Confidence intervals, Statistical significance

6.100 LECTURE 10

SPRING 2026

Announcements

- Pset 2 checkoff due this evening
- Pset 3 due next Mon 3/9
- Scores page to be released Friday

Random seeds

A quick word on randomness and seeds

- A stochastic program, by definition, gives a different result each time
  - e.g., random splits will yield different r^2 values on validation data, and hence, possibly different best fit degree
- For development/debugging purposes, often advantageous to have deterministic behavior
  - **`random.seed(`*`number`*`)`** sets an internal state for the Python's random number generator
  - results in all subsequent random outputs being deterministic
  - however, still hard to predict if you don't know / aren't considering how Python actually generates random numbers

More info about random number generation

- Actually pretty non-trivial to generate truly random numbers
- Common strategy is to use the **Mersenne Twister**
  - outputs a sequence of numbers between 0 and $2^{32}-1$
  - given initial state, sequence is deterministic and repeats every $2^{19937}-1$ steps
  - for most pratical purposes, output looks random ***(pseudo-random)***
  - setting **`random.seed()`** determines the initial state
- Documentation
  - https://docs.python.org/3/library/random.html
  - https://docs.python.org/3/library/random.html#random.seed
  - https://en.wikipedia.org/wiki/Mersenne_Twister

Confidence intervals

Review: estimate mean of X through samples

- Run an experiment consisting of $\boldsymbol{n}$ samples of random variable $\boldsymbol{X}$
  - as $\boldsymbol{n}$ increases, samples approximate $\boldsymbol{X}$'s distribution
- Estimate mean of $\boldsymbol{X}$ using mean of samples $\boldsymbol{M}$, also a random variable
  - $\boldsymbol{M = (X_1 + X_2 + \cdots + X_n) \ / \ n}$
- **Central Limit Theorem**
  - as $\boldsymbol{n}$ increases, $\boldsymbol{M}$'s distribution approaches a normal
  - Same mean as $\boldsymbol{X}$'s mean: $\boldsymbol{\mu_M = \mu_X}$
  - Variance gets scaled down by $\boldsymbol{n}$: $\boldsymbol{\sigma_M^2 = \sigma_X^2 \ / \ n}$

Review: verifying CLT

- Run k experiments to get k samples of random variable M
  - as k increases, samples approximate M's distribution

- In practice, often expensive enough to run a single experiment
  - each experiment contains n samples of X
  - can we infer something useful about M from a single experiment?

Empirical rule

- We can exploit the fact that M should be close to normally distributed
- Percentile properties of the normal distribution
  - within $\pm\sigma$ 68.3%
  - **within $\pm 1.96\sigma$ 95.0%**
  - **within $\pm 2\sigma$ 95.4%**
  - within $\pm 3\sigma$ 99.7%
  - within $\pm 4\sigma$ 99.99%
  - within $\pm 5\sigma$ 99.99 99 4%
  - within $\pm 6\sigma$ 99.99 99 99 8%

Expressing confidence from one sample of M

- Given one sample/estimate of $\boldsymbol{M}$
  - composed of $\boldsymbol{n}$ samples of $\boldsymbol{X}$
  - **95% chance** it will lie within $\pm 2\boldsymbol{\sigma_M}$ of the true mean $\boldsymbol{\mu_M = \mu_X}$
- So what can we infer about the true mean $\boldsymbol{\mu_M}$?
  - well, keep in mind we can never truly know it
  - we only know what our sample mean $\boldsymbol{m}$ is (sampled from $\boldsymbol{M}$'s distribution)
  - but we can relate the two
- Imagine $\boldsymbol{k}$ samples of $\boldsymbol{M}$, $\boldsymbol{m_1}$, ..., $\boldsymbol{m_k}$
  - about 95% of them should lie within $\boldsymbol{\mu_M \pm 2\sigma_M}$
  - so about 95% of the intervals $\boldsymbol{m_i \pm 2\sigma_M}$ should contain $\boldsymbol{\mu_M}$
- Back to a single sample $\boldsymbol{m}$ of $\boldsymbol{M}$
  - the range $\boldsymbol{m \pm 2\sigma_M}$ is a **95% confidence interval** on the true mean

Common misconceptions

- Given a 95% confidence interval $[a, b]$:

1. there is a 95% chance the true mean lies within the interval
   - ok, this one kind of depends
   - **frequentist interpretation:** the true mean is already fixed, it either lies within or outside an interval, we just don't know it
   - **Bayesian interpretation:** probabilities represent beliefs, we're 95% sure about that statement
2. 95% of the data lies within the interval
3. if we repeat the experiment (i.e., get a sample of M), there is a 95% chance its mean would lie within the interval

Statistical significance

Another perspective: hypothesis testing

- Confidence intervals express where we believe a true statistic might lie
  - but what does that mean for our experiment? was it "successful"?
- Measure "success" with respect to a **hypothesis**
  - **null hypothesis:** "under some assumptions, this is where we'd expect the result to be"
- What does "where we'd expect the result to be" mean?
  - imagine running the experiment many times (under those assumptions), build up a distribution of results
  - if original result is an **"outlier",** then that's **statistically significant**
  - leads us to believe the assumptions for the null hypothesis aren't true, **reject the null hypothesis**

Example of 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

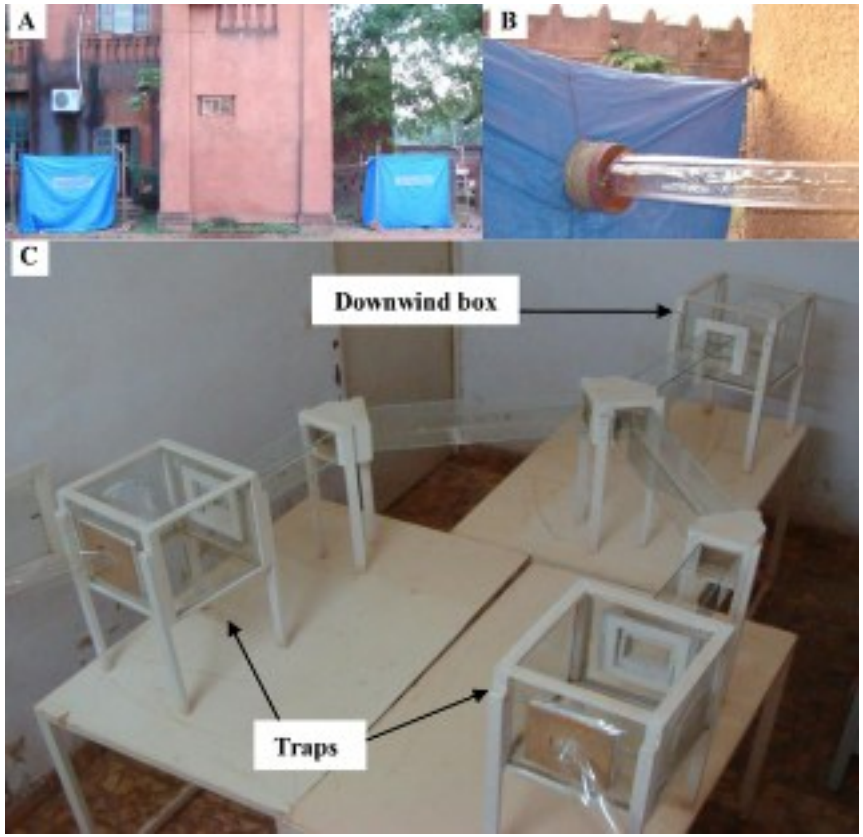
Mean: 23.6

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?



Beer Consumption #Human Attractiveness to Malaria Mosquitoes
Thierry Lefèvre , Louis-Clément Gouagna, Kounbobr Roch Dabiré, Eric Elguero, Didier Fontenille, François Renaud, Carlo Costantini, Frédéric Thomas

Example of hypothesis testing

- **Null hypothesis:** Beer consumption **does not increase** human attractiveness to malaria mosquitoes
  - **Alternative hypothesis:** Beer consumption **increases** human attractiveness to malaria mosquitoes
- Define a **significance level for rejecting** the null hypothesis
  - e.g., $\boldsymbol{\alpha = 0.05}$
  - if original outcome falls outside $\boldsymbol{1 - \alpha}$ percentile of the experiment distribution under the null hypothesis, reject
  - portion of results more extreme than the original outcome is the $\boldsymbol{p}$**-value**
- **Issue:** how to get that distribution?
  - **ideal:** conduct many more experiments (so many health waivers…)
  - **proxy:** swap around (i.e., permute) the original data

