190F Foundations of Data Science

Fall 2018

Lecture 21/22

Midterm Exam Review

Announcements

Histograms

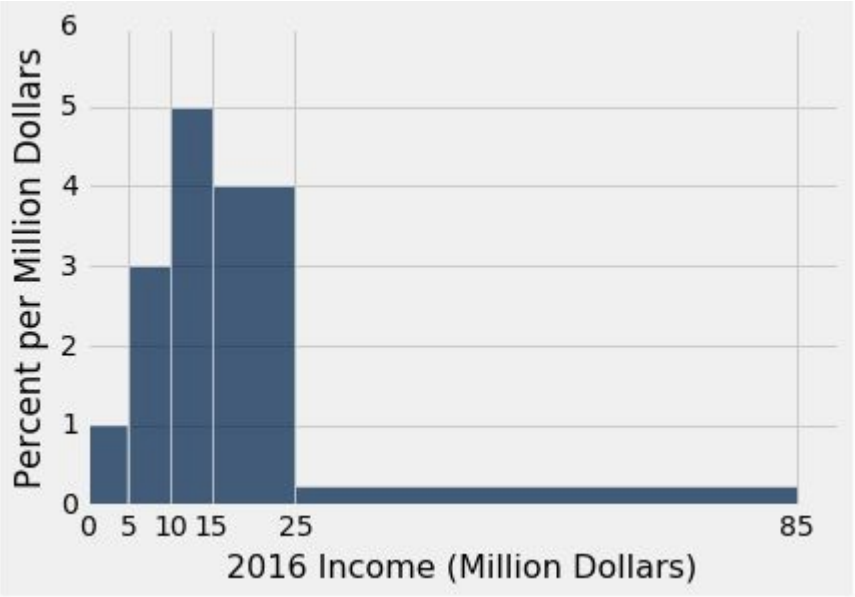
Using the Density Scale

1. Which bin has more people: [10, 15) or [15, 25)?

2. What percent of incomes are in the [25,85) bin?

3. If you draw one bar over [10,25), how tall will it be?

4. Find (or give bounds for) the median income.



Answers

(a) [15, 25)

(b) 15%

(c) 4.33 percent per million dollars

(d) At least 15 and less than 25

Probability

Exercise 1

I pick one of the 12 months at random. Independently, you pick one of the 12 months at random.

What is the chance that we both pick the same month?

(i) (1/12) * (1/12) (ii) (1/12) + (1/12) (iii) 1/12

(iii) = (12/12) * (1/12)

Exercise 2

Marbles: G, G, G, G, R, R, R, B, B, Y. Draw 4 at random.

P(no G) = ?

If with replacement:
(6/10)*(6/10)*(6/10)*(6/10)

If without replacement:
(6/10)*(5/9)*(4/8)*(3/7)

P(all G) = ?

If with replacement:
(4/10)*(4/10)*(4/10)*(4/10)

If without replacement:
(4/10)*(3/9)*(2/8)*(1/7)

Exercise 3

Marbles: G, G, G, G, R, R, R, B, B, Y. Draw 4 times at random with replacement.

1 - (6/10)*(6/10)*(6/10)*(6/10) is the chance of:
at least one G

(4/10)**4 + (3/10)**4 + (2/10)**4 + (1/10)**4 is the chance of:
all four are the same color

Testing Hypotheses

Before You Compute Anything

- Figure out the viewpoint the question wants to test, and formulate:
  - Null hypothesis: Completely specified chance model under which you can simulate data
  - Alternative hypothesis: Viewpoint comes from the question
  - Test statistic: to help you choose one viewpoint
- Say what kind of values of the statistic will make you lean towards each alternative

Permutation

Hypothesis Testing

- Is there an association between (label) and (statistic of measurement)?
- Permute label order, compute statistic, compare distribution under null to observed

Bootstrap

Estimation

- What is the population value of a parameter based on this sample?
- Resample (with replacement), compute statistic that targets parameter, compute percentiles of bootstrap statistics for CIs

