Stats 250 Final Exam

Introduction

This summary is not meant to be comprehensive. By using this document, you acknowledge you have read the disclaimer.

Intrepetations Look back at Exam One and Exam Two for the interpretations of things like p-values and other vocabulary. Remember to be concise when presenting these on an exam - sometimes saying more than necessary can hurt more than it helps.

The Formula Card Knowing how to quickly utilize the formula card is helpful for many of the problems you may encounter. The notation on the formula card is correct, and a good habit is to always check the formula card whenever you feel you have hit a dead-end. Often, a formula or hint on the card can provide the next step to finish the problem.

Name that scenerio Being able to identify which test you are looking at quickly and accurately is invaluable. The assumptions, interpretations, distributions, and conclusions all hinge on selecting the correct test, so be sure to study name-that-scenario questions

Broad Concepts The final will test your understanding of broad concepts in this course. Make sure you have a grasp of the big picture as well as the methodology behind each test.

Difference in Population Means

This test compares the means of two independent populations. We say that this is a "difference in population means" and not a "population mean difference". When you notice that you are comparing two population means and the subgroups have a different size, this is a good indication that you will be using this test and not a paired test. However, this doesn't imply that if the sample sizes are the same that it is automatically a paired test. Make sure to do lots of name-that-scenario to get a good understanding of this distinction.

Assumptions:

- $H_0 : \mu_1 = \mu_2$
- $H_a : \mu_1 - \mu_2 \neq 0$
- Random Samples from each population are independent of each other
- Each population is normally distributed (check with TWO QQ-plots)
- Pooled Test
  - Equal Population Variances (checking below)
- Unpooled Test
  - Unequal Population Variances (checking below)

QuickFacts:

Distribution of Test Statistic t_{n-1}

Common Mistakes Not checking variance assumptions, assuming this is a paired test, error using the wrong t-value from the t-table, not using $n-1$ degrees of freedom in calculations.

Do we use pooled or unpooled variance?:

- Population variances equal $\implies$ pooled
- Population variances unequal $\implies$ unpooled
- There are three ways of checking equality in variance:

1. Side-by-Side boxplots
   - Are IQRs of the sample data similar?
   - Remember, boxes don't need to line up, but lengths of boxes should be similar (within two times the size of each other)
2. Check on Sample Standard Deviations
   - Assumption is valid if sample standard deviations are similar. Each standard deviation is no more than 2x the other.
3. **Levene's test** for homogeneity of variances
   - Hypothesis test of equal variances. Use $\alpha = \mathbf{0.10}$
   - $H_0 : \sigma_1^2 = \sigma_2^2$ v. $H_a : \sigma_1^2 \neq \sigma_2^2$
   - Reject $H_0 \implies$ not reasonable to assume equal variances. You must run an unpooled test (Welch's)
   - Fail to reject $\implies$ reasonable to assume equal variances. You can run a pooled test

Example:

Is the mean number of people who progress to AIDS each year in Tanzania different from the mean number of people who progress to AIDS each year in South Africa over the past 10 years?

- There are two independent populations
  - Population 1: People who progress to AIDS in *Tanzania*
  - Population 2: People who progress to AIDS in *South Africa*
- We are comparing the *means* of two populations where
  - μ_1 = population mean number of people who progress to AIDS in Tanzania
  - μ_2 = population mean number of people who progress to AIDS in South Africa

Example:

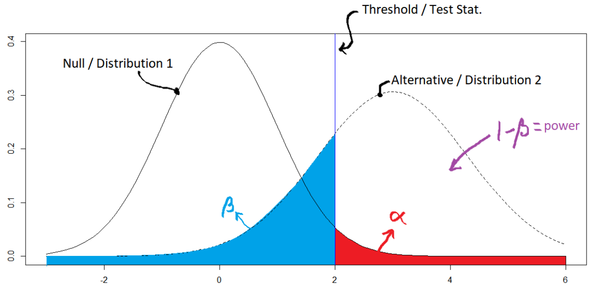
A researcher is curious to find out if study time is higher for older students (group 1) versus younger students (group 2). A sample of 47 older students and 45 younger students was taken. A pooled independent t-test was run and the difference in sample mean, $\bar{X}_1 - \bar{X}_2$ was computed. This difference is **3 standard errors** above the hypothesized difference of 0 for $\mu_1 - \mu_2$.

