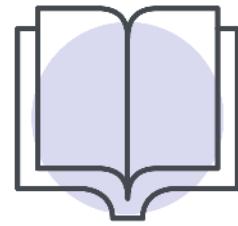

U of A

STAT151

FINAL EXAM

STUDY GUIDE



Textbook Notes

Chapter 19 - Notes (1 Proportion z-Intervals)

Standard error is just standard deviation that is estimated.

$$SE(\hat{p}) = \sqrt{\frac{(\hat{p})(1-\hat{p})}{n}} \quad \text{from } \beta$$

Therefore reaching out 2 SE's, we would be 95%. Sure that we have trapped the true proportion (p) even though we don't know its true value.

This sureness is called a confidence interval (C.I.). Formally we state: there is ____% confidence that the true proportion lies in our interval.

The extent of the interval on either side of \hat{p} is the margin of error (ME)

$$\text{Estimate} \pm \text{ME}$$

as the CI becomes larger, ie we become more confident we have captured the true proportion (p), the ME increases.

The # of SE's away from \hat{p} is also known as the critical value (z^*)

ex. z^* for 95% CI = 1.96 \therefore 95%

of a normal model is found within ± 1.96 SD's of the mean.

To select a sample size the equation is:

$$ME = z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \boxed{\hat{p} \text{ if unknown is 0.5}}$$

SQUARE Root n!

Example:

$$CI = 95\%$$

$$ME = 0.001$$

$$\hat{p} = 0.005$$

$$n = ?$$

$$ME = z^* \sqrt{\frac{p(1-p)}{n}}$$

$$0.001 = 1.96 \sqrt{\frac{0.005(0.995)}{n}}$$

$$0.001^2 = 1.96^2 \left[(0.005)(0.995) \right] / n$$

$$n = 1.96^2 \left[(0.005)(0.995) \right] / 0.001^2$$

$$n = 19.11196 \approx 19.112$$

To Construct a typical confidence interval we use

$$CI: \text{point estimate} \pm (\text{critical value}) \times (SE)$$

Textbook Summary on pg. 522

Chapter 19 - Questions

1. The $ME = \pm 4\%$. This means that in repeated random sampling at the same confidence level the estimates would come within 0.04 of the true proportion.
2. With an ME of 3% means that in repeated random sampling at the same CI the estimates of children who came back expected to 'bad taste' would be within 0.03 of the true proportion.

Chapter 20 - notes (Hypothesis Testing)

The null hypothesis, H_0 , specifies a population model parameter of interest and proposes a value for that parameter the typical form of H_0 :

$$H_0: \text{parameter} = \text{hypothesized value.}$$

The alternative hypothesis, H_A , contains the value of the parameter that we consider plausible if we reject H_0 . typical form of H_A :

$$H_A: \text{parameter} >< \text{hypothesised value.}$$

The Standard Deviation can be found using:

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

we know the \hat{p} value unlike chapter 1a so it is SD not SE b/c no estimation involved!

The next step is to find out how likely it would be to see the observed rate so, we use the Normal model if all assumptions are satisfied and find a Z-score.

$$Z = \frac{y - \mu}{\sigma}$$

P = probability
P-value = Prob of H_0 being true
 p = proportion parameter
 \hat{p} = observed proportion

The P-value is the probability of seeing data like this given that H_0 is true. When the P-value is high we haven't seen anything surprising and we have no reason to reject H_0 . When the p-value is low we have little evidence that H_0 is true therefore we ~~fail to~~ reject H_0 .

STEPS in Hypothesis Analysis:

1. Assumptions

- State and accept.

		2 tail		upper	lower
H_0		=	= or \geq	= or \leq	
H_A		\neq	<	>	
Reject areas		both tails	left tail	right tail	

2. Hypotheses

H_0 : parameter = hypothesized value
 H_A : parameter $<$ or $>$ hypothesized value.
 \neq "

3. Test Statistic

$$Z = \frac{Y - \mu}{\sigma} \text{ AKA } Z = \frac{\hat{P} - p_0}{SD(\hat{P})}$$

4. p-value

exact value or a range.

$$SD(\hat{P}) = \sqrt{p_0(1-p_0)}$$

5. Conclusions

given alpha \uparrow SLA: P-value $\leq \alpha \therefore$ reject H_0 ^{is significant evidence}
 $P\text{-value} \geq \alpha \therefore$ fail to reject H_0

at given alpha \uparrow JA: $0 < \text{Pvalue} \leq 0.01$ convincing to strong against H_0
 \uparrow $0.01 < \text{Pvalue} \leq 0.05$ strong to moderate against H_0
 \uparrow $0.05 < \text{Pvalue} \leq 0.1$ moderate to suggestive but inconclusive
 \uparrow $0.1 < \text{Pvalue} \leq 1$ weak evidence

Hilary

Chapter 21 - Notes (More About Tests).

