
Statistical significance, Hypothesis testing

Lecture 24

Slides courtesy of John Guttag, Tim Kraska

- Reminder about Pset 6 checkoff and Pset 7 due dates
- Finger exercise for today's lecture is optional (will not count towards your total score) and will be released later tonight
- Course evaluations are open until Mon, Dec 15 at 9 am
 - <https://eduapps.mit.edu/subjeval/studenthome.htm>

Been Talking a Lot About Statistics

- Statistics is about finding ways to analyze data that allow us to draw conclusions despite uncertainty
- Allows us to derive conclusions about the world from limited samples
- **Descriptive statistics**
 - Describing the sample—these are factual statements
 - Plots, mean, median, variance, etc.
- **Inferential statistics**
 - Extrapolate from samples to assertions about the populations from which the samples are drawn
 - No guarantee that these are correct

“All models are wrong, but some are useful.”
--George Box

- There are methods to quantify probability that they are correct
- Already looked at putting confidence intervals around estimates
- Today, we will talk (briefly) about **hypothesis testing**

Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes

Beer (25):

27 20 21 26 27 31 24 21 20 19
23 24 28 19 24 29 18 20 17 31
20 25 28 21 27

Mean: 23.6

Water (18):

21 22 15 12 21 16 19 15 22 24
19 23 13 22 20 24 18 20

Mean: 19.2

**Is a difference of 4.4
significant?**



- What do we mean by “statistically significant”?
- Hypothesis testing
 - A principled way to “test” a **hypothesis about data**
 - **Intuition:** “What is the likelihood my observation would occur just by random chance?”
 - If the likelihood is “very low”: we believe the hypothesis is true
 - Otherwise: We cannot draw a conclusion about the hypothesis because our observation could be explained by pure chance

- Alternative hypothesis (H_a)
 - Something that you think is implied by the data
 - E.g., Beer consumption increases human attractiveness to Malaria mosquitos
- Null hypothesis (H_0)
 - The data does not support the alternative hypothesis
 - E.g., Beer consumption does not increase human attractiveness to Malaria mosquitos

1. State H_a and H_0 unambiguously
2. Choose a **statistical significance threshold (α)**
 - Common choices: 0.05, 0.01
3. Compute a **test statistic**
 - E.g., The difference in number of mosquitoes between beer consumers and non-consumers is 4.4
4. Ask, **if H_0 were true**, how likely (probability) are we to see a difference **as extreme as in the data?**
 - This is called a **p-value (p)**
5. If **$p \leq \alpha$** , we **reject the null hypothesis**
 - The p-value is a measure of how surprising the data are if H_0 were true. The smaller, the more surprising.

- There are multiple ways of computing p-values
- They all take into account
 - Magnitude of effect
 - Size of samples
 - Variance of samples
- In practice, important to know under what circumstances each is appropriate
 - Take a statistics subject to learn this
- For now, we will use something call a **permutation test**
- See reading for other examples (Chapter 21.1 – 21.6)

Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes

Beer (25):

27 20 21 26 27 31 24 21 20 19
23 24 28 19 24 29 18 20 17 31
20 25 28 21 27

Mean: 23.6

Water (18):

21 22 15 12 21 16 19 15 22 24
19 23 13 22 20 24 18 20

Mean: 19.2

**Is a difference of 4.4
significant?**



- **Null hypothesis:** Beer consumption does not increase human attractiveness to malaria mosquitoes
- **Alternative hypothesis:** Beer consumption increases human attractiveness to malaria mosquitoes
- $\alpha = 0.01$

Permutation Test

Beer (25)

27	23	20	31	29
20	24	25	24	18
21	28	28	21	20
26	19	21	20	17
27	24	27	19	31

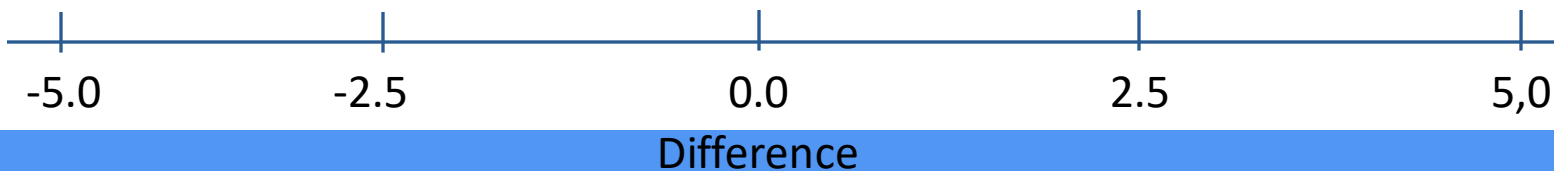
Water (18)

21	19	16	24
22	23	19	18
15	13	15	20
12	22	22	
21	20	24	

Difference: 4.4 (Difference of means)

Observation from the data (test statistic):

The beer group has **4.4** more bites on average.



Permutation Test

Beer (25)

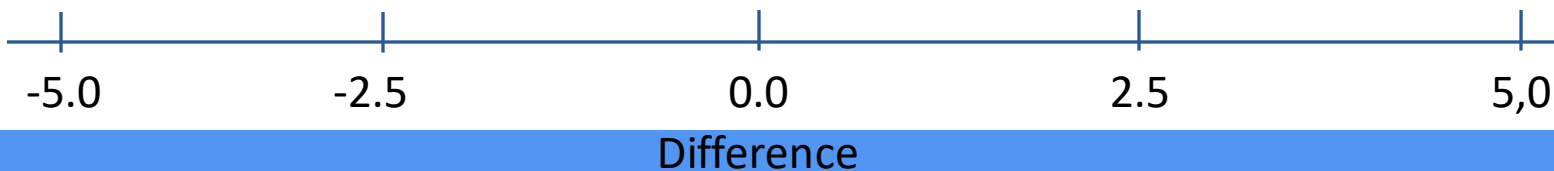
27	23	20	31	29
20	24	25	24	18
21	28	28	21	20
26	19	21	20	17
27	24	27	19	31

Water (18)

21	19	16	24
22	23	19	18
15	13	15	20
12	22	22	
21	20	24	

Difference: 4.4 (Difference of means)

Key idea: If the null hypothesis is true, “Beer” vs. “Water” shouldn’t matter, so it shouldn’t matter who is in each group

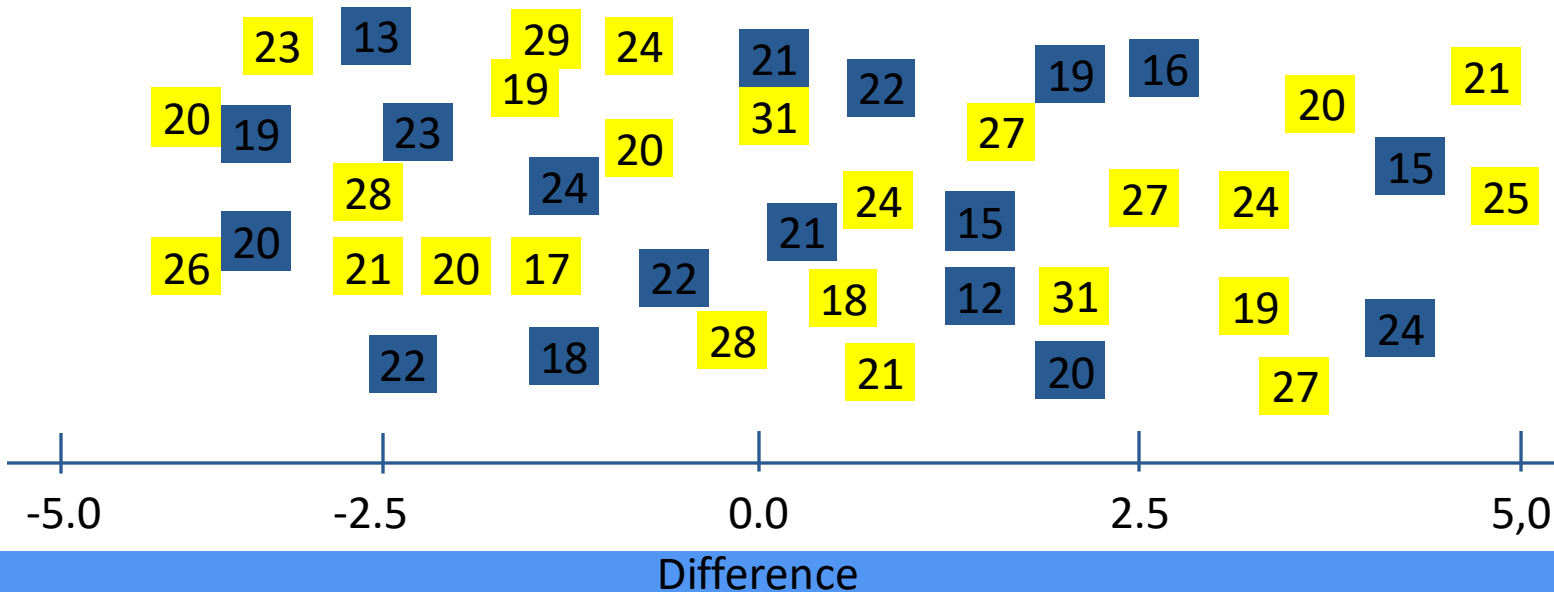


Permutation Test

Beer (25)

