ST517 Note Outline 6: Hypothesis Testing

Lecture 6.1: Introduction and Foundations

Recall: Process of Data Analysis

1. Identify the research question(s)
2. Conduct background research
3. Form a hypothesis or make a prediction
4. Experiment / Collect data
5. Explore, summarize and analyze data
    - Explore and summarize data = exploratory data analysis (EDA)
    - Analyze data = apply tools of statistical inference
        - Last outline: Confidence intervals (CI)
            - Used to estimate what is true in a population (e.g. the true value of a parameter)
        - This outline: Hypothesis tests (HT)
            - Used to support (or not) a specific hypothesis about the population
6. Make conclusions

Motivating Example: Suppose the mean caffeine content in a cup of regular coffee is 120mg. I bought a "half-caff" blend but am still sleepy in the morning. I suspect that the mean caffeine content in a cup of half-caff is really less than 60mg.

a. What is the population of interest?

b. What is the response variable?

c. What is the parameter of interest?

d. What is the research question?

Example (continued): Suppose I take a random sample of 10 cups of coffee from the blend and find the mean caffeine content is 58.5mg.

a. Based on this, can we say that the average caffeine content for this brand of half-caff is truly less than 60mg?

b. If the true average is in fact 60mg, how unusual would my observed sample mean be? Assume the true standard deviation is 2.5mg.

c. Based on your previous answer, can we say that the average caffeine content for this brand of half-caff is truly less than 60mg?

Important Points

- Sometimes estimating what is true in the population (via a confidence interval) is not enough; we need to be able to make a yes/no decision
- Decision based on how surprising sample results are after making certain assumptions about the population
    - Determine how consistent sample results are with expected sampling variability
    - Would not be able to describe sampling variability if we did not make assumptions about the population!
- When sample results are not consistent with expected sampling variability, we say those results are **statistically significant**
    - Other ways to say "not consistent with expected sampling variability" are "not likely to have occurred just by random chance" or "not likely to have occurred due to natural variability between subjects"
- Results that are statistically significant indicate that there might be something interesting going on
    - Perhaps assumptions about population that were used to define the sampling distribution were incorrect
    - Perhaps there is an effect
- Results that are statistically significant do NOT indicate that an effect is *real* or *meaningful* (more on this later)
- Establishing statistical significance occurs via a statistical hypothesis test

Connection between the Sampling Distribution and Hypothesis Testing

- Sampling distribution serves a frame of reference to evaluate sample results against
    - Established by making assumptions about what is true in the population
- Look to see if sample results are consistent with expected sampling variability under those assumptions
    - If consistent, we should not refute assumptions about the population
    - If not consistent, then we can refute assumptions about the population

Basic steps to hypothesis testing

1. Identify research question, population, and parameter of interest
2. Establish null and alternative hypotheses
3. Identify type of hypothesis test and check conditions
4. Calculate test statistic
5. Identify null distribution and calculate p-value
6. Make a decision about null hypotheses
7. State conclusion in context of alternative hypothesis

Step 1: Identify research question, population, and parameter of interest

- In practice: This is first step of Process of Data Analysis
    - Followed by conducting background research to establish/refine hypothesis
- Coursework: Research question will be provided; you will need to identify it and also the relevant population, response variable, and corresponding parameter

Step 2: Establish null and alternative hypothesis

- Null Hypothesis: Beginning claim

    - Allows establishment of sampling distribution, in this context called the:

    - Notation:

- Alternative Hypothesis: Another theory

    - Notation:

Example: Vitamin E and Prostate cancer

- Null hypothesis: Vitamin E does not make a difference in prostate cancer

- Alternative Hypothesis: Vitamin E does make a difference in prostate cancer

Step 3: Identify type of hypothesis test and check conditions

- Type of test that is appropriate based on
    - Type of response variable:
    - Number of variables*:
    - Number of categories for explanatory variable*:
    - * Can also be thought of as number of populations under consideration
    - This outline: Tests for 1 population
    - Future outlines: other cases
- Conditions:
    - Every hypothesis test has certain conditions or assumptions that need to hold in order for that test to be valid
    - Which conditions are necessary depend on type of test

Step 4: Calculate test statistic

- **Test statistic**: Numeric measure of distance from sample value to what is expected under null hypothesis
    - General form:
    - Interpretation:

Step 5: Identify null distribution and calculate p-value

- **Null distribution**: Distribution of the test statistic assuming

    - Can be created by simulation or by theoretical model (i.e. a probability distribution)
    - Sets frame of reference so we can see how the sample statistics would be expected to behave

- **P-value**: Proportion of the null distribution that is as or more extreme than the test statistic

    - Notes:

Step 6: Make a decision about null hypotheses

- Basic idea:
    - If the p-value is too small, then the statistic is unlikely given the null distribution. Perhaps something else is going on.

