Midterm II Review

Sta 104 - Summer 2018, Term 1

Duke University, Department of Statistical Science

1. Housekeeping

2. Review

- Project Proposal is due Friday 11:55 pm
- Midterm 2 is Thursday in class
- PS 5 is due Saturday (11:55 pm)
- PA 5 is due Saturday (11:55 pm)
- RA 6 is Friday

- When: Thursday June 14, 11 am - 12:15 pm
- What to bring:
  - Scientific calculator (graphing calculator ok, No Phones!)
  - Writing utensils
  - Cheat sheet (can be typed!)
- Tables will be provided in the same way as last time (Piazza link and Sakai course resources)

- Covers from last midterm on. HT - end of Unit 3, Unit 4, and Unit 5
- 2 ”written” questions: 22 points
- 11 multiple choice questions - 2 points each
- 44 points (as compared to 67 from last midterm)

1. Housekeeping

2. Review

What should you know?

- Two mean testing problems
  - Independent means
  - Paired (dependent) means
- Conditions
  - Independence
  - Skew or Approximate Normality

All other details of the inferential framework is the same...

$$HT: test\ statistic = \frac{point\ estimate - null}{SE}$$

$$CI: point\ estimate \pm critical\ value \times SE$$

One mean:

$df = n - 1$

HT:

$H_0 : \mu = \mu_0$

$T_{df} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

CI:

$\bar{x} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}$

Paired means:

$df = n_{diff} - 1$

HT:

$H_0 : \mu_{diff} = 0$

$T_{df} = \frac{\bar{x}_{diff} - 0}{\frac{s_{diff}}{\sqrt{n_{diff}}}}$

CI:

$\bar{x}_{diff} \pm t^{\star}_{df} \frac{s_{diff}}{\sqrt{n_{diff}}}$

Independent means:

$df = min(n_1 - 1, n_2 - 1)$

HT:

$H_0 : \mu_1 - \mu_2 = 0$

$T_{df} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

CI:

$\bar{x}_1 - \bar{x}_2 \pm t^{\star}_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Clicker question

A study examining the relationship between weights of school children and absences found a 95% confidence interval for the difference between the average number of days missed by overweight and normal weight children ($\mu_{overweight} - \mu_{normal}$) to be 1.3 days to 2.8 days. According to this interval, we are 95% confident that overweight children on average miss

1. 1.3 days fewer to 2.8 days more
2. 1.3 to 2.8 days more
3. 1.3 to 2.8 days fewer
4. 1.3 days more to 2.8 days fewer

than children with normal weight.

Clicker question

A study examining the relationship between weights of school children and absences found a 95% confidence interval for the difference between the average number of days missed by overweight and normal weight children ($\mu_{overweight} - \mu_{normal}$) to be 1.3 days to 2.8 days. According to this interval, we are 95% confident that overweight children on average miss

1. 1.3 days fewer to 2.8 days more
2. ***1.3 to 2.8 days more***
3. 1.3 to 2.8 days fewer
4. 1.3 days more to 2.8 days fewer

than children with normal weight.

- Bootstrapping works as follows:

- Bootstrapping works as follows:
  1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample

- Bootstrapping works as follows:
  1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  2. calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples

- Bootstrapping works as follows:
  1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  2. calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
  3. repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics

- Bootstrapping works as follows:
  1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  2. calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
  3. repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics
- The XX% bootstrap confidence interval can be estimated by

- Bootstrapping works as follows:
  1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  2. calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
  3. repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics
- The XX% bootstrap confidence interval can be estimated by
  - the cutoff values for the middle XX% of the bootstrap distribution,

- Bootstrapping works as follows:
  1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  2. calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
  3. repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics
- The XX% bootstrap confidence interval can be estimated by
  - the cutoff values for the middle XX% of the bootstrap distribution,

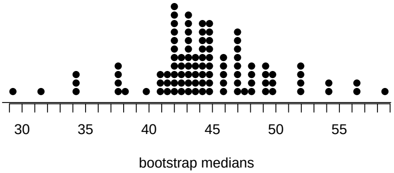
    OR

- Bootstrapping works as follows:
  1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
  2. calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
  3. repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics
- The XX% bootstrap confidence interval can be estimated by
  - the cutoff values for the middle XX% of the bootstrap distribution,

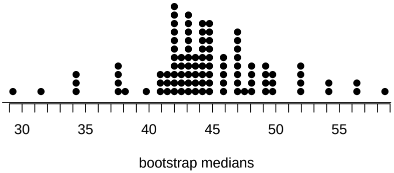
    OR

  - $point\ estimate \pm t^\star SE_{boot}$

For a random sample of 20 Horror movies, the dot plot below shows the distribution of 100 bootstrap medians of the Rotten Tomatoes audience scores. The median of the original sample is 43.5 and the bootstrap standard error is 4.88. Estimate the 90% bootstrap confidence interval for the median RT score of horror movies using the standard error method.

