STAT 220 Slides
Summary and Review

Yibi Huang
Department of Statistics
University of Chicago

Common Misunderstandings About Hypothesis Testing

We have mentioned a number of common misunderstanding about hypothesis testing when we first introduced hypothesis testing in L16.pdf.

- Rejecting H_0 doesn't means we are 100% that H_0 is false. We might make Type 1 errors. Setting a significance level just guarantee we won't make Type 1 error too often
- P-value is not $P(H_0 \ is \ true \mid data)$ but it is $P(data \mid H_0 \ is \ true)$.

We are going to talk about more common misunderstanding about hypothesis testing here.

Failing to Reject H_0 Does Not Prove H_0 to Be True

Another mistake is to conclude from a high p-value that the H_0 is probably true

- We have said that low p-value is evidence that the H_0 may not be true
- If our p-value is high, can we conclude that H_0 is true?
  - No, we could make a type 2 error when failing to reject H_0
  - Moreover, unlike type 1 error rate is controlled at a low level, type 2 error rate is usually quite high. It is quite often that the data fail to reject H_0 even though H_0 is not true.
- When we fail to reject H_0, often it just means the data are not able to distinguish between H_0 and H_A (because the data are to noisy, etc)

Real Example

- As an example, the Women's Health Initiative found that low-fat diets reduce the risk of breast cancer with a p-value of 0.07
- The *New York Times* headline: "*Study finds low-fat diets won't stop cancer*"
- The lead editorial claimed that the trial represented "*strong evidence that the war against fats was mostly in vain*" and sounded "*the death knell for the belief that reducing the percentage of total fat in the diet is important for health*"
- Failing to prove the effect of low-fat diets doesn't prove that low-fat diets have no effect

http://www.nytimes.com/2006/02/07/health/study-finds-lowfat-diet-wont-stop-cancer-or-heart-disease.html

Don't Take the 0.05 Significance Level Too Seriously

- A p-value of 0.049 and a p-value of 0.051 give nearly the same strength of evidence against H_0
- For example, in the highly publicized 2009 study involving a vaccine that may protect against HIV infection, the two-sided p-value is 0.08, and the one-sided p-value of is 0.04
- Much debate and controversy ensued, partially because the two ways of analyzing the data produce p-values on either side of 0.05
- Much of this debate and controversy is fairly pointless; both p-values tell you essentially the same thing — that the vaccine holds promise, but that the results are not yet conclusive

Hypothesis Testing Cannot Tell Us...

Hypothesis testing cannot tell us

- whether the design of a study is flawed
- whether the data is appropriately collected

So we cannot conclude from a small P-value about whether one variable has a causal effect on another variable or whether the conclusion can be generalized to a bigger population.

Garbage In $\rightarrow$ Garbage Out

Statistical Significance Does Not Mean Practical Importance

Another mistake is reading too much into the term “statistically significant”

- Saying that results are statistically significant informs the reader that the findings are unlikely to be due to chance alone
- However, it says nothing about the practical importance of the finding.
- E.g., rejecting the H_0: $\mu_1 = \mu_2$ just tells us $\mu_1 \neq \mu_2$, but not how big and how important $\mu_1 - \mu_2$ is. It is possible that the difference is too small to be relevant even if it is significant.
- Remedy: *Attach a confidence interval* for the parameter so that people can decide whether the difference is big enough to be relevant.

Recap: Common Misunderstandings about Hypothesis Testing

- Rejecting H_0 doesn't mean we are 100% sure that H_0 is false. We might make Type 1 errors
- P-value is not the probability that the H_0 is true
- Failing to reject H_0 does not prove H_0 to be true
- Don't take the 0.05 significance level too seriously
- Hypothesis testing cannot tell us if data were collected properly or if the design of a study was flawed
- Statistical significance does not mean practical importance

Summary and Review

In the second half of STAT 220, we covered

- CLT and Sampling Distributions
- Overview of Confidence Intervals
- Overview of Hypothesis Testing
- Inference about population mean(s)
    - one-sample data
    - two-sample data
    - paired data
- Inference about population proportion(s)
    - one-sample data
    - two-sample data
- Correlation and Regression

CLT and Sampling Distributions

Let $X_1, X_2, \ldots, X_n$ be **i.i.d.** random variables (discrete or continuous) with **mean μ and variance σ^2**. Then, when *n is large*, the distribution of the **sample mean**

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

is approximately

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

even when the population distribution is not normal.

