

## Exam #3 STA-209 Sections 03, 05

Name: Key

### Directions

- You **do not** need to fill all the space provided. In many cases, large amounts of white space only exist to correct for general spacing of the exam
- You **do not** need to write in complete sentences: bullet points are completely acceptable and even preferred
- Use your critical value sheet to find critical values appropriate for the distribution at hand. When  $df > 40$ , use the standard normal approximation
- Have fun

### Formulas:

$$\chi^2 = \sum_{j=1}^k \frac{(\text{Observed}_j - \text{Expected}_j)^2}{\text{Expected}_j}$$

$$F = \frac{\text{MSG}}{\text{MSE}} = \left( \frac{n-k}{k-1} \right) \frac{\text{SSG}}{\text{SSE}}$$

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

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

### Question 1

Below I will describe a number of different study designs. You will be asked to

- (1) Identify the null hypothesis
- (2) Identify the correct statistical test for this hypothesis

**Part A:** For this study, I am interested in determining if a student's major (Humanities/STEM/Social Sciences) is associated with their final exam score in STA-209.

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad (\text{all means equal})$$

ANOVA

**Part B:** A two-day workshop for learning basic R has been created, where attendees are tested in their R skills both prior to the workshop and after the workshop has been completed. For each of these tests, a numeric score is given. We wish to determine whether or not the workshop has been effective in improving the R skill of the attendees.

$H_0$ : No difference

paired t-test

**Part C:** Binge drinking is defined as a pattern of drinking that involves consuming 5 or more standard drinks within 2 hours. Respondents of a survey were asked for their sex and whether or not they have engaged in binge drinking more than twice in the previous week. We wish to determine whether or not there is a difference in binge drinking patterns between men and women.

$\chi^2$  test

$H_0$ : Independent.

## Question 2

Cocaine addicts have been reported to have a significant depletion of stimulating neurotransmitters and thus continue to use cocaine to avoid feelings of depression and anxiety. A 3-year study with 72 chronic cocaine users compared an antidepressant drug called desipramine with lithium and a placebo (lithium is the standard treatment for cocaine addiction). One third of the subjects were randomly assigned to each treatment group with the following results:

	Relapse	No Relapse	
Desipramine	10	14	24
Lithium	18	6	24
Placebo	20	4	24
	48	24	72

Part A: What type of plot would you use to visually display these results

Conditional BAR CHARTS

Part B: Describe the null hypothesis of this study and construct a table of expected counts under the assumption of the null hypothesis.

Independence

$$P(\text{Desipramine and Relapse}) =$$

$$P(\text{Desipramine}) \times P(\text{Relapse}) \Rightarrow$$

$$= \left( \frac{24}{72} \right) \times \left( \frac{48}{72} \right)$$

Expected Probability

	R	NR
D	16	8
L	16	8
P	16	8

Part C: Compute the  $\chi^2$  statistic for the null hypothesis in B. Based on this, what conclusion would you reach if you were testing at the  $\alpha = 0.05$  level?

$$\chi^2 = \sum \frac{(\text{Exp} - \text{obs})^2}{\text{Exp}} = 10.5$$

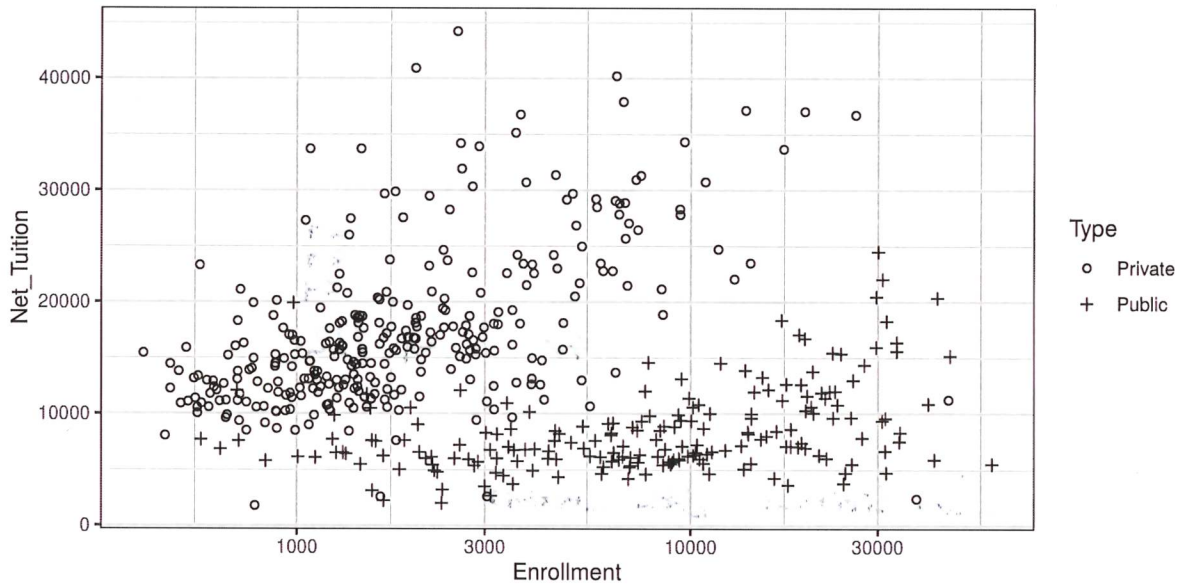
in.,  $df = (3-1) \times (2-1) = 2$

For  $\alpha = 0.05$   $\chi^2_{0.05} = 5.99$

Since  $10.5 > 5.99$ , we reject.