Water (18)

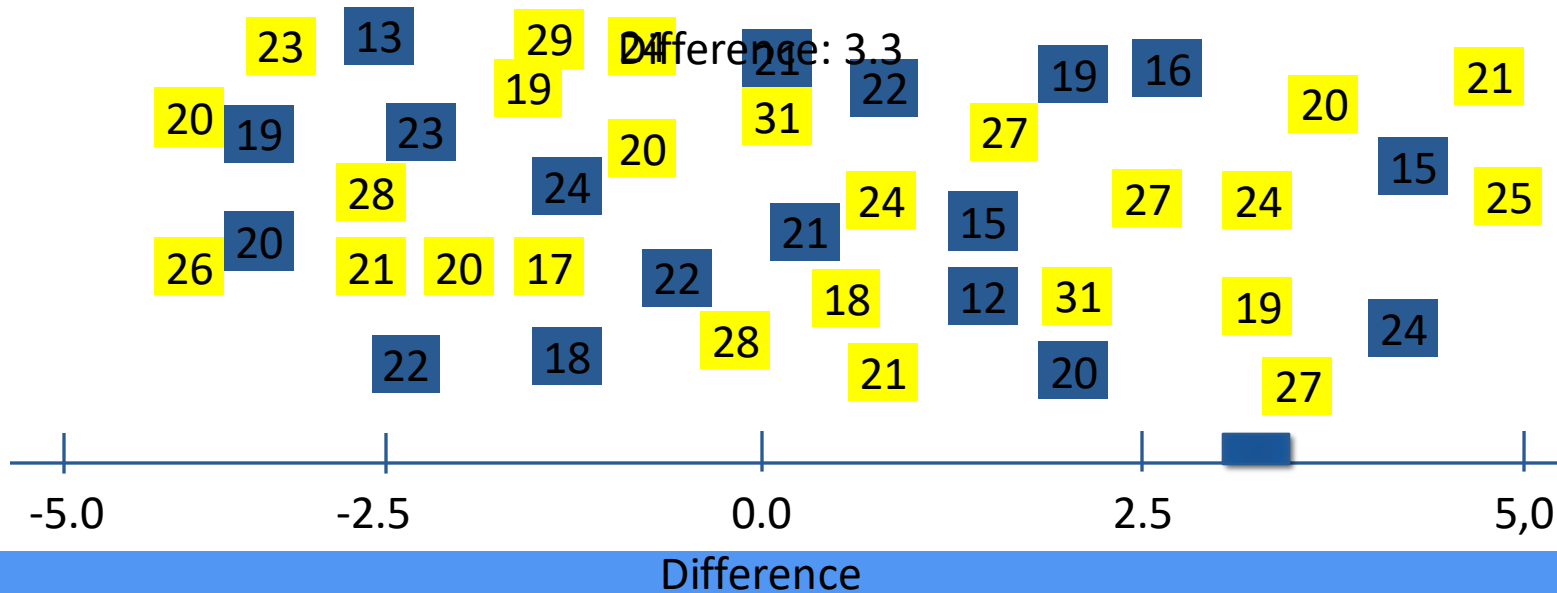
So let's simulate different groupings (**permutations**) and see what differences we get!



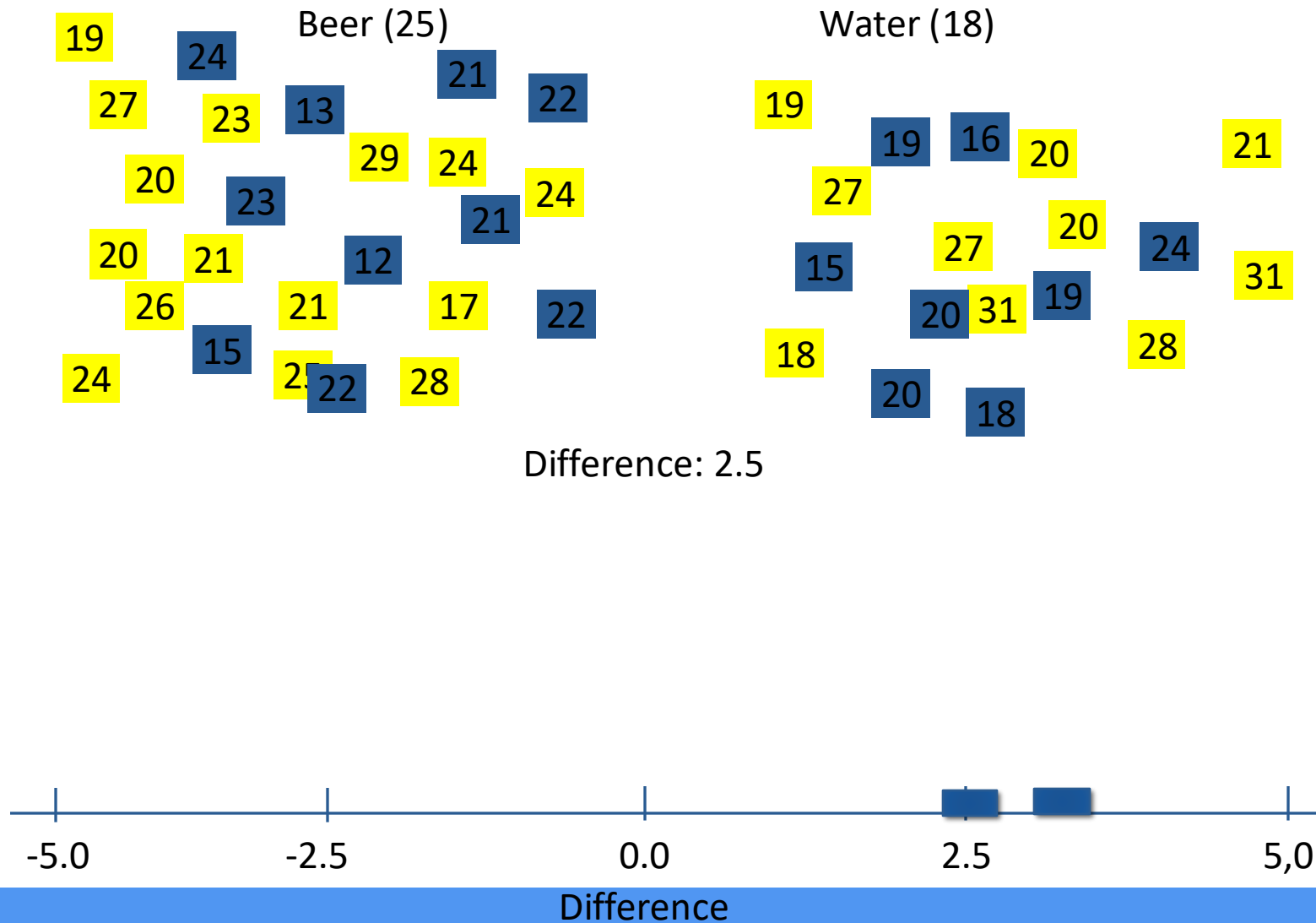
Permutation Test

Beer (25)

Water (18)