- The sample size required to use CLT depends on how irregular and how skewed the population distribution is

Sample Problems (CLT and Sampling Distributions)

Each year about 1500 students take the introductory statistics course at a large university. This year scores on the final exam are distributed with a median of 74 points, a mean of 70 points, and a standard deviation of 10 points. There are no students who scored above 100 (the maximum score attainable on the final) but a few students scored below 20 points.

- Can we calculate the probability that a randomly chosen student scored above 75 using the normal distribution?
- Can we calculate the probability that the average score for a random sample of 40 students is above 75 using the normal distribution? Or should we use a t-distribution
- T or F: the histogram of the score of 100 randomly selected students is roughly normal.

Student's t-Distributions

- is bell-shaped, symmetric about 0, has a heavier tail than $N(0, 1)$
- The larger the df, the lighter the tails, the closer the t-curve to the $N(0, 1)$ curve
- As df $= \infty$, $t = N(0, 1)$

Student's t-Distributions

If $X_1, X_2, \ldots, X_n$ are i.i.d. from some population distribution with mean μ and SD σ, the *t-statistic* defined as

$$t = \frac{\overline{X} - \mu}{(\text{sample SD})/\sqrt{n}} \sim t_{n-1}$$

- true when the population distribution is normal
- approx. true when the population is not normal but n is large. The less normal the population, the larger n needs to be to use the approximation

Inference About Population Means

- one-sample data:

  H_0: $\mu = \mu_0$: test statistic $t = \dfrac{\bar{x} - \mu_0}{\text{SE}}$ where $\text{SE} = s/\sqrt{n}$

  CI for μ: $\bar{x} \pm t^*$ SE $df = n - 1$

- two-sample data:

  H_0: $\mu_1 = \mu_2$: test statistic $t = \dfrac{\bar{x}_1 - \bar{x}_2}{SE}$

  CI for $\mu_1 - \mu_2$: $(\bar{x}_1 - \bar{x}_2) \pm t^*$ SE, where

$$SE = \begin{cases} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} & \text{if } \sigma_1 \neq \sigma_2, \quad df = \min(n_1 - 1, n_2 - 1) \\ s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} & \text{if } \sigma_1 = \sigma_2 \text{ where } s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}, \ df = n_1 + n_2 - 2 \end{cases}$$

- paired data

  H_0: $\mu_1 = \mu_2$: test statistic $t = \dfrac{\bar{d}}{SE}$ where $d_i = x_{1i} - x_{2i}$

  CI for $\mu_1 - \mu_2$: $\bar{d} \pm t^*$ SE SE $= s_d/\sqrt{n}$, df = # of pairs -1

Conditions For Using t-Tests or t-Intervals

- Observations (or differences for paired data) must be independent (use your judgement)
- For small sample $(n < 10)$, the population must be fairly normal to use t-tests or t-CIs
- For moderately large sample ($n > 20$ or 30), t-tests and t-CIs can be safely used even when the population is not normal as long as it's not too skewed, no outlier

Inference About Population Proportions

- one-sample

  test statistic for H_0: $p = p_0$ is $z = \dfrac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$.

  CI for p: $\widehat{p} \pm z^* \times \sqrt{\widehat{p}(1-\widehat{p})/n}$

  sample size required control the margin of error of a CI at m:

$$n \geq \left(\frac{z^*}{m}\right)^2 \widehat{p}(1-\widehat{p})$$

- two-sample

  test statistic for H_0: $p_1 = p_2$ is

$$z = \frac{(\widehat{p}_1 - \widehat{p}_2) - 0}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \text{ where } \widehat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

  CI for $p_1 - p_2$: $(\widehat{p}_1 - \widehat{p}_2) \pm z^* \times \sqrt{\dfrac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \dfrac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$

Inference About Population Proportions

Conditions:

- independence of observations
- number of successes and number of failures both ≥ 10

A Larger Population Does NOT Require a Larger Sample

- All the SEs depend on the sample size only, not the population size.
- The relative size of a sample to the population size doesn't matter. It is the absolute size of a sample that matters.
- A larger population does NOT require a larger sample!

Which Tests/CIs to Use?

- Is the questions about a numerical variable (means) or categorical data (proportions or counts)?
- Are we considering a *single sample*, comparing *two samples*, or using *paired data*?

For each problem below, identify the type of problem.

1. What is the height difference between 5th grade boys and 5th grade girls? To find out, you sample students from the local school district and construct a confidence interval.
2. Social Security benefits depend on there being enough younger working adults in the population, relative to the number of older adults receiving the benefits. Is the proportion of the population receiving Social Security benefits higher in Illinois than in Indiana?