Climbers

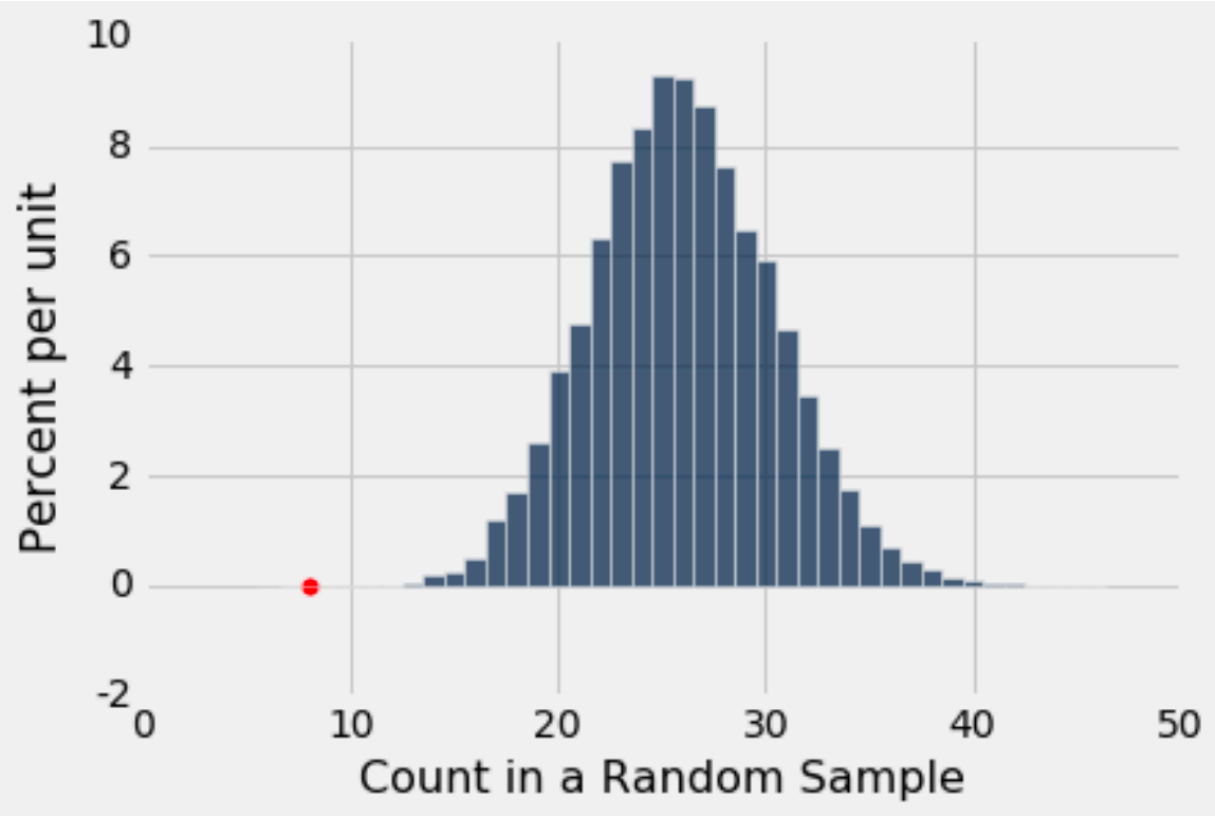
- Suppose we have “C+” climbers that climb faster on average when drinking black tea before a climb compared to climbing without tea.
- What is the null and alternative hypothesis?
- How would you test the hypothesis that they climb better after drinking tea?
- How would you estimate the average climbing speed of the climbers that drank tea (with confidence)?

One Data Sample
Simulating Random Null Model

Swain v. Alabama

- **Null Hypothesis:** Swain's jury panel was drawn at random from a population that had 26% black men
- **Alternative:** There were too few black men on the panel for it to look like a random sample
- **Data:** Observed number of black jurors in the Swain trial was 8/100 jurors.
- **Null Simulation:** Repeatedly simulate samples of panels of 100 jurors according to population proportions.
- **Test statistic:** Number of black men in panel.
- **Favors Alternative:** Low value of statistic in observed data versus simulations.

