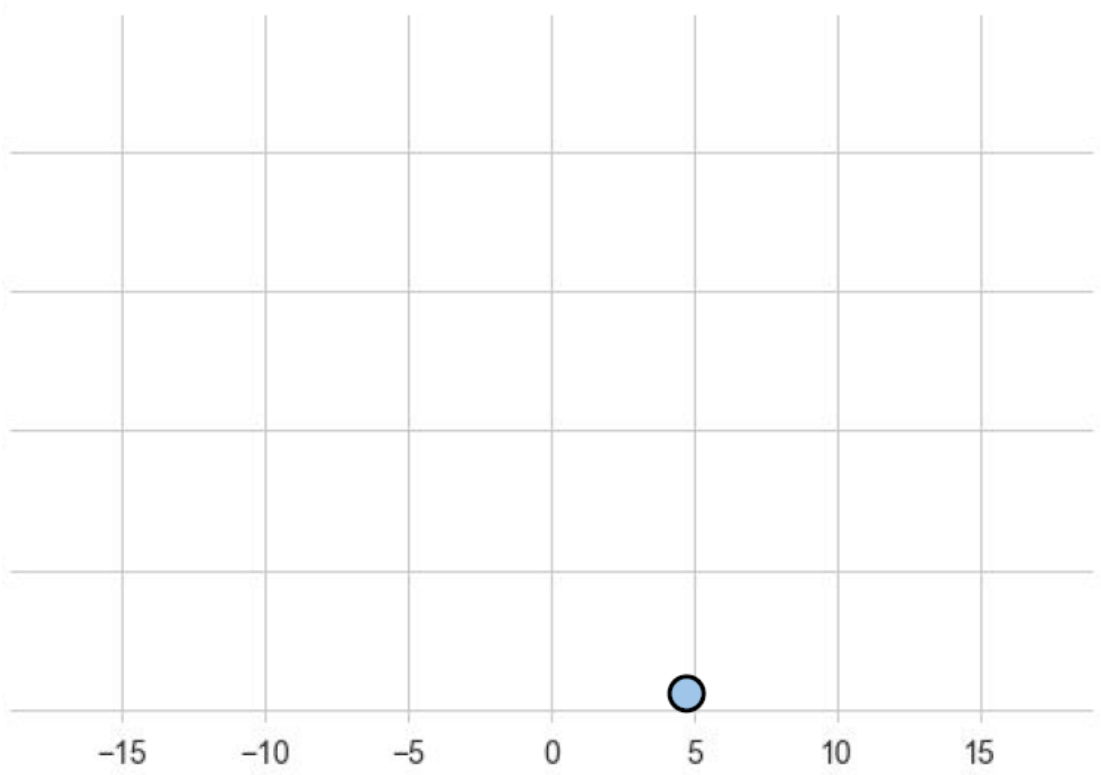
★ mean: 72.4

× mean: 67.6

difference: 4.8

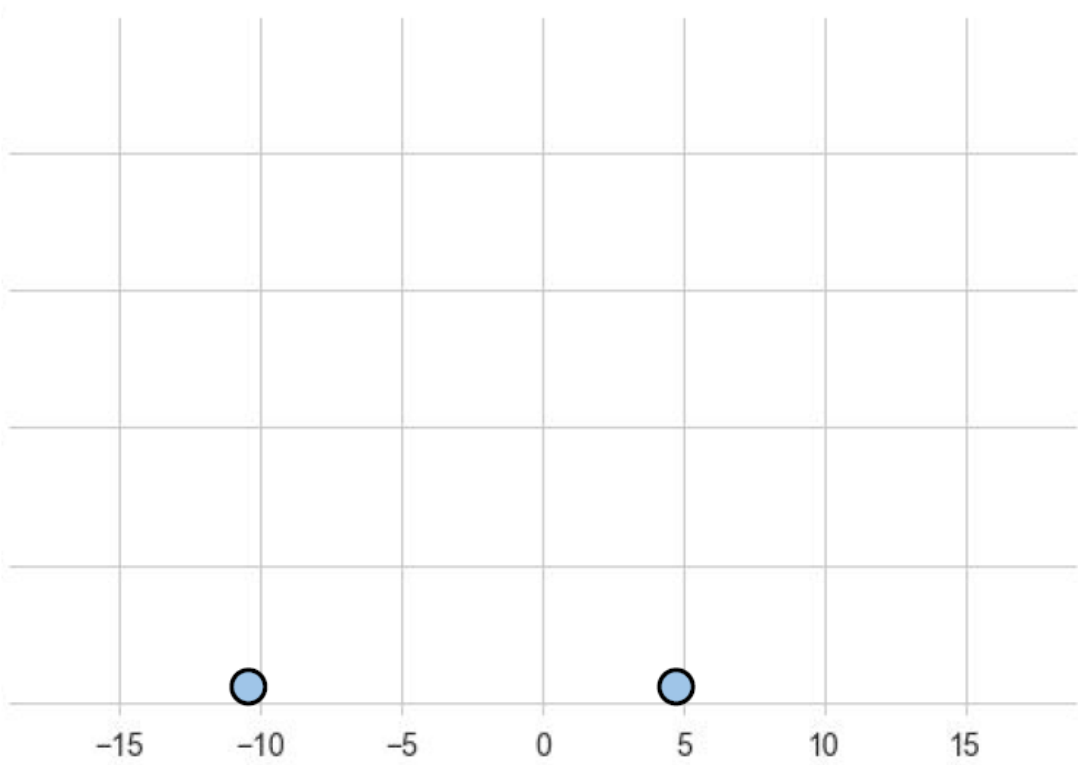
★		×	
84	81	72	69
61	69	74	57
65	76	56	87
99	44	46	63
		66	91
		62	69