$H_0 : \quad \mu_1 - \mu_2 = 0$
$H_a : \quad \mu_1 - \mu_2 > 0$
Test statistic value: $t = 3$
Pooled t-test df: $n_1 + n_2 - 2 = 45 + 47 - 2 = 90$

Determination of α and β (Type I & Type II Error)

You may get questions on determining α and β (probability of a type I or type II error), or you may be given two different distributions (think about the homework problem where we had two farmers and we were trying to decide whether a pumpkin belonged to Farmer A or Farmer B).

Always begin by drawing the null and the alternative distributions. They may look something like this



Note: *They do NOT need to be normally distributed - they can be normal, uniform, or any other distribution*: Most problems will have a decision criteria. They will say, "Greater than value X we choose the alternative, else we choose the null". This X is the vertical line in the distribution drawn. From here, just think about what each question is asking and how to go about finding the p-values that they are wondering about.

Linear Regression

Linear regression is all about finding relationships between a response Y and a predictor X.

Assumptions:

- There is a linear relationship between the population of responses and the population of the predictor
- True Errors are Normally Distributed
- True Errors have a constant variance (homoscedasticity)

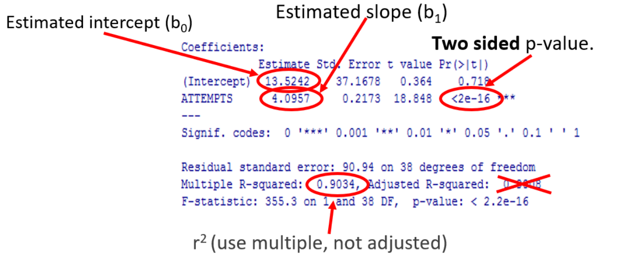
QuickFacts:

When we **extrapolate**, we make a prediction of y for an x that is outside of the range that we fit the data on. Imagine that we are predicting the blood pressure of adult males aged 18-40 using a model. We recruit males aged 18 to 40 and then fit our model, but then we are asked to run this model on a 60 year old male or on a female. These points are outside of our sample, and thus we need to be very careful with predictions from our model.

Distribution of Test Statistic If the test is a two sided test, we can use either the T Test Statistic on the $b_1$coefficient or the F Test Statistic at the bottom of the test. If it is one-sided, we have to use the T Test Statistic. The F-statistic is related to the T Test Statistic by: $F = t^2$

Common Mistakes Always be sure to be on the lookout for questions that are testing that you know that *correlation does not imply causation*. Also, always try to write out the form of the regression equation in words - this tends to help out a lot. It is often better to write $Final = b_0 + b_1 \times Exam2$ than $\hat{y} = b_0 + b_1 x$ to keep them straight on the exam.

Interpreting the R Output and Plots:


Estimated intercept (b_0)
Estimated slope (b_1)
Two sided p-value.
r^2 (use multiple, not adjusted)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.5242    37.1678   0.364    0.718
ATTEMPTS      4.0957     0.2173  18.848   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 90.94 on 38 degrees of freedom
Multiple R-squared:  0.9034,	Adjusted R-squared:  ~~0.9008~~
F-statistic: 355.3 on 1 and 38 DF,  p-value: < 2.2e-16
```

Estimated equation ($\hat{y} = b_0 + b_1x$): $\hat{y} = 13.5242 + 4.0957x$

In this class, our interpretation of β_1 is the true change in mean for the response variable for every additional unit increase in the explanatory variable. R^2 is the percent of the variation in the response variable explained by the linear relationship between response and predictor

- A scatter plot can tell us a lot about the relationship between two variables
- We can view a scatterplot to look for:

1. Shape/form
   - Linear or non-linear relationship (does it look like a cup or a rainbow?)
2. Direction
   - Positive or negative correlation
3. Strength of relationship
   - What's the magnitude of the correlation?
4. Potential Outliers
   - Be sure to comment on outliers if you feel there are some

ANOVA

Performed when we have k *independent* populations, and we are trying to test if any of their means differ from one another. This should give you the exact same result as the Tukey test, but it will only provide one p-value instead of each of the pairwise confidence intervals. Note: This test uses an $F_{k-1,N-k}$ distribution, whereas the Tukey test does not (it uses a special distribution similar to a t-distribution). Always remember the Golden Rule: *Add Down → Divide Across*

Hypothesis & Assumptions:

- $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
- H_a : at least one of the μ_i is different from the others
- Each sample is independent of one another
- Each *population* is normally distributed (check with k QQplots)
- All population variances have to be equal

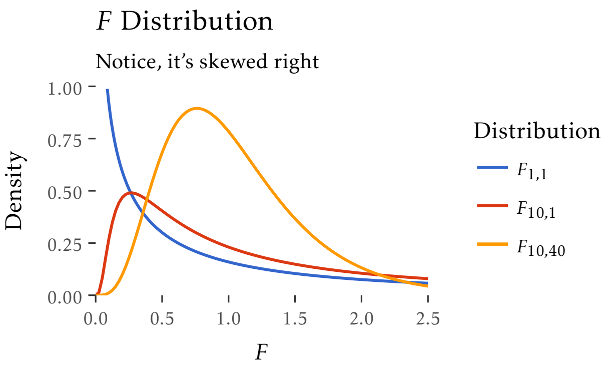
QuickFacts:

Distribution of Test Statistic $F_{k-1,N-k}$ Remember - there are two separate degrees of freedom!

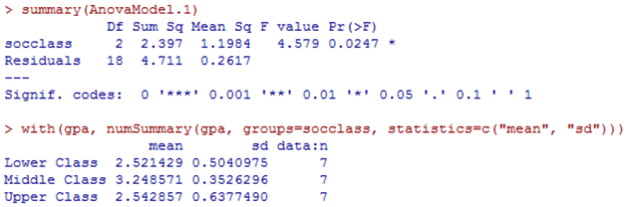
$$F = \frac{\text{Variation among sample means}}{\text{Variation within groups}} = \frac{\text{MS Groups}}{\text{MSE}}$$

If the null hypothesis is true, then we expect an F-statistic of approximately one. The larger the F-statistic, the more evidence we have against the hypothesis that all of the means are zero.

Common Mistakes Not checking the variance is equal, not remembering the relationships between MSE and the standard deviation of the population s_p, table manipulation errors (make sure you know how to compute the table! - see example below).



Example:

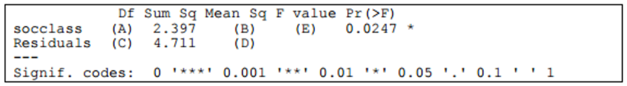


```
> summary(AnovaModel.1)
            Df Sum Sq Mean Sq F value Pr(>F)
socclass     2  2.397  1.1984   4.579 0.0247 *
Residuals   18  4.711  0.2617
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(gpa, numSummary(gpa, groups=socclass, statistics=c("mean", "sd")))
                 mean        sd data:n
Lower Class  2.521429 0.5040975      7
Middle Class 3.248571 0.3526296      7
Upper Class  2.542857 0.6377490      7
```

- MSE is 0.2617
- MS Groups is 1.1984
- $F = \frac{1.1984}{0.2617} = 4.579$
- $d_1 = k - 1 = 3 - 1 = 2$
- $d_2 = N - k = N - 3 = 18$
- P-value is 0.0247

Example:



```
             Df  Sum Sq Mean Sq F value Pr(>F)
socclass    (A)  2.397    (B)     (E)    0.0247 *
Residuals   (C)  4.711    (D)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. We know there are $k = 3$ groups, so the numerator degrees of freedom (A) must be: 2
2. We divide SS Groups: 2.397 by the numerator df (A) 2 to obtain the MS Groups (B): 1.1984
3. We know there are $N = 21$ total observations, so the denominator df (C) must be: $N - k = 18$
4. We divide the SS Error by the denominator df (C) to obtain the MS Error (D) of: 0.2617
5. We divide the MS Groups (B): 1.1984 by the MS Error (D) 0.2617 to get the F statistic: 4.57

Chi-Squared

There are three different types of Chi-Squared tests.

- Test of Independence
- Test of Homogeneity
- Goodness of Fit

Each test uses the same test statistic given by:

$$\chi^2 = \sum_{i=1}^{n} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \;\Big|\; \text{Contribution} = \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

The further away our observation is away from what we would expect, the larger the test statistic.