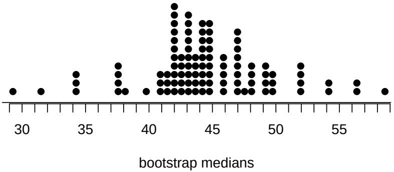


For a random sample of 20 Horror movies, the dot plot below shows the distribution of 100 bootstrap medians of the Rotten Tomatoes audience scores. The median of the original sample is 43.5 and the bootstrap standard error is 4.88. Estimate the 90% bootstrap confidence interval for the median RT score of horror movies using the standard error method.



two tails		0.200	0.100	0.050
df	16	1.34	1.75	2.12
	17	1.33	1.74	2.11
	18	1.33	1.73	2.10
	19	1.33	1.73	2.09

For a random sample of 20 Horror movies, the dot plot below shows the distribution of 100 bootstrap medians of the Rotten Tomatoes audience scores. The median of the original sample is 43.5 and the bootstrap standard error is 4.88. Estimate the 90% bootstrap confidence interval for the median RT score of horror movies using the standard error method.



two tails		0.200	0.100	0.050
df	16	1.34	1.75	2.12
	17	1.33	1.74	2.11
	18	1.33	1.73	2.10
	19	1.33	1.73	2.09

$$43.5 \pm (1.73 \times 4.88) = (35.1, 51.9)$$

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error, α*
	H_A true		

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	*Power, $1-\beta$*

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- *Power* of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1-\beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs.

- ANOVA tests for some difference in means of many different groups

- ANOVA tests for <u>some</u> difference in means of many different groups
- Conditions
  1. *Independence*:
     (a) within group: sampled observations must be independent, i.e., SRS + 10% rule
     (b) between group: groups must be independent of each other
  2. *Approximate normality*: distribution should be nearly normal within each group
  3. *Equal variance*: groups should have roughly equal variability

ANOVA tests for some difference in means of many different groups

Null hypothesis:

$$H_0 : \mu_{\text{placebo}} = \mu_{\text{purple}} = \mu_{\text{brown}} = \ldots = \mu_{\text{peach}} = \mu_{\text{orange}}.$$

Clicker question

Which of the following is a correct statement of the alternative hypothesis?

- (a) For any two groups, including the placebo group, no two group means are the same.
- (b) For any two groups, not including the placebo group, no two group means are the same.
- (c) Amongst the jelly bean groups, there are at least two groups that have different group means from each other.
- (d) Amongst all groups, there are at least two groups that have different group means from each other.

ANOVA tests for <u>some</u> difference in means of many different groups

Null hypothesis:

$$H_0 : \mu_{\text{placebo}} = \mu_{\text{purple}} = \mu_{\text{brown}} = \ldots = \mu_{\text{peach}} = \mu_{\text{orange}}.$$

Clicker question

Which of the following is a correct statement of the alternative hypothesis?

(a) For any two groups, including the placebo group, no two group means are the same.

(b) For any two groups, not including the placebo group, no two group means are the same.

(c) Amongst the jelly bean groups, there are at least two groups that have different group means from each other.

(d) ***Amongst all groups, there are at least two groups that have different group means from each other.***

F-statistic:

$$F = \frac{SSG \ / \ (k-1)}{SSE \ / \ (n-k)} \quad = \quad \frac{MSG}{MSE}$$

k: # of groups; n: # of obs.