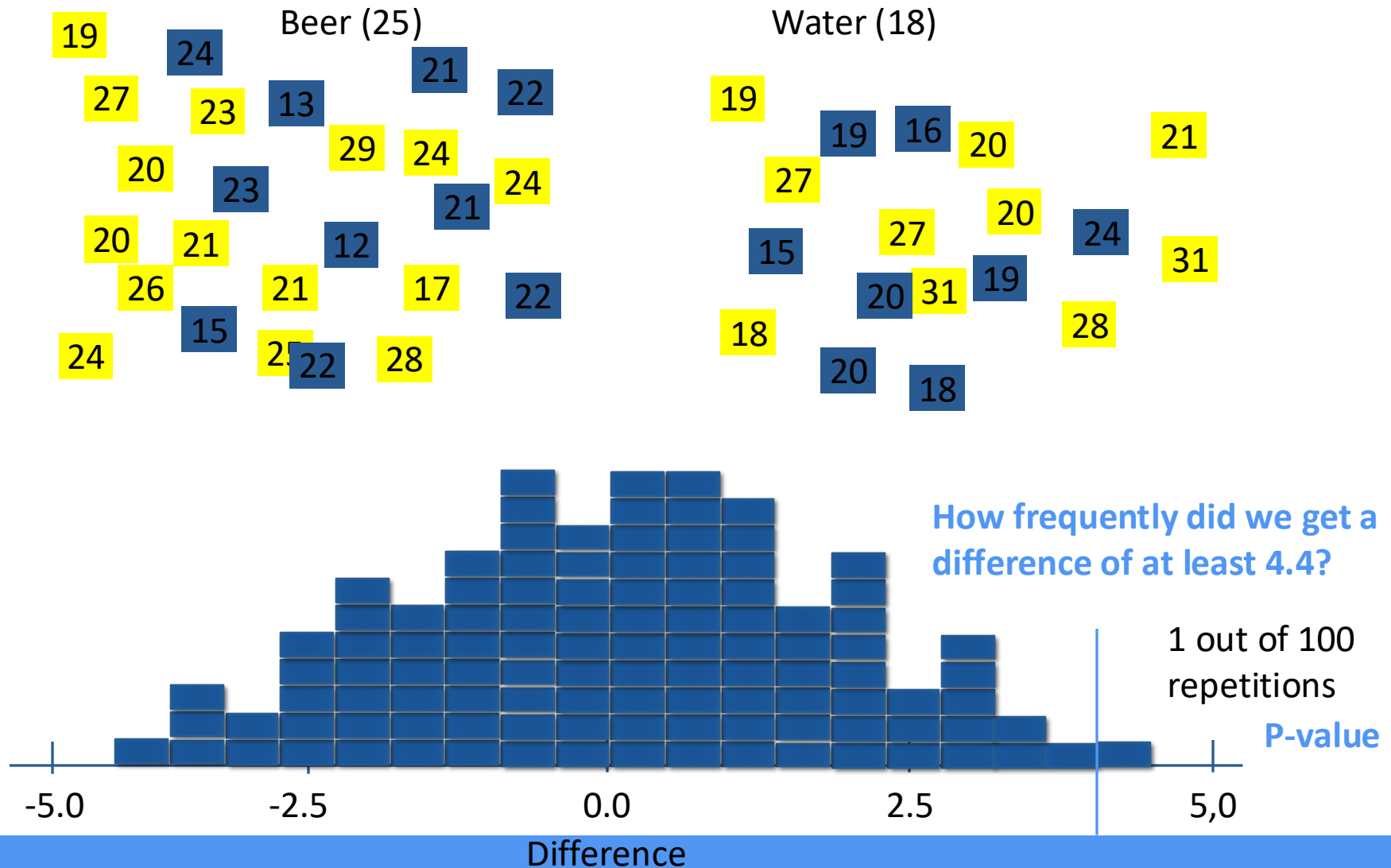
Permutation Test



Permutation Test

Observation from the data (test statistic):

The beer group has 4.4 more bites on average.

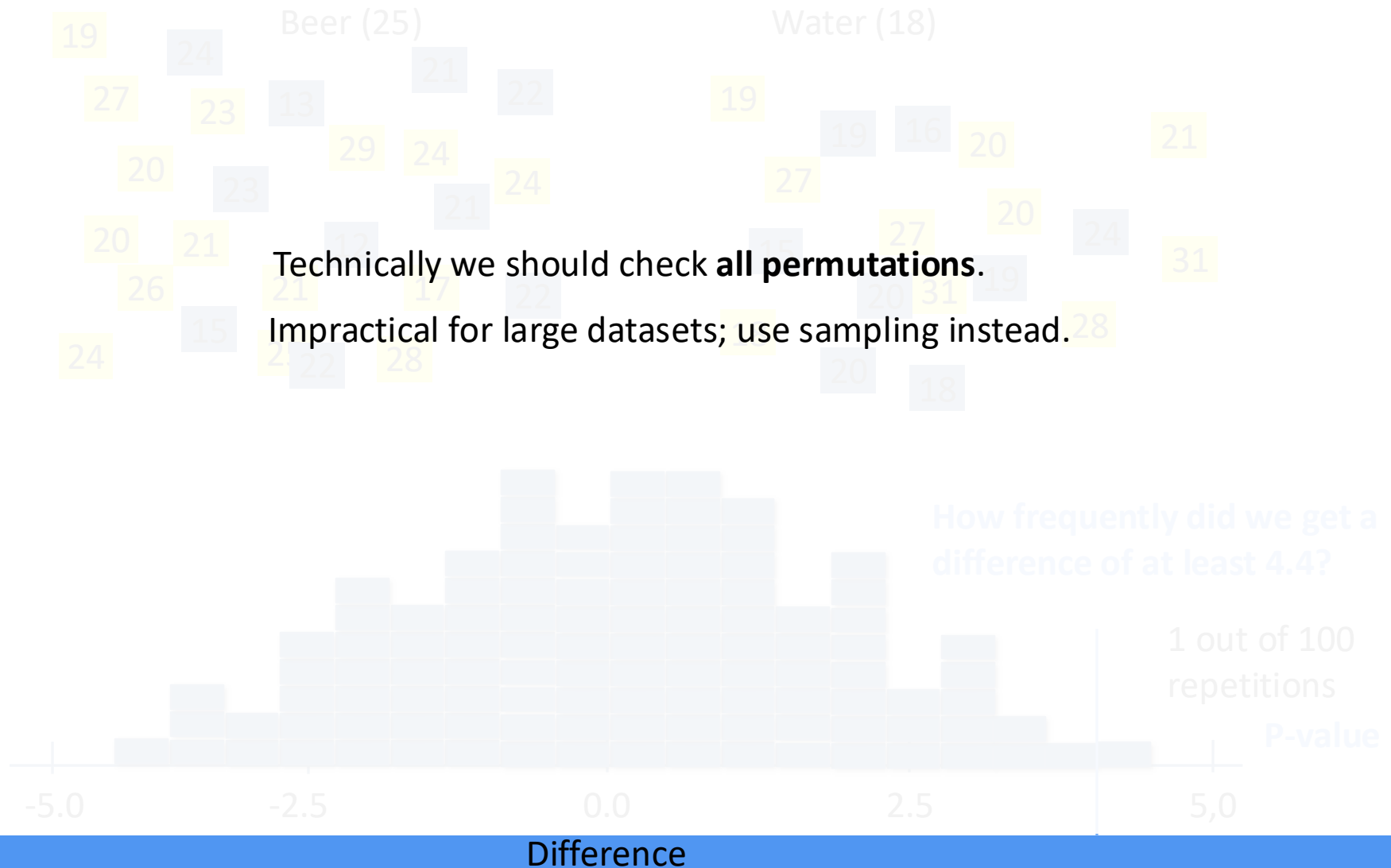


Permutation Test

Observation from the data (test statistic):

The beer group has 4.4 more bites on average.

Technically we should check **all permutations**.
Impractical for large datasets; use sampling instead.



Let's see the code

Checking the Hypothesis

```
def permutation_test(num_permutations, seed):
    random.seed(seed)

    # Combine the data
    combined_data = beer + water
    observed_diff = sum(beer) / N_beer - sum(water) / N_water

    # Try out permutations
    count = 0
    for _ in range(num_permutations):
        random.shuffle(combined_data)
        perm_beer = combined_data[:N_beer]
        perm_water = combined_data[N_beer:]
        perm_diff = sum(perm_beer) / N_beer - sum(perm_water) / N_water
        if perm_diff >= observed_diff:
            count += 1

    p_value = count / num_permutations
    print("Permutation test")
    print("=====")
    print("Number of sampled permutations:", num_permutations)
    print(f"Observed difference: {observed_diff:.4f}")
    print(f"Number of differences >= observed difference ({observed_diff:.4f}): {count}")
    print(f"P value: {p_value:.6f}")
    return p_value
```

Difference in means = 4.4, p-value = 0.00049

Difference in means = 4.4, P-value = 0.00049

If there were actually no difference in the means of the two groups (i.e., if the null hypothesis were true)...

The probability of observing a difference of **at least 4.4** is approximately **0.00049**.



It does not imply that the probability of the null hypothesis being true is 0.00049.

Since we have our **p-value $\leq \alpha$** ($0.00049 \leq 0.01$), we **reject the null hypothesis**.

What This Means (continued)

- **Null hypothesis:** Beer consumption does not increase human attractiveness to malaria mosquitoes
- **Alternative hypothesis:** Beer consumption increases human attractiveness to malaria mosquitoes