    - How small is too small?

- **Significance level:** Cutoff point for the p-value; indication of small

    - Notation:

    - Most common value:

    - Other common values:

    - Guidelines for choosing value:

    - Decision rule of thumb:

Step 7: State conclusion in context of alternative hypothesis

- Recall: This is final step of Process of Data Analysis
- Necessary because just telling someone "We reject the null hypothesis" or "The results are statistically significant" does not give them useful information about the problem at hand
- Conclusion needs to give enough information in the context of the problem that the reader can make an informed decision
    - Needs to be about the alternative hypothesis, because this is the hypothesis driving the research (the one we are interested in learning about)
- Conclusion of a hypothesis test:

- Caution! Be careful that your language is not too strong.

Lecture 6.2: Hypothesis Test for a Population Proportion

- When to use:

- In general, parameter is:

1. Identify research question, population, and parameter of interest
    - Recall: Identifying the population requires describing (in context!) the large group of units we are interested in learning about
    - Recall: Identifying the parameter requires communicating (in context!)…
        - What is the response variable of interest
        - How response variable is summarized (with a proportion in this case)
        - That the summary is for the entire population

2. Establish null and alternative hypotheses

    - Null hypothesis:

        - Notation: p_0 – null value; a specific proportion of interest (e.g. 0.75)

    - Choice of alternatives:

3. Identify type of hypothesis test and check conditions

    - Type of test:

    - Conditions:

4. Calculate test statistic

    - If conditions are met, appropriate test statistic is:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

5. Identify null distribution and calculate p-value

    - Null dist: Under above conditions, test stat follows

    - P-value found as probability under H_0, but in the direction of H_A

6. Make decision about null hypothesis

    - Decision Rule of Thumb: If p-value $\leq \alpha$, reject H_0

7. State conclusion in the context of alternative hypothesis

    - If you rejected H_0: Conclude that there is enough evidence to support the alternative hypothesis.
    - If you did <u>not</u> reject H_0: Conclude that there is <u>not</u> enough evidence to support the alternative hypothesis.
    - Describe alternative hypothesis in context!

P-values and the Alternative Hypothesis

- P-value is found in the direction of the alternative hypothesis

    - Greater than: p-value = proportion of the null distribution that is

    - Less than: p-value = proportion of the null distribution that is

    - Not equal: p-value = proportion of null distribution that is

Example: Consider the following—$H_0: p = 0.5$ and $z = 1.58$. We will consider the p-values that correspond to each alternative hypothesis.

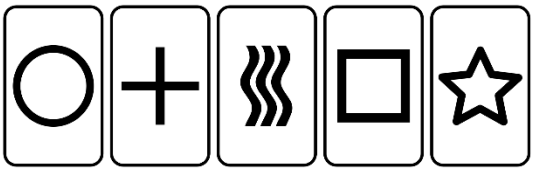
$H_A: p > 0.5$

$H_A: p < 0.5$

$H_A: p \neq 0.5$

Important Point: P-value (and thus decision and conclusion) depend on

Example: One method used to evaluate psychic ability is to test the subjects with Zener cards, which have one of five shapes on them. In a psychic experiment the subject is asked to guess which of the shapes is on a card. Many cards are presented to the subject. In the long run we would expect subjects to get around 1/5 = 0.2 of the answers correct just by guessing, but that proportion may vary by chance. Suppose a subject was presented with 100 cards and correctly identifies 30 of these. Does this provide significant evidence of psychic abilities? Use a significance level of 1%



Lecture 6.3: Hypothesis Test for a Population Mean

- When to use:

- In general, parameter is:

1. Identify research question, population, and parameter of interest
    - Recall: Identifying the population requires describing (in context!) the large group of units we are interested in learning about
    - Recall: Identifying the parameter requires communicating (in context!)…
        - What is the response variable of interest
        - How response variable is summarized (with a mean in this case)
        - That the summary is for the entire population

