190F Foundations of Data Science

Fall 2018

Lecture 17

A/B Testing

Announcements

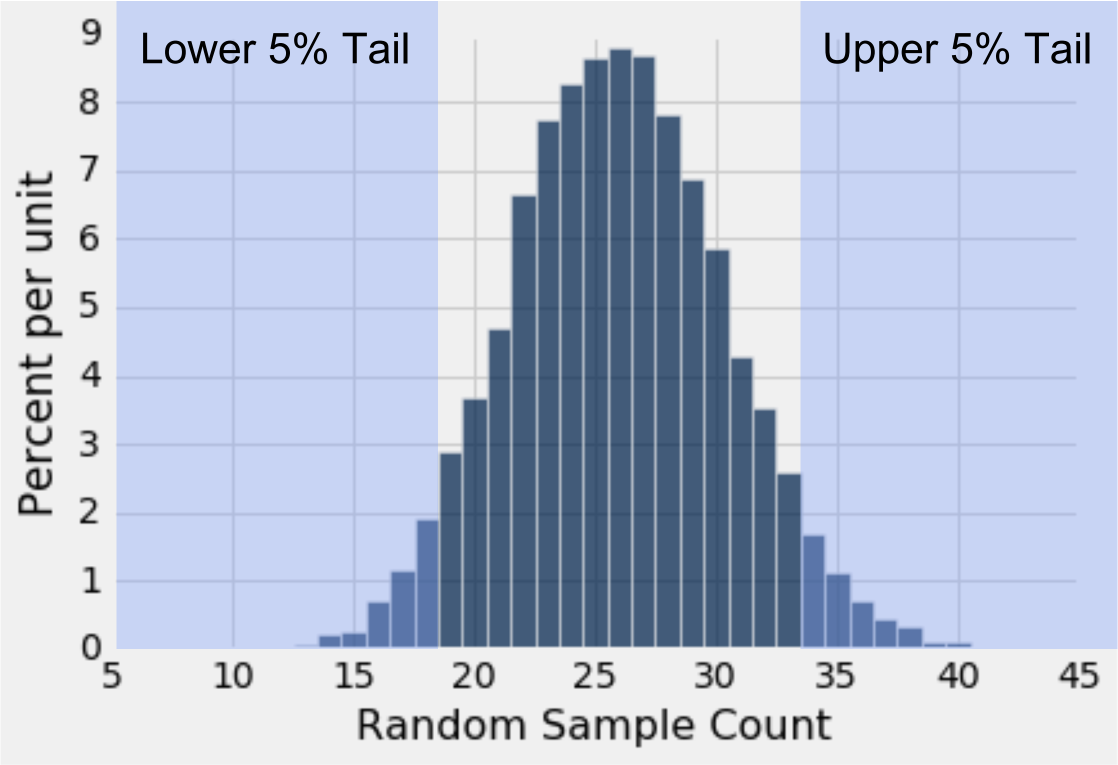
Statistical Significance

Conventions About Inconsistency

- **“Inconsistent”:** The test statistic is **in the tail** of the empirical distribution under the null hypothesis

- **“In the tail,” first convention:**
  - In the upper (or lower) 5% of the distribution
  - The result is “statistically significant”