1. **Shuffle Labels**
2. Rearrange
3. Compute means



★		×	
84	56	72	69
61	63	74	57
65	66	81	87
62	44	46	69
		76	91
		99	69

1. Shuffle Labels
2. Rearrange
3. **Compute means**



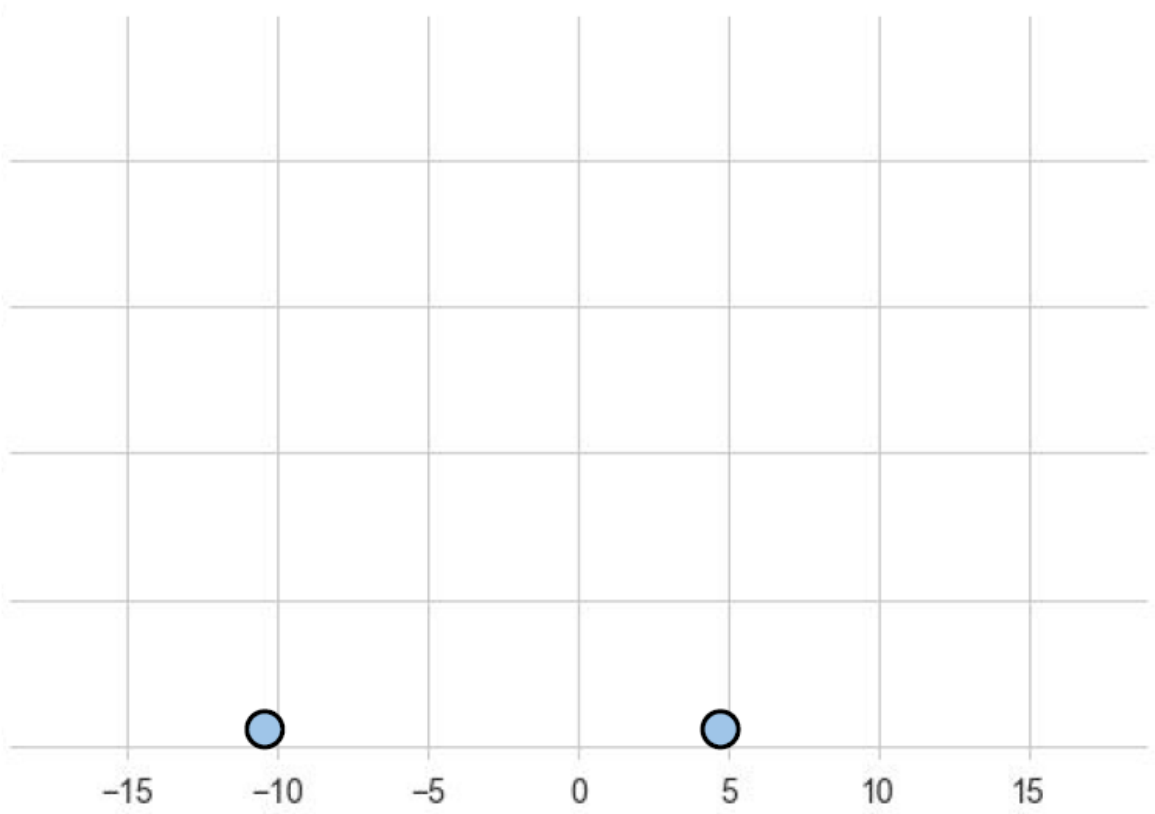
★ mean: 62.6

× mean: 74.1

difference: -11.6

★		×	
84	56	72	69
61	63	74	57
65	66	81	87
62	44	46	69
		76	91
		99	69

1. **Shuffle Labels**
2. Rearrange
3. Compute means



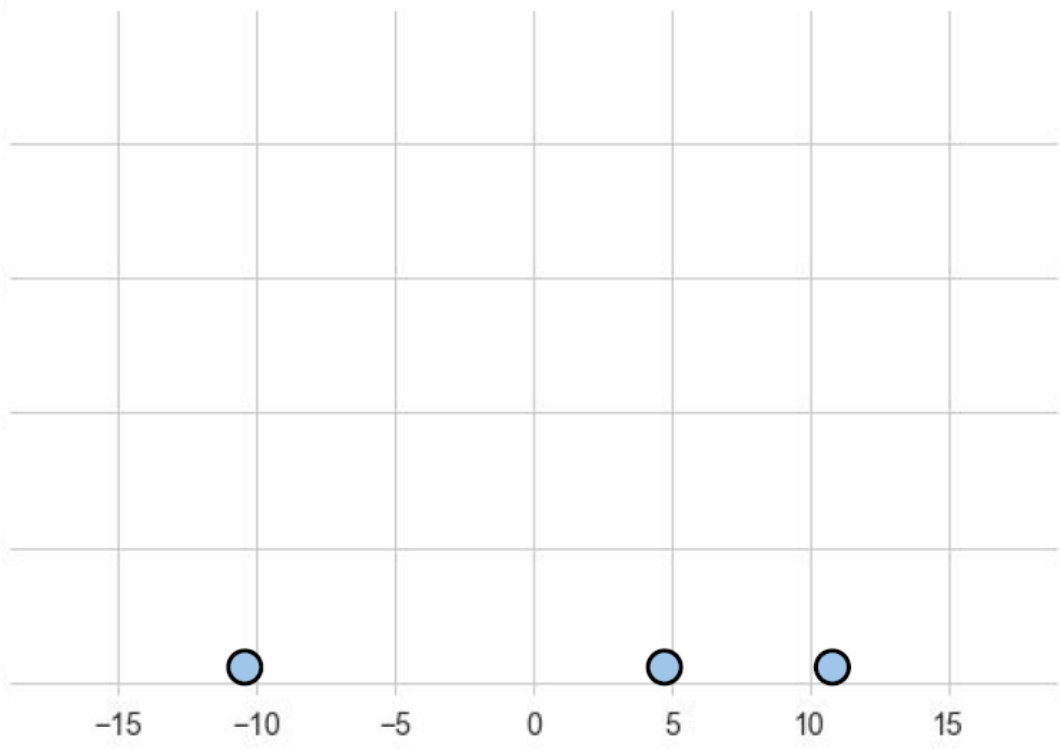
★		×	
74	56	72	69
61	63	84	57
87	76	81	65
91	99	46	69
		66	62
		44	69