A P-value is actually a conditional probability. It tells us the probability of getting results at least as unusual as the observed statistic, GIVEN that the null hypothesis is true.

P value = $P(\text{obs statistic value (or more extreme)} \mid H_0)$ NOT the probability that H_0 is true.

The lower the P-value the more comfortable we are with rejecting H_0 . The higher the P-value the more likely we are to fail to reject H_0 .

Example P-value = 0.03 : Given the null hypothesis, there is a 3% chance of observing the statistic value we actually observed.

When our P-value is small, it tells us that our data are "rare", given the null hypothesis. A rare event is defined as setting a threshold for our P-value. If our P-value falls below this then we'll reject H_0 . These results would be called statistically significant, and the threshold is called an alpha level:

$$\text{P-value} \leq \alpha \rightarrow \text{reject } H_0$$

$$\text{P-value} \geq \alpha \rightarrow \text{fail to reject } H_0.$$

The alpha level is also known as the test's significance level. We say that the test is "significant at that level".

When using CI keep in mind we CANNOT REJECT H_0 if the observed value is in the interval.

No test is perfect and sometimes errors are made, for H_0 testing we have 2 very specific types:

Type I error = H_0 is true, but we mistakenly reject it.

OR
Type II error - H_0 is false, but we fail to reject it.

Since we always start by assuming H_0 is true it makes sense that type I has a true H_0 .

When we choose an α value we set the $P(\text{Type I error}) = \alpha$. We also have β which equals:

$$1 - \alpha = \beta$$

So we could reduce β aka type II error by increasing α .

The Power of a test is the probability that it correctly rejects a false H_0 .

Power ↑ when Significance ↑

The Effect Size is defined as the distance between the H_0 value, p_0 , and the truth, p .

Chapter 21 - Questions

1. a) $H_0: p = .50$ $H_A: p \neq .50$
2 sided.

b) $H_0: p = .50$ $H_A: p > .50$
1 sided. rejection in right tail

c) $H_0: p = .67$ $H_A: p > .67$
1 sided. rejection in right tail

d) $H_0: p = .50$ $H_A: p \neq .50$
2 sided.

3. P-value = 0.047 ∵ Given the null hypothesis there is a 4.7% chance of observing the statistical value we observed. That would indicate no change, 95.3% is due to natural sampling distribution.

5. Same decision crit $.10 = \alpha$ but different at $\alpha = .01$

7. a) The p-value = 0.011 mean given the null hypothesis there is a 1.1% chance of observing the statistical value observed ∵ Reject H_0 b/c $P\text{value} \leq \alpha = .10$

Hilary

Chapter 23 - Notes (Inferences about means)

Since we have the CLT and can assume approximate normality the SD of a sample mean is calculated by estimating the population parameter σ with s :

$$SE(\bar{y}) = \frac{s}{\sqrt{n}} * \text{square root} *$$

The t-models create a whole bunch of distributions that relate to one another that depend on parameter called degrees of freedom.

The standardized sample mean is found using:

$$t = \frac{\bar{y} - \mu_0}{SE(\bar{y})} \quad \text{where } SE(\bar{y}) = \frac{s}{\sqrt{n}}$$

with $n-1$ degrees of freedom.

The confidence interval of the one-sample t-interval for the mean, μ , is:

$$\bar{y} \pm t_{n-1}^* \times SE(\bar{y})$$

where $SE(\bar{y}) = \frac{s}{\sqrt{n}}$

The one sample t-test for the mean is just like the other tests we have

Hypo

learned, using Student's t with $n-1$ degrees of freedom.

When finding t values with $1-n$ degrees of freedom their may not be exact entries so get as close as possible.

