Final Review Session Spring 2025

Recap - confidence intervals and testing

How confidence intervals behave

Recall the form of a CI:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

Where $z^* \frac{\sigma}{\sqrt{n}}$ is the margin of error.

How confidence intervals behave

The margin of error gets smaller when:

- z* is smaller (i.e., you change to a smaller confidence level). Thus, there is a trade-off between the confidence level and the margin of error.
- σ is smaller. You might be able to reduce σ if there is measurement error. Often times, the σ can't be reduced, it is just a characteristic of the population
- n is larger.

How hypothesis tests behave

- Statistical significance depends on sample size (since sample size determines the standard error of the sampling mean)
- Recall the form of the z-test:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{\text{magnitude of observed effect}}{\text{size of chance variation}} = \frac{\textit{signal}}{\textit{noise}}$$

- The numerator quantifies the distance between what you observe in the sample and the null hypothesized parameter.
- The denominator represents the size of chance variations from sample to sample

How hypothesis tests behave

- Statistical significance depends on:
  - The size of the observed effect ($\bar{x} - \mu$)
  - The variability of individuals in the population (σ)
  - The sample size (n)
  - Your criteria for rejection the null (α)

If you obtain a small p-value it is not necessarily because the effect size is large.

Type I error, and Type II error in hypothesis tests

	H_a is true	H_0 is true
Reject H_0	Correct decision	Type I error (α)
Fail to reject H_0	Type II error (β)	Correct decision

This table should remind you of something we have seen before....

Power

- The power is the chance of making the correct decision when the alternative hypothesis is true.
- Thus, it is the complement of β
- Power $= 1 - \beta$

	H_a is true	H_0 is true
Reject H_0	Correct decision	Type I error (α)
Fail to reject H_0	Type II error (β)	Correct decision

However, there are an infinite number of possible values that μ could assume that are not $= \mu_0$

Thus we must choose a value at which to evaluate the β and power for an alternative hypothesis...

When we evaluate β we do so at a single such value μ_1

Overview of Part III

part III: basic goal

Our overarching goal in part III is really to take our two "recipes" for statistical inference

1) Hypothesis testing
2) Confidence intervals

and figure out which ingredients to add in different situations.

choose the ingredients

We want to answer the questions:

1) What kind of an outcome variable are we working with?
2) How many groups/categories do we have data from that we want to compare?
3) Are the groups independent from each other or are observations inherently related/paired?
4) Do we meet the assumptions for a parametric test?

choose the ingredients

Based on the answers to the previous questions, we choose our "ingredients"

1) the effect/difference we want to examine
2) the variability we have in the data
3) a distribution we will be using to draw a critical value from based on:
   - our desired alpha
   - one or two tailed hypothesis

Decision tree:

Decision tree:

Note: the prezi presentation also linked on the resources page may be helpful

Recipe 1: Hypothesis testing

1) Check the assumptions for the theoretical distribution
2) Create a ratio of the effect/difference to the variability in the data
3) Generate the probability of observing that difference or greater if the null is true
4) Make a decision to reject or not reject the null hypothesis

Recipe 2: Confidence intervals

1) Generate your estimate
2) Calculate the critical value associated with your theoretical distribution
3) multiply the critical value by the variability
4) create your upper and lower bounds

For each test know:

- When to use it
- Any important assumptions that can be checked using data
- Appropriate visualization for the data
- What theoretical distribution are we using for inference
- How to construct the statistical test
  - what are the null and alternative hypotheses for the test
- How to construct the confidence interval
- relevant syntax in R
- any special notes/considerations

for example:

One sample T - used when we have 1 sample of a continuous outcome that we are comparing to a hypothesized value - assuming SRS, normality of the outcome, independence of outcomes - we might look at a histogram, density plot or qq plot - compared to a t distribution with n-1 df

for example:

One sample T - null: the mean is = hypothesized mean - alternative: could be one or two sided

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

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

for example:

One sample T relevant syntax:
pt() t.test(variable, alternative =“ “, mu=)

Notes: think about when the test is robust to violations of the assumptions and why

Some examples - what tests?