Difference in means = 4.4, p-value = 0.00049

$\alpha = 0.01$

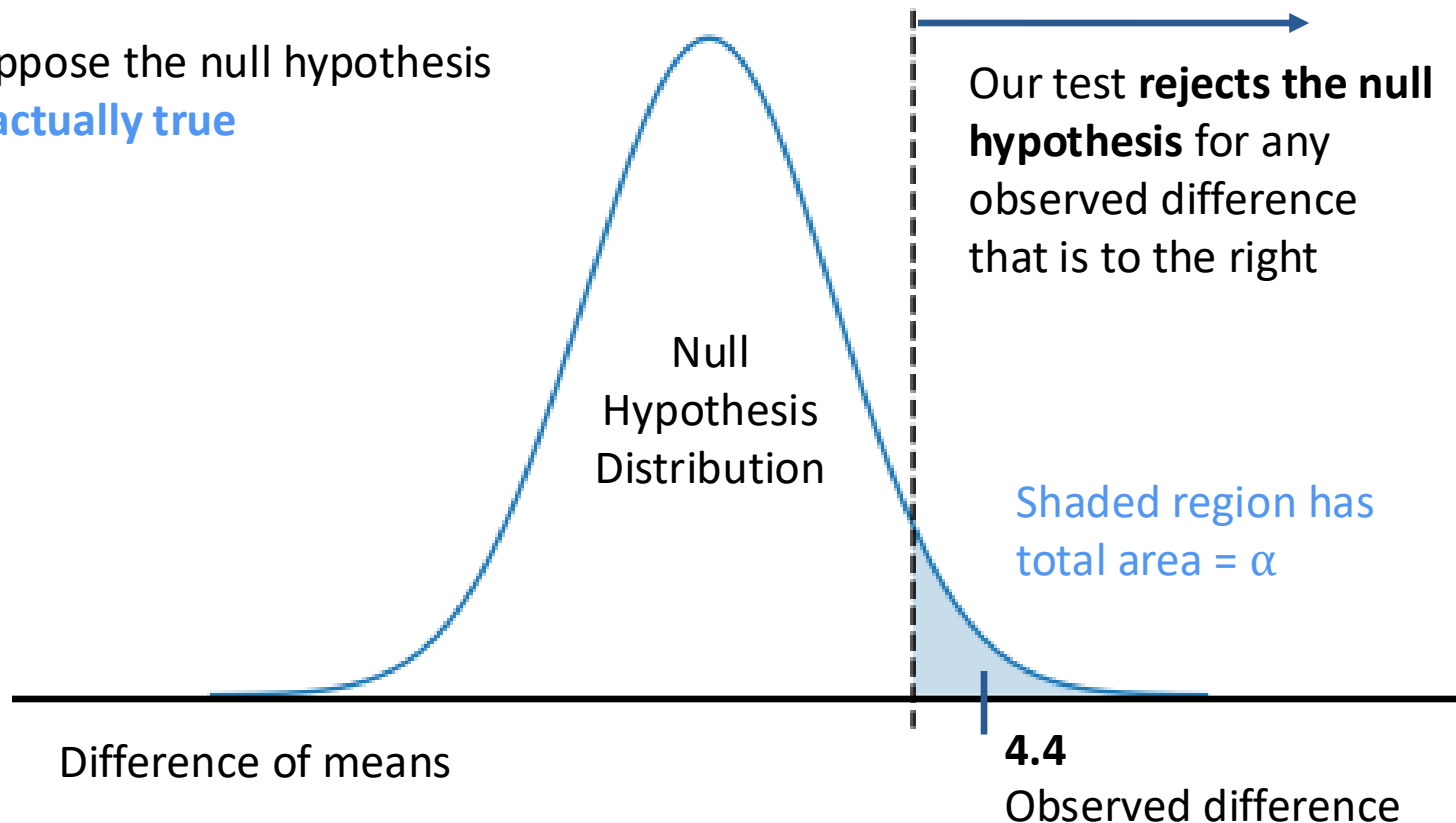
Since we have our **p-value $\leq \alpha$** ($0.00049 \leq 0.01$), we **reject the null hypothesis**.

Can we be *certain* that the alternative hypothesis is true?

Type I errors (false positives)

- It's still possible for the difference of means to have occurred by random chance!

Suppose the null hypothesis is **actually true**








Type I errors (false positives)

- It's still possible for the difference of means to have occurred by random chance!
- Our statistical test rejects the null hypothesis when $p \leq \alpha$
- If we mistakenly reject: Type I error (“false positive”)
- Probability of a type I error = α

- It's also possible that **we mistakenly fail to reject** the null hypothesis
- Type II error (“false negative”)
- Probability of type II error defined as “ β ”
 - Value determined by statistical power (not covered in this subject)

Type I and II errors

		Null Hypothesis Is	
		True	False
Decision about null hypothesis	Accept	 Correct (true negative) (Probability = $1 - \alpha$)	 Type II Error (false negative) (Probability = β)
	Reject	 Type I Error (false positive) (Probability = α)	 Correct (true positive) (Probability = $1 - \beta$)

 These are not probabilities of the null hypothesis being true/false.

■ **Sensitivity (a.k.a. recall)**

- $\text{True positives} / (\text{True positives} + \text{False negatives})$
- “Of all the actually positive samples, how many do we identify?”

■ **Specificity**

- $\text{True negatives} / (\text{True negatives} + \text{False positives})$

■ **Positive predictive value (a.k.a. precision)**

- $\text{True positives} / (\text{True positives} + \text{False positives})$
- “When we predict positive, how often are we actually right?”

■ **Negative predictive value**

- $\text{True negatives} / (\text{True negatives} + \text{False negatives})$

What is the interpretation of $p < 0.05$?

- A) The chances are greater than 1 in 20 that a difference would be found if the study were repeated.
- B) The probability is less than 1 in 20 that a difference this large or larger could occur by chance alone.
- C) The probability is greater than 1 in 20 that a difference this large could occur by chance alone.
- D) The chance is 95% that the study is correct
- E) None of the above

Misconception 1

“In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

What is wrong with this?

Misconception 1

“In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

This is an understandable but categorically wrong interpretation because the *P* value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true. “

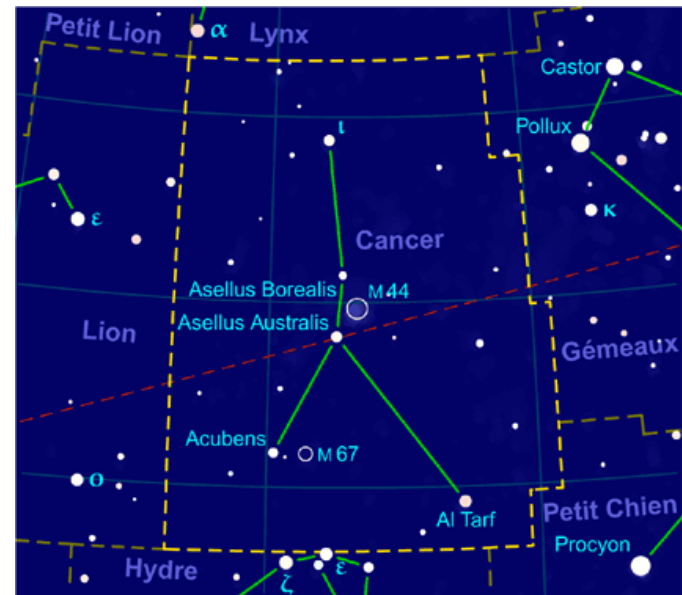
A non significant difference (e.g., $p = 0.5$) means there is no difference between groups.

What is wrong with this?

A non significant difference (e.g., $p = 0.5$) means there is no difference between groups.

- A non significant difference only means the null effect is statistically consistent with the observation
- It does not make the null effect most likely

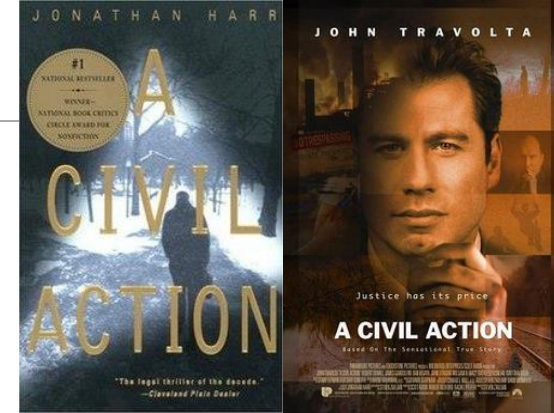
- A **cancer cluster** is defined by the CDC as “a greater-than-expected number of cancer cases that occurs within a group of people in a geographic area over a period of time”
- About 1000 “cancer clusters” per year are reported to health authorities in the U.S.
- Vast majority, but not all, are deemed not significant



A Local Example

By **Michael Weisskopf**
January 29, 1987

W.R. Grace & Co. was indicted yesterday on charges of falsifying statements to the Environmental Protection Agency in connection with the poisoning of drinking-water wells in Woburn, Mass. The dumping of cancer-causing solvents into the wells has been blamed for the city's high rate of leukemia among children.

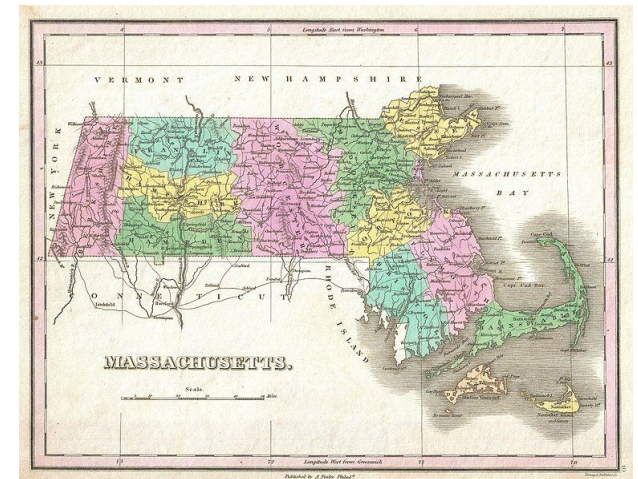


About 11 miles
from here



A Hypothetical Example

- Massachusetts is about 10,000 square miles
- About 36,000 new cancer cases per year
- An attorney partitioned state into 1000 regions of 10 squares miles each, and looked at distribution of cases
 - Expected number of cases per year per region: 36
- Discovered that region 111 had 30% more new cancer cases than expected over a 3 year period!
- How worried should residents be?



https://commons.wikimedia.org/wiki/File:1827_Finley_Map_of_Massachusetts_-_Geographicus_-_Massachusetts-finley-1827.jpg

How Likely Is it Just Bad Luck?

```
#Initialize constants
num_cases_per_year = 36000
num_years = 3
state_size = 10000
community_size = 10
num_communities = state_size//community_size
multiple = 1.3 #because 30% more cases than expected
cases_per_sim = num_cases_per_year*num_years
```

How Likely Is it Just Bad Luck?

```
def find_prob(num_communities, community,
              cases_per_sim, multiple, num_sims):
    num_times_over = 0
    threshold = (cases_per_sim/num_communities)*multiple
    for t in range(num_sims):
        cases = [0]*num_communities
        for i in range(cases_per_sim):
            cases[random.choice(range(num_communities))] += 1
        if cases[community] > threshold:
            num_times_over += 1
    return 1 - num_times_over/num_sims

num_trials = 10
num_sims_per_trial = 20
probs = []
for t in range(num_trials):
    print('Starting trial', t)
    probs.append(find_prob(num_communities, 111, cases_per_sim,
                           multiple, num_sims_per_trial))
print('Est. prob. of being a random event =',
      round(1 - probs[-1], 4))
print('Standard deviation of trials =',
      round(np.std(probs), 4))
```

Est. prob. of being a random event = 0.0

Standard deviation of trials = 0.015

- Seems highly unlikely that it was a random event
- Time to check the water supply and air quality in area 111?