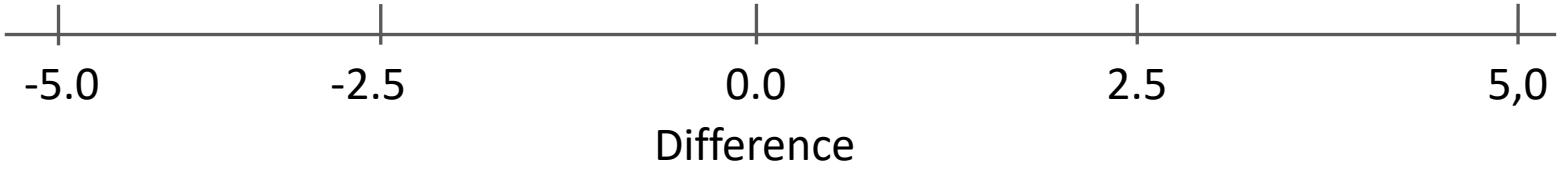
Permutation test

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 of means: 4.4

Observation from the data (test statistic):
The beer group has **4.4** more bites on average.



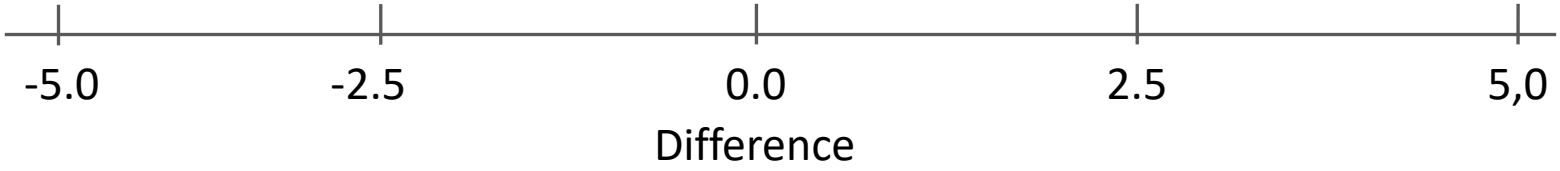
Permutation test

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 of means: 4.4

