Session 7

Administrivia

- Course recordings
- Final examination logistics:
  - December 16, 4:30pm-7pm Eastern time
  - Makeup: December 17, 4:30am-7am Eastern time
    TOTALLY different examination
  - **<u>If you are taking the MAKEUP examination, the absolute deadline for informing me is THIS FRIDAY NOON (Eastern Time)</u>**
- Final examination: Structure
  - There *may* be some computation involved (e.g., via Excel), but you will not be required to run an R script
  - Two numbers worth memorizing: 1.96 and 1.645
  - If you need to make an assumption in order to answer a question, state the assumption you need to make, as part of your response to the question
- Di will be holding VOH as usual up through December 15th
- No further quizzes
- I will be available to meet individually with class members upon request

Quiz 5

Suppose you wish to test the alternative hypothesis that the population proportion is different than 0.50. You draw a random sample of 100 people, for which the sample proportion is .55.

What is the alternative hypothesis (in words)?
Please provide a value for the standard error:
Using an α level of .05, please provide a value for z_{crit} (the critical value) accurate to two places to the right of the decimal point:
Based on a z-value of 1.0, and using an α level of .05, would you:

a. Reject the null hypothesis (Yes/No): If "No", why not?

b. Fail to reject the null hypothesis (Yes/No): If "No", why not?

c. Accept the null hypothesis (Yes/No): If "No", why not?

d. Fail to accept the null hypothesis (Yes/No): If "No", why not?

e. Reject the alternative hypothesis (Yes/No): If "No", why not?

f. Fail to reject the alternative hypothesis (Yes/No): If "No", why not?

g. Accept the alternative hypothesis (Yes/No): If "No", why not?

h. Fail to accept the alternative hypothesis (Yes/No): If "No", why not?

Assignment 5

You have been asked to analyze the data from a recent poll of "fast casual" restaurant-goers. The data for this study comprise a random sample from the population of all patrons of "fast-casual" restaurant-goers, and are contained in file "Restaurant.txt", which has a header line and contains two variables: Restaurant Preference ("Pref": B=Burger, Tap, & Shake; C=Cava; P=Panera Bread; R=Roti; S=Shake Shack); Age Group of the respondent ("Agegroup": 1=18-27, 2=28-37, 3=38-47, 4=48-57, 5=Over 57). Note that Agegroup is a categorical variable: do not treat this variable as continuous.

1. **Based on other media reports, it is hypothesized that, among those between 28 and 37, more than 20% prefer Burger, Tap, & Shake. What light can you shed on this hypothesis?**

2. **It is hypothesized that the proportion who prefer Cava among the agegroup 28-37 population is higher than the proportion who prefer Cava among the agegroup 38-47 population. Can we be reasonably sure that this is correct?**

3. **Question 3: Roti claims that the number of people in all agegroups combined who prefer Roti is more than 50% higher than the number of people in all agegroups combined who prefer Burger, Tap, & Shake. Can we be reasonably sure that this is correct? (NOTE: this question can, and must, be answered using a z-test of a single population proportion. One of your challenges is to recognize how the problem can be recast in that form.)**

Prelude to Today's Discussion

Inferential statistics consists of two subdomains: estimation and hypothesis testing;

The estimation subdomain consists of point estimation and interval estimation

While interval estimation can, in some cases, be used to test a hypothesis, this is not generally a good practice – especially in the context of directional alternative hypotheses

We learned how to calculate a confidence interval for a single population proportion and for the difference in two population proportions

We learned how to test a hypothesis about a population proportion (both non-directional and directional) and about a difference in two population proportions

The advantage of focusing on the alternative hypothesis when summarizing the results of a statistical test

Which types of statements are (and are NOT) defensible in reaching a conclusion about a statistical test

Today, we will learn about:

Inferences about the difference in two population proportions when the data are paired

Three forms of the chi-square test: the Goodness of Fit Test, the Test of Homogeneity, and the Test of Independence, as well as about the logic which underlies the chi-square test

The assumptions associated with multinomial tests

(Provisional)A model which allows for the prediction of a *discrete* dependent variable (multiple discriminant analysis)

Homework: Which Statements Are Defensible?