Today's fun fact

The length of your hand from your wrist to your elbow is the same as the length of your foot from your heel to your big toe. If I want to show that these two measures are almost perfectly correlated how might I do that?

You could do a correlation test, a linear regression, a paired t, you could show a scatterplot. . .

Today's fun fact

What kind of plot would I expect to see for these data?

Perfectly correlated scatterplot. Observations falling on a straight line with increasing slope.

Example 1: Staph infections

Researchers recruited 917 patients who had tested positive for staphylococcus Aureus and randomly assigned them to a staph-killing nasal ointment or placebo. They were interested in testing weather this drug was associated with a reduction in post-surgical infections. In the active treatment group 17 of 504 patients developed infections, in the placebo group 32 of 413 patients developed infections.

- What are the exposure and outcome variables?
- What kind of a test would you use for these data?
- What is the null hypothesis of this test?

Staph infections: ingredients

1) the effect/difference we want to examine
2) the variability we have in the data
3) a distribution we will be using to draw a critical value from based on:
   - our desired alpha
   - one or two tailed hypothesis

Staph infections ingredient 1: effect

Difference between two proportions:

$$\hat{p_1} - \hat{p_2} = \frac{17}{504} - \frac{32}{413} = .0337 - .0775 = -.0438$$

Staff infections ingredient 2: variability

- If the null hypothesis is true, then p_1 is truly equal to p_2. In this case, our best estimate of the underlying proportion that they are both equal to is

$$\hat{p} = \frac{\text{no. successes in both samples}}{\text{no. individuals in both samples}} = \frac{17 + 32}{504 + 413} = 0.0534$$

Staff infections ingredient 2: variability

- Our best guess at the SE for $\hat{p}$ is:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

$$\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}$$

This is the formula for the SE for the difference between two proportions but we have substituted $\hat{p}$ for p_1 and p_2.

$$\sqrt{0.0\hat{5}34 * (0.9466)(\frac{1}{504} + \frac{1}{413})} = 0.01492$$

Staph infections ingredient 3: distribution

Here for two sample testing, we have more than 10 successes and 10 failures in each group so we feel comfortable using a normal approximation to the binomial.

We will use a Z distribution, and an alpha of 0.05

The test is one sided because we are only interested in the left side of the distribution - decreases

Staph infections: test statistic

$$z = \frac{.0337 - .07748}{\sqrt{.0534 * 0.9466\left(\frac{1}{504} + \frac{1}{413}\right)}} = -2.936$$

Staph infections

```
## [1] 0.001662372
```

In this case we are only interested in a reduction in infections - so we will only look at the left tail of the distribution.

Staph infections

In R?

```r
prop.test(x = c(17,32), # x is a vector of number of successes
          n = c(504,413), alternative="less" , correct=F) # n is a vector of
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(17, 32) out of c(504, 413)
## X-squared = 8.5906, df = 1, p-value = 0.00169
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.01839005
## sample estimates:
##     prop 1     prop 2
## 0.03373016 0.07748184
```

Staph infections

In R?

```
binom.test(x=c(17,32),n=c(504,413), alternative="less")
```

```
##
##  Exact binomial test
##
## data:  c(17, 32)
## number of successes = 17, number of trials = 49, p-value = 0.02219
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.4738246
## sample estimates:
## probability of success
##              0.3469388
```

Staph infections: Confidence interval

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$-.0438 \pm 1.96\sqrt{\frac{.0\hat{3}37(1-.0\hat{3}37)}{504} + \frac{.0\hat{7}75(1-.0775)}{413}} = -.0438 \pm 0.01542$$

95% CI is -0.0592 to - 0.04226

Example 2

August 8, 2019 Vitamin D Supplementation and Prevention of Type 2 Diabetes

BACKGROUND Observational studies support an association between a low blood 25-hydroxyvitamin D level and the risk of type 2 diabetes. However, whether vitamin D supplementation lowers the risk of diabetes is unknown.