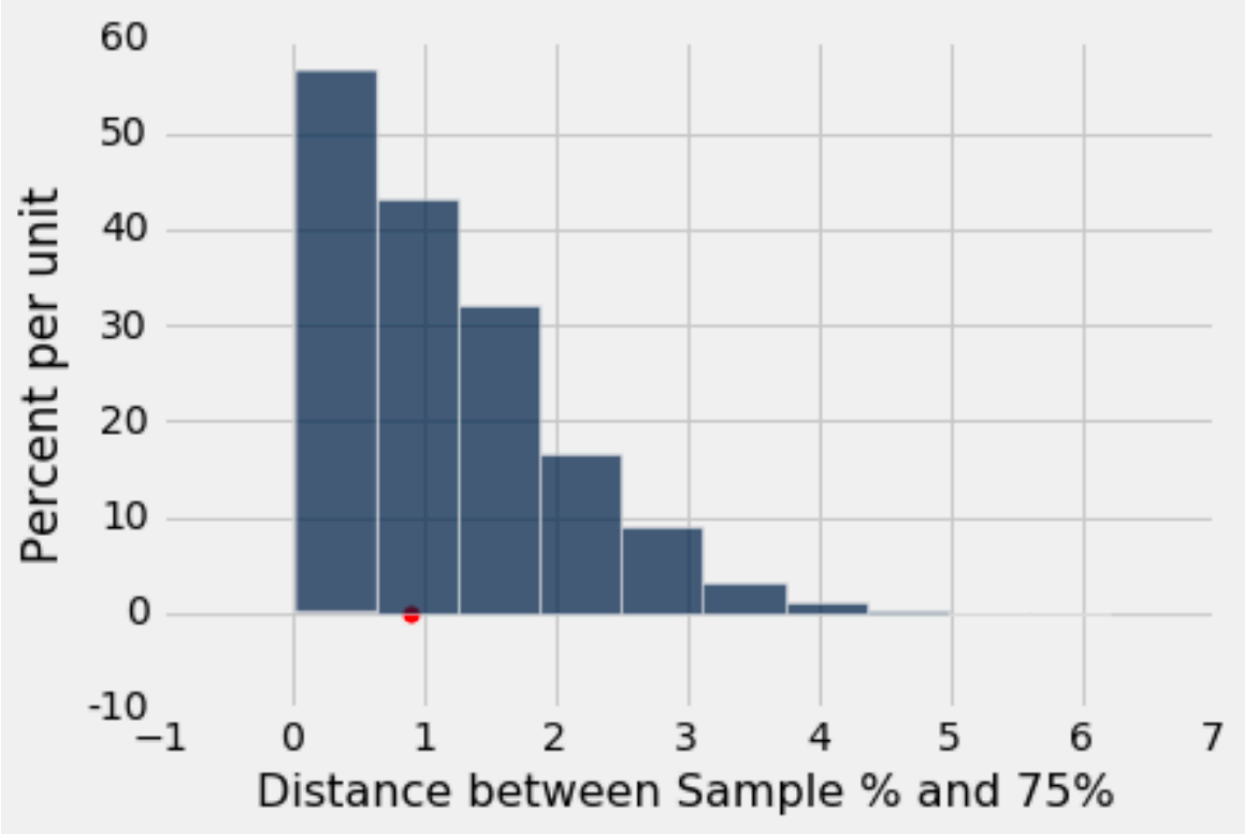
Swain v. Alabama



Mendel's Model

- **Null Hypothesis:** Each pea plant has a 75% chance of being purple flowering, independent of other plants.
- **Alternative:** The model isn't correct.
- **Data:** Number of plants with purple flowers was 705/929.
- **Null Simulation:** Repeatedly simulate samples of 929 plants with 75% chance of purple, 25% chance of white.
- **Test statistic:** | percent purple in sample - 75 |
- **Favors Alternative:** High values of statistic in the data sample compared to the simulated samples.

Mendel's Model



Alameda County Jury Panels

- **Null Hypothesis:** The Alameda County jury panels were drawn at random from all eligible jurors.
- **Alternative:** The panels were not drawn at random.
- **Data:** Observed proportion of jurors of each ethnicity within 1,453 jurors called for service.

Ethnicity	Eligible	Panels
Asian	0.15	0.26
Black	0.18	0.08
Latino	0.12	0.08
White	0.54	0.54
Other	0.01	0.04

Alameda County Jury Panels

- **Null Simulation:** Repeatedly sample 1,453 individuals from the distribution of eligible jurors.
- **Test statistic:** Total variation distance (TVD) between the distribution of ethnicity in a sample and in the population of eligible jurors.
- **Favors Alternative:** High values of the TVD computed for the observed data compared to the simulations.