★ mean: 75.9

× mean: 65.3

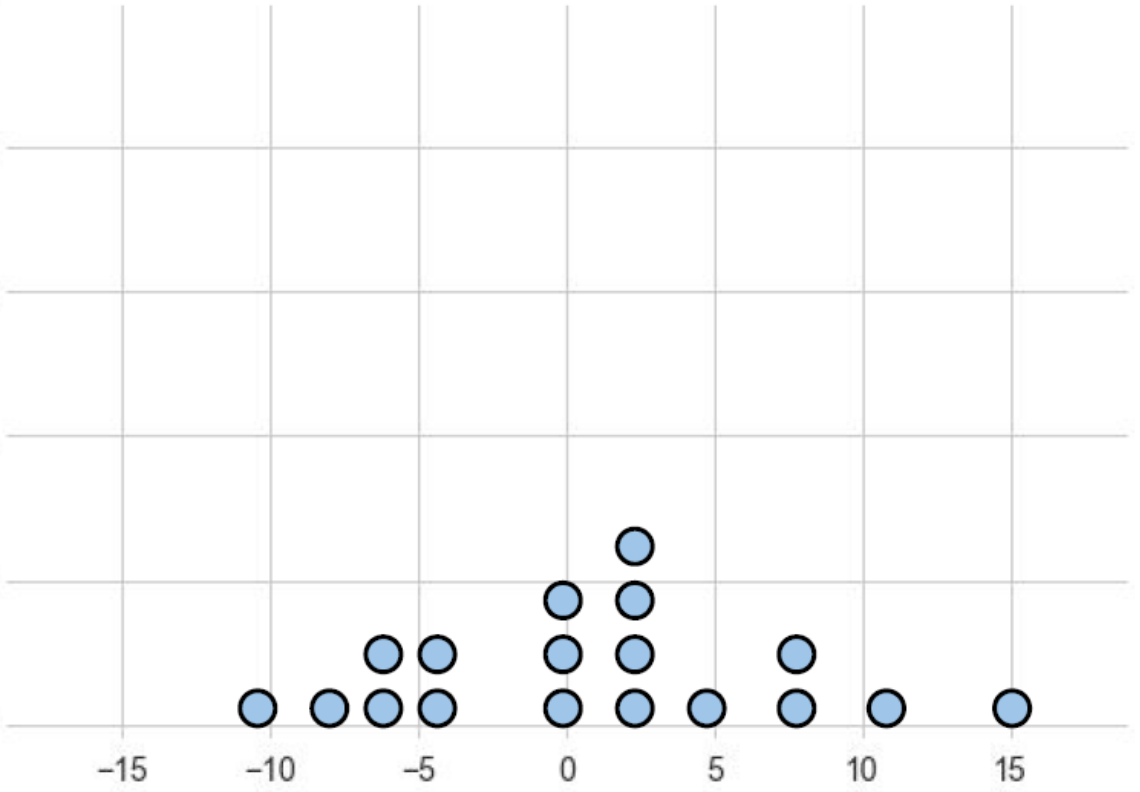
difference: 10.6

1. Shuffle Labels
2. Rearrange
3. **Compute means**



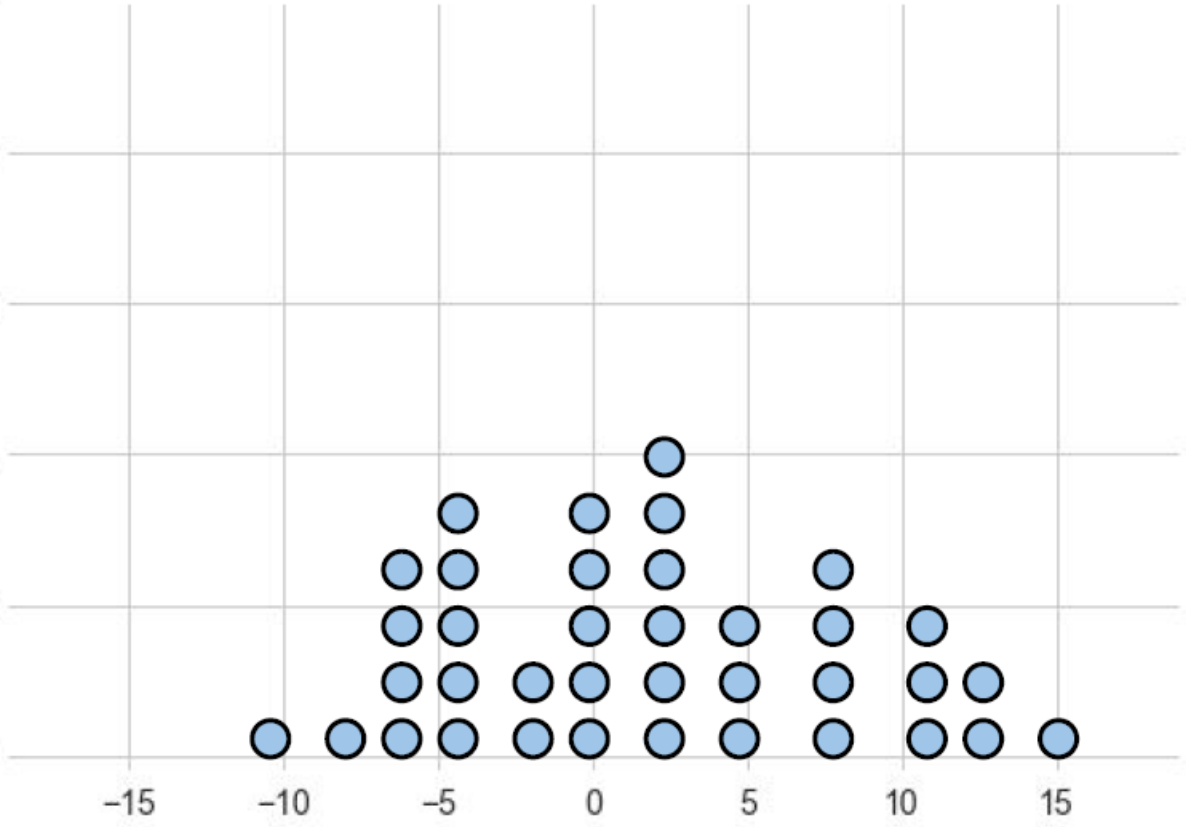
★		×	
84	56	72	69
61	63	74	57
65	66	81	87
62	44	46	69
		76	91
		99	69