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

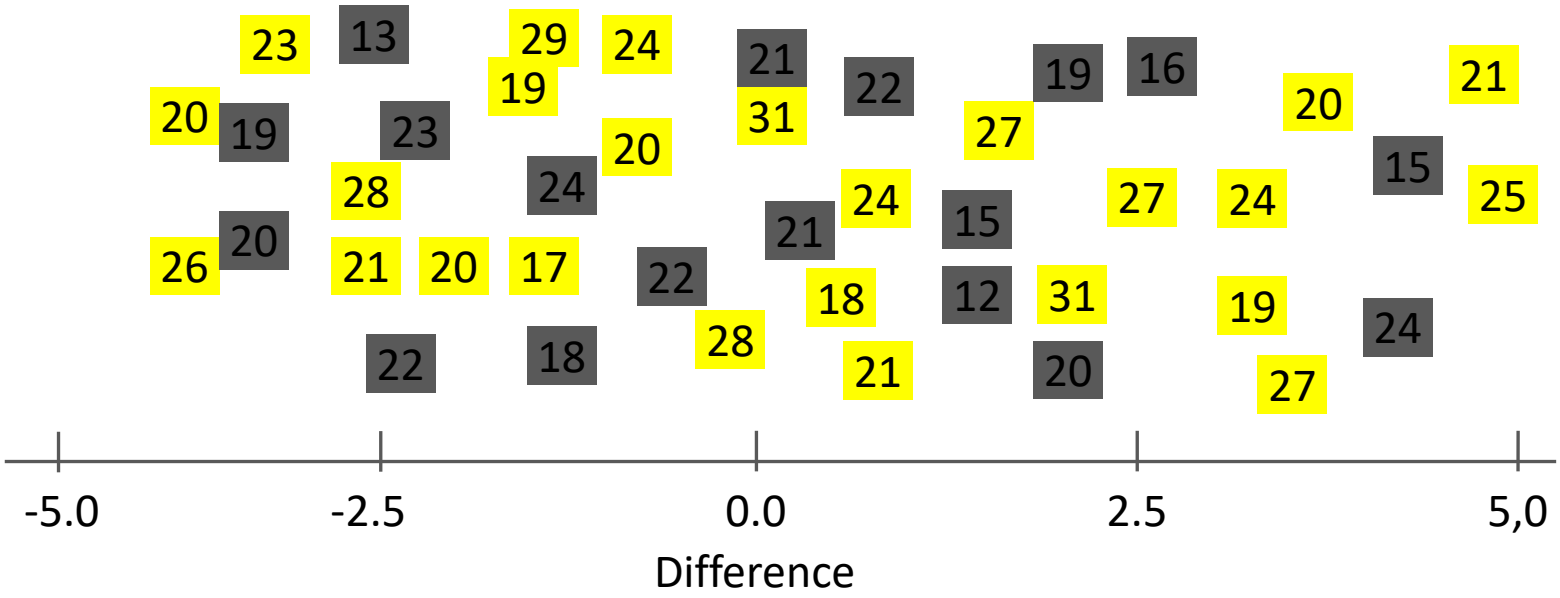


Permutation test

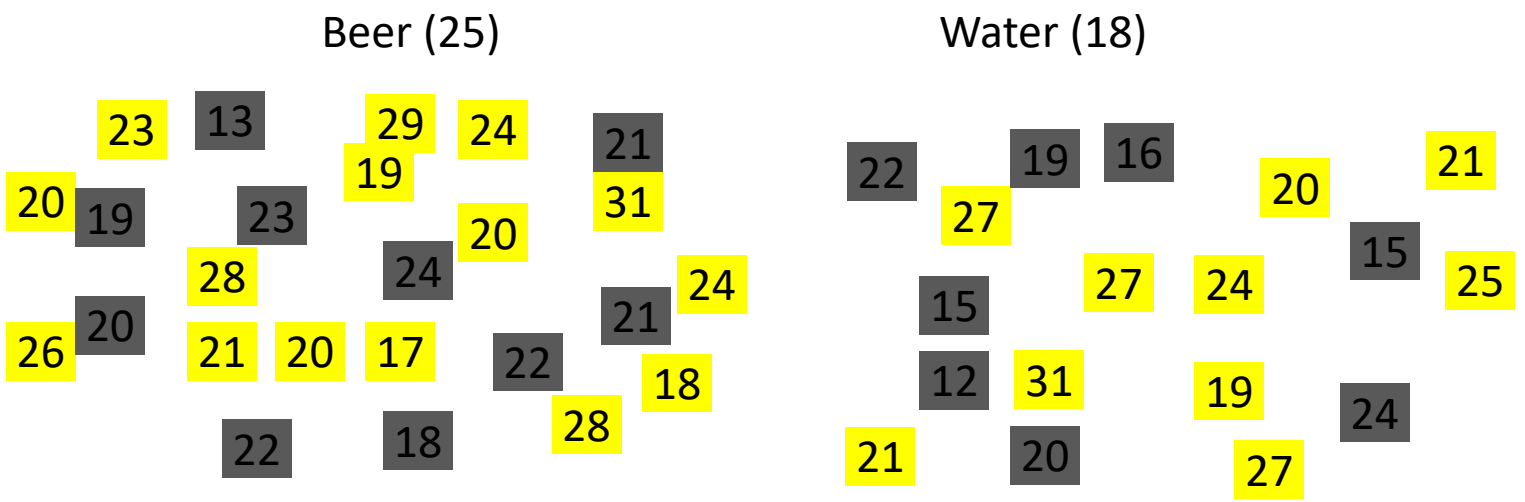
Beer (25)

Water (18)

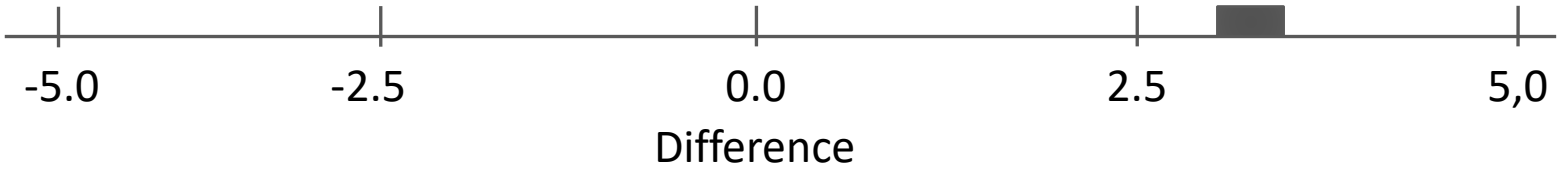
So let's simulate different groupings (permutations) and see what differences we get!