Do you buy my analysis?

Does Code Answer the Right Question?

```
def find_prob(num_communities, community,
              cases_per_sim, multiple, num_sims):
    num_times_over = 0
    threshold = (cases_per_sim/num_communities)*multiple
    for t in range(num_sims):
        cases = [0]*num_communities
        for i in range(cases_per_sim):
            cases[random.choice(range(num_communities))] += 1
    → if cases[community] > threshold:
        num_times_over += 1
    return 1 - num_times_over/num_sims
```

Code answers what is the probability that a **SPECIFIC** community had that many cases

Attorney asked is there **ANY** community that had that many cases

A variant of cherry picking called
multiple hypothesis testing

Fixing the Code



```
def find_prob(num_communities, community,
              cases_per_sim, multiple, num_sims):
    num_times_over = 0
    threshold = (cases_per_sim/num_communities)*multiple
    for t in range(num_sims):
        cases = [0]*num_communities
        for i in range(cases_per_sim):
            cases[random.choice(range(num_communities))] += 1
    → if max(cases) > threshold:
        num_times_over += 1
    return 1 - num_times_over/num_sims
```

Est. prob. of being a random event = 0.75

Standard deviation of trials = 0.0955

LIES
DAMNED LIES
and
STATISTICS

It's True, But ...

- “less than 10 percent of documented transmission [of Covid-19], in many studies, have occurred outdoors”
 - Dr. Rochelle Walensky, CDC Director, April 27, 2021



Poll: What do you think actual fraction is?

10%

5%

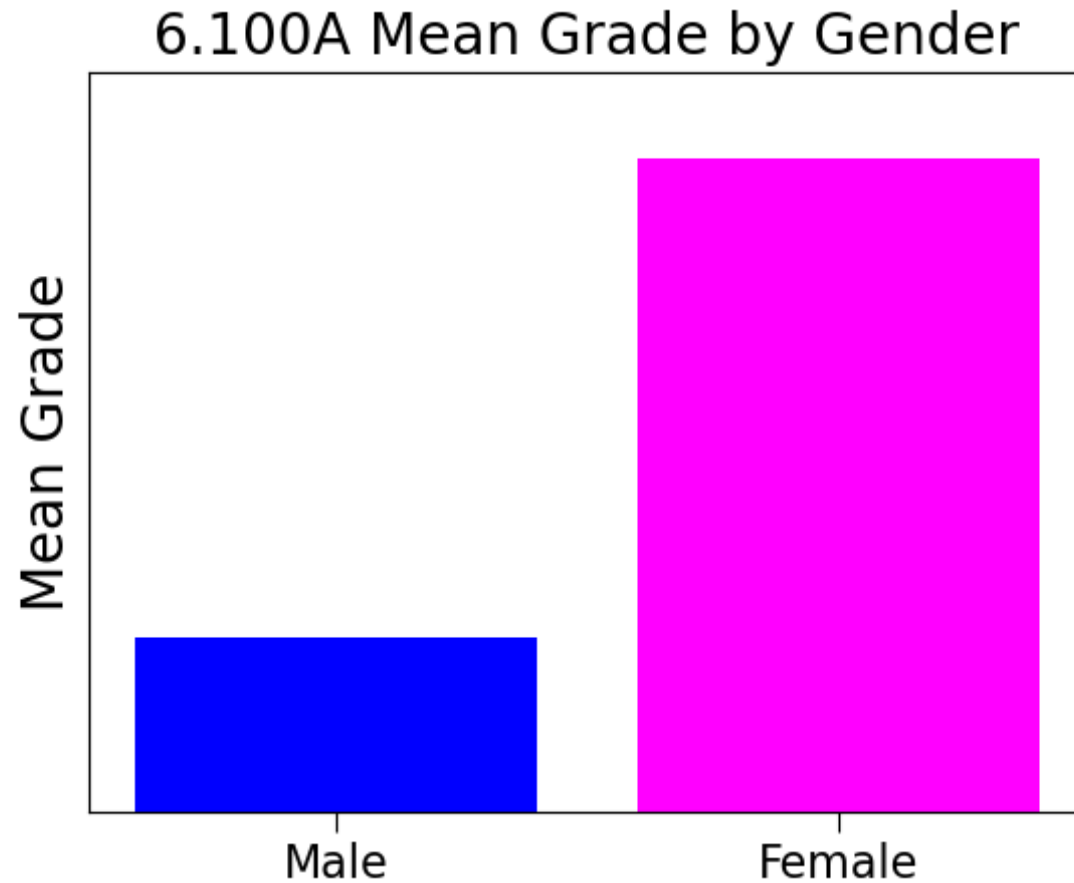
2%

1%

Estimate is <1%

Moral: Spurious baselines are useful for constructing statements that are both **true** and **deceiving**

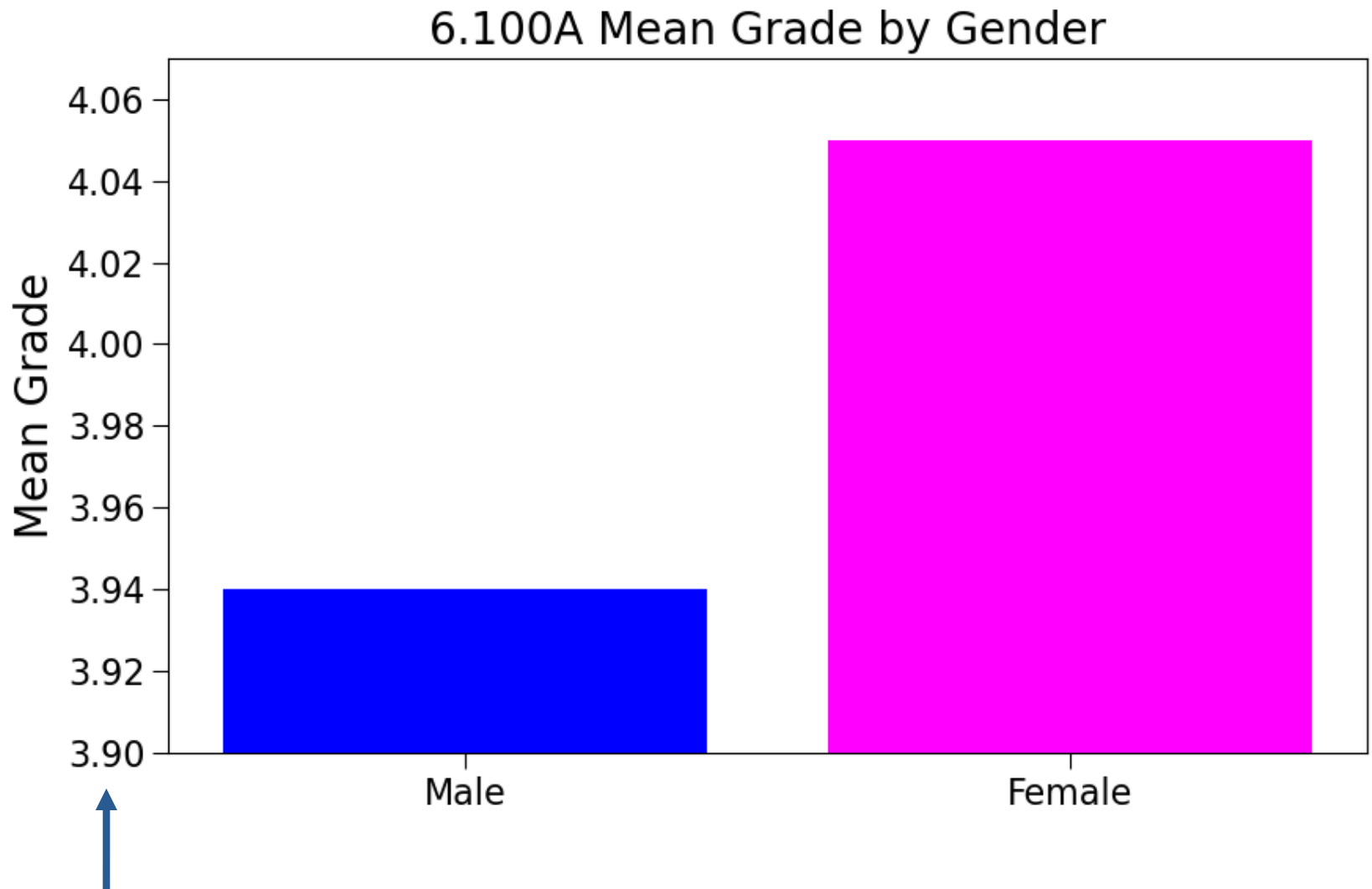
Is a Picture Always Worth 1000 Words?