Important to remember that statistically important doesn't actually mean it needs attention

A level C confidence interval contains ALL plausible null hypothesis values that wouldn't be rejected by a 2 tailed test at $+ - C$

How big of a sample size is needed? Well we can calculate that using:

$$ME = t_{n-1} \left(\frac{s}{\sqrt{n}} \right) * \text{square root } n *$$

Chapter 24 - Notes (Comparing Means)

unequal variance

We know for independent random samples the variance of their difference is the sum of their variances $\text{Var}(\bar{Y} - \bar{X}) = \text{Var}\bar{Y} + \text{Var}\bar{X}$:

$$\begin{aligned} \text{SD}(\bar{Y}_1 - \bar{Y}_2) &= \sqrt{\text{Var}(\bar{Y}_1) + \text{Var}(\bar{Y}_2)} \\ &= \sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2} \\ &= \sqrt{\frac{\sigma_{\bar{Y}_1}^2}{n_1} + \frac{\sigma_{\bar{Y}_2}^2}{n_2}} \end{aligned}$$

Of course since we don't know the true standard deviations we are using s as an estimate and are finding SE:

$$\text{SE}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{s_{\bar{Y}_1}^2}{n_1} + \frac{s_{\bar{Y}_2}^2}{n_2}}$$

The two sample t-interval (difference of means) is:

$$(\bar{Y}_1 - \bar{Y}_2) \pm \text{ME}$$

$$\text{where } \text{ME} = t^* \times \text{SE}(\bar{Y}_1 - \bar{Y}_2)$$

So overall we see:

$$df = (\min\{n_1 - 1, n_2 - 1\}) \quad (\bar{Y}_1 - \bar{Y}_2) \pm t^* \times \sqrt{\frac{s_{\bar{Y}_1}^2}{n_1} + \frac{s_{\bar{Y}_2}^2}{n_2}}$$

↑
choose
smallest
result.

Hilary

When the assumptions are met
standardized value is:

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)}$$

Where $SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

The two sample t-test for the difference between means has:

$$H_0: \mu_1 - \mu_2 = \Delta_0 \leftarrow \text{usually } 0$$

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}$$

Pooled t test for the difference of means comes into play if we are willing to assume equal variances.

The similar spread condition comes into play for pooling data and can be checked by looking at the box plots.

The common variance is found:

$$S^2_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Chapter 26 - Notes (Comparing Counts)

The goodness-of-fit test involves testing hypotheses to see if they "fit" our distribution. This is used when we have 1 variable but 3 or more levels with categorical variables.

Test Statistic / Goodness of fit / Chi squared / $\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$

The χ^2 distribution is:

- 1) only one parameter, $df = \text{any } (+) \text{ integer}$
- 2) small $df = \text{skewed right}$ but as $df \uparrow$ becomes more symmetric

$$df = \# \text{ categories} - 1$$
$$= k - 1$$

P-values taken from χ^2 distribution table.

When conditions hold

- $$H_0: \text{All proportions are equal}$$
- $$H_A: \text{At least one } p_i \text{ isn't as claimed}$$

Chi squared Calculations are carried out by :

- 1) Find Exp values. H_0 value / # observations
- 2) Find residuals. Observed - expected
- 3) Square residuals.
- 4) Calculate the components of each cell using $\frac{(Obs - Exp)^2}{Exp}$ or $\frac{(residual)^2}{Exp}$.
- 5) Find the sum of all components.
This is $\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$
- 6) Find df. # of categories - 1
- 7) Test hypothesis.

Test of homogeneity is basically the χ^2 test for 2 proportions. The obs are presented in a 2 way frequency table w marginal totals at the end of each row and at the bottom of each column. A grand total is also shown.

The test of homogeneity tests if category proportions are the same for all populations by:

$$\frac{\text{row marginal total} \times \text{Column}^{\text{marg}}}{\text{grand total}}$$

where the rows are our samples and columns are variables.

The df is found using :

$$df = (\text{Row} - 1)(\text{Column} - 1)$$

A test of homogeneity has several samples, a single variable and several categories. If assumptions hold then:

H_0 : homogeneity of proportions

H_A : Some absence of homogeneity

If H_0 is REJECTED we want to check residuals. First we have to standardize them using:

$$c = \frac{(Obs - Exp)}{\sqrt{Exp}} \quad \text{or} \quad \sqrt{\text{components}}$$

The Chi test for independence does just that, tests for independence. If assumptions hold:

H_0 : variables are independent

H_A : variables are dependent.

The test stat and df are the same as the test of homogeneity but in this H_0 testing NOT REJECTING H_0 is favorable

Chapter 26 - Summary

Goodness of Fit test (Chi Squared):

- 1) Single Sample
- 2) Single Variable
- 3) Several Categories.