2. Establish null and alternative hypotheses

    - Null hypothesis:

        - Notation: μ_0 – null value; a specific mean of interest (e.g. 60)

    - Choice of alternatives:

3. Identify type of hypothesis test and check conditions

    - Type of test:

    - Conditions:

4. Calculate test statistic

    - If conditions are met, appropriate test statistic is:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

        - Note: formula standardizes sample mean so it is on a scale appropriate for the t-distribution

5. Identify null distribution and calculate p-value

    - Null dist: Under above conditions, test stat follows

    - P-value found as probability under H_0, but in the direction of H_A

6. Make decision about null hypothesis

    - Decision Rule of Thumb: If p-value $\leq \alpha$, reject H_0

7. State conclusion in the context of alternative hypothesis

    - If you rejected H_0: Conclude that there is enough evidence to support the alternative hypothesis.
    - If you did <u>not</u> reject H_0: Conclude that there is <u>not</u> enough evidence to support the alternative hypothesis.
    - Describe alternative hypothesis in context!

Example: A company produces candy covered chocolates. Specifications indicate that the candies should average 0.85 grams. A quality control manager for the company took a random sample of 38 candies from the production line. The resulting values were used to produce the summary statistics below. Is there evidence the average weight differs from the specification? Conduct an appropriate hypothesis test. Use a significance level of 5%.

Column	Mean	Std. Dev.	Min	Max	Q1	Q3	n	Median
weight	0.873	0.034	0.81	0.98	0.85	0.89	38	0.87

Lecture 6.4: Errors and Statistical Power

Errors that can occur

- What if we make the wrong decision?

- Type I error

    - Interpretation:

- Significance Level (α)

- Type II error

    - Interpretation:

	H_0 Actually True	**H_0 Actually False**
Say "Reject H_0"		
Say "Do not reject H_0"		

Power of a statistical test

- Power

    - Interpretation:

Example: Malaria is a serious health concern. Early diagnosis and treatment are important tools in the fight to control this disease. We can think of a test for malaria as a hypothesis test with the following hypotheses:

Ho: patient does not have malaria vs. Ha: patient has malaria

If a patient tests positive for malaria, the therapy recommended by the World Health Organization is artemisinin-combination therapy or ACT, which has a known side effect of vomiting. If a patient exhibits symptoms of malaria but tests negative for the disease, this may result in the patient being treated for a different condition. Untreated malaria can lead to more severe conditions and, in most cases, death.

a. Define Type I and Type II Errors in the context of this problem.

b. Describe a real-world consequence of making each type of error in this context.

Notes:

- In practice, we want to protect the "status quo"

- Most tests have the

- For a fixed sample size *n*, there is a

- Ideally want probabilities of making a mistake to be small and power to be large
    - These probabilities are properties of the procedure

    - They are not applicable to the decision once it is made

Increasing Power

- Power increases if we increase the sample size.

- Power increases if we increase the significance level

- Power increases if there is a bigger effect to find

Calculating Power or the Probability of Making an Error

- Finding power requires setting several other values

- Recall: Power = 1 – P(Type II Error) = 1 – β
- There are several formulas for calculating β based on normal distribution
- However, for this course we will focus on understanding rather than calculation
    - Reason through finding probabilities based on the context of the problem
    - Or use computer software

Example: There is a basket with 10 marbles inside. The marbles are either red or white, but we do not know the number of each color. We want to decide between the hypotheses:

H_0: Basket has 9 Red and 1 White vs. H_A: Basket has 4 Red and 6 White

We will select a single marble from the basket. What is the most reasonable Decision Rule?

- Reject the null hypothesis if the ball is

With this rule, what are the chances of making a mistake?

- P(Type I error) =

- P(Type II error) =

- What is the power of the test?

Example (continued): Suppose a ball is now selected from the basket and it is observed and found to be white.

- What would be the decision?
- Could a mistake have been made?
- If so, which type?
- What is the probability that this type of mistake was made?

Determining Sample Size

- Issues of error and power should be considered before a study is conducted
- Limit type I error by setting the significance level the test will be conducted at
- Set an acceptable level for power (and type II error)
- Determine minimum sample size that would be needed to achieve these values
    - Depends on setting several other specifications:
        - Direction of test (based on the alternative hypothesis)
        - *Hypothesized* value of σ
        - *Hypothesized* effect size

Example: Determine the sample size that would be necessary to achieve 80% power with a 5% chance of making a Type I Error for a 1-sided hypothesis test of $H_0: \mu = 2$. To get a sense of how the size of the effect (difference between mean under null and mean under alternative) impacts sample size, consider alternative mean values of: 5, 10, and 15. To get a sense of how the hypothesized standard deviation impacts sample size, consider σ values of: 30 and 50. [Note: SAS code that produced the output below is posted to Moodle.]