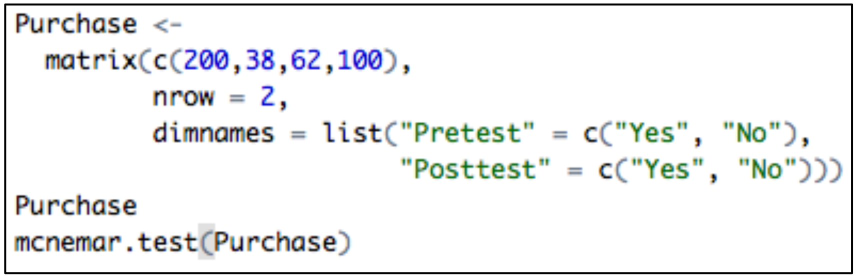
Consider the following attempts, all drawn from prior classes, to express in words the concept of "failing to reject the null hypothesis." Which ones are defensible?

1. **"I can be 95% confident that one supplier has a different proportion of defective batteries than the other"** ■
2. **"I can be 95% confident that the two suppliers have equal proportions of defective batteries"** ■
3. **"I can be 95% confident that neither supplier has a higher proportion of defective batteries than the other"** ■
4. **"The data are insufficient to allow me to say anything about the comparative proportions of defective batteries supplied by the two suppliers"** ■
5. **"I cannot be 95% confident that one supplier has a different proportion of defective batteries than the other"** ■
6. **"I cannot be 95% confident that the two suppliers have equal proportions of defective batteries"** ■
7. **"I can, with 95% confidence, fail to reject the null hypothesis"** ■
8. **"I can fail to (reject the null hypothesis with 95% confidence)"** ■
9. **"I can (fail to reject the null hypothesis) with 95% confidence"** ■
10. **"I can fail to reject the null hypothesis with 95% confidence"** ■

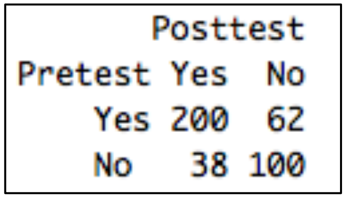
Inferences About the Difference in Two Population Proportions $\pi_1 - \pi_2$ When the Data are Paired

Your firm distributes romaine lettuce. There has recently been a recall of romaine lettuce that was distributed by a competitor. You create an advertisement that you think will convince people to purchase your lettuce, but you want to test it out to see if you are correct. You randomly select 400 people and ask them if they will or will not purchase your lettuce (Pretest). You then show them the advertisement, and ask again (Posttest). You use the McNemar test to determine whether there has been a change in the percentage of people who will purchase your product.

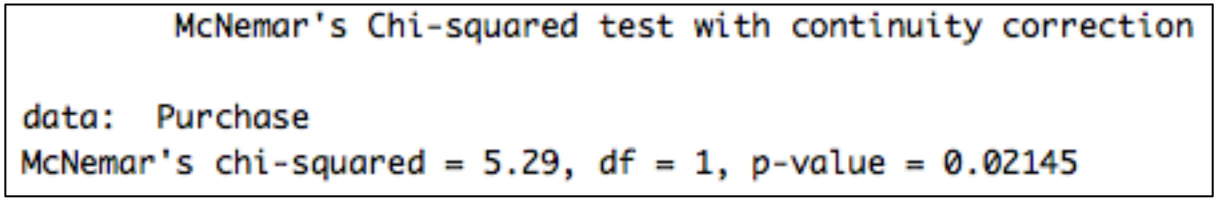
$H_0 : \pi_{pretest} = \pi_{posttest} \quad H_A : \pi_{pretest} \neq \pi_{posttest}$



```
Purchase <-
  matrix(c(200,38,62,100),
         nrow = 2,
         dimnames = list("Pretest" = c("Yes", "No"),
                         "Posttest" = c("Yes", "No")))
Purchase
mcnemar.test(Purchase)
```



```
       Posttest
Pretest Yes  No
    Yes 200  62
    No   38 100
```



```
        McNemar's Chi-squared test with continuity correction

data:  Purchase
McNemar's chi-squared = 5.29, df = 1, p-value = 0.02145
```