Discerning the Three Types of Tests:

Goodness of Fit This test answers the question, “Do the data fit well compared to a specified distribution?” It considers one categorical response, and assesses whether the proportion of sampled observations falling into each category matches well to a specified distribution. The null hypothesis specifies this distribution which describes the population proportion of observations in each category. Ex: Imagine that one grabs a bag of MM's. MM's lists the color distribution as: According to it, each package of Milk Chocolate M&M's should contain 24% blue, 14% brown, 16% green, 20% orange, 13% red, and 14% yellow M&M's. We could take a few packs of MM's and compare our observed percentages to what we would expect to see if we are getting results consistent with MM's website.

Test of Homogeneity This test answers the question, “Do two or more populations have the same distribution for one categorical variable?” It considers one categorical response, and assesses whether the model for this response is the same in two (or more) populations. The null hypothesis is that the distribution of the categorical variable is the same for the two (or more) populations. Ex. I have five TV shows and I am trying to see if each of the proportions of men and women are the same across the five shows.

Test of Independence This test answers the question, “Are two factors (or variables) independent for a population under study?” It considers two categorical variables (sometimes one is a response and the other is explanatory), and assesses whether there appears to be a relationship between these two variables for a single population. The null hypothesis is that the two categorical variables are independent (not related) for the population of interest. Ex: Imagine that I was trying to determine if someones school year (Freshman, Sophomore, Junior, Senior), is independent of where they live on campus (in the dorms, off campus, at home, fraternity / sorority)

Variables and Populations:

Test	Number of Variables	Number of Populations
Goodness of Fit	1	1
Homogeneity	1	2+
Independence	2	1

Null Hypothesis:

Goodness of Fit $H_0 : p_1 = \langle PopulationProp_1\rangle$, $p_2 = \langle PopulationProp_2\rangle, \ldots, p_k = \langle PopulationProp_k\rangle$.

Homogeneity H_0 : The distribution of $\langle VariableNames\rangle$ is the same for populations of $\langle Namesof thePopulations\rangle$

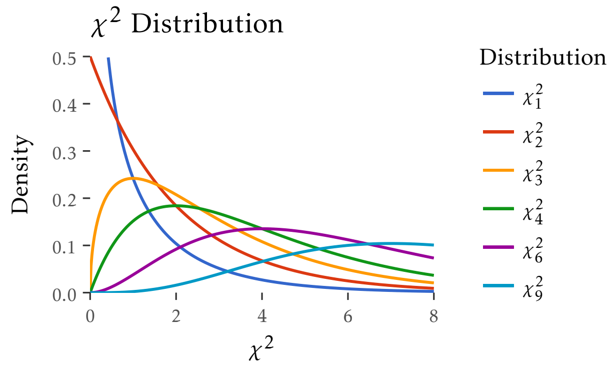
Independence H_0 : There is no relationship between $\langle Variable1\rangle$ and $\langle Variable2\rangle$

Independence vs. Homogeneity:

You'll see that the test for independence and homogeneity are easily confused. Some quick tips:

- Check the sampling method and the problem setup
- In tests for Homogeneity, we have two or more populations and only one variable. In independence testing we have one population and more than one variable.
- In tests for Homogeneity, we have two or more samples. In independence tests, we have one sample and record two variables
- In tests for Homogeneity, the question is asking if the populations are *different*. In tests for independence, we are asking if the *variables* are *independent* or *associated*
- What if...
  - I took a random sample of wedding guests and recorded if they were from the bride's side or the groom's side and what type of meal they wanted. (Independence)
  - I took a random sample of wedding guests from the bride's side and a second random sample from the groom's side and recorded what type of meal they wanted. (Homogeneity)

The Chi-Squared Distribution:



Disclaimer

Mark has made every attempt to ensure the accuracy and reliability of the information provided in this study material. However, the information is provided ”as is” without warranty of any kind. Mark does not accept any responsibility or liability for the accuracy, content, completeness, legality, or reliability of the information contained in this study material. No warranties, promises and/or representations of any kind, expressed or implied, are given as to the nature, standard, accuracy, content, completeness, reliability, or otherwise of the information provided in this study material nor to the suitability or otherwise of the information to Stats 250.

Version: 1.618
Please email typos to mtkurzej@umich.edu
Copyright Mark Kurzeja 2019