1. Shuffle Labels
2. Rearrange
3. Compute means



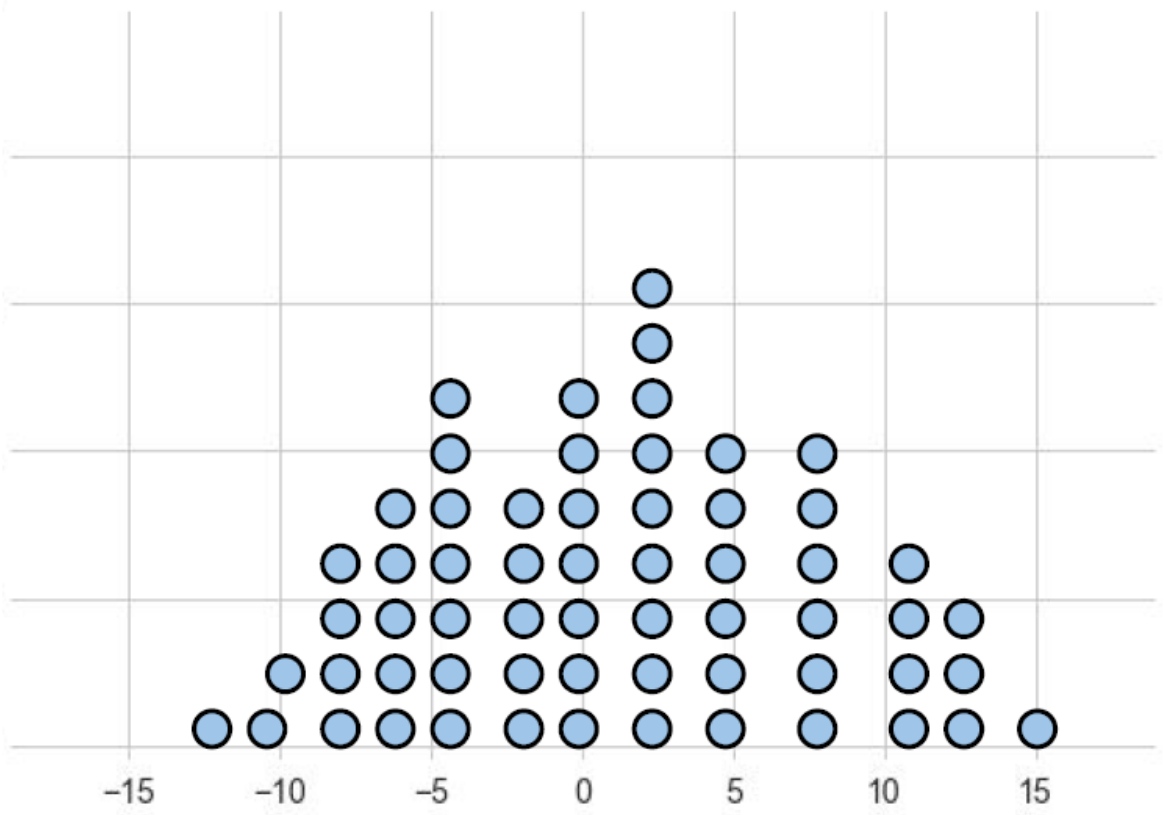
★		×	
84	81	69	69
61	69	87	74
65	76	56	57
99	44	46	63
		66	91
		62	72

1. Shuffle Labels
2. Rearrange
3. Compute means



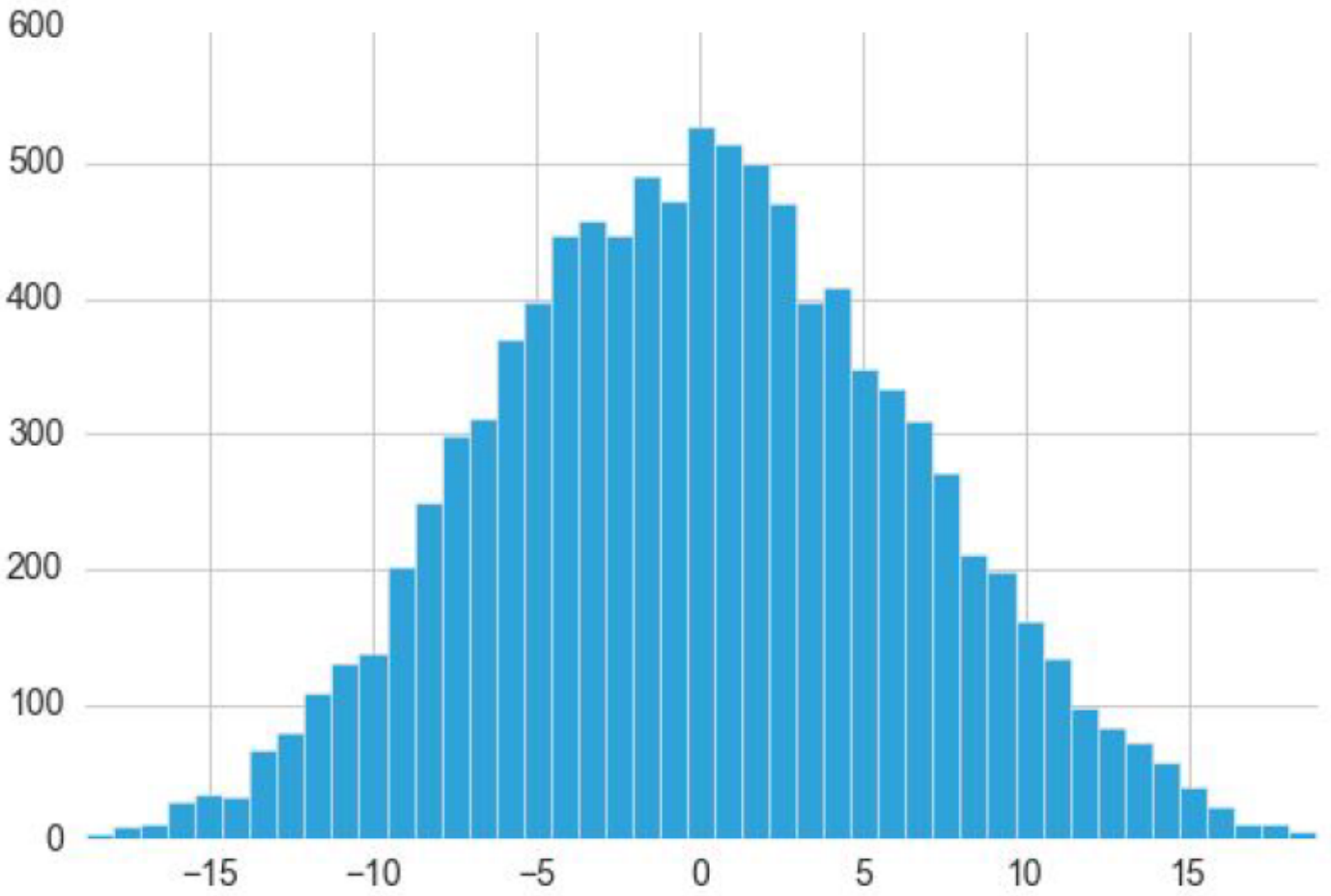
★		×	
74	62	72	57
61	63	84	69
87	81	76	65
91	99	46	69
		66	56
		44	69