Reject  $H_0$ . (use critical Value sheet)

### Question 3



Model 1:

```
lm(formula = Net_Tuition ~ Enrollment, data = college)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	14225.3137	272.8034	52.1	<0.000000000000002 ***
Enrollment	-0.0820	0.0265	-3.1	0.002 **

Residual standard error: 7180 on 1093 degrees of freedom

Multiple R-squared: 0.00869, Adjusted R-squared: 0.00779

F-statistic: 9.58 on 1 and 1093 DF, p-value: 0.00201

Model 2:

```
lm(formula = Net_Tuition ~ Enrollment + Type, data = college)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5746.1019	377.2481	15.2	<0.000000000000002 ***
Enrollment	0.2533	0.0239	10.6	<0.000000000000002 ***
TypePrivate	10808.5970	398.6370	27.1	<0.000000000000002 ***

Residual standard error: 5550 on 1092 degrees of freedom

Multiple R-squared: 0.408, Adjusted R-squared: 0.406

F-statistic: 376 on 2 and 1092 DF, p-value: <0.000000000000002



**Part A:** For this part, consider **Model 1** from above. What is the null hypothesis in linear regression? Based on the summary output, how would you describe the relationship between enrollment and tuition?

No linear association b/w Net Tuition and Enrollment.

Based on output, we see evidence of negative linear association.

**Part B:** Now consider **Model 2**, which includes an indicator for whether or not a college is private. How would you interpret the intercept in this model? Is this a meaningful value in this model?

Intercept: Net Tuition at Public College w/ 0 Enrollment.

Not meaningful.

**Part C:** Compare the coefficient for Enrollment between **Model 1** and **Model 2**. What has changed? In other words, what impact has adding an indicator for Private had on this value, and why did it result in such a drastic change?

1.) ~~Pos.~~ Negative to Positive

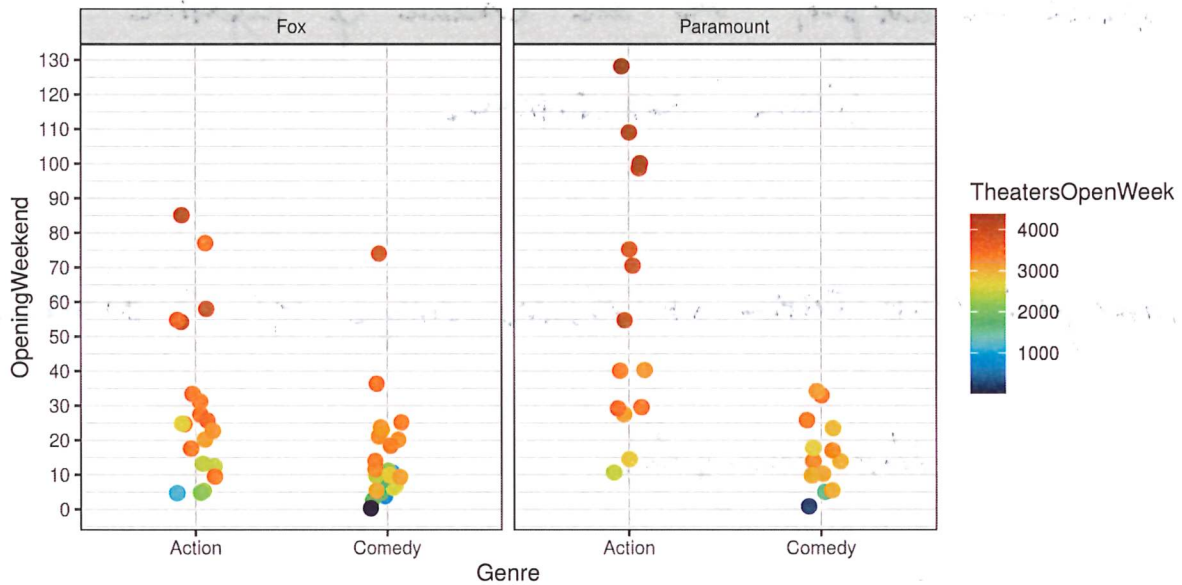
2.) By adding indicator, it allowed us to find lines of best fit that take Type into account.

\* Ignore question as to why.

## Question 4

(This page can be ripped off)

Included below are data from 70 Hollywood films released between 2007 and 2001. Movies in this dataset include Action and Comedy films from two major studios, Fox and Paramount. The plot below illustrates the total sales over a film's opening weekend, with a color aesthetic to indicate the number of theaters in which the film was shown: dark red corresponds to a film showing in a large number of theaters, while dark blue indicates that it was shown in relatively few theaters.