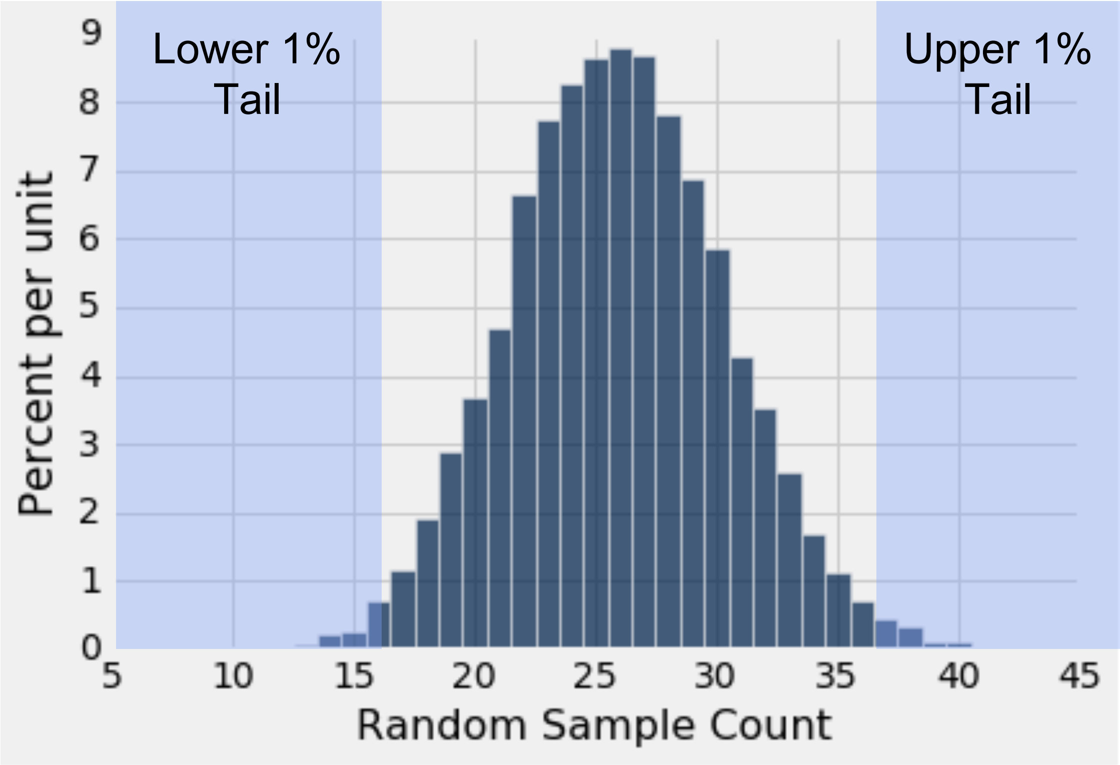
Tail Areas


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

Conventions About Inconsistency

- **“Inconsistent”:** The test statistic is **in the tail** of the empirical distribution under the null hypothesis
- **“In the tail,” first convention:**
  - In the upper (or lower) 5% of the distribution
  - The result is “statistically significant”
- **“In the tail,” second convention:**
  - In the upper (or lower) 1% of the distribution
  - The result is “highly statistically significant”

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

Conventions About Inconsistency

- Which tail do you look at?

- The tail that corresponds to values of the statistic that favor the alternative.

- **This is why you generally don’t want a statistic where *both tails* indicate support for the alternative hypothesis.**

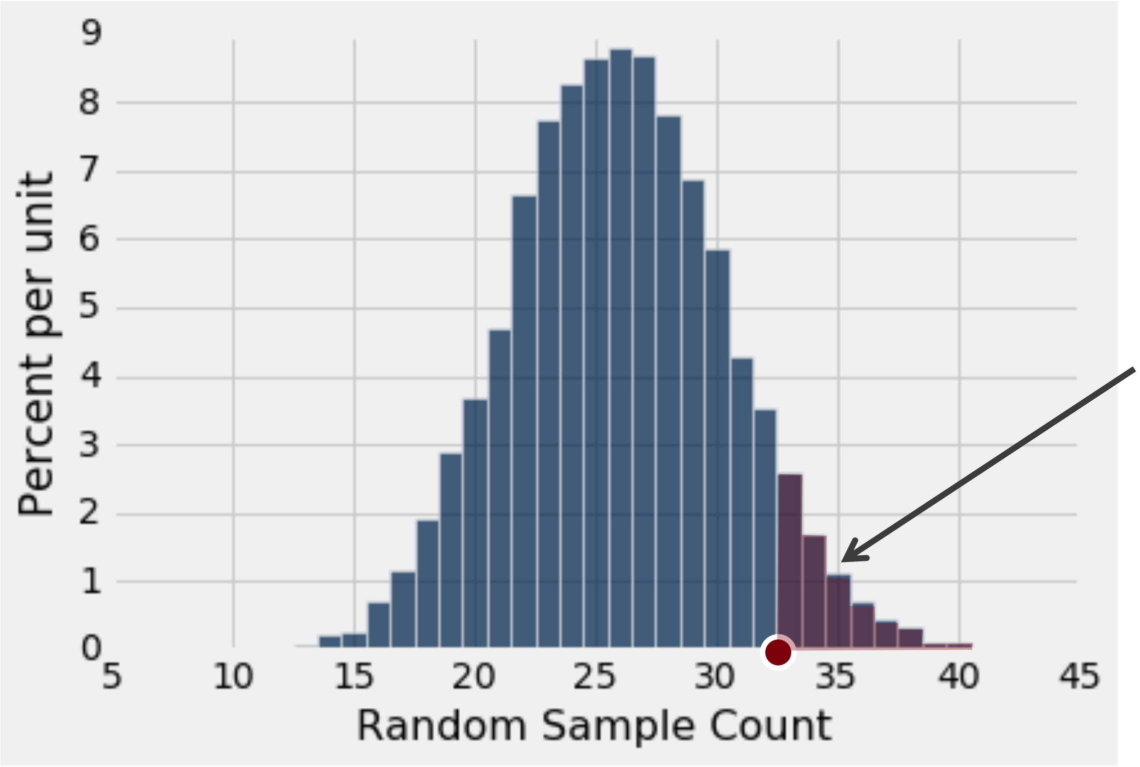
Definition of the *P*-value

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.