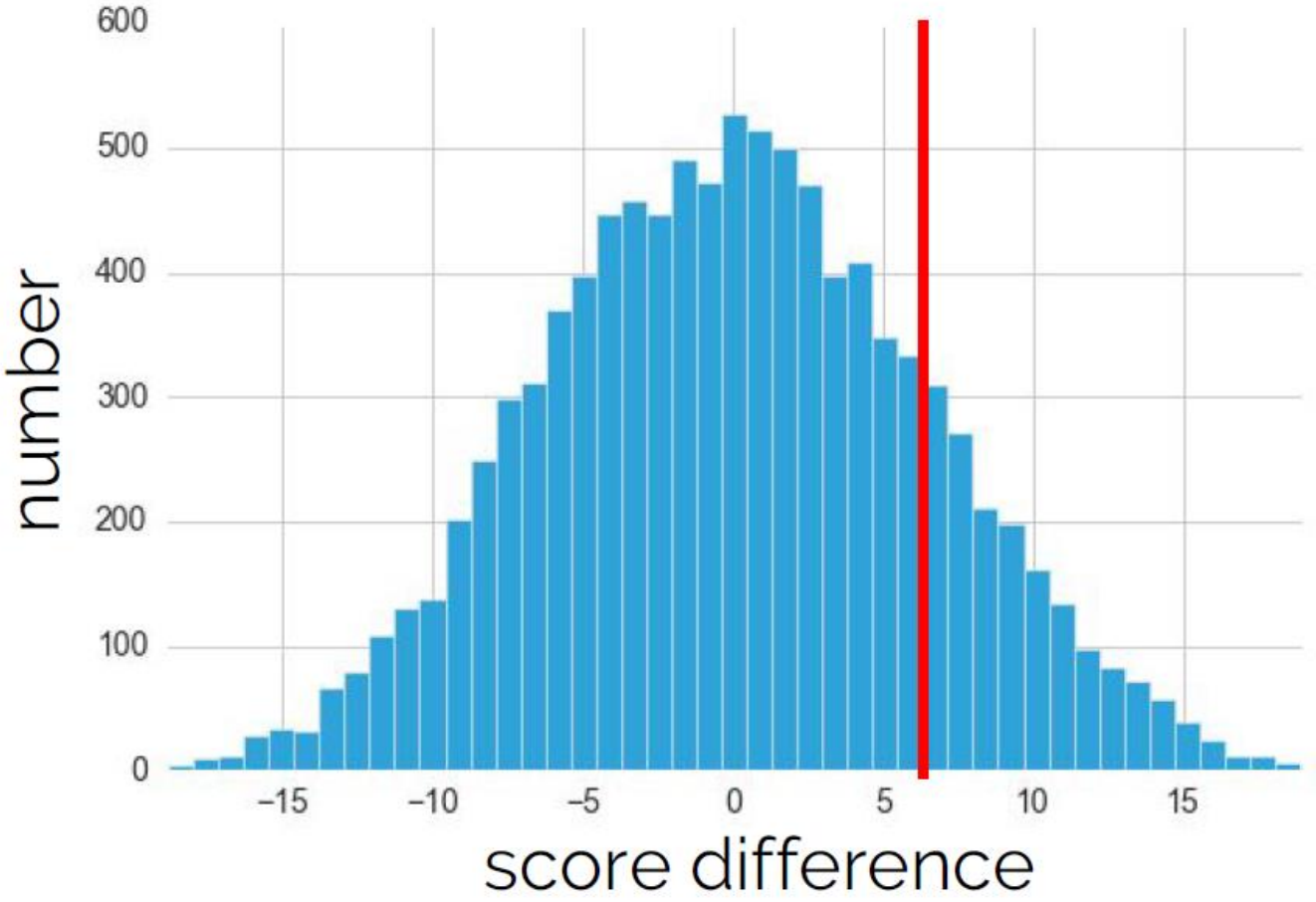
1. Shuffle Labels
2. Rearrange
3. Compute means