The POWER Procedure
One-sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Number of Sides	1
Null Mean	2
Alpha	0.05
Nominal Power	0.8

Computed N Total				
Index	**Mean**	**Std Dev**	**Actual Power**	**N Total**
1	5	30	0.800	620
2	5	50	0.800	1719
3	10	30	0.803	89
4	10	50	0.800	243
5	15	30	0.807	35
6	15	50	0.801	93

Which sample size would you recommend?

Determining Predicted Power

- Alternatively, you could determine highest power possible for a given sample size (and other conditions specified as before)

Example: Predict the power for a hypothesis test of $H_0: \mu = 2$ vs. $H_a: \mu > 2$, with a 5% chance of making a Type I Error and assuming the mean under the alternative hypothesis is 10 and true standard deviation is 30. Due to budget constraints, the sample size is limited to 62 units. [Note: SAS code that produced the output below is posted to Moodle.]

The POWER Procedure
One-sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Number of Sides	1
Null Mean	2
Alpha	0.05
Mean	10
Standard Deviation	30
Total Sample Size	62

Computed Power
Power
0.667

Lecture 6.5: Important Points about Hypothesis Tests

- Statistical tests are based on randomness and assumptions.
    - All types of statistical inference assume the sample is random; if not

    - Other conditions (e.g. large sample, normal population): Different analysis methods may be available
        - Ex: If sample size is not large enough to use the large sample z-test for a proportion from Lecture 6.2, then you could use a small sample test for the count of successes where the p-value is calculated using the Binomial distribution with parameters n and p_0
        - Ex: If the population is highly skewed and the sample size is not large enough for the CLT to apply so that you cannot use the t-test for a mean from Lecture 6.3, then you could use a non-parametric test for the median such as the Sign test or the Wilcoxon Signed Rank test

- Null hypothesis significance testing has come under fire recently
    - Researchers have routinely misunderstood p-values and statistical significance
    - Growing frustration, some calls to ban p-values or statistical significance
    - American Statistical Association (ASA) statement responding to concerns
        - Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "$p < 0.05$", The American Statistician, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913
        - Highlights several common misuses of statistical hypothesis testing
        - Several direct quotes from the article are in italics below, with some discussion points / considerations after each

- *Don't believe that an association or effect exists* [only] *because it was statistically significant. Don't believe that an association or effect is absent* [only] *because it was not statistically significant.*
    - Note: I have added the word *only* to emphasize that p-values are only one tool that can be used to help us make decisions

    - Recall: as sample size increases, standard deviation of a statistic decreases; this leads to a larger test statistic, which corresponds to smaller p-value
    - Thus, having more subjects in a study makes it more likely the p-value will be below the cutoff for establishing statistically significant results.
    - Statistically significant results may not be meaningful in context!

    - Results that are not statistically significant may be meaningful in context!

Example: Do families in Wake County, NC make more average income then the state average of \$59,481? Based on a rs of 100 families in Wake, $\bar{y} = \$59{,}850$ and $s = \$4{,}000$. Calculate the test statistic and p-value for testing H_0: $\mu = 59481$ vs. H_A: $\mu > 59481$.

Example: What if n had been larger? Use the same $\bar{y}$ and s to conduct another test, now using $n = 500$. Calculate test stat and p-value for testing H_0: $\mu = 59481$ vs. H_A: $\mu > 59481$.

- *Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.*
    - P-value is probability we would observe the data we did (or something more extreme) if we were living in the world where the null hypothesis were true
        - We assume the null hypothesis is true so we can use it as a frame of reference to evaluate sample data against
    - We cannot prove or disprove the null hypothesis.
        - These decisions mean we did (or did not) find enough evidence to support the alternative
        - Though unlikely, we could have made a Type I or Type II error

- *Don't conclude anything about scientific or practical importance based* [solely] *on statistical significance (or lack thereof).*
    - Note: Again, the word *solely* was added to emphasize that p-values are only one tool that can be used to help us make decisions; they give useful information but other metrics should be considered for decision making
    - Statistical significance occurs when the p-value is less than or equal to the significance level.