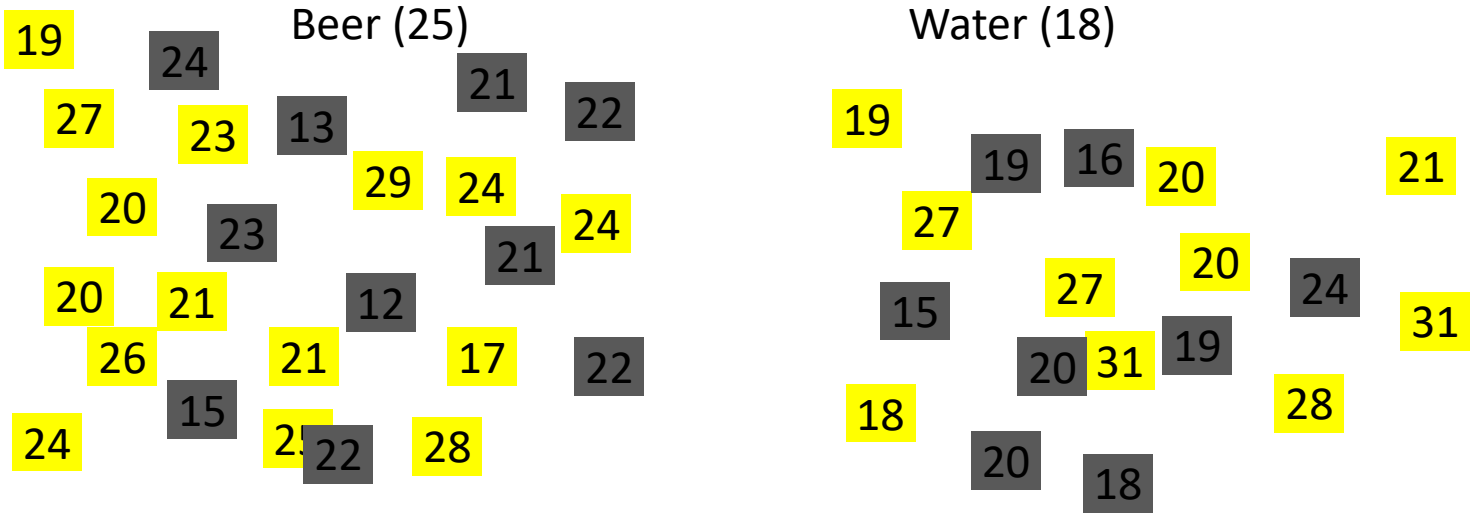
Permutation test



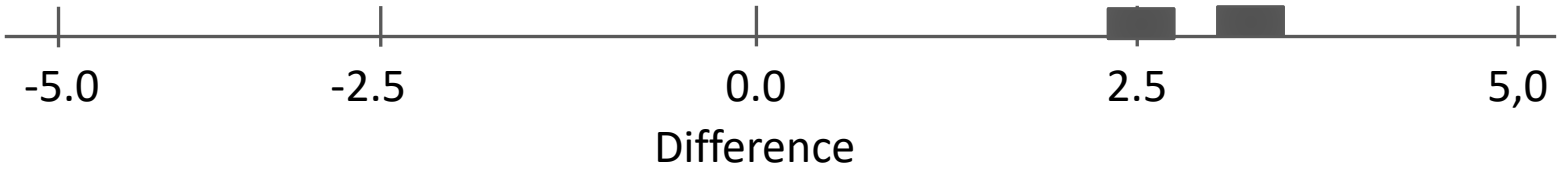
Difference of means: 3.3



Permutation test

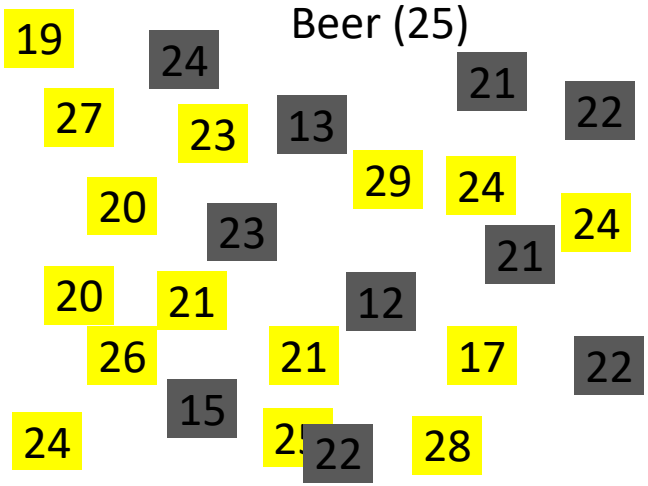