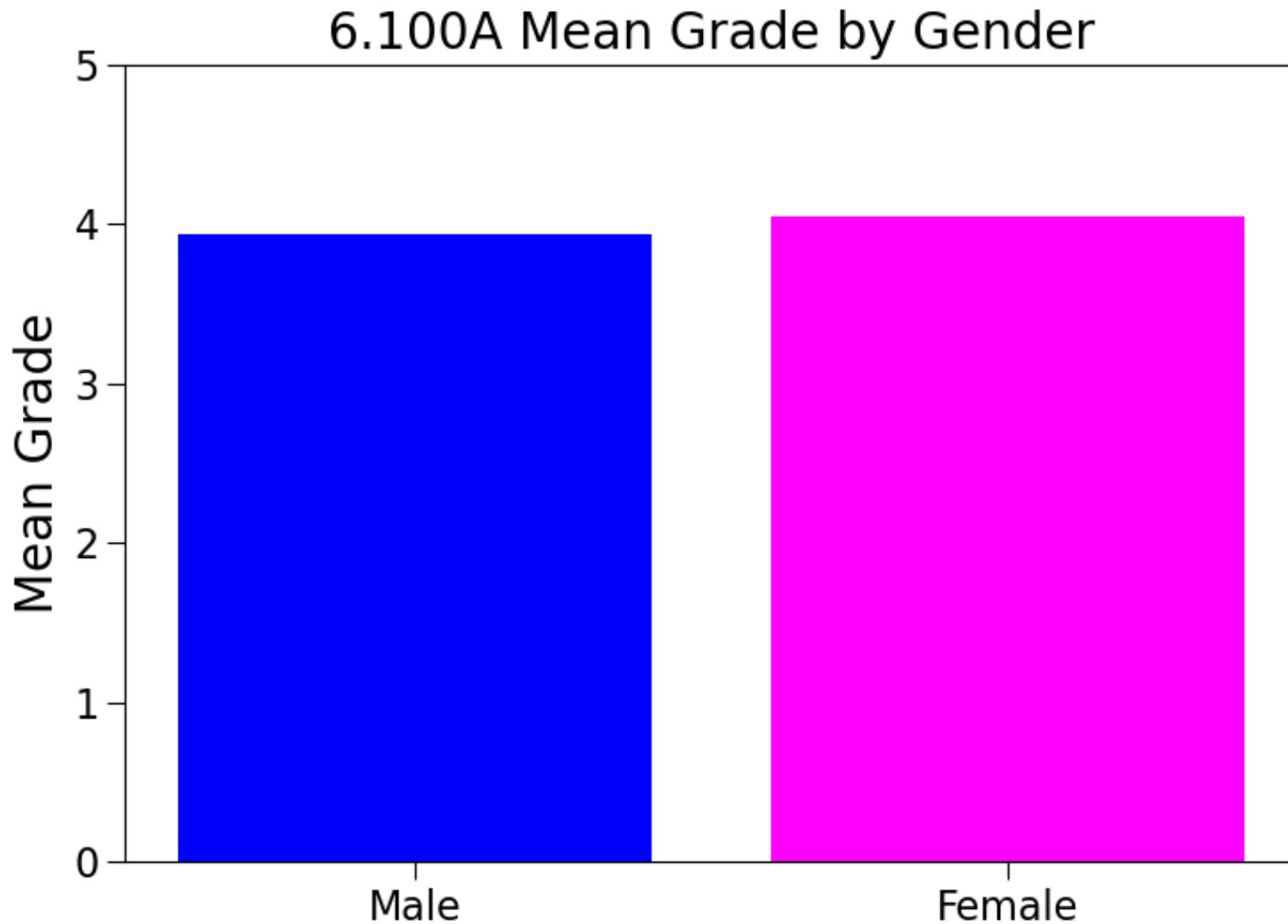
This plot is best explained by

1. Women are better coders
2. The plot is fabricated
3. Bias in grade assignment
4. None of the above

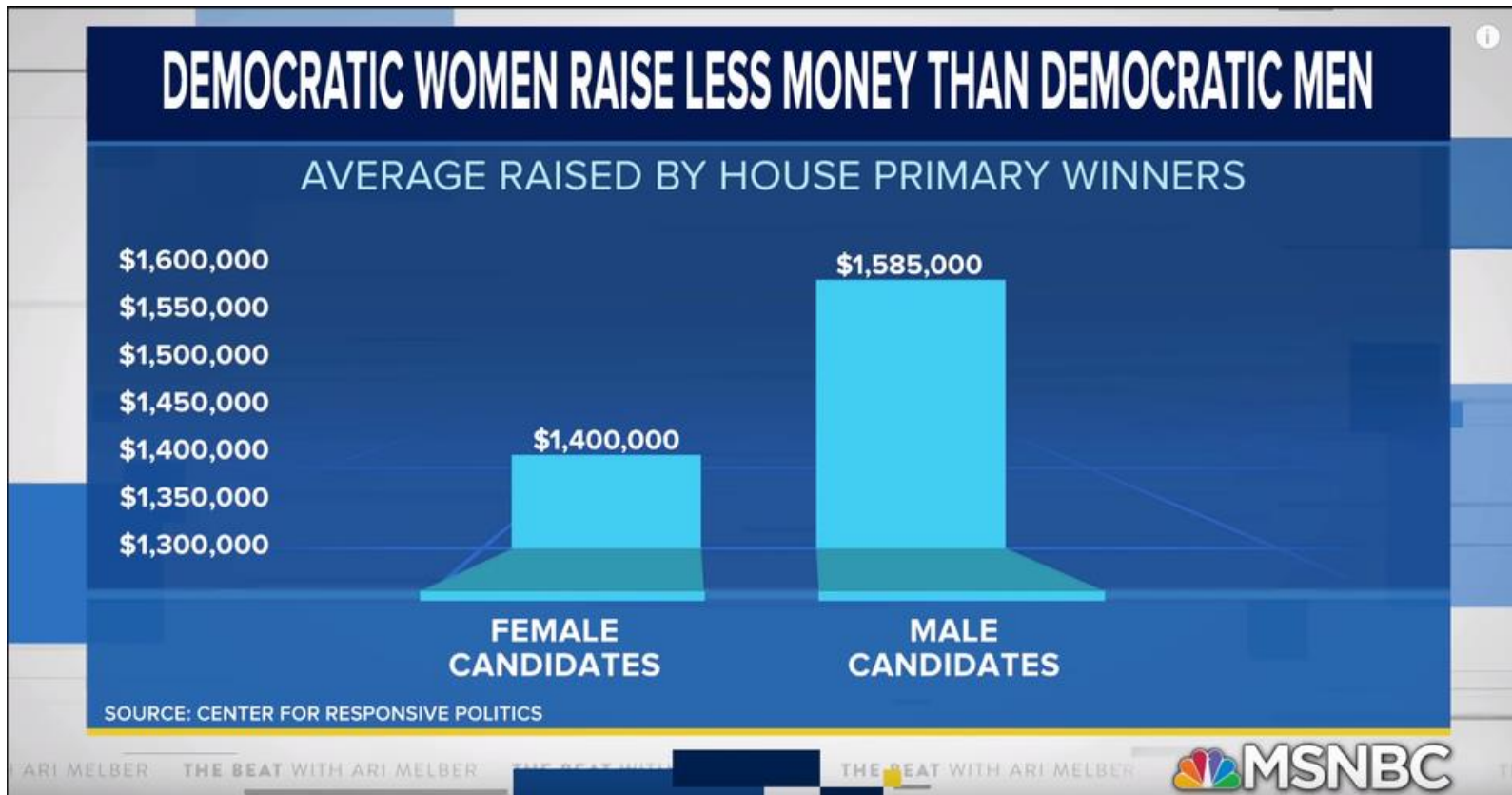
Is a Picture Always Worth 1000 Words?