Inferences About Distributions: χ^2 Goodness of Fit Test

Iota Steel, Inc., wishes to know whether accidents at its largest plant are more common at certain times of the day. They tabulate last month's accidents by time period: 8am-10am (31 accidents), 10am-12pm (30 accidents) 1pm-3pm (41 accidents), 3pm-5pm (58 accidents). Can we be reasonably sure that accidents at the plant are NOT equally likely to occur across these four time periods?

8am-10am	10am-12pm	1pm-3pm	3pm-5pm
31	30	41	58

$$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$$

$$H_A : Not(\pi_1 = \pi_2 = \pi_3 = \pi_4)$$

$$O_1 = 31, O_2 = 30, O_3 = 41, O_4 = 58, O_{Total} = 160$$

$$E_1 = E_2 = E_3 = E_4 = 160 / 4 = 40$$

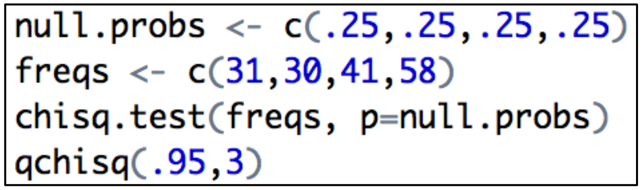
$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(31-40)^2}{40} + \ldots + \frac{(58-40)^2}{40} = 12.65$$

$$df = 4 - 1 = 3$$

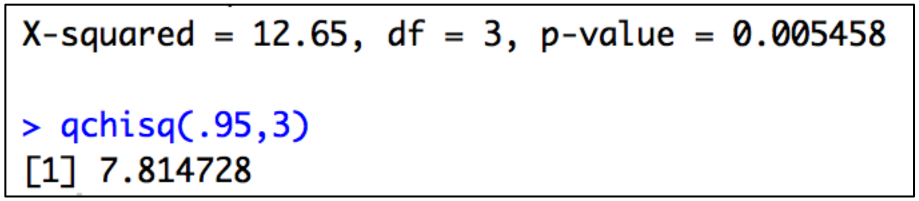
$$\chi^2_{crit} = 7.814728$$

$12.65 > 7.814728$ ==> Reject H_0, Accept H_A

$Also, p = 0.005458 < .05$ ==> Reject H_0, Accept H_A



```
null.probs <- c(.25,.25,.25,.25)
freqs <- c(31,30,41,58)
chisq.test(freqs, p=null.probs)
qchisq(.95,3)
```



```
X-squared = 12.65, df = 3, p-value = 0.005458

> qchisq(.95,3)
[1] 7.814728
```

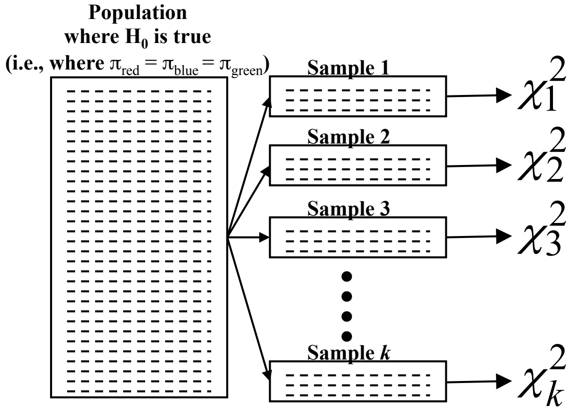
The Logic Underlying the Chi-Square Goodness-of-Fit Test

- Statistical theory tells us that if we drew a very large number of random samples from a population where the null hypothesis is true, and calculated the χ^2 value for each of those samples, then the χ^2 value would exceed 7.815 in only 5% of those samples.

- This means that a χ^2 value of 7.815 or larger is unlikely to occur if the null hypothesis is true.