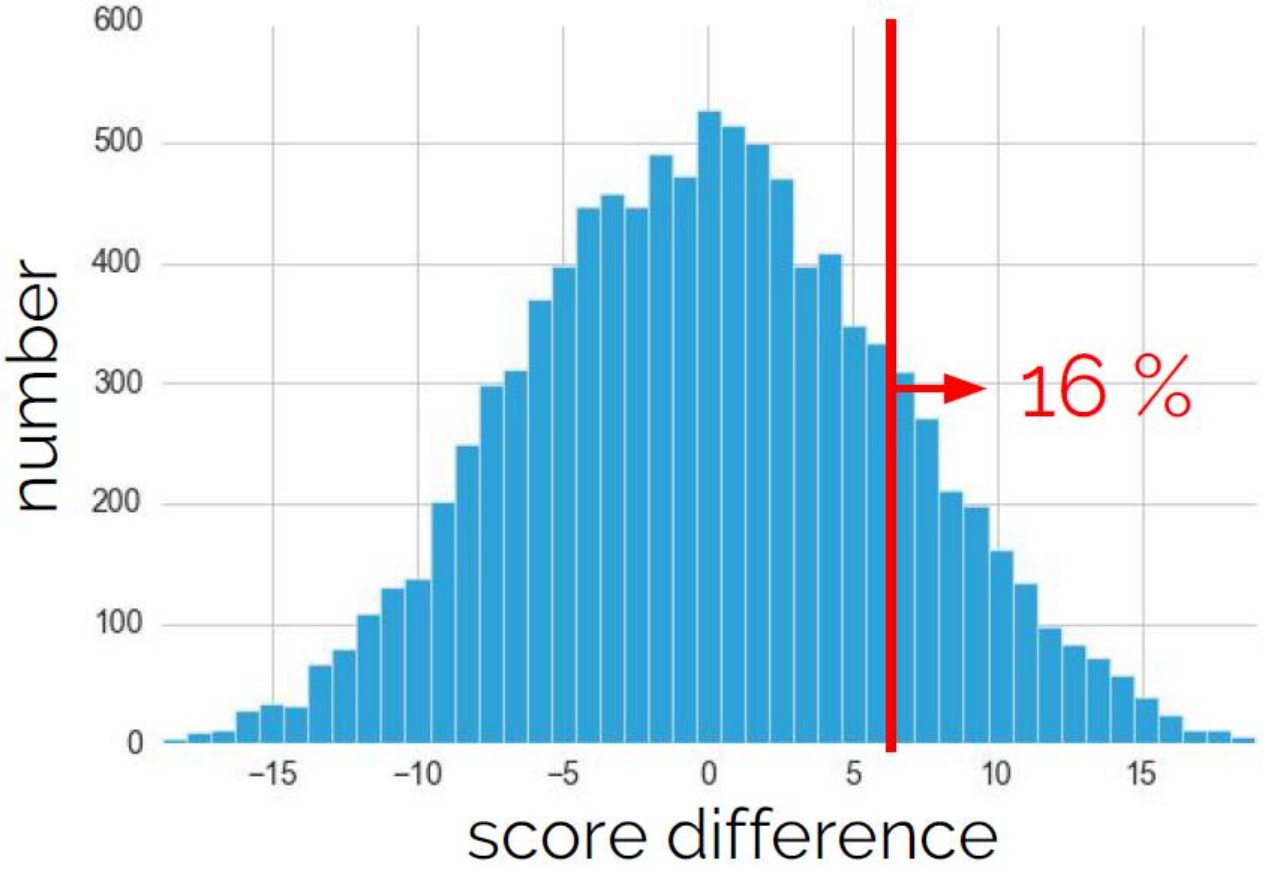


★		×	
84	81	72	69
61	69	74	57
65	76	56	87
99	44	46	63
		66	91
		62	69

1. Shuffle Labels
2. Rearrange
3. Compute means

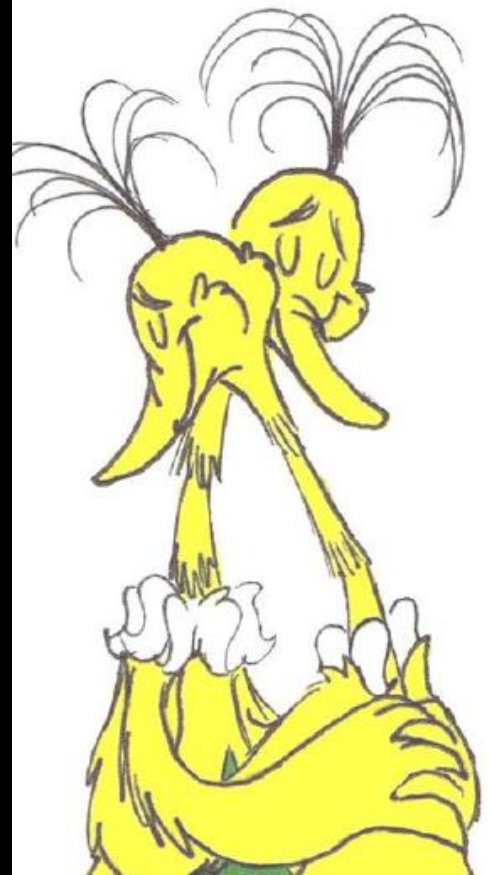



600
500
400
300
200
100
0
number
−15
−10
−5
0
5
10
15
score difference



$$\frac{N_{>6.6}}{N_{tot}} = \frac{1608}{10000} = 0.16$$

“A difference of 6.6 is not significant at p = 0.05.”



That day, all the Sneetches
forgot about stars
And whether they had one,
or not, upon thars.

Two groups

Same weight?

How do we sample?

Don’t know the
true distribution of
apple OR orange weights!

Target statistic

How do we compare?