Tail Areas



The probability associated with **this area** is the (approximate) P-value for the observed statistic

Error Probability of a Test

Can the Conclusion be Wrong?

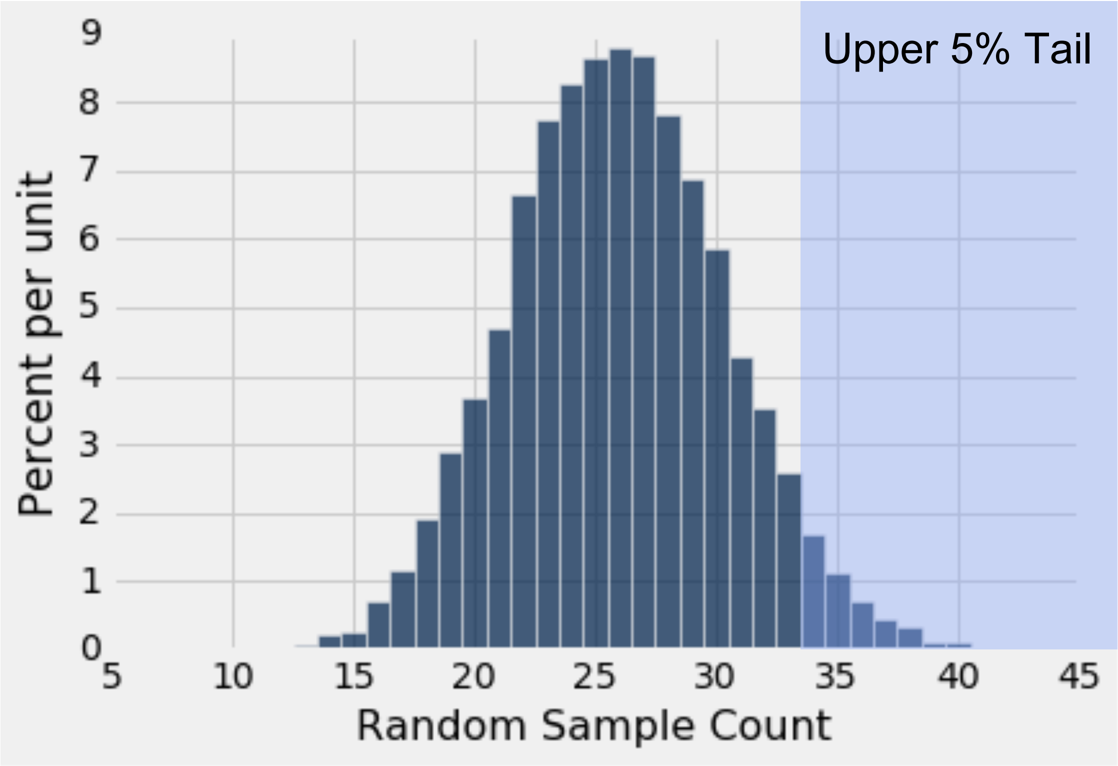
Yes.

	Null is true	Alternative is true
Test rejects the null	❌	✅
Test doesn’t reject the null	✅	❌