- Thus, if the χ^2 value in *our* sample is greater than 7.815, we can conclude with reasonable certainty that the null hypothesis is not true (i.e., is false), since a χ^2 value as large or larger than the one we got in our sample would be an unlikely value if H_0 was true.

- Formally, "if H_0 was true, we would see a χ^2 value this larger or larger fewer than 5 samples out of 100. Since this is less than the alpha level set prior to our analysis (α=.05), we reject H_0 and accept H_a."



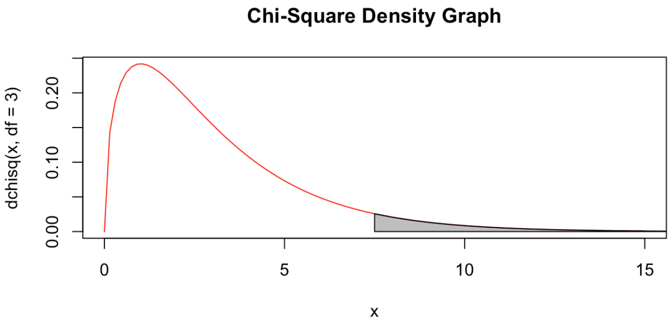
Population
where H_0 is true
(i.e., where $\pi_{red} = \pi_{blue} = \pi_{green}$)

Sample 1 → χ_1^2

Sample 2 → χ_2^2

Sample 3 → χ_3^2

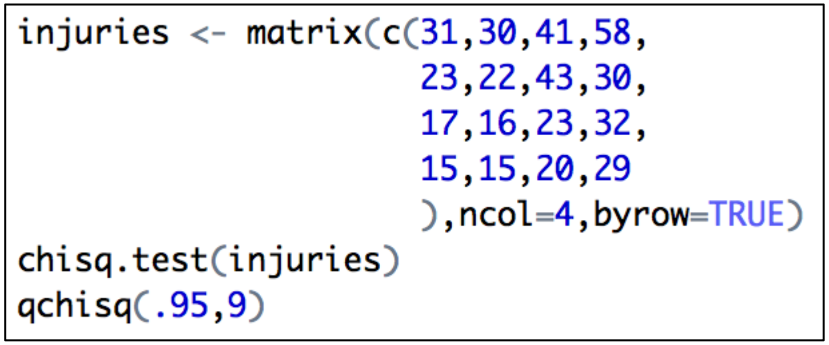
Sample *k* → χ_k^2



Inferences About Distributions: χ^2 Test of Homogeneity

Iota Steel, Inc., wishes to know whether the distribution of accidents by time of day differs across its four plants. They tabulate last month's accidents by time period at each plant. Can we be reasonably sure that the distribution of accidents by time of day is not identical across its four plants?

	8am-10am	10am-12pm	1pm-3pm	3pm-5pm
Plant 1	31	30	41	58
Plant 2	23	22	43	30
Plant 3	17	16	23	32
Plant 4	15	15	20	29



```
injuries <- matrix(c(31,30,41,58,
                     23,22,43,30,
                     17,16,23,32,
                     15,15,20,29
                     ),ncol=4,byrow=TRUE)
chisq.test(injuries)
qchisq(.95,9)
```

$$H_0 : \pi_{(i,j)|i} = \pi_{(k,j)|k} \qquad i \neq k,\ \ 1 \le i,j,k \le 4$$

$$H_A : Not(\pi_{(i,j)|i} = \pi_{(k,j)|k}) \quad i \neq k,\ \ 1 \le i,j,k \le 4$$