3. Is the proportion of the population receiving Social Security benefits higher in Illinois than the national average, which is known to be 19%?
4. To test the claims of a famous mind reader, you draw a random card and ask the mind reader to guess the card. Then you replace it back in the deck. Out of 1000 trials, he correctly identifies the card correctly 30 times. Is his performance better than random guessing, where there's a 1 out of 52 chance of guessing correctly?

For each problem below, identify the type of problem.

1. What is the height difference between 5th grade boys and 5th grade girls? To find out, you sample students from the local school district and construct a confidence interval. 2-sample mean
2. Social Security benefits depend on there being enough younger working adults in the population, relative to the number of older adults receiving the benefits. Is the proportion of the population receiving Social Security benefits higher in Illinois than in Indiana?

3. Is the proportion of the population receiving Social Security benefits higher in Illinois than the national average, which is known to be 19%?
4. To test the claims of a famous mind reader, you draw a random card and ask the mind reader to guess the card. Then you replace it back in the deck. Out of 1000 trials, he correctly identifies the card correctly 30 times. Is his performance better than random guessing, where there's a 1 out of 52 chance of guessing correctly?

For each problem below, identify the type of problem.

1. What is the height difference between 5th grade boys and 5th grade girls? To find out, you sample students from the local school district and construct a confidence interval. 2-sample mean
2. Social Security benefits depend on there being enough younger working adults in the population, relative to the number of older adults receiving the benefits. Is the proportion of the population receiving Social Security benefits higher in Illinois than in Indiana? 2-sample proportion
3. Is the proportion of the population receiving Social Security benefits higher in Illinois than the national average, which is known to be 19%?
4. To test the claims of a famous mind reader, you draw a random card and ask the mind reader to guess the card. Then you replace it back in the deck. Out of 1000 trials, he correctly identifies the card correctly 30 times. Is his performance better than random guessing, where there's a 1 out of 52 chance of guessing correctly?

For each problem below, identify the type of problem.

1. What is the height difference between 5th grade boys and 5th grade girls? To find out, you sample students from the local school district and construct a confidence interval. 2-sample mean
2. Social Security benefits depend on there being enough younger working adults in the population, relative to the number of older adults receiving the benefits. Is the proportion of the population receiving Social Security benefits higher in Illinois than in Indiana? 2-sample proportion
3. Is the proportion of the population receiving Social Security benefits higher in Illinois than the national average, which is known to be 19%? 1-sample proportion
4. To test the claims of a famous mind reader, you draw a random card and ask the mind reader to guess the card. Then you replace it back in the deck. Out of 1000 trials, he correctly identifies the card correctly 30 times. Is his performance better than random guessing, where there's a 1 out of 52 chance of guessing correctly?

For each problem below, identify the type of problem.

1. What is the height difference between 5th grade boys and 5th grade girls? To find out, you sample students from the local school district and construct a confidence interval. 2-sample mean
2. Social Security benefits depend on there being enough younger working adults in the population, relative to the number of older adults receiving the benefits. Is the proportion of the population receiving Social Security benefits higher in Illinois than in Indiana? 2-sample proportion
3. Is the proportion of the population receiving Social Security benefits higher in Illinois than the national average, which is known to be 19%? 1-sample proportion
4. To test the claims of a famous mind reader, you draw a random card and ask the mind reader to guess the card. Then you replace it back in the deck. Out of 1000 trials, he correctly identifies the card correctly 30 times. Is his performance better than random guessing, where there's a 1 out of 52 chance of guessing correctly? 1-sample proportion

Exercise 5.18. Paired or Not

In each of the following scenarios, determine if the data are paired?

1. We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days, and record Intel's and Southwest's stock on those same days.
2. We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.
3. A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

Exercise 5.18. Paired or Not

In each of the following scenarios, determine if the data are paired?

1. We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days, and record Intel's and Southwest's stock on those same days. Paired data
2. We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items.
3. A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

Exercise 5.18. Paired or Not

In each of the following scenarios, determine if the data are paired?

1. We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days, and record Intel's and Southwest's stock on those same days. Paired data
2. We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items. Paired data
3. A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school.

Exercise 5.18. Paired or Not

In each of the following scenarios, determine if the data are paired?

