YData: An Introduction to Data Science

Lecture 34: Classification

Jessi Cisewski-Kehe
Statistics & Data Science, Yale University
Spring 2020

Credit: data8.org

Announcements

Review:
Hypothesis testing
Regression Inference

Back to Lecture 18: hypothesis testing

- **Null hypothesis**
  - A well defined chance model about how the data were generated
  - We can simulate data under the assumptions of this model – "under the null hypothesis"
- **Alternative hypothesis**
  - A different view about the origin of the data

Prediction Under the Null Hypothesis

- Simulate the test statistic under the null hypothesis; draw the histogram of the simulated values
- This displays the **empirical distribution of the statistic under the null hypothesis**
- It is a prediction about the statistic, made by the null hypothesis
  - It shows all the likely values of the statistic
  - Also how likely they are (assuming the null hypothesis is true)

Resolve choice between null and alternative hypotheses

- Compare the **observed test statistic** and its empirical distribution under the null hypothesis
- If the observed value is not consistent with the distribution, then the test favors the alternative – "rejects the null hypothesis"

Discussion question

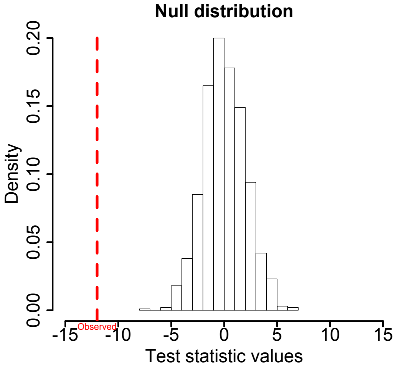
In a hypothesis test about an unknown parameter, the test statistic...

a. is the value of the unknown parameter under the null hypothesis.
b. measures the compatibility between the null and alternative hypotheses.
c. is the value of the unknown parameter under the alternative hypothesis.
d. measures the compatibility between the null hypothesis and data.
e. None of the above.

Hypothesis testing: illustrations

- Null hypothesis: **Population average = 0**
  (Or some other assumption about the population)
  Recall Swain vs. Alabama, Mendel purple flowering plan (Lec 16), or Jury Selection in Alameda County (Lec 17)

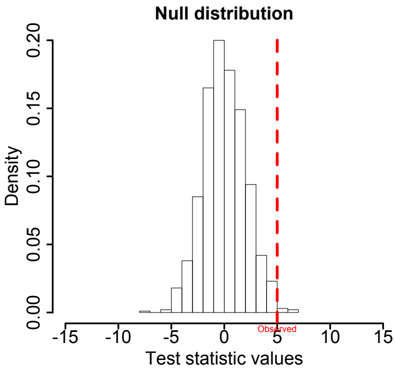
- Alternative hypothesis: **Population average $<$ 0**



Hypothesis testing: illustrations

- Null hypothesis: **Population average = 0**
  (Or some other assumption about the population)
  Recall Swain vs. Alabama, Mendel purple flowering plan (Lec 16), or Jury Selection in Alameda County (Lec 17)

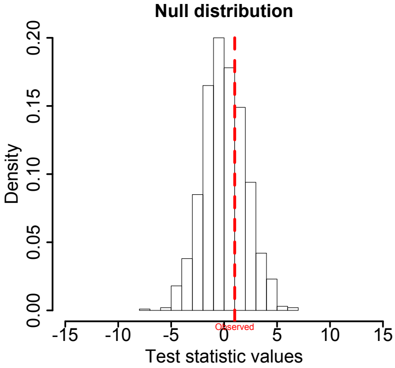
- Alternative hypothesis: **Population average > 0**



Hypothesis testing: illustrations

- Null hypothesis: **Population average = 0**
  (Or some other assumption about the population)
  Recall Swain vs. Alabama, Mendel purple flowering plan (Lec 16), or Jury Selection in Alameda County (Lec 17)

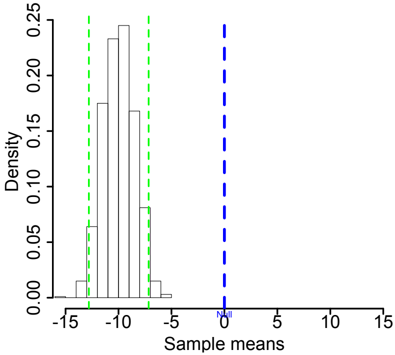
- Alternative hypothesis: **Population average > 0**



Hypothesis testing with confidence intervals

- Null hypothesis: **Population average = 0**
- Alternative hypothesis: **Population average $\neq$ 0**

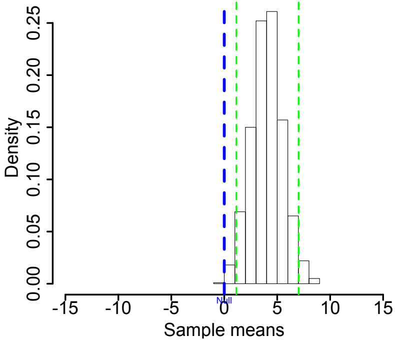
Vertical green dashed lines indicate approximate 95% confidence bounds using bootstrap samples.