Alameda County Jury Panels

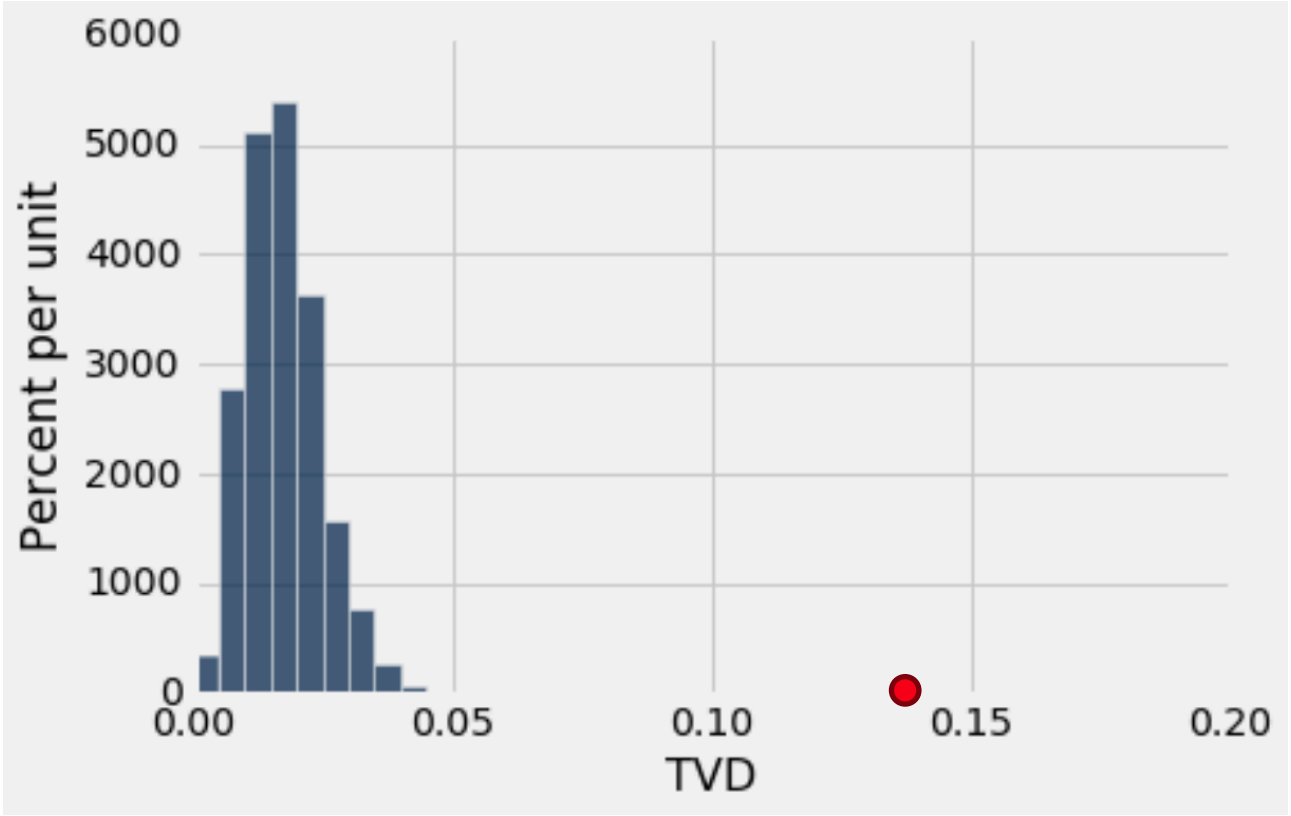
```
# Simulate total variation distance between
# distribution of sample selected at random
# and distribution of eligible population

eligible_population = jury.column('Eligible')
panel_size = 1453

tvds = make_array()

repetitions = 5000
for i in np.arange(repetitions):
    sample_distribution = sample_proportions(panel_size, eligible_population)
    new_tvd = total_variation_distance(sample_distribution, eligible_population)
    tvds = np.append(tvds, new_tvd)
```