F-statistic:

$$F = \frac{SSG \ / \ (k-1)}{SSE \ / \ (n-k)} \quad = \quad \frac{MSG}{MSE}$$

k: # of groups; n: # of obs.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Between groups	$k-1$	SSG	MSG	F_{obs}	p_{obs}
Within groups	$n-k$	SSE	MSE		
Total	$n-1$	SSG+SSE			

Note: F distribution is defined by two dfs: $df_G = k-1$ and $df_E = n-k$

F-statistic:

$$F = \frac{SSG \ / \ (k-1)}{SSE \ / \ (n-k)} \quad = \quad \frac{MSG}{MSE}$$

k: # of groups; n: # of obs.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Between groups	$k-1$	SSG	MSG	F_{obs}	p_{obs}
Within groups	$n-k$	SSE	MSE		
Total	$n-1$	SSG+SSE			

Note: F distribution is defined by two dfs: $df_G = k-1$ and $df_E = n-k$

The p-value will be given on exam, compare with the standard α level.

To identify which means are different, use t-tests and the Bonferroni correction

- If the ANOVA yields a significant results, next natural question is: "Which means are different?"
- Use t-tests comparing each pair of means to each other,
  - with a common variance (MSE from the ANOVA table) instead of each group's variances in the calculation of the standard error,
  - and with a common degrees of freedom (df_E from the ANOVA table)
- Compare resulting p-values to a modified significance level

$$\alpha^\star = \frac{\alpha}{K}$$

where $K = \frac{k(k-1)}{2}$ is the total number of pairwise tests

To identify which means are different, use t-tests and the Bonferroni correction

You will not be asked to perform the actual tests, but you should know:

- How to compute the adjusted Bonferonni significance level α^*.
- How to compute the standard error for this test.
- The associated degrees of freedom for the test statistic.

Application Exercise 4.4

	Df	Sum Sq	Mean Sq	F	p-value
Rank	2	1.59	0.795	2.74	0.066
Residuals	460	135.07	0.29		
Total	462	136.66			

Application Exercise 4.4

	Df	Sum Sq	Mean Sq	F	p-value
Rank	2	1.59	0.795	2.74	0.066
Residuals	460	135.07	0.29		
Total	462	136.66			

What percent of the total variability in evaluation scores is explained by instructor rank?

Application Exercise 4.4

	Df	Sum Sq	Mean Sq	F	p-value
Rank	2	1.59	0.795	2.74	0.066
Residuals	460	135.07	0.29		
Total	462	136.66			

What significance level should be used for a pair-wise post hoc test comparing the evaluation scores of teaching professors and tenured professors?

Distribution of $\hat{p}$

Central limit theorem for proportions: Sample proportions will be nearly normally distributed with mean equal to the population mean, p, and standard error equal to $\sqrt{\frac{p\,(1-p)}{n}}$.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p\,(1-p)}{n}}\right)$$

Conditions:

- Independence: Random sample/assignment + 10% rule
- At least 10 successes and failures

HT vs. CI for a proportion

- Success-failure condition:
  - CI: At least 10 *observed* successes and failures
  - HT: At least 10 *expected* successes and failures, calculated using the null value
- Standard error:
  - CI: calculate using observed sample proportion: $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - HT: calculate using the null value: $SE = \sqrt{\frac{p_0(1-p_0)}{n}}$

If the S-F condition is not met

- HT: Randomization test – simulate under the assumption that H_0 is true, then find the p-value as proportion of simulations where the simulated $\hat{p}$ is at least as extreme as the one observed.
- CI: Bootstrap interval – resample with replacement from the original sample, and construct interval using percentile or standard error method.

Clicker question

A report on your local TV station says that 60% of the city's residents support using limited city funds to hire and train more police officers. A second local news station has picked up this story, and they claim that certainly less than 60% of residents support the additional hiring and training. In order to test this claim the second news station takes a random sample of 100 residents and finds that 57 of them (57%) support the use of limited funds to hire additional police officers.

Clicker question

Which of the following is the correct set-up for calculating the p-value for this test?

(a) Roll a 10-sided die (outcomes 1-10) 100 times and record the proportion of times you get a 6 or lower. Repeat this many times, and calculate the proportion of simulations where the sample proportion is 57% or less.

(b) Roll a 10-sided die (outcomes 1-10) 100 times and record the proportion of times you get a 6 or lower. Repeat this many times, and calculate the proportion of simulations where the sample proportion is 60% or less.

(c) In a bag place 100 chips, 57 red and 43 blue. Randomly sample 100 chips, with replacement, and record the proportion of red chips in the sample. Repeat this many times, and calculate the proportion of samples where 57% or more of the chips are red.

