Hypothesis Testing and Single Sample Statistical Inference

Lectures 13, 14, & 15

Readings: GW 8, *SS* 8 (except 8.4)

Where Am I?

- Wake up after a rough night in unfamiliar surroundings
- Still in Boulder?

Expected if in Boulder (large probability)

Surprising but not impossible (moderate probability)

Couldn't happen **IF** in Boulder (probability near zero) → Reject that you're in Boulder

Conceptual Steps of Hypothesis Testing

1. State clearly the two hypotheses regarding a statistic of interest (e.g., the mean). Determine which is the null hypothesis (H_0) and which is the alternative hypothesis (H_1)
2. Find the sampling distribution of the statistic (e.g., the shape, mean, and SD of the distribution of means) ASSUMING THAT NOTHING IS GOING ON (the null is true).
3. Compute a relevant test statistic (e.g., z-statistic) from your sample
4. Compare your test statistic (e.g., a z-statistic) to those expected if the null is true.
5. Inference: If your test statistic is surprising if the null is true, then conclude the null isn't true.

Practical Steps of Hypothesis Testing of Means

1. Determine H_0 vs. H_1 (should regard the population mean)
2. Choose alpha level (α): how willing you are to abandon null (usually .05)
3. Find the shape (normal if n>30), mean (μ, defined by H_0), and SD ($SD/\sqrt{n}$,or SEM) of the hypothetical sampling distribution of means assuming H_0 is true
4. Compute the mean and then the z-value: $z = \frac{\bar{X} - \mu}{SD/\sqrt{n}}$
5. Find the p-value: the probability of observing a sample statistic this far or further from the mean that is expected if H_0 is true.
6. If $p \leq \alpha$, then reject H_0 (because it is unlikely to be true)

Single Sample Statistical Inference

- Do CU students have SAT math scores higher than the national average (μ= 500 σ=100)?
- Null Hypothesis:
  - "CU student average is the same as the national average."
  - (A sample of CU students is just like a sample randomly drawn from the population of all students taking the SAT test.)
- Alternative Hypothesis:
  - "CU students have average SAT scores that are different than the national average."
- Assuming the null (that CU students are just a sample randomly drawn from the national population), what is the sampling distribution of sample means if *n*=64?
  - 1) Shape? *Normal if a) original scores normal or b) n>30*
  - 2) Mean? *500 (defined from the null!)*
  - 3) Standard Deviation? *100/sqrt(64) = 100/8*

$$normal; E(\bar{X}) = 500; \sigma_{\bar{X}} = \frac{100}{\sqrt{n}}$$

To reject the null or not reject the null?

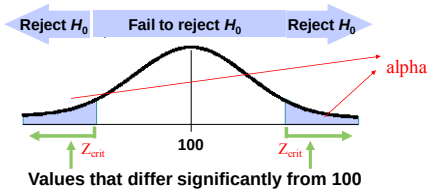
- If my sample is sufficiently surprising, given the sampling distribution we constructed (by assuming H_0 is true), then I reject the null hypothesis and go with the alternative hypothesis.
  - Unsurprising: Middle-most 95% of the sampling distribution if H_0 true
  - Surprising: Outer 5% of sampling distribution if H_0 true. Would get a sample mean this or more extreme <5% (α) of the time if H_0 true
- For a standard normal distribution, 5% of scores are as far or further than z=1.96 away from the mean. Thus, we can call +1.96 and -1.96 our critical values.
  - If z > 1.96 or z < -1.96, we can reject the null if the sampling distribution is normally distributed.
  - Such values occur less than 1 time in 20 if the null hypothesis is true.

Alpha for a two-tailed test

H_0: μ = 100

H_1: μ ≠ 100

Means less than OR greater than



Values that differ significantly from 100

p-values

- p-value
  - Def: Probability of a statistic **equal to or more extreme** than what you actually got if the null is true.
  - Measure of how consistent data are with H_0
  - Large p-value → H_0 is a good explanation of the data
  - Small p-value → H_0 is a poor explanation of the data
- p > α: Retain null hypothesis
- p ≤ α: Reject null hypothesis; accept alternative hypothesis
- Researchers generally report p-values, because reader can choose own alpha level
  - E.g. "p = .03"
  - If willing to allow 5% error rate, then accept result as reliable
  - If more stringent, say 1% (α = .01), then remain skeptical
  - Alternative (optional) definition of p-value: the smallest α at which the null would be rejected given the data. Thus, p-values can be thought of as type-I error rates given the data at hand.