Alameda County Jury Panels

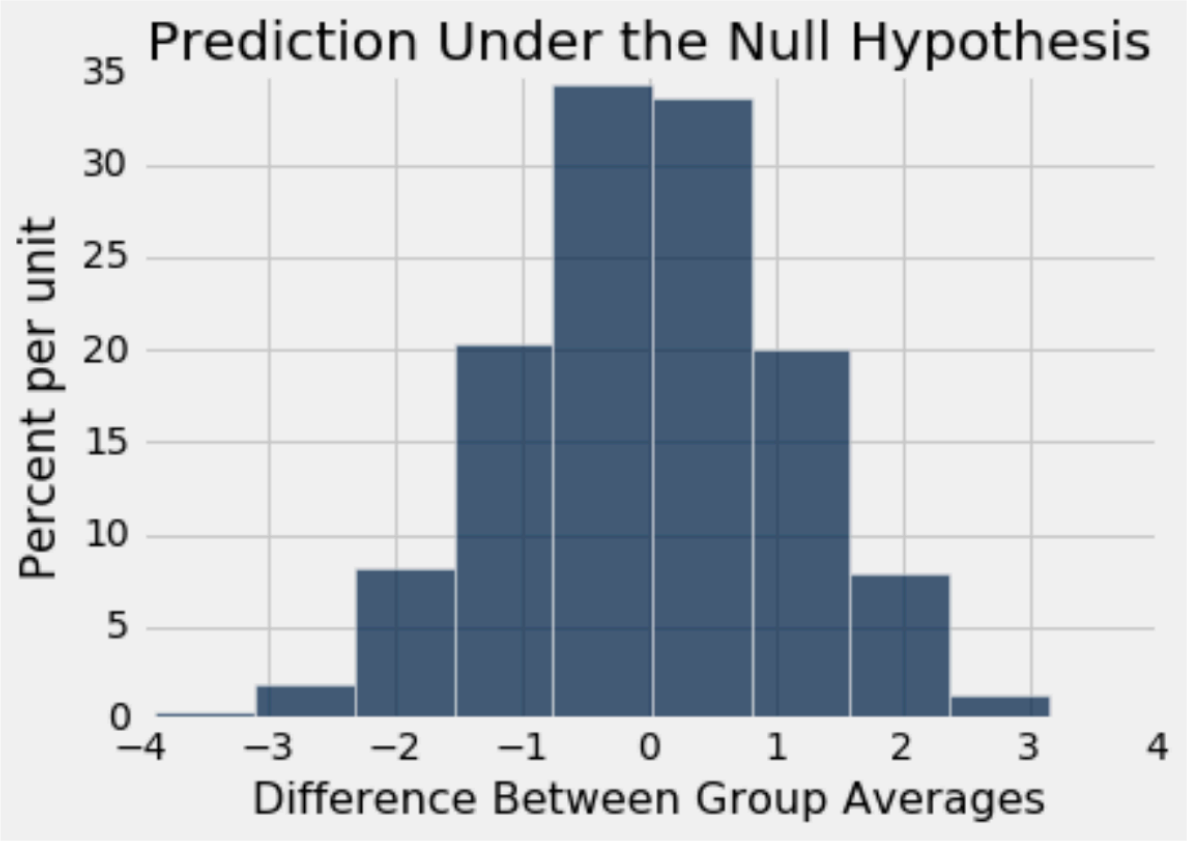


Comparing Two Samples

Birthweights

- **Null:** In the population, the distributions of the birth weights of the babies in the two groups are the same.
- **Alternative:** In the population, the babies of the mothers who didn't smoke (B) were heavier, on average, than the babies of the smokers (A).
- **Data:** Birthweights for each baby in each group.
- **Null Simulation:** Repeatedly re-assign babies to groups A and B at random.
- **Test statistic:** Group B average - Group A average
- **Favors the Alternative:** High observed value of the statistic relative to the simulated samples.