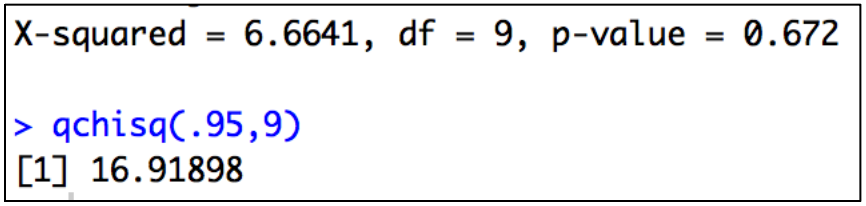
$$\chi^2 = \sum \frac{(O-E)^2}{E} = 6.6641$$

$$df = (4-1)*(4-1) = 9$$

$$\chi^2_{crit} = 16.91898$$

$6.6641 < 16.91898$ ==> Fail to reject H_0, Fail to accept H_A

$Also, p = 0.672 > .05$ ==> Fail to reject H_0, Fail to accept H_A



```
X-squared = 6.6641, df = 9, p-value = 0.672

> qchisq(.95,9)
[1] 16.91898
```

Independence *vs.* Homogeneity: The Difference

- ✧ With independence, you have one population and two categorical variables of interest (e.g., plant and time of injury)
- ✧ With homogeneity, you have *multiple* populations and one categorical variable of interest (e.g., time of injury)
- ✧ Difference lies largely with the way the sampling was done
- ✧ Both tests produce exactly the same result; thus, in practice the distinction is often irrelevant

Assumptions Associated With Multinomial Tests

- The probability that a single "trial" will result in outcome *i* is π_i, i=1, 2, ..., *k*, and remains constant from "trial" to "trial"
  - Tenability: Typically cannot be formally tested, and must be assumed
- The "trials" are independent
  - Tenability: Typically guaranteed by selection of a *random* sample

Violation: Data collection extends over a long period of time, and the probabilities change

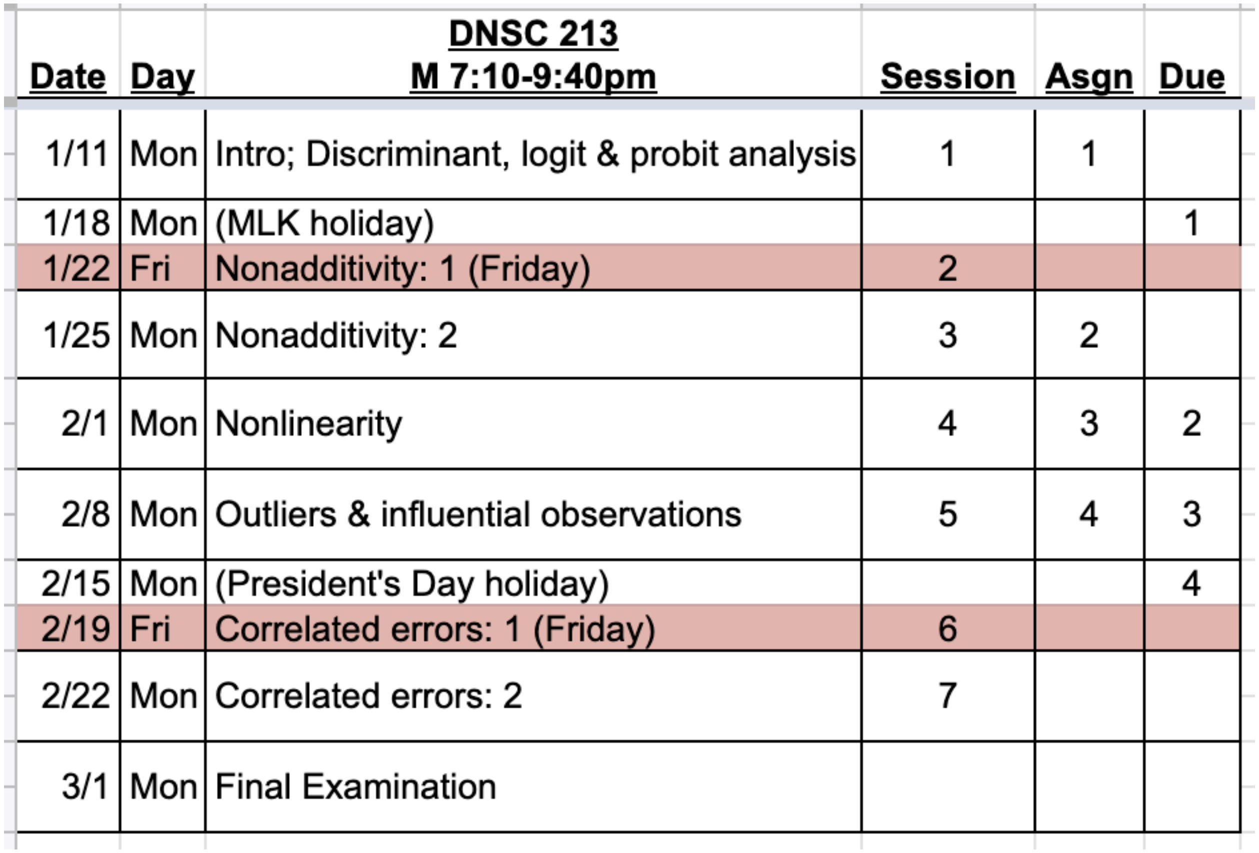
Violation: Data collection process includes each person as well as his/her "significant other"