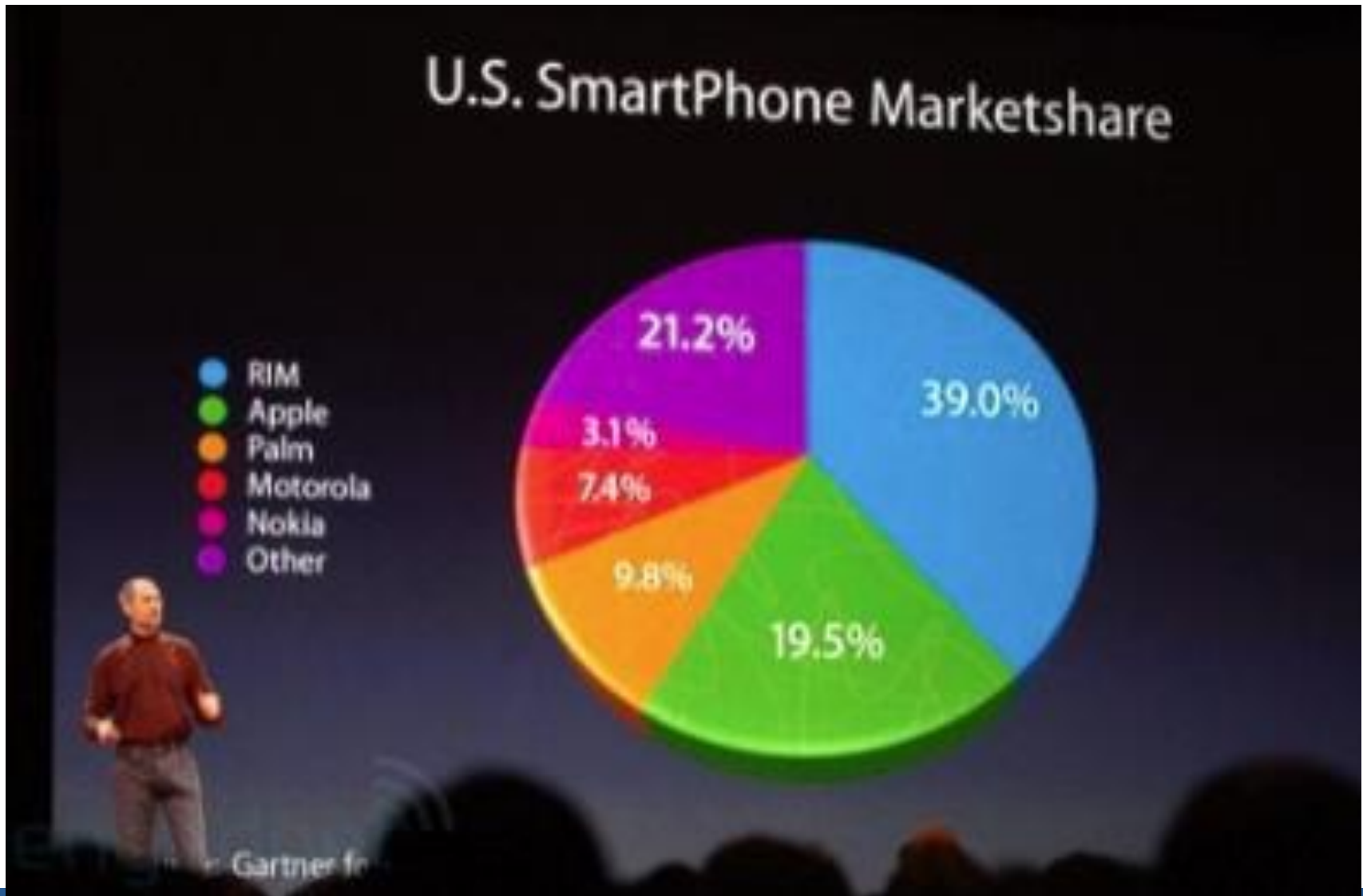
The Same Data, Different Picture



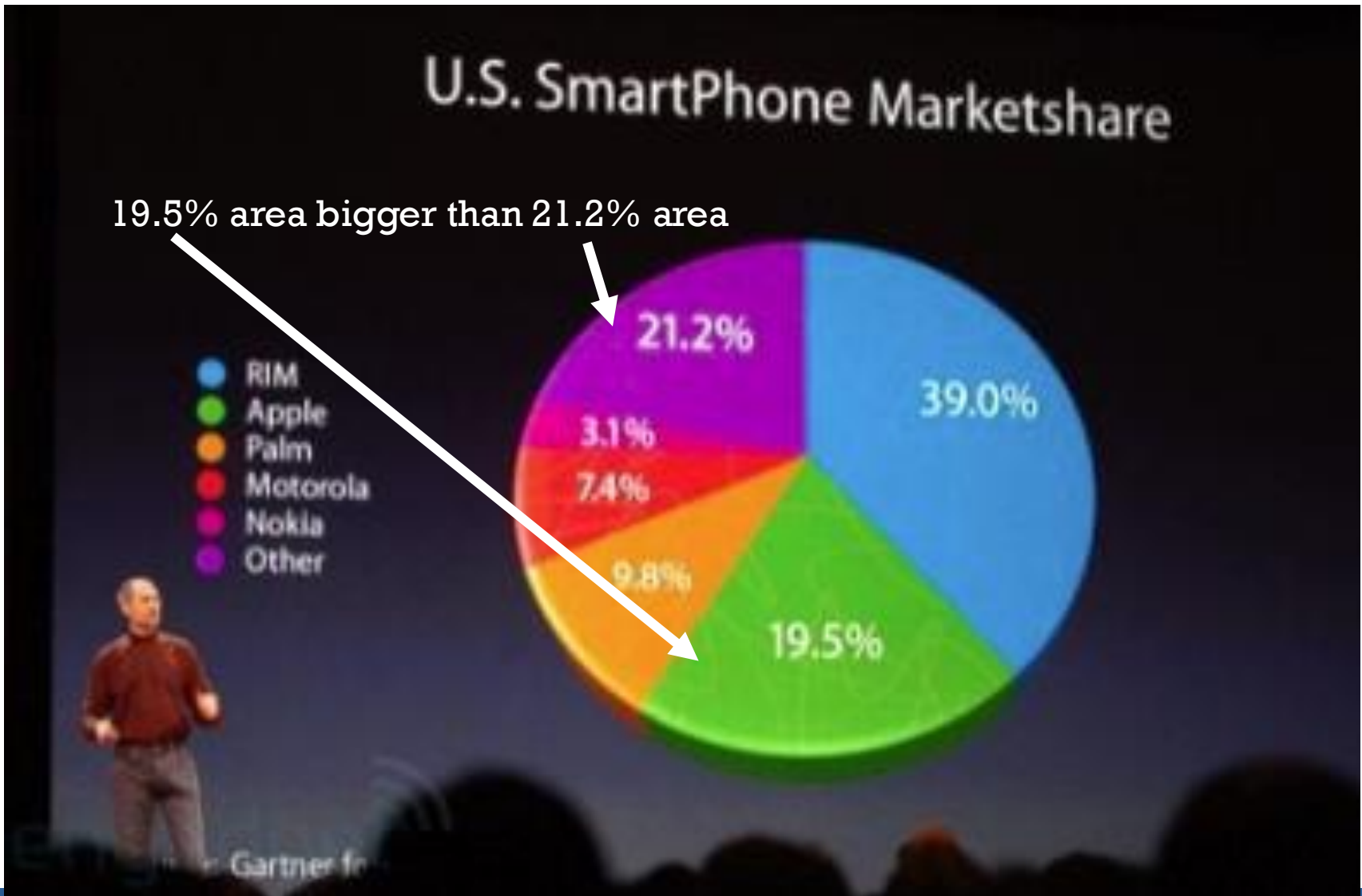
Example of Truncated Y-axis



Apple WWDC 2008



APPLE WWDC 2008

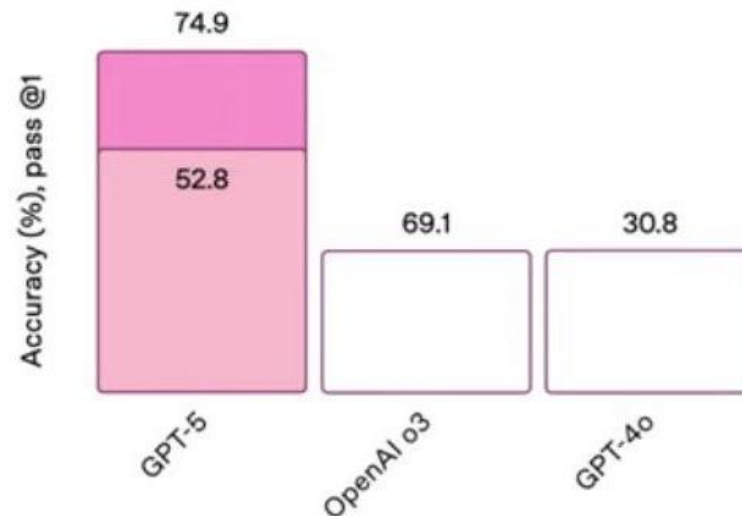


Academic

SWE-bench Verified

Software engineering

Without thinking With thinking



* Later clarified as an unintentional error

Some Other Ways to Deceive with Plots



Is it linear?