Birthweights

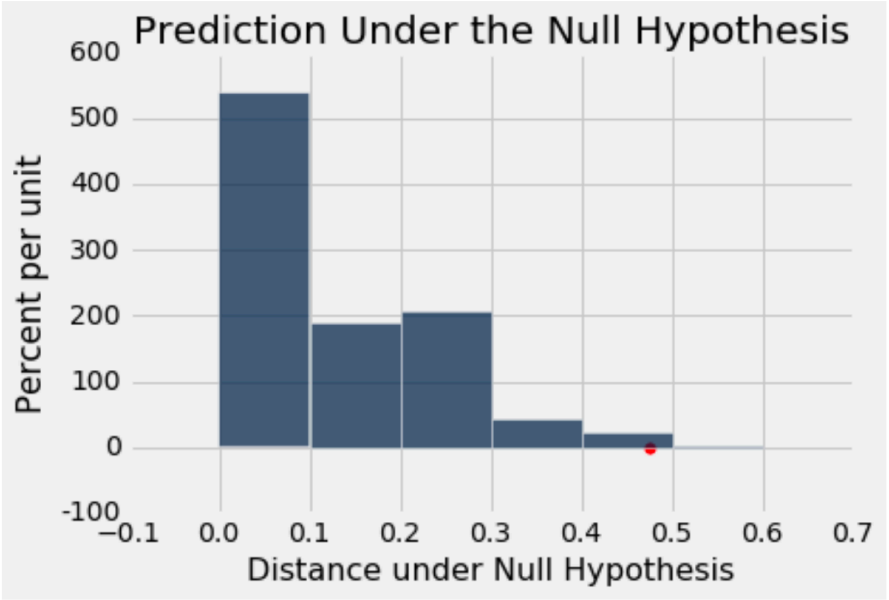


Back Pain RCT

- **Null:** The distribution of all the potential control scores is the same as the distribution of all the potential treatment scores.
- **Alternative:** The distribution of all the potential control scores is different from the distribution of all the potential treatment scores.
- **Data:** Scores for treatments and controls.
- **Null Simulation:** Repeatedly re-assign individuals to treatment or control at random.

Back Pain RCT

- **Test statistic:** abs(control group average - treatment group average)
- **Favors the Alternative:** Large values of the statistic on the observed data compared to the simulated samples.



Permutation Test Code

```
def permuted_sample_average_difference(table, label, group_label, repetitions):

    tbl = table.select(group_label, label)

    differences = make_array()
    for i in np.arange(repetitions):
        shuffled = tbl.sample(with_replacement = False).column(1)
        original_and_shuffled = tbl.with_column('Shuffled Data', shuffled)

        shuffled_means = original_and_shuffled.group(group_label, np.average).column(2)
        simulated_difference = shuffled_means.item(1) - shuffled_means.item(0)

        differences = np.append(differences, simulated_difference)

    return differences
```