Example 2 cont.

METHODS We randomly assigned adults who met at least two of three glycemic criteria for prediabetes (fasting plasma glucose level, 100 to 125 mg per deciliter; plasma glucose level 2 hours after a 75-g oral glucose load, 140 to 199 mg per deciliter; and glycated hemoglobin level, 5.7 to 6.4%) and no diagnostic criteria for diabetes to receive 4000 IU per day of vitamin D3 or placebo, regardless of the baseline serum 25-hydroxyvitamin D level.

What type of study is this?

What type of variable is the predictor (how many groups)?

Example 2 cont.

RESULTS By month 24, the mean serum 25-hydroxyvitamin D level in the vitamin D group was 54.3 ng per milliliter (from 27.7 ng per milliliter at baseline), as compared with 28.8 ng per milliliter in the placebo group (from 28.2 ng per milliliter at baseline). After a median follow-up of 2.5 years, the primary outcome of diabetes occurred in 293 participants in the vitamin D group and 323 in the placebo group (9.39 and 10.66 events per 100 person-years, respectively).

What kinds of tests would you use here for the vitamin D comparison? for the outcome?

Example 2

Other considerations: why might we not see the result we were expecting?

Per the discussion in the article:

"Because vitamin D supplements are used increasingly in the U.S. adult population,29 approximately 8 of 10 participants had a baseline serum 25-hydroxyvitamin D level that was considered to be sufficient according to current recommendations ($\geq$ 20 ng per milliliter) to reduce the risk of many outcomes,23,30 including diabetes.6 The high percentage of participants with adequate levels of vitamin D may have limited the ability of the trial to detect a significant effect.""

Example 3:

A study on the effects of vaping classifies people as “never vapers”, “occasional vapers”, “frequent vapers”. You interview a sample of 150 people in each group and ask a questionnaire to derive a quantitative score (between 0 and 100) on stress levels.

What kind of an outcome is this? What test is appropriate here?

Example 4:

The amygdala is a brain structure involved in the processing of memory of emotional reactions. Ten subjects were shown emotional video clips and non emotional video clips in random order. They then had their memory of the clips assessed. Recall accuracy was scored from 1 to 100.

What type of data do you have? What kind of a test is appropriate?

Example 5:

A random sample of 700 births from local records shows this distribution across the days of the week. Do these data give evidence that local births are not equally likely on all days of the week?

Day	Births
Monday	110
Tuesday	124
Wednesday	104
Thursday	94
Friday	112
Saturday	72
Sunday	84

What test would we use here?

What is the null hypothesis?

Example 5: expectation

Day	Births
Monday	100
Tuesday	100
Wednesday	100
Thursday	100
Friday	100
Saturday	100
Sunday	100

Example 6

You have heard that the grading scale is harsher at UC Berkeley than at other California universities. You want to test this rumor with data. You have data from a random sample of 100 transcripts from students at Berkeley who took PH142 and data on the letter grade distribution for undergraduate statistic courses in general from a California wide survey.

What kind of outcome? How many groups/samples?

What test would you consider here?

Example 6

You have heard that the grading scale is harsher at UC Berkeley than at other California universities. You want to test this rumor with data. You have data from a random sample of 100 transcripts from students at Berkeley who took PH142 and data on the letter grade distribution for undergraduate statistic courses in general from a California wide survey.

This is a catgorical outcome, with one sample, so we would use a chi-squred goodness of fit test

Example 6

You find a source that gives the distribution for UC intro biostat courses as: A=50%, B=30%, C=15%, Fail=5%

Grade	N	Expected?
A	224	?
B	99	?
C	17	?
F	10	?
Total	350	350

Example 6

Grade	N	Expected?
A	224	175
B	99	105
C	17	52.5
F	10	17.5
Total	350	350

statistic$= 13.72 + 0.34 + 24.00 + 3.214 = 41.28$

```
pchisq(41.28, df=3, lower.tail=FALSE)
```

```
## [1] 5.703407e-09
```