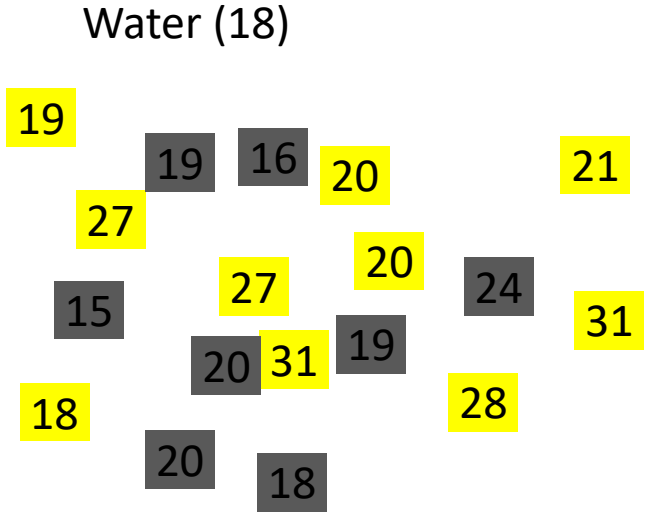
Difference of means: 2.5

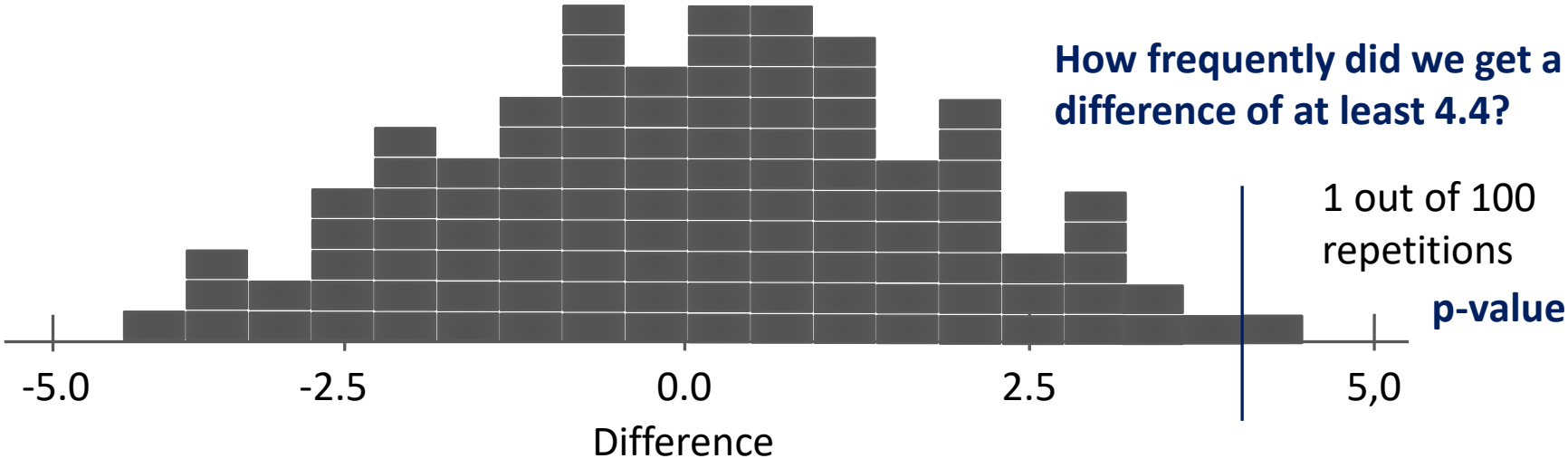


Permutation test

Observation from the data (test statistic):
The beer group has **4.4** more bites on average.