Some Other Ways to Deceive with Plots



Spacing on x-axis
9, 6, 15 months

6 month period occupies more
horizontal space than 15
month period

Spacing on y-axis doesn't
match numbers

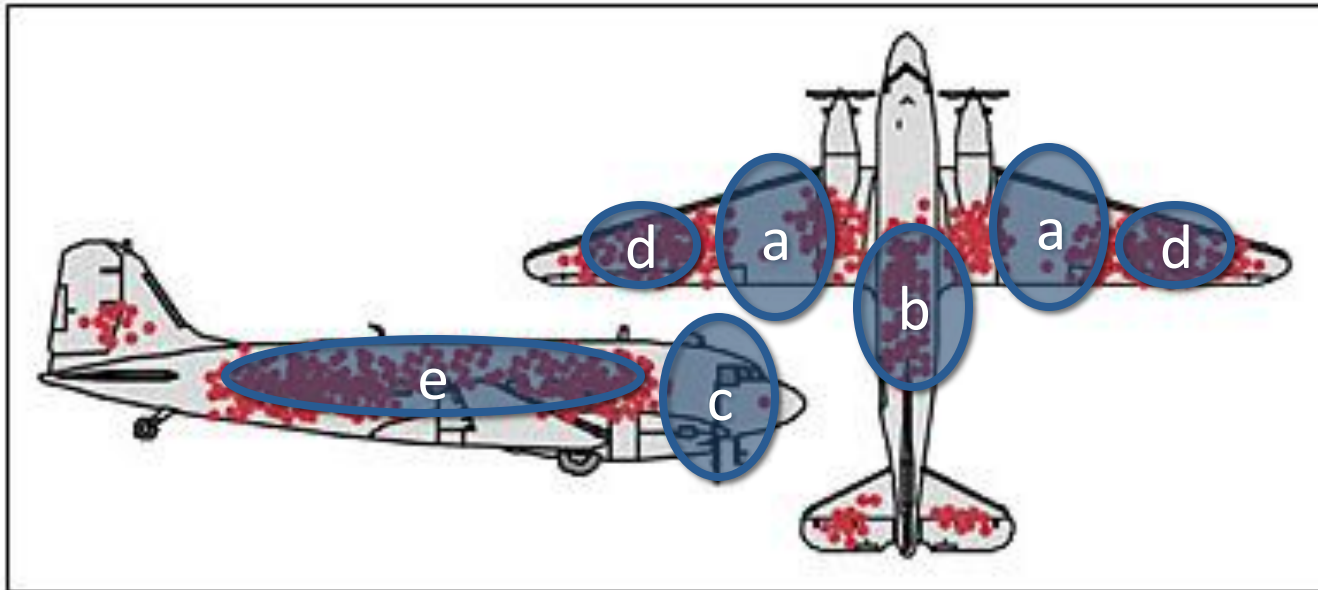


Same data

**Moral: Look carefully at
the axes labels and scales**

What do you think?

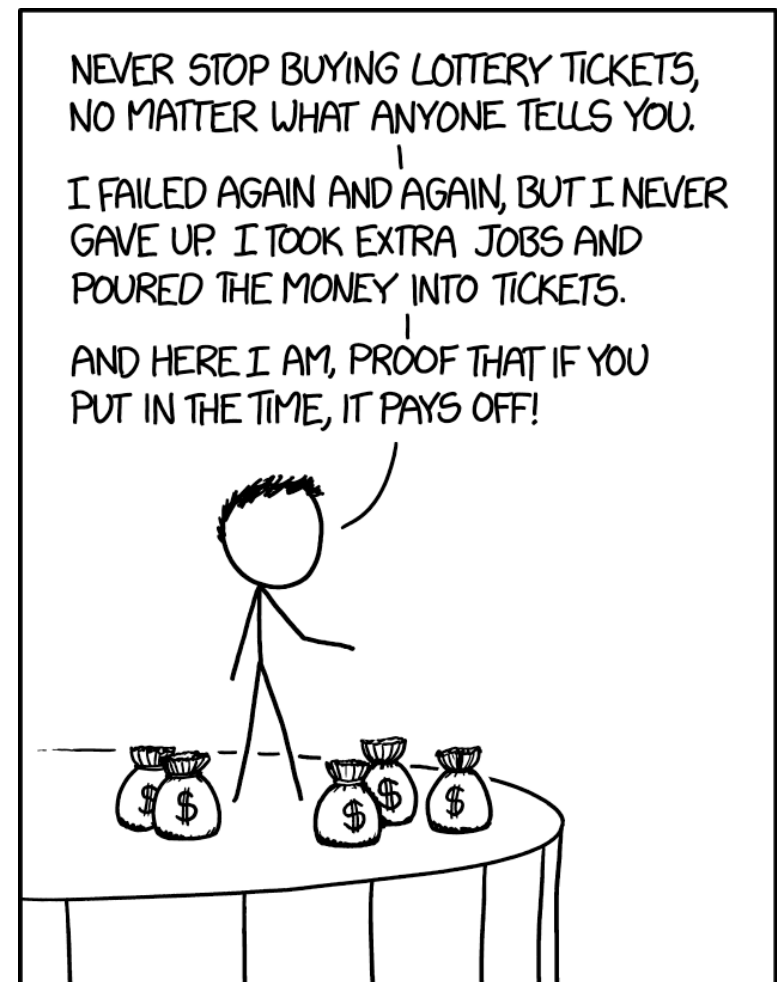
Back during World War II, the Air Force lost a lot of planes to German anti-aircraft fire. So they decided to armor them up. But where to put the armor?



Credit: Cameron Moll

Non-representative Sampling

- “Convenience sampling” not usually random, e.g.,
 - **Survivor bias**, e.g., course evaluations at end of course
 - **Non-response** bias, e.g., opinion polls conducted by mail or online

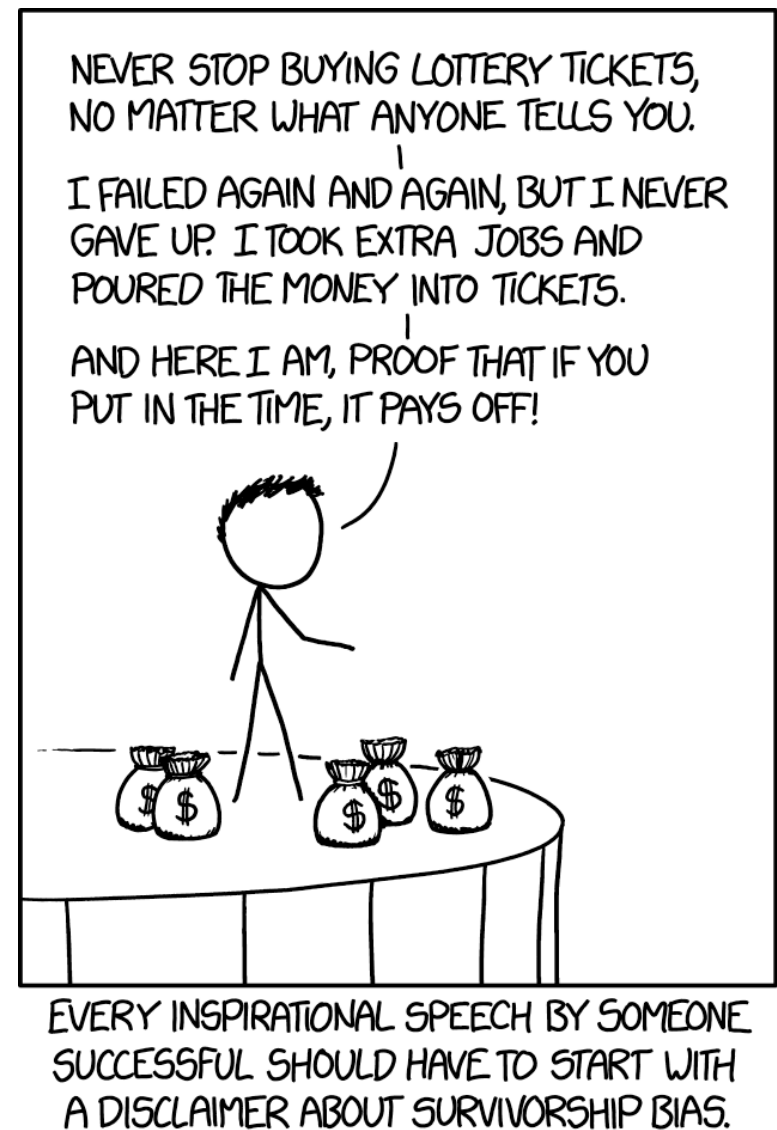


Non-representative Sampling

- “Convenience sampling” not usually random, e.g.,
 - **Survivor bias**, e.g., course evaluations at end of course
 - **Non-response** bias, e.g., opinion polls conducted by mail or online

**Course evaluations are open until
Mon, Dec 15 at 9 am**

[https://eduapps.mit.edu/subjeval/
studenthome.htm](https://eduapps.mit.edu/subjeval/studenthome.htm)



- When samples not random and independent, we can still do things like compute means and standard deviations
- **But we should be very wary of using statistical techniques to draw conclusions from them**
- **Moral: Understand how data was collected, and whether assumptions used in the analysis are satisfied**

- When drawing inferences from data, **skepticism** is merited.
- But remember, skepticism and denial of facts is different.
- “Doubt, indulged and cherished, is in danger of becoming denial; but if honest, and bent on thorough investigation, it may soon lead to full establishment of the truth.” – Ambrose Bierce