Single Sample Statistical Inference: Completed example 1

- Null Hypothesis: $H_0 : \mu = 500$
- Alternative Hypothesis: $H_1 : \mu \neq 500$
- My data; 540, 620,490, …, 520 $\bar{X} = 531; n = 64$
- Will convert my sample mean to a Z statistic in the sampling distribution if H_0 true.

$$Z_{\bar{x}} = \frac{\bar{X} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{531 - 500}{100/\sqrt{64}} = \frac{31}{12.5} = 2.48$$

- 2.48 is more extreme than the critical value of Z = ±1.96
- Say p < .05 (the probability of observing a mean as far or further from the hypothesized mean under the null is <5%)
- Therefore reject the null hypothesis: Unlikely that I got a sample mean this extreme if the null hypothesis is true.
- I conclude that CU students on average have higher SAT scores.

4-Sentence Summary of Our Study

- 4 sentences:
  - Intro – what is it that we're interested in studying?
  - Method – how did we go about studying the question just identified?
  - Results – what did we find (with statistics in parentheses)?
  - Discussion – what can we conclude?
- We were interested in determining whether students at the University of Colorado at Boulder (CU) have higher SAT scores than the national average of 500. To investigate this, we ascertained the SAT scores of 64 randomly selected students at CU. The average of the 64 students was 531, or thirty-one points higher than the national average (z=2.48, p<.05). We conclude that the typical CU student scores higher on the SAT than the national average.

Single Sample Statistical Inference: Completed example 2

- Same question by a researcher at the University of Nebraska
- Null hypothesis: $H_0 : \mu = 500$
- Alternative hypothesis:: $H_1 : \mu \neq 500$
- His data; 510, 490,540, …, 520 $\bar{X} = 519; n = 64$

$$Z_{\bar{x}} = \frac{\bar{X} - \mu_0}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{519 - 500}{100/\sqrt{64}} = \frac{19}{12.5} = 1.52$$

- 1.52 is less extreme than the critical value of Z = ±1.96
- Say *ns* (for *not significant*, rather than p > .05)
- Therefore do not reject the null hypothesis: Such a sample is not sufficiently unlikely given the null hypothesis
- NU students could have true mean not different from national average.

Common Misunderstandings of inferential tests

- "Failure to reject means the Null Hypothesis is true"
  - NO! Indeed, the null (being a single point) is rarely true. Rather, we don't have evidence to reject it.
- "The p-value is the probability that the null is true"
  - NO! We cannot know the probability of the null. The p-value is the probability of observing a statistic as or more extreme if the null were true
- "Rejecting the null means my research hypothesis is correct"
  - NO! It just means the null hypothesis is unlikely – for WHATEVER reason

Single Sample Statistical Inference: Example 3

- We know the scores on a extraversion test have a standard deviation of 20 and a mean of 50. The scores are right skewed (not normally distributed). We are interested in testing whether people who are from the deep South are more extraverted than the population at large. We randomly select 100 individuals from the deep South and find that their mean is 51. What do we conclude?

1. H_0 vs. H_1?
2. What is alpha level (α)?
3. Shape, mean, and SD (i.e., SEM) of the hypothetical sampling distribution of means assuming H_0 is true?
4. The mean and then the z-value?
5. The p-value?
6. If $p \leq \alpha$, then reject H_0

Single Sample Statistical Inference: Example 4

- We know the scores on a extraversion test have a standard deviation of 20 and a mean of 50. The scores are right skewed (not normally distributed). We are interested in testing whether people who are from the deep South are more extraverted than the population at large. We randomly select 10,000 individuals from the deep South and find that their mean is 51. What do we conclude?

1. H_0 vs. H_1?
2. What is alpha level (α)?
3. Shape, mean, and SD (i.e., SEM) of the hypothetical sampling distribution of means assuming H_0 is true?
4. The mean and then the z-value?
5. The p-value?
6. If $p \leq \alpha$, then reject H_0

Effect size

- Effect size
  - Measure magnitude of a finding or effect, irrespective of statistical significance
- E.g., how much more extraverted are Southerners?
  - Example 3: mean difference of 1, ns
  - Example 4: mean difference of 1, p << .05
- Both have the same effect size: "mean difference of 1" is an effect size
  - But how big is "1"?
  - Answer: it depends on the standard deviation
  - "1" with small SD = "big" effect
  - "1" with large SD = "small" effect
  - So standardize by the SD!