Below is summary information for a linear regression model with revenue from the opening weekend (`OpeningWeekend`) serving as the *dependent variable* and with film studio (`LeadStudio`) and genre (`Genre`) serving as the *independent variables*.

```
> lm(OpeningWeekend ~ Genre + LeadStudio, movies) %>% summary()
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	36.15	4.81	7.52	0.0000000018	***
GenreComedy	-25.69	5.78	-4.44	0.00003427689	***
LeadStudioParamount	14.69	5.94	2.47	0.016	*

Residual standard error: 24.2 on 67 degrees of freedom

Multiple R-squared: 0.288, Adjusted R-squared: 0.266

F-statistic: 13.5 on 2 and 67 DF, p-value: 0.0000116

You will use these plots and summary data to answer the following questions:

[This potential canvas intentionally left blank]

*[Faint, illegible handwritten text, possibly bleed-through from the reverse side of the page]*

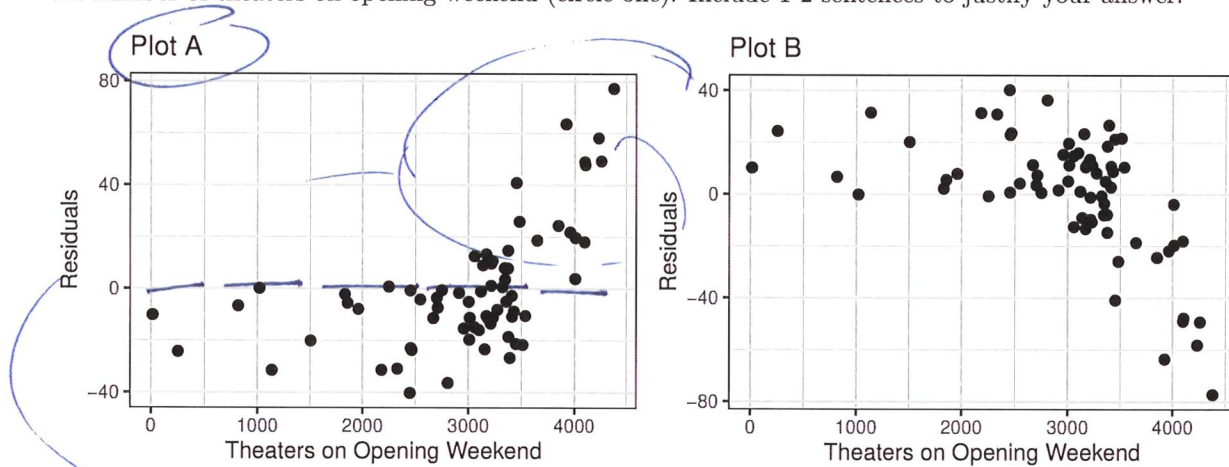
Part A: Give an interpretation of the intercept of this model. Is this meaningful?

*Predicted opening weekend for Fox Action movie  
yes meaningful*

Part B: Again using the summary information, find the predicted opening weekend revenue for each genre/studio combination (i.e., predicted opening weekend revenue for a Comedy film from Fox). With these predictions, mark horizontal lines on the plot above indicating the predicted opening weekend revenue for each category.

*Fox Comedy 10.46      Paramount Comedy 25.15  
Fox Action 36.15      Paramount Action 50.84*

Part C: We are now interested in determining if the variable for the number of theaters showing a film on opening weekend (*TheatersOpenWeek*) should be included in our model. We will do this by plotting the residuals of the model above against the missing variables. Using the changes you made to the plot above in Part B, determine which of the plots below shows the correct association between the model residuals and the number of theaters on opening weekend (circle one). Include 1-2 sentences to justify your answer.



*Positive residuals w/  
high opening weekend.*



Part D: Below is the updated model for predicting revenue on opening weekend, now including the variable for the number of theaters:

```
> lm(OpeningWeekend ~ Genre + LeadStudio + TheatersOpenWeek, movies) %>% summary()
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-21.48761	9.95081	-2.16	0.0345 *
GenreComedy	-13.69726	4.99129	-2.74	0.0078 **
LeadStudioParamount	8.74991	4.82823	1.81	0.0745 .
TheatersOpenWeek	0.01804	0.00287	6.28	0.000000031 ***

Residual standard error: 19.3 on 66 degrees of freedom

Multiple R-squared: 0.554, Adjusted R-squared: 0.534

F-statistic: 27.3 on 3 and 66 DF, p-value: 0.0000000000133

Given an interpretation of the intercept in this model. Is this meaningful?

*Box Office w/ 0 theaters opening weekend.*

*Not meaningful.*

Part E: Consider the presented linear models, both with and without the variable TheatersOpenWeek. Based on the summary information, which would you prefer to use to predict revenue on opening weekend? Briefly justify your answer.

*Second model*

*Higher  $R^2$  value so larger proportion of variance explained.*

### Question 5

Need something cute to do for ANOVA