P-Values

Definition of the *P*-value

Formal name: **observed significance level**

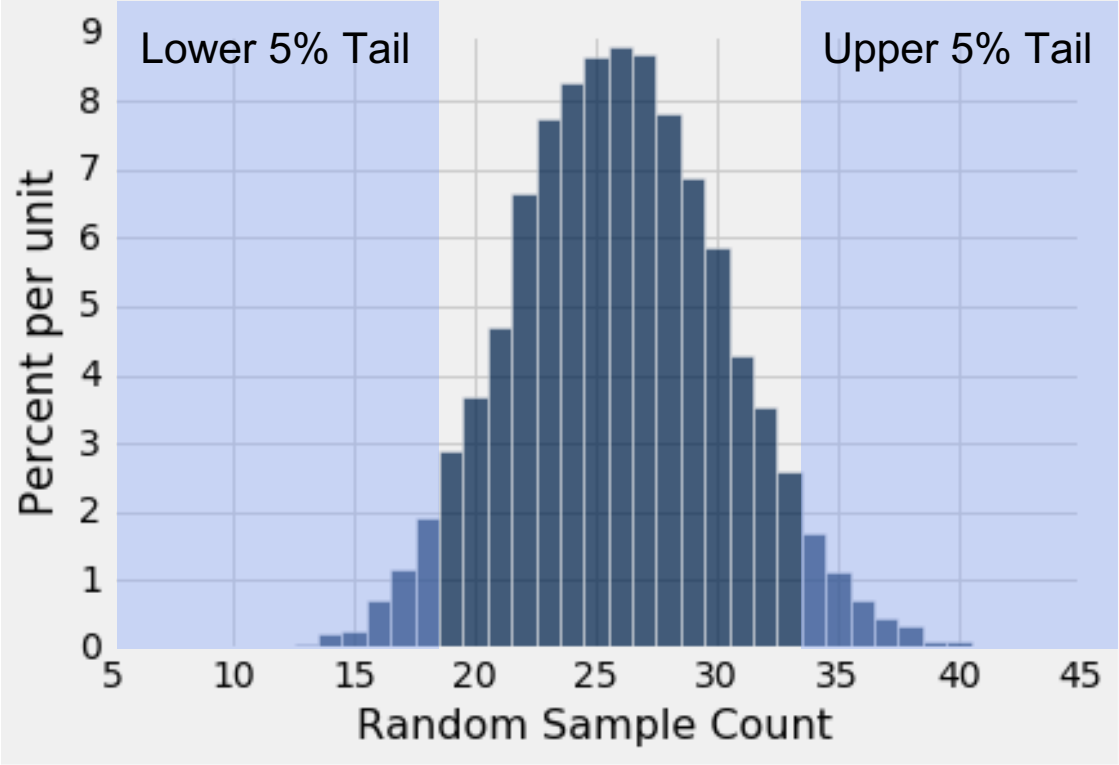
The *P*-value is the chance,

- under the null hypothesis,
- that the test statistic
- is equal to the value that was observed in the data
- or is even further in the direction **of the alternative**.

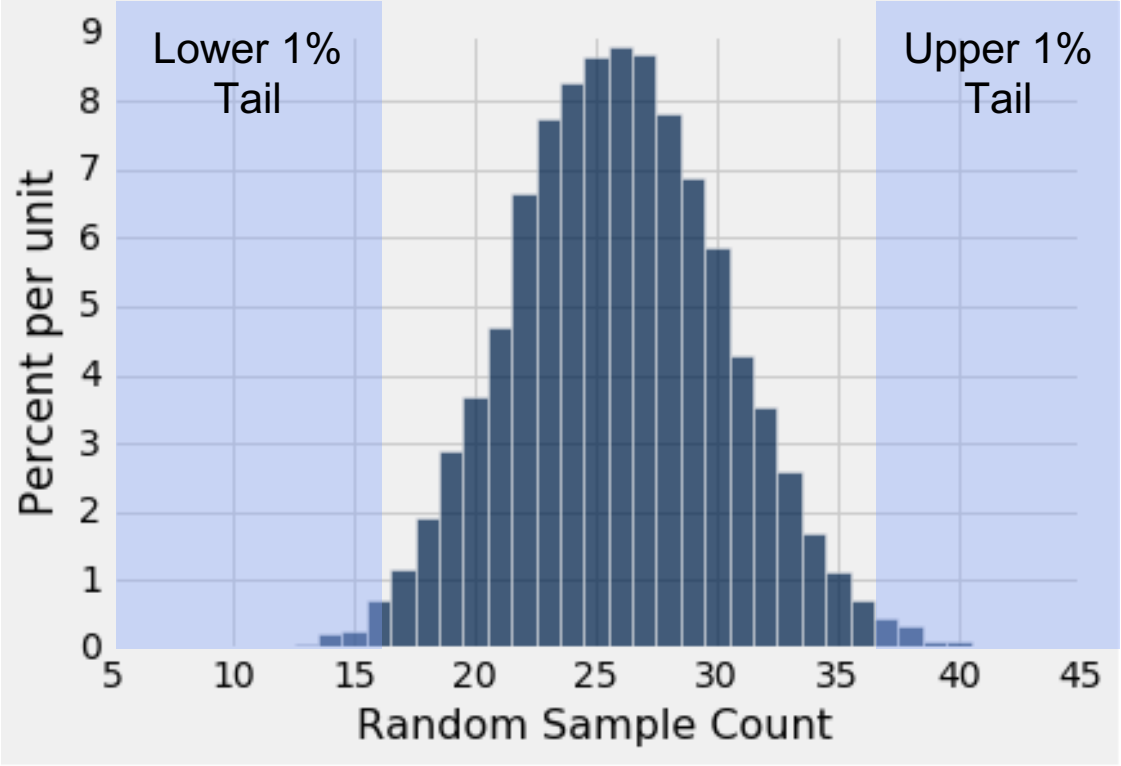
Conventions About Inconsistency

- **“Inconsistent”:** The test statistic is **in the tail** of the empirical distribution under the null hypothesis
- **“In the tail,” first convention: P<0.05**
  - The area in the tail is less than 5%
  - The result is “statistically significant”
- **“In the tail,” second convention: P<0.01**
  - The area in the tail is less than 1%
  - The result is “highly statistically significant”