Cohen's d: Measure of effect size

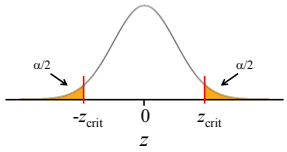
- Cohen's *d*: standardized mean difference
- Cohen's *d* = $\frac{\bar{X} - \mu}{SD}$
- How far sample mean is from hypothesized null mean in units of SD (exactly like our old z-score formula!)
- NOT the same (!) as the inferential z-score formula:

$$\frac{\bar{X} - \mu}{SD / \sqrt{n}}$$

- The inferential z-score formula tells you how unlikely a mean is, given the null. Cohen's d tells you how far a mean is from the null hypothesized mean, in SD units

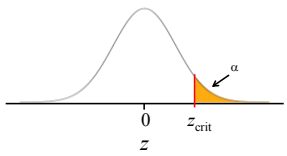
Two-Tailed Tests

- Usually we want to be able to detect effects in either direction
  - Drugs that help or drugs that hurt
  - Students take too few or too many starbursts
- Formalized in alternative hypothesis
  - $\mu \neq \mu_0$ (i.e., $\mu < \mu_0$ **or** $\mu > \mu_0$)
- Two critical values, one in each tail
- Type I error rate is sum from both critical regions
  - Each gets $\alpha/2$ (2.5%)

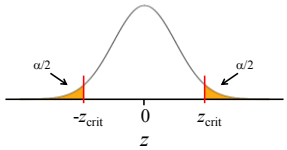


One-Tailed vs. Two-Tailed Tests

One-tailed



Two-tailed



Problem with one-tailed tests

- We conduct a one-tailed test, but we actually find a large effect in the *opposite* direction from that predicted. Do we report it? Most would!
- If a scientist *would report* results in the opposite direction, then the actual type-I error rate is two times larger than we say it is in a one-tailed test.
- Therefore, such tests are more liberal than they purport to be
- For this reason, one-sided tests are looked down upon and rarely used

Statistical Decision Matrix

Decision	Reality: Null Hypothesis True $\mu = 500$	Reality: Null Hypothesis False $\mu \neq 500$
Reject H_0		
Do Not Reject H_0		

Statistical Decision Matrix

Decision	Reality: Null Hypothesis True $\mu = 500$	Reality: Null Hypothesis False $\mu \neq 500$
Reject H_0	Type I error FALSE ALARM Prob=α	Correct HIT
Do Not Reject H_0	Correct ALL QUIET	Type II error MISS Prob=β

What error could the researcher have made for CU? With what probability?

For NU?

Real Life Decision Matrix

Decision	Reality: H0 true Suspect innocent	Reality: H1 true Suspect guilty
Say "guilty"	False Alarm: Cost? innocent person in jail	Correct conviction
Say "Not guilty"	Correct acquittal	Miss: Cost? criminal walks free

Trade off between Misses and False Alarms:
How do you avoid misses? – always say "guilty".
How do you avoid false alarms – always say "not guilty".

Scientific Decision Matrix

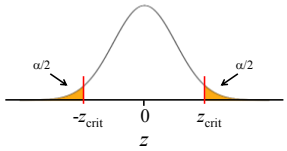
Decision	Reality: H0 true Nothing's going on	Reality: H1 true Something's going on
Reject null	False Alarm Cost?	Hit
Do not reject null	All quiet	Miss Cost?

Types of Errors

- Goal: Reject null hypothesis when it's false; retain it when it's true
- Two ways to be wrong
  - Type I Error: Null is correct but you reject it
  - Type II Error: Null is false but you retain it
- Type I Error rate
  - **IF** H_0 is true, probability of mistakenly rejecting H_0
  - Proportion of false theories we conclude are true
    - Proportion of useless drugs that are deemed effective
- Logic of hypothesis testing is founded on controlling Type I Error rate
  - Set critical value to give desired Type I Error rate

Alpha Level

- Choice of acceptable Type I Error rate
  - Usually .05 in psychology
  - Higher → more willing to abandon null hypothesis
  - Lower → require stronger evidence before abandoning null hypothesis
- Determines critical value
  - Under the sampling distribution of the test statistic according to the null hypothesis, the probability of a result beyond the critical value is α



Power

- Type II Error rate
  - **IF** H_0 is false, probability of failing to reject it
  - E.g., fraction of cheaters that don't get caught
- Power (see *SS 8.6.4*)
  - **IF** H_0 is false, probability of correctly rejecting it
  - Equal to one minus Type II Error rate
- Power increases with
  - Larger "effect sizes"
  - Larger sample sizes
  - Smaller standard deviations in original scores