(d) Randomly sample 100 residents of a nearby city, record how many of the them who support the hiring and training of additional police officers. Repeat this many times and calculate the proportion of samples where at least 57% of the residents support additional hiring and training.

Clicker question

Which of the following is the correct set-up for calculating the p-value for this test?

(a) ***Roll a 10-sided die (outcomes 1-10) 100 times and record the proportion of times you get a 6 or lower. Repeat this many times, and calculate the proportion of simulations where the sample proportion is 57% or less.***

(b) Roll a 10-sided die (outcomes 1-10) 100 times and record the proportion of times you get a 6 or lower. Repeat this many times, and calculate the proportion of simulations where the sample proportion is 60% or less.

(c) In a bag place 100 chips, 57 red and 43 blue. Randomly sample 100 chips, with replacement, and record the proportion of red chips in the sample. Repeat this many times, and calculate the proportion of samples where 57% or more of the chips are red.

(d) Randomly sample 100 residents of a nearby city, record how many of the them who support the hiring and training of additional police officers. Repeat this many times and calculate the proportion of samples where at least 57% of the residents support additional hiring and training.

CLT also describes the distribution of $\hat{p}_1 - \hat{p}_2$

$$(\hat{p}_1 - \hat{p}_2) \sim N\left(mean = (p_1 - p_2), SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Conditions:

- Independence: Random sample/assignment + 10% rule
- Success-failure condition: At least 10 successes and failures

For HT where $H_0 : p_1 = p_2$, pool!

As with working with a single proportion,

- When doing a HT where $H_0 : p_1 = p_2$ (almost always for HT), use expected counts / proportions for S-F condition and calculation of the standard error.
- Otherwise use observed counts / proportions for S-F condition and calculation of the standard error.

For HT where $H_0 : p_1 = p_2$, pool!

As with working with a single proportion,

- When doing a HT where $H_0 : p_1 = p_2$ (almost always for HT), use expected counts / proportions for S-F condition and calculation of the standard error.
- Otherwise use observed counts / proportions for S-F condition and calculation of the standard error.

Expected proportion of success for both groups when $H_0 : p_1 = p_2$ is defined as the *pooled proportion*:

$$\hat{p}_{pool} = \frac{total\ successes}{total\ sample\ size} = \frac{suc_1 + suc_2}{n_1 + n_2}$$

Type	Parameter	Estimator	SE	Sampling Dist.
One mean	μ	$\bar{x}$	$s/\sqrt{n}$	t_{n-1}
Two means Paired data	μ_{diff}	$\bar{x}_{diff}$	$s_d/\sqrt{n}$	t_{n-1}
Two means Independent	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	t_{df} for df use $\min\{n_1 - 1, n_2 - 1\}$
One prop	p	$\hat{p}$	**C.I.** $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ **H.T.** $\sqrt{\frac{p_0(1-p_0)}{n}}$	Z
Two prop	$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	**C.I.** $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ **H.T.** $\sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_1} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{n_2}}$	Z

Categorical data with more than 2 levels $\rightarrow \chi^2$

Categorical data with more than 2 levels $\rightarrow \chi^2$

- one variable: χ^2 *test of goodness of fit*, no CI
- two variables: χ^2 *test of independence*, no CI

Categorical data with more than 2 levels $\rightarrow \chi^2$

- one variable: χ^2 *test of goodness of fit*, no CI
- two variables: χ^2 *test of independence*, no CI

Conditions for χ^2 testing

1. *Independence:* In addition to what we previously discussed for independence, each case that contributes a count to the table must be independent of all the other cases in the table.
2. *Sample size / distribution:* Each cell must have at least 5 *expected* cases.

χ^2 statistic: When dealing with counts and investigating how far the observed counts are from the expected counts, we use a new test statistic called the *chi-square (χ^2) statistic*:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$ where k = total number of cells

Important points:

- Use **counts** (not **proportions**) in the calculation of the text statistic, even though we're truly interested in the proportions for inference
- Expected counts are calculated assuming the null hypothesis is true

The χ^2 distribution has just one parameter, *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

- For χ^2 GOF test: $df = k - 1$
- For χ^2 independence test: $df = (R - 1) \times (C - 1)$

The χ^2 distribution has just one parameter, *degrees of freedom (df)*, which influences the shape, center, and spread of the distribution.

- For χ^2 GOF test: $df = k - 1$
- For χ^2 independence test: $df = (R - 1) \times (C - 1)$