    - Practical significance or practical importance occurs when an effect is meaningful in context.

        - Ex: How large would average reduction in blood pressure need to be before doctors would start prescribing new medication?
        - Ex: How large would improvement in graduation rates need to before a school district would adopt a new policy?
    - For some studies, results will be both statistically significant and practically significant; for other studies, results will only be one or the other.

- Hypothesis testing is one tool in your decision making arsenal
    - ASA discouraged dichotomous thinking via statistical significance

    - Other metrics include: likelihood ratios, confidence intervals, credible intervals, prediction intervals

- Hypothesis testing and confidence intervals should be used together!
    - Concepts presented separately to allow you to focus on core ideas for each
    - These tools have different primary goals
        - CI: Estimate what is true in a population (e.g. true value of parameter)
        - HT: Explore support for specific hypothesis about a population
    - But these goals complement each other

    - Recall: interval provides a range of reasonable estimates for the parameter
        - What support do these estimates provide for the hypothesis?

Example: A company manufactures large sugar cookies that are then sold to a chain of coffee shops. The cookies are made by a machine set to produce cookies with an average weigh 75 grams. The company wants to know if this machine is working correctly. They take a random sample of 30 cookies and used their weights to produce the output below. Based on this output, do we have evidence the machine is working properly? Explain.

95% confidence interval results:

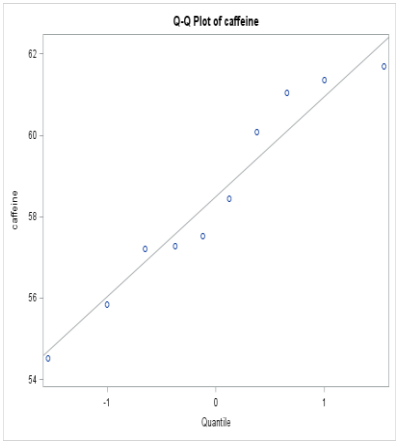
Variable	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
Oatmeal	75.26	1.63	29	71.93	78.60

Lecture 6.6: Additional Examples

Use separate sheets of paper to work on the examples below.

Example 1: Of 3,865 full-time employees surveyed by the Gallup Organization, 42% said that the ability to use their computers and mobile devices to stay connected to the workplace outside of their normal working hours was a 'strongly positive' development. Conduct the appropriate hypothesis test to determine if more than 40% of full-time employees feel that the ability to use their computers and mobile devices to stay connected to the workplace outside of their normal working hours was a 'strongly positive' development. Use a 1% significance level.

Example 2: Suppose the mean caffeine content in a cup of regular coffee for a certain brand is 120mg. I bought a "half-caff" blend from this brand but am still sleepy in the morning. I suspect that the mean caffeine content in a cup of half-caff is really less than 60mg. I take a random sample of 10 cups. I find the mean caffeine content for this sample is 58.5mg with a standard deviation of 2.5mg. I also produce the normal probability plot (qq-plot) at the right.



a. Conduct a hypothesis test of my suspicion using $\alpha = 0.05$.
b. Conduct a hypothesis test of my suspicion using $\alpha = 0.01$. What changes from part (a)?
c. Based on the decision in part (b):
    i. It is possible a Type I Error was made.
    ii. It is possible a Type II Error was made.
    iii. It is possible either a Type I or a Type II Error was made.
    iv. It is not possible either error was made if the test was implemented correctly.

    Select one answer and explain.

Example 3: Suppose the average height of American adults is 67 inches. I want to know if the average height of students in my large introductory statistics class is different.

a. Use SAS to determine the sample size necessary to achieve 85% power for a hypothesis test using a 10% significance level when I want to be able to detect at least a 2 inch difference from the national average (hint: think about which value(s) for the true mean under the alternative this would correspond to). Based on historical data, I believe the true standard deviation is 5 mg.
b. Write sentences interpreting alpha, beta, and power in this context.
c. In a random sample of 50 students, the average height is found to be 68.5 inches with a standard deviation of 10 inches. Conduct a test with a 10% significance level.

Example 4: The head librarian at a large university would like to determine if at least 80% of students have used the university library's website to find resources for a class. A random sample of 100 undergraduate students this university found that 78 of them have done so. Does this provide evidence to support the librarian's hypothesis?