An Error Probability

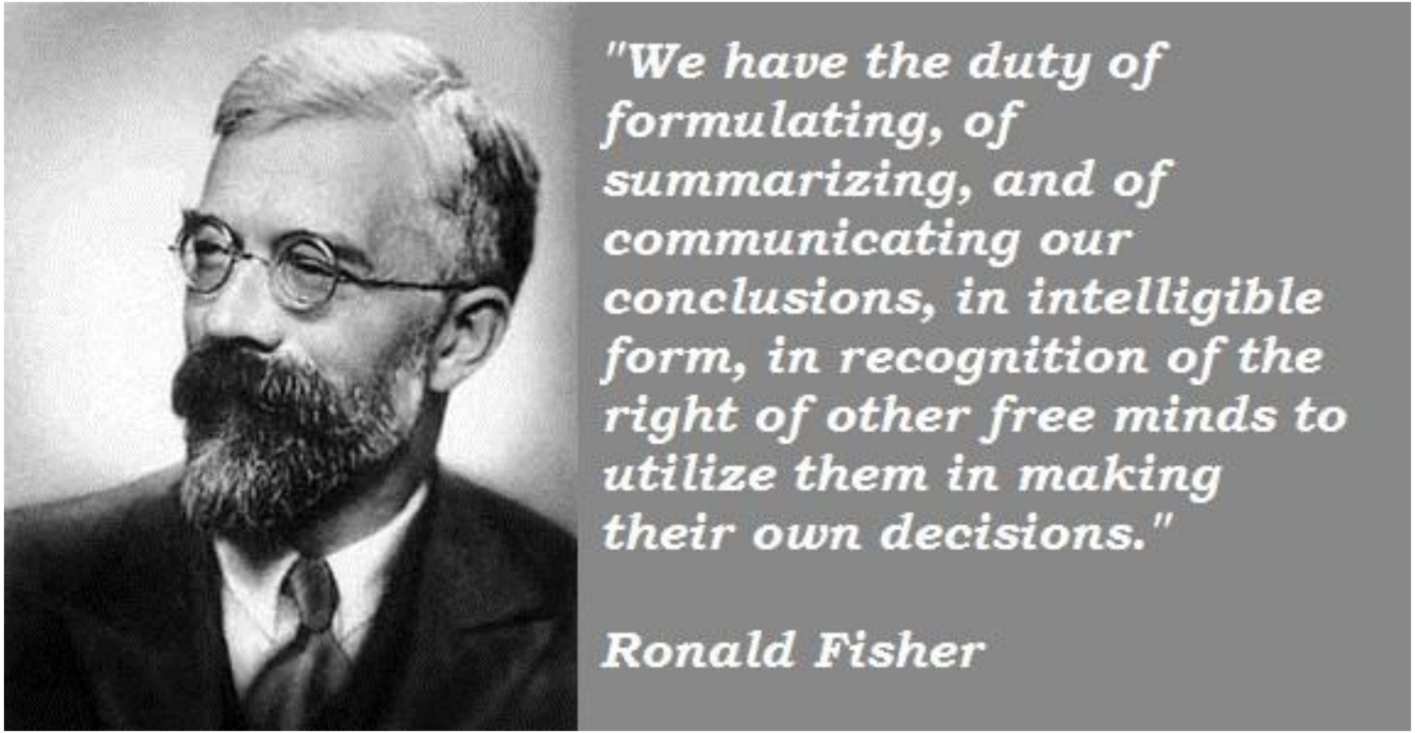
- The significance cutoff you require to reject the null is an **error probability**.
- If you require significance at the 5% level, and the null hypothesis is true, then…
- **there is about a 5% chance that your test will reject the null hypothesis, even though it shouldn’t.**
- **Why?**

Tail Areas


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

Origin of the Conventions

Sir Ronald Fisher, 1890-1962



"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."

Ronald Fisher

Sir Ronald Fisher, 1925

“It is convenient to take this point [5%] as a limit in judging whether a deviation is to be considered significant or not.”

— *Statistical Methods for Research Workers*

Sir Ronald Fisher, 1926

“If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the author prefers to set a low standard of significance at the 5 percent point …”

A/B Testing

Comparing Two Samples

- Compare values of sampled individuals in Group A with values of sampled individuals in Group B.

- Question: Do the two sets of values come from the same underlying distribution?

- Answering this question by performing a statistical test is called **A/B testing**.

(Demo)

The Groups and the Question

- Random sample of mothers of newborns. Compare:
  - (A) Birth weights of babies of mothers who smoked during pregnancy
  - (B) Birth weights of babies of mothers who didn’t smoke
- Question: Could the difference be due to chance alone?

Hypotheses

- Null:
    - In the population, the distributions of the birth weights of the babies in the two groups are the same. (They are different in the sample just due to chance.)
- Alternative:
    - In the population, the babies of the mothers who smoked weighed less, on average, than the babies of the non-smokers.

Test Statistic

- Group A: smokers
- Group B: non-smokers

- Statistic: Difference between average weights

  Group B average - Group A average

- Large values of this statistic favor the alternative

Simulating Under the Null

Non-smoker

120 oz

Non-smoker

113 oz

Smoker

128 oz

Non-smoker

136 oz

...

Smoker

108 oz

Simulating Under the Null

Smoker

120 oz

Non-smoker

113 oz

Non-smoker

128 oz

Smoker

136 oz

...

Non-smoker

108 oz

Simulating Under the Null

- If the null is true, all rearrangements of the birth weights among the two groups are equally likely
- Plan:
  - Shuffle all the birth weights
  - Assign some to “Group A” and the rest to “Group B”, maintaining the two sample sizes
  - Find the difference between the averages of the two shuffled groups
  - Repeat

(Demo)