Roadmap of DNSC 6203

1. **Modeling business analytics problems. Simple Linear Regression. Introduction to regression equations, dependent vs. independent variables, prediction, expected values, Coefficients of Determination. Samples vs populations, null vs. alternative hypotheses, sampling distributions, statistical tests, degrees of freedom, critical values, Type I vs. Type II errors, p-values. Statistical significance vs. strength of relationship measures.**
2. **Multiple regression Part I. Multicollinearity, statistical significance, coefficient of multiple determination, Global F, semi-partial effects, coefficient of partial determination.**
3. **Multiple regression Part II: "controlling for" (holding constant), correlation, t-test, Simpson's Paradox, suppressor effect, regression and "big data".**
4. **Discrete independent variables. Independent samples t-test. Analysis of Variance: I. One-way Analysis of Variance. A-posteriori tests (Scheffe). A-priori tests. F-test for ANOVA. Coefficient of multiple determination (eta-squared) and coefficients of partial determination in ANOVA. Confounding. Stevens' Levels of Measurement Taxonomy. Three primary forms of the General Linear Model.**
5. **Analysis of Variance: II. Two-Way ANOVA models. Analysis of covariance. Least squares means. Coefficient of Partial Determination. Eta squared.**
6. **Focus: count data. Review of inferential statistical reasoning: point estimation, confidence interval estimation, null hypotheses, alternative hypotheses, directional vs. non-directional alternative hypotheses. Confidence interval estimation: single population proportion, difference in two population proportions. Hypothesis testing: single population proportion, difference in two population proportions (directional, non-directional).**
7. **Chi-square Goodness of Fit test, test of homogeneity, and test of independence. Requisite assumptions for multinomial statistical tests. Discriminant analysis. Course Review.**

Pre-Exam Questions

A Preview of Things to Come (Provisional)



		DNSC 213			
Date	Day	M 7:10-9:40pm	Session	Asgn	Due
1/11	Mon	Intro; Discriminant, logit & probit analysis	1	1	
1/18	Mon	(MLK holiday)			1
1/22	Fri	Nonadditivity: 1 (Friday)	2		
1/25	Mon	Nonadditivity: 2	3	2	
2/1	Mon	Nonlinearity	4	3	2
2/8	Mon	Outliers & influential observations	5	4	3
2/15	Mon	(President's Day holiday)			4
2/19	Fri	Correlated errors: 1 (Friday)	6		
2/22	Mon	Correlated errors: 2	7		
3/1	Mon	Final Examination			

In Conclusion

Final examination:

- **Wed. December 16, 4:30pm-7pm Eastern time**
- **Makeup: December 17, 4:30am-7am Eastern time TOTALLY different examination**

I look forward to everyone earning an "A" on the Final Examination:

- **Di will be holding VOH as usual up through December 15th**
- **No further quizzes**
- **I will be available to meet individually with class members upon request**

HAVE A GREAT WINTER BREAK!
SEE YOU IN JANUARY!!!