Tail Areas


Lower 5% Tail
Upper 5% Tail
Percent per unit
0
1
2
3
4
5
6
7
8
9
5
10
15
20
25
30
35
40
45
Random Sample Count

Tail Areas

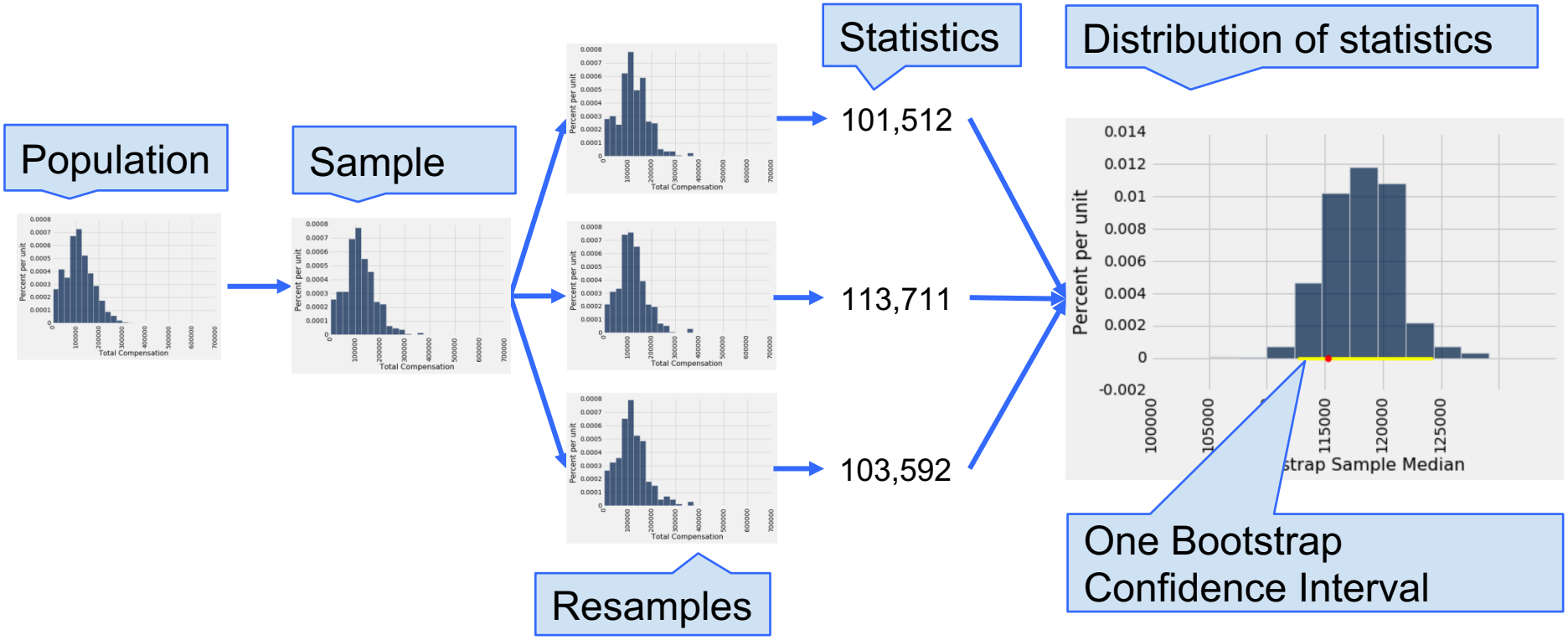

Lower 1%
Tail
Upper 1%
Tail
Percent per unit
0
1
2
3
4
5
6
7
8
9
5
10
15
20
25
30
35
40
45
Random Sample Count

Confidence Intervals

95% Confidence Interval

- Interval of **estimates of a parameter**
- Based on random sampling
- 95% is called the confidence level
  - Could be any percent between 0 and 100
  - Bigger means wider intervals
- The **confidence is in the process** that generated the interval:
  - This process generates an interval containing the population parameter about 95% of the time.
  - Can generalize to any %

Bootstrap Confidence Intervals



Bootstrap Confidence Intervals