Hypothesis testing with confidence intervals

- Null hypothesis: **Population average = 0**
- Alternative hypothesis: **Population average $\neq$ 0**

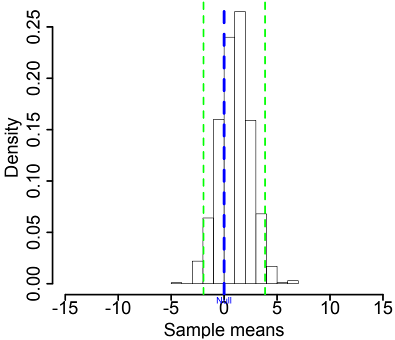
Vertical green dashed lines indicate approximate 95% confidence bounds using bootstrap samples.



Hypothesis testing with confidence intervals

- Null hypothesis: **Population average = 0**
- Alternative hypothesis: **Population average $\neq$ 0**

Vertical green dashed lines indicate approximate 95% confidence bounds using bootstrap samples.



Using a CI for Testing (Lecture 24)

- Null hypothesis: **Population average = x**
- Alternative hypothesis: **Population average $\neq$ x**
- Cutoff for P-value: p%
- Method:
  - Construct a (100-p)% confidence interval for the population average
  - If x is not in the interval, reject the null
  - If x is in the interval, can't reject the null

Discussion question

If we only have a 90% confidence interval for the population mean (μ), which is $(-.2, .8)$. Based on this interval, we wish to test the hypotheses $H_0 : \mu = 0 \quad vs. \quad H_a : \mu \neq 0$ at a p-value cutoff of $\alpha = .05$. Determine which of the following statement is true.

a. We cannot make any decision since the confidence level we used to calculate the confidence interval is 90%, and we would need a 95% confidence interval.
b. We do not reject H_0, because the value 0 falls in the 90% confidence interval.
c. We reject H_0, because the value 0 falls in the 90% confidence interval.
d. We cannot make a decision since the confidence interval is so wide.
e. None of the above

Discussion question

A physics instructor is convinced that every test he writes has a population mean score of 78 ($\mu = 78$). Students who have enrolled in the course do not believe him, but are not sure if the population mean score is above or below 78. Suppose a random sample of students was taken from his large lecture course, and a 95% confidence interval was found to be [70.864, 77.136].

(a) State a null and alternative hypothesis test.

(b) Given the confidence interval provided, what would be your conclusion to the hypothesis test specified at the $\alpha = 0.05$ level of significance?

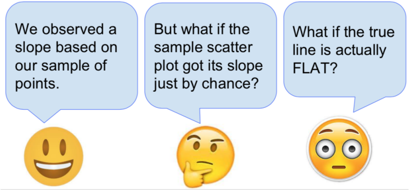
Discussion question

If you reject the null hypothesis $H_0 : \mu = \mu_0$ vs. $H_0 : \mu \neq \mu_0$ at a p-value cutoff of $\alpha = 0.05$, then μ_0 would fall in the 90% confidence interval for μ.

True or False or Not Enough Information

Test Whether There Really is a Slope

- **Null hypothesis**: The slope of the true line is 0.
- **Alternative hypothesis**: No, it's not.
- Method:
  - Construct a bootstrap confidence interval for the true slope.
  - If the interval doesn't contain 0, reject the null hypothesis.
  - If the interval does contain 0, there isn't enough evidence to reject the null hypothesis.



Confidence Interval for True Slope

- **Bootstrap the scatter plot**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the "middle 95%" interval.
- That's an approximate 95% confidence interval for the slope of the true line.

(DEMO)

A/B testing: Comparing Two Samples (Lec 19,20)

- Previously, we only considered data from a single group

- Compare values of sampled individuals in Group A with values of sampled individuals in Group B.
  $\rightarrow$ Question: Do the two sets of values come from the same underlying distribution?
  $\rightarrow$ Answering this question by performing a statistical test is called A/B testing.

  Examples:
  (A) Birth weights of babies of mothers who smoked during pregnancy
  (B) Birth weights of babies of mothers who didn't

  (A) Control group
  (B) Treatment group

  Deflategate

A/B testing: 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

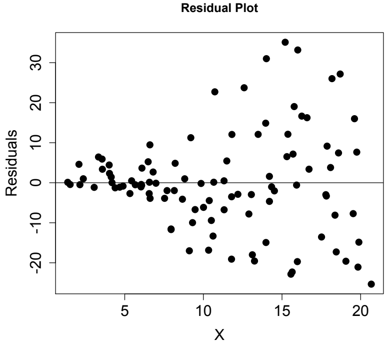
Discussion question

A study on the effect of caffeine involved asking subjects to take a memory test 20 minutes after drinking cola. Some subjects were randomly assigned to drink caffeine-free cola, and some to drink regular cola (with caffeine). For each subjects, a test score (the number of items recalled correctly) was recorded. The subjects were not told which type of cola they had been given.

a. The memory test had a total of 25 items on it. The average number of items recalled was 15 for the caffeine-free group and 16 for the regular cola group. Are the values 15 and 16 statistics or parameters?
b. Can an A/B hypothesis testing framework be used here? How?

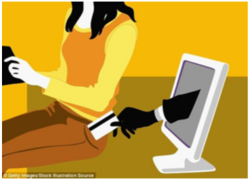
Discussion question

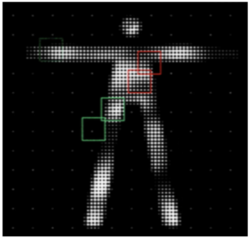
Suppose a Least-squares linear model was fit on explanatory variable X and response variable Y, with the residuals plotted in the figure below against X. What linear model assumption appears to be violated given the residual plot below?



Classification

Classification Example





(DEMO)

Classifiers

Training a Classifier