1. We would like to know if Intel's stock and Southwest Airlines' stock have similar rates of return. To find out, we take a random sample of 50 days, and record Intel's and Southwest's stock on those same days. Paired data
2. We randomly sample 50 items from Target stores and note the price for each. Then we visit Walmart and collect the price for each of those same 50 items. Paired data
3. A school board would like to determine whether there is a difference in average SAT scores for students at one high school versus another high school in the district. To check, they take a simple random sample of 100 students from each high school. Two-sample data

Correlation r

- $-1 \leq r \leq 1$, measures the *direction* and *strength* of *linear* association
- $r = 1$ or -1 if and only if all points lie on a straight
- r is unit-free, not affected by the units used
- r doesn't change if x & y are swapped
- r is very sensitive to outlier
- r cannot reflect the strength of a non-linear association
- r can be misleading if the data are clustered
- correlation does not imply causation

Least Square Regression Line

- The *least-square regression line* is the line $y = b_0 + b_1 x$ that minimizes the sum of squared errors:

  $$\sum_i e_i^2 = \sum_i (y_i - \widehat{y_i})^2 = \sum_i (y_i - b_0 - b_1 x_i)^2$$

  It can be shown by math that the slope and intercept of the least-square regression line are

  $$b_1 = \text{slope} = r \cdot \frac{s_y}{s_x}, \quad b_0 = \text{intercept} = \bar{y} - \text{slope} \cdot \bar{x}$$

- The least-square regression line *always passes through* $(\bar{x}, \bar{y})$
- Interpretation of b_1: For observations with $x = x_0 + 1$, their y values are *on average* b_1 higher than the y values of those observations with $x = x_0$.
- Interpretation of b_0: the predicted value of response when $x = 0$, which might not have a practical meaning if $x = 0$ is not a possible value

Least Square Regression

- Prediction: The regression line can be used to make predictions of one variable from another. But if you have to extrapolate far from the data, or to a different group of subjects, be careful.
- One cannot plug in y values into the regression line for predicting y from x and solve for x to predict the x values. The regression line that predicts y from x is different from the one that predicts x from y

Residuals and R-squared

- Residual $e_i = y_i - \widehat{y_i}$ = observed y– predicted y
- If predicted with a least regression line,
  - the residuals add up to 0
  - residuals have zero correlation with the explanatory variable
- R-squared

$$R^2 = r^2 = \frac{s^2_{\hat{y}}}{s^2_y} = \frac{\text{Variance of predicted } y\text{'s}}{\text{Variance of observed } y\text{'s}}$$

$$= \text{proportion of variation in the response explained by the explanatory variable}$$

$$1 - r^2 = \frac{s^2_e}{s^2_y} = \frac{\text{Variance of Residuals}}{\text{Variance of } y}$$

$$= \text{proportion of variation in the response not included in the model or by inherent randomness in the data}$$

Simple Linear Regression Models

- Assumptions: independence, linearity, constant variability, normality
- Checking model assumptions using residual plots and histograms of residuals
- CI for β_i: $b_i \pm t^*$ SE(b_i), df $= n - 2$
- To test $H_0 : \beta_i = c$, use t-statistic $t = (b_i - c)/SE(b_i)$ with df $= n - 2$
- SE(b_i) for the intercept and slope can be obtained from R summary output
- estimate for $\sigma = \sqrt{\sum_i e_i^2/(n-2)}$ can be obtained from R summary output
- outlier, influential point, high leverage point

Overview of Confidence Intervals

- It's wrong to say a specific 95% CI (L, U) (where L and U are values) has a 95% probability to enclose a parameter.
- Correction interpretation: About 95% of the intervals constructed following the procedure (taking a SRS (or SRSs) and then calculating (estimate) $\pm$ (critical value)SE) will cover the true population parameter
- A confidence interval is for covering some population parameter, not for covering a sample mean, a sample proportion, or 95% of the entire population
- marginal of error = half of the width of CI
- Factors that affect margin of error: sample size, confidence level, size of noise (population SDs)

Overview of Hypothesis Testing

- H_0 and H_a are always statements about population (parameters), not about samples, e.g., Exercise 4.19
- Type 1 error = falsely rejecting a true H_0,
  Type 2 error = failing to reject a false H_0
- Significance level = chance of making a Type 1 error
- A parameter value is rejected in a two-sided test at level α if and only if the value is not in the $100(1 - \alpha)\%$ CI for that parameter.

Overview of Hypothesis Testing

- Rejecting H_0 doesn't mean we are 100% sure that H_0 is false. We might make Type 1 errors
- P-value is not the probability that the H_0 is true
- Failing to reject H_0 doesn't prove H_0 to be true
- Statistical significance doesn't mean practical importance. Always report a CI so that people can gauge whether the difference is important
- Don't take the 0.05 significance level too seriously. A P-value of 0.049 or 0.051 do not differ much in the strength of evidence against H_0
- Hypothesis testing cannot tell us if data were collected properly or if the design of a study was flawed