Additional examples for review

ANOVA

The `melanoma` data set contains data on 205 patients from Denmark with malignant melanoma. You have joined a lab in which the principal investigator is interested in determining whether mean tumor thickness (mm) differs by patient status (1 = died from melanoma, 2 = alive, 3 = died from other causes).

```
head(melanoma)
```

```
##   time status sex age year thickness ulcer status2
## 1   10      3   1  76 1972      6.76     1       3
## 2   30      3   1  56 1968      0.65     0       3
## 3   35      2   1  41 1977      1.34     0       2
## 4   99      3   0  71 1968      2.90     0       3
## 5  185      1   1  52 1965     12.08     1       1
## 6  204      1   1  28 1971      4.84     1       1
```

ANOVA

```
anova <- aov(thickness ~ status2, data = melanoma)
tidy(anova)
```

```
## # A tibble: 2 x 6
##   term         df sumsq meansq statistic    p.value
##   <chr>     <dbl> <dbl>  <dbl>     <dbl>      <dbl>
## 1 status2       2  180.   90.2       11.3  0.0000216
## 2 Residuals   202 1606.    7.95      NA   NA
```

Your task is to help your PI analyze the results of this ANOVA run in R. For the ANOVA above, what are the null and alternative hypotheses?

ANOVA

```
tidy(anova)
```

```
## # A tibble: 2 x 6
##   term          df sumsq meansq statistic    p.value
##   <chr>      <dbl> <dbl>  <dbl>     <dbl>      <dbl>
## 1 status2        2  180.  90.2       11.3  0.0000216
## 2 Residuals    202 1606.   7.95      NA   NA
```

Based on the results shown, what would you conclude?

ANOVA

Based on these results what would be your next step in the analysis process? Justify your answer in 1-2 sentences. (would you continue your analysis with an additional test and if so, what test would you use)

ANOVA

What would you conclude from these results?

```
TukeyHSD(anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = thickness ~ status2, data = melanoma)
##
## $status2
##           diff        lwr       upr     p adj
## 2-1 -2.0663511 -3.1192614 -1.013441 0.0000191
## 3-1 -0.5931955 -2.5792549  1.392864 0.7606857
## 3-2  1.4731557 -0.3970052  3.343316 0.1531399
```

Swimming

A group in the athletic department is working with the swim team. They implement a new training program and want to know if there has been an improvement in the 50m swim time (in seconds) at 12 weeks following the start of the program.

Swimming

They have collected the following data:

Pre program	Post Program
24.23	24.26
24.12	24.09
24.15	24.11
24.12	24.13
24.16	24.15
24.18	24.19
24.51	24.42
24.69	24.69
24.88	24.82
25.01	24.94
25.58	25.55
25.47	25.45
25.66	25.67

Swimming

pre	post	diff	abs	sign	rank
24.69	24.69	0 0	0	na	na
24.12	24.13	0.01	0.01	+	2.5
24.16	24.15	-0.01	0.01	-	2.5
24.18	24.19	0.01	0.01	+	2.5
25.66	25.67	0.01	0.01	+	2.5
25.47	25.45	-0.02	0.02	-	5
25.58	25.55	-0.03	0.03	-	7
24.23	24.26	0.03	0.03	+	7
24.12	24.09	-0.03	0.03	-	7
24.15	24.11	-0.04	0.04	-	10
24.88	24.82	-0.06	0.06	-	11
25.01	24.94	-0.07	0.07	-	12
24.51	24.42	-0.09	0.09	-	13

swimming

Calculate the appropriate test statistic. Show your work by writing the formula needed to calculate with values plugged in.

$$Z_T = \frac{T - \mu_T}{\sigma_T}$$

Where

$$\mu_T = \frac{n(n+1)}{4}$$

and

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

swimming

Write the line of code that would give you the appropriate p-value for this test statistic.:

swimming

```
pnorm(-1.922, mean = 0, sd = 1)
```

```
## [1] 0.02730288
```

swimming

Based on your findings would you recommend that the athletic department continue this training program? Why or why not?