Chi Squared Test of Homogeneity:

- 1) Several Samples
- 2) Single Variable
- 3) Several Categories

Chi Squared Test of Independence:

- 1) Single Sample
- 2) 2 Variables
- 3) Several Categories.

Hilary

Chapter 28 - Notes (Analysis of Variance).

Looking at multiple groups we have to analyze their box plots to find if all means are equal.

For this chapter:

H_0 : all means are equal

H_a : at least one mean is unequal.
OR

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$

H_a : $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu$.

We have two different estimates for the underlying variance in the groups:

1) MST (mean square treatment / between)
which is based on the differences between the groups.

a) MSE (mean square error / within)
is based only on the variation within the groups around each of their own means.

MST depends on the null hypothesis being true where MSE doesn't. When H_0 is true MST / MSE ratio is close to 1 and if H_0 is false usually the ratio will be > 1 .

The ratio we previously measured MST / MSE = the F-statistic. Using the F-stat with the appropriate F-distribution we can find a P-value. In practice big F Stats go with small P-values and visa versa.

This chapter also deals with 2 diff types of degrees of freedom. The MST deals with variance between groups so its $df = k - 1$ or groups - 1. The MSE is variation within the groups so its $df = n - 1$ for each group. That means #of groups ($n - 1$) or $k(n - 1)$ gives us the df of variation of the all the groups.

MST = Between
 $df = k - 1$

MSE = Within
 $df = \frac{k(n-1)}{N-k}$

The SS_T can be divided by $(k-1)$ to find MST . Also $SSE / (N - k)$ will find MSE .

Then to find total SS just add up SS_T and SSE . Total $df = (k-1) + (N-1)$

To find residual standard deviation simply square root MSE :

$$Sp = \sqrt{MSE} = \sqrt{\frac{SSE}{N-K}}$$

Hilroy

Confidence intervals can be found by:

$$ME = t^* \times Sp \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$



Exam Notes

Mode: the value that occurs most in a sample.

Median: the center of the data (literally).

The probability of an outcome is the proportion of times it occurs in the long run. The probability of each outcome must fall between 0 and 1. The total of all the probabilities of the possible outcomes equals 1.

For an event A, its complement consists of all outcomes not in that event. The probability of the complement of A relates to the probability of A by $P(AC) = 1 - P(A)$.

- An event and its complement are disjoint events, because they have no common elements.

Independent events: one does not depend on the other

- When A and B are independent $P(A \text{ and } B) = P(A) \times P(B)$.

Not Independent: rely on one another

- When A and B are not independent, $P(A \text{ and } B) = P(A) \times P(B | A) = P(B) \times P(A | B)$.

Whether or not A and B are independent

- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$.

Conditional probability: The P of an event given the P of another event
The conditional probability of the event A, given that event B has occurred is denoted $P(A | B)$;

- it can be calculated by $P(A|B) = P(A \text{ and } B) / P(B)$

A **probability distribution** specifies the probabilities for all the possible values of a random variable.

- The distribution is summarized by a mean μ and a standard deviation σ to describe its center and spread.

The **probability distribution of a continuous random variable** is described by a curve which has total area 1 under the curve and above the x-axis.

The **normal distribution** is the probability distribution of a continuous random variable that has a certain symmetric bell-shaped graph. For any z, the probability falling within z standard deviations of the mean is the same for every normal distribution.

A **z-score** tells us the number of standard deviations that an observation falls from the mean. For a value x from a probability distribution with mean μ and a standard deviation σ ,

- the z-score is $z = \bar{x} - \mu / \sigma$

A **sampling distribution** is the probability distribution of a sample statistic such as \bar{y} or p . It is the distribution we would obtain if we repeatedly took random samples of a fixed size n , calculated the statistic each time, and formed a histogram of the values of the statistic.

The sampling distribution for p with sample size n from a population with population proportion p has mean $\mu_p = p$ and standard error $\sigma_p = \sqrt{p(1-p) / n}$

Normality conditions depend on instructor.

The **sampling distribution** for \bar{y} with sample size n from a population with mean μ and standard deviation σ has mean $\mu_{\bar{y}} = \mu$ and standard error $\sigma_{\bar{y}} = \sigma / \sqrt{n}$

When the population distribution is normal, the sampling distribution of \bar{Y} is also normal for any sample size n .

The **Central Limit Theorem** states that for relatively large random samples ($n \geq 30$), the sampling distribution for \bar{y} is approximately normal, even if the population distribution is not.