Common misconceptions

1. Given a p-value of 0.05, we can conclude there is a 95% chance the null hypothesis is false
   - p-value was calculated assuming the null hypothesis was true
   - it only states the "outlier significance" in the situation/world/universe where the null hypothesis holds

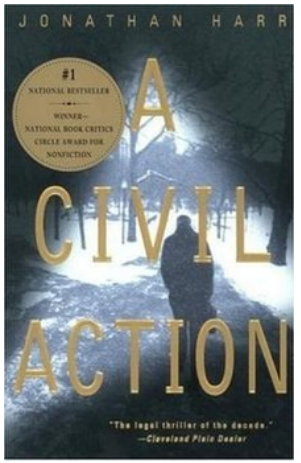
2. Given a p-value of 0.5, we can conclude the null hypothesis is most likely true
   - a non-significant p-value only means the null hypothesis is is statistically consistent with the original measurement
   - we can only say we don't rule out the null hypothesis

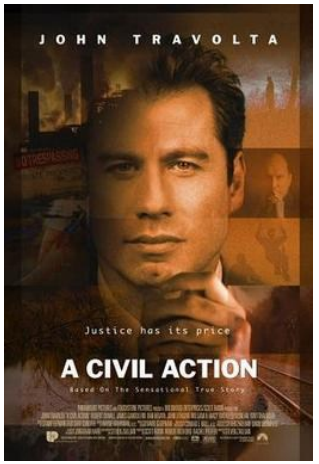
Another example: asking the right question

About 11 miles from here

By **Michael Weisskopf**
January 29, 1987

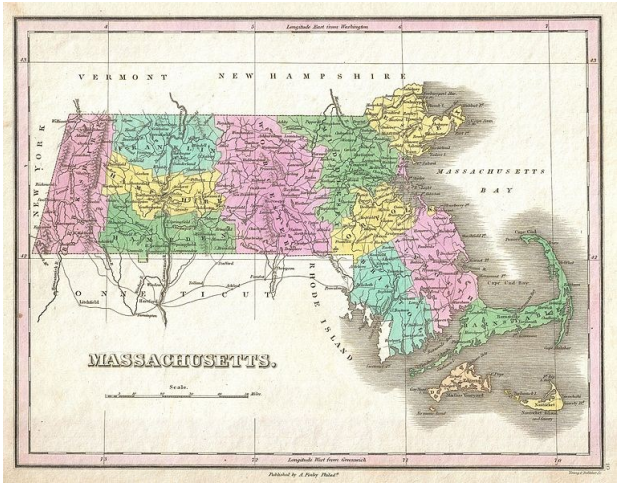
W.R. Grace & Co. was indicted yesterday on charges of falsifying statements to the Environmental Protection Agency in connection with the poisoning of drinking-water wells in Woburn, Mass. The dumping of cancer-causing solvents into the wells has been blamed for the city's high rate of leukemia among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





Hypothetical example: cancer clusters

- 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?



Run a hypothesis test

- **Null hypothesis:** the likelihood of developing cancer in any region is the same
- **Simulation code:** under the null hypothesis, what is the likelihood that **region 111** develops 30% cases more than the state-wide average?
  - reported output

```
Est. prob. of being a random event = 0.0000
Standard deviation of trials = 0.0150
```

  - highly unlikely region 111's original outcome is consistent with the null hypothesis? time to check water supply and air quality in region 111?

Actual hypothesis test

- **Null hypothesis:** the likelihood of developing cancer in any region is the same
- **Attorney's question:** under the null hypothesis, what is the likelihood that **ANY region** develops 30% cases more than the state-wide average?
  - reported output

```
Est. prob. of being a random event = 0.7500
Standard deviation of trials = 0.0955
```

  - **"any region"** includes region 1, region 2, … ,region 1000
  - effectively testing against **multiple alternate hypotheses,** then **cherry-picking** any one that's activated

Analogy to birthday problem

- Recall: what is the likelihood that two people in a room have the same birthday?
  - generalize to n people sharing a birthday

- 366 possible days
  - → 1000 regions
- total number of people in a room
  - → total number of cases over three years
- n people who share the same birthday
  - → 30% more cases in any region
- n people who were born on November 1
  - → 30% more cases in region 111

Takeaways on stochastic programs

- Stochastic simulation leverages computing speed to generate large samples/distributions of data
  - can simulate any process, e.g.,. **random walks**
  - can estimate **deterministic** quantities, e.g., pi
- Perform statistical analysis to make claims about data
  - **Central Limit Theorem** characterizes distribution of mean
  - **confidence intervals** and **p-values**
  - **linear regression** explains relationships in data, need to **validate** separately from **training**
- Have implemented in code using only **loops, lists, functions**
  - plotting with **`matplotlib`**
  - using **`random.seed()`**

Next week

- New topic: graphs and graph search algorithms
  - new Python features: dictionaries and tuples
- Two more weeks and then spring break!