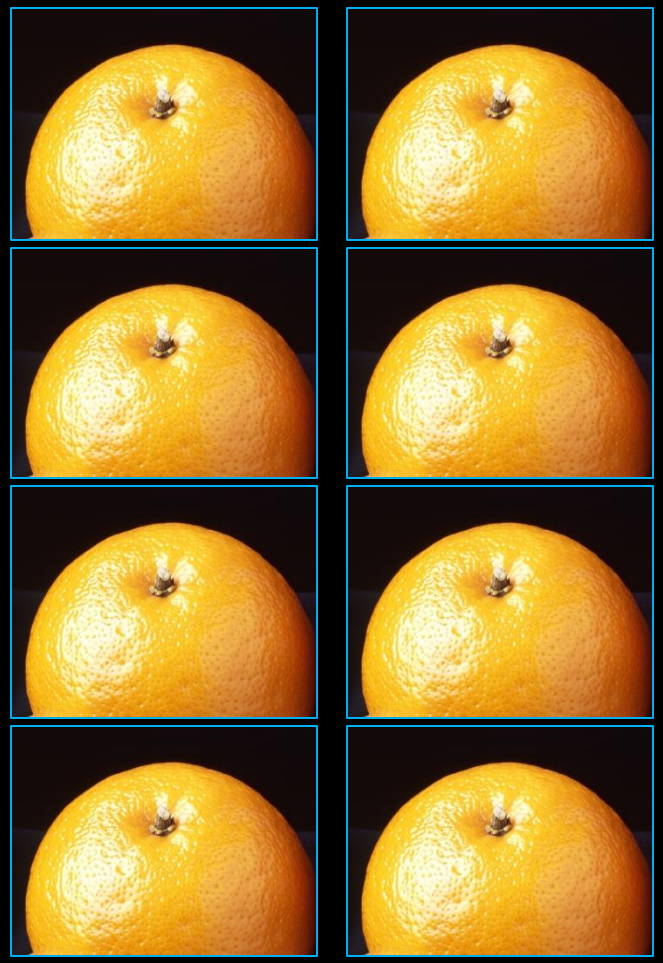
Target statistic

How do we compare? Difference of means

How do we sample from the null

Can’t get fake fruit with same means

Two Groups

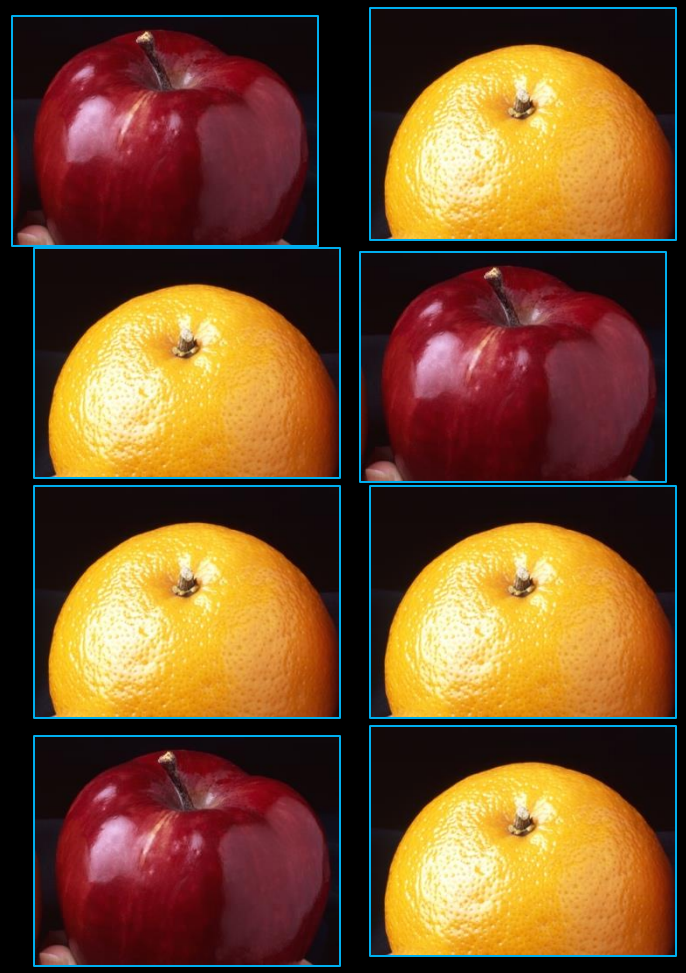


Mean 1

Mean 1.5

Is this a big enough difference?

If there was no difference …..

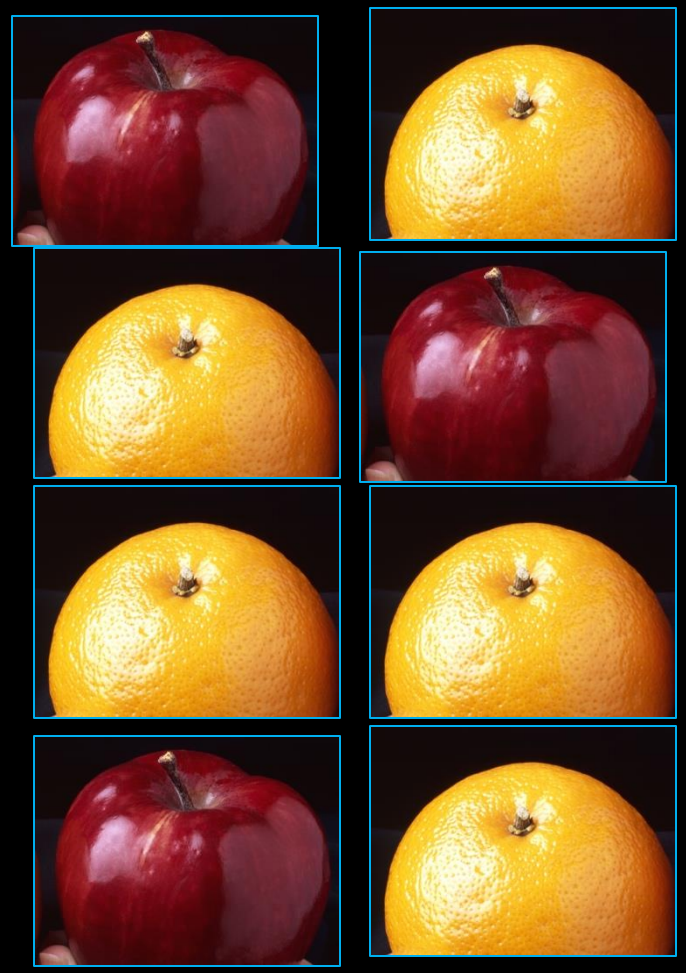


Mean 1.2

Mean 1.1

0.1 < 0.5
But that's just one new data point

Keep Shuffling: Permutation Test



Mean 1.2

Mean 1.1

0.1 < 0.5
But that's just one new data point

Boiler Plate

In [1]:
```
import numpy as np
import pandas as pd
from pandas import DataFrame, Series, read_csv
from matplotlib import pyplot as plt
import matplotlib as mpl
%matplotlib inline
mpl.style.use('ggplot')
mpl.rcParams['figure.figsize'] = (10, 6)
```

Birth Weight and Smoking

```
In [2]: data_url = "https://bit.ly/smokebaby"
```

```
In [3]: births = read_csv(data_url)
        births.head()
```

Out[3]:

	Birth Weight	Gestational Days	Maternal Age	Maternal Height	Maternal Pregnancy Weight	Maternal Smoker
0	120	284	27	62	100	False
1	113	282	33	64	135	False
2	128	279	28	64	115	True
3	108	282	23	67	125	True
4	136	286	25	62	93	False

Always get info and ….

```
In [4]: births.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1174 entries, 0 to 1173
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Birth Weight            1174 non-null   int64
 1   Gestational Days        1174 non-null   int64
 2   Maternal Age            1174 non-null   int64
 3   Maternal Height         1174 non-null   int64
 4   Maternal Pregnancy Weight  1174 non-null   int64
 5   Maternal Smoker         1174 non-null   bool
dtypes: bool(1), int64(5)
memory usage: 47.1 KB
```

Describe for summary stats

```
In [5]: births.describe()
```

Out[5]:

	Birth Weight	Gestational Days	Maternal Age	Maternal Height	Maternal Pregnancy Weight
count	1174.000000	1174.000000	1174.000000	1174.000000	1174.000000
mean	119.462521	279.101363	27.228279	64.049404	128.478705
std	18.328671	16.010305	5.817839	2.526102	20.734282
min	55.000000	148.000000	15.000000	53.000000	87.000000
25%	108.000000	272.000000	23.000000	62.000000	114.250000
50%	120.000000	280.000000	26.000000	64.000000	125.000000
75%	131.000000	288.000000	31.000000	66.000000	139.000000
max	176.000000	353.000000	45.000000	72.000000	250.000000

More non-smokers

In [6]:
```
smoking_and_birthweight = births[['Maternal Smoker', 'Birth Weight']]
smoking_and_birthweight.groupby('Maternal Smoker').count()
```

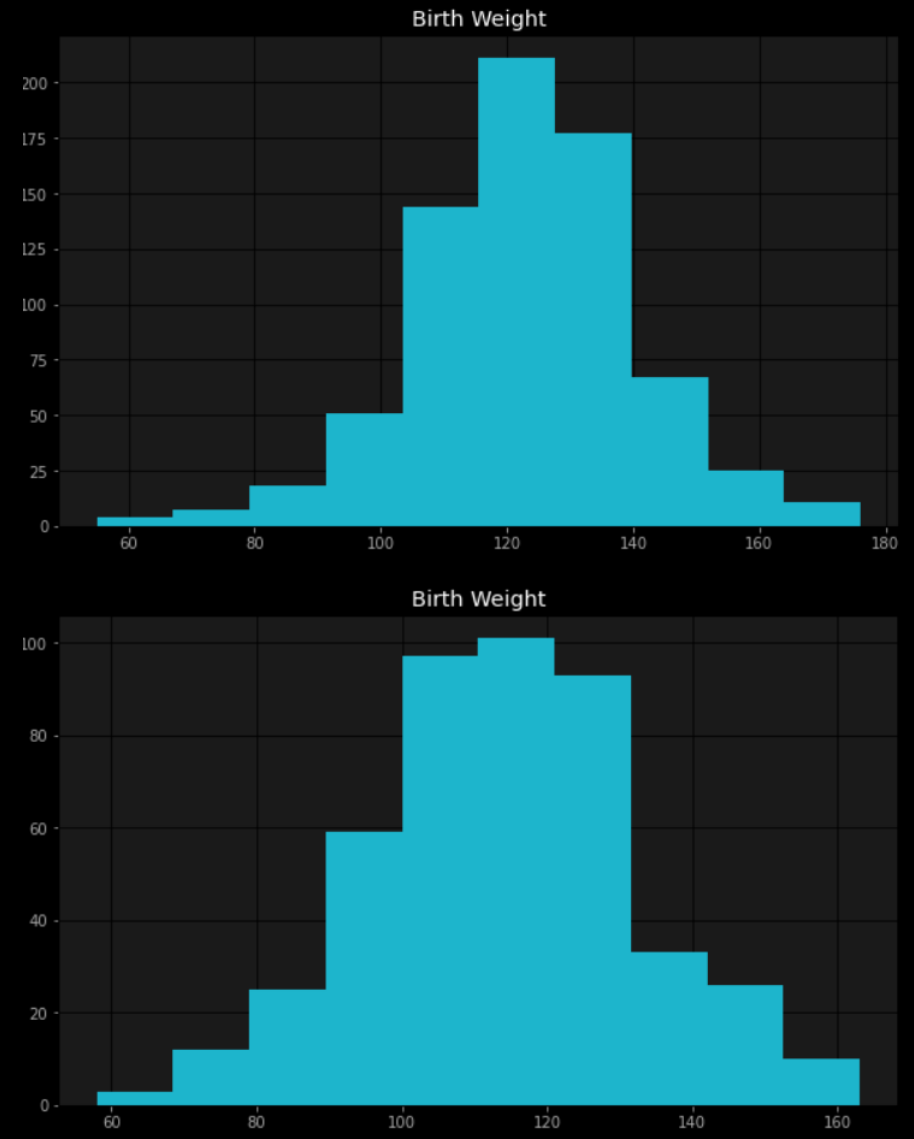
Out[6]:

Maternal Smoker	Birth Weight
False	715
True	459

Looking at the differences

```
In [7]: smoking_and_birthweight.groupby("Maternal Smoker").hist()

Out[7]: Maternal Smoker
        False    [[AxesSubplot(0.125,0.125;0.775x0.755)]]
        True     [[AxesSubplot(0.125,0.125;0.775x0.755)]]
        dtype: object
```



Lets make it look nicer

In [24]:
```
fig, axs = plt.subplots(2,1)
hist1 = smoking_and_birthweight.groupby("Maternal Smoker").get_group(False).hist(
    ax=axs[0], color='blue')
hist2 = smoking_and_birthweight.groupby("Maternal Smoker").get_group(True).hist(ax=axs[1])
axs[0].set_title("Maternal Smoker: False")
axs[1].set_title("Maternal Smoker: True")
for ax in axs:
    ax.set_xlim([50,180])
    ax.set_ylim([0,220])
    ax.set_xlabel("Birth Weight")
    ax.set_ylabel("Count")
```