STA101 Final Review

Fan Bu

June 21, 2021

Summer Session I 2021

Department of Statistical Science, Duke University

Logistics of the Final

About the Final

- All chapters will be covered
- With emphasis on Chapters 5-9
- “Tests & Quizzes” on Sakai
- Open at 00:00am June 23 and close at 12pm June 24 (36-hour window)
- **2-hour** time window (you can start any time, but once started you must finish within 2 hours)
- 3 types of questions: (1) True or false, (2) Fill in the blank, and (3) Multiple choices
- 2 extra-credit questions
- The “sample exam” is still open for you to try out

About the Final (cont'd)

- **Individual** exam - you should **NOT** consult other people or the Internet
- Open-book (sort of) - you can consult the textbook, your notes, lecture slides, etc.
- Find a place with a steady Internet connection
- If you encounter **technical** issues (or if you believe certain questions are mis-worded or wrong), you should email Fan (`fan.bu1@duke.edu`); emails will be replied 9am-6pm EDT on June 23 and 9am-12pm on June 24.

A brief review of the topics

Chapters 1-4

- Observational studies v.s. experiments (Chp. 1)
- Summary statistics for numerical data (Chp. 2) - robust v.s. non-robust
- Reading plots (histograms, boxplots, scatterplots, etc.) (Chp. 2)
- Independence & conditional probabilities (Chp. 3)
- Normal distribution - shape, z-score, percentiles, quantiles (Chp. 4)

Chapter 5

- Point estimates and sampling variability
- Confidence intervals: the "$PE \pm CV \times SE$ formula for constructing CIs interpretation of CIs ("we're xx% confident that ...")
- Hypothesis testing framework: how to set up H_0 and H_A, Type 1 and 2 errors, significant level α, p-values (and interpretations)

Chapter 5 - example questions

How large should the expected number of successes and failures be for us to apply the CLT to approximate the sample proportion distribution under the null?

Chapter 5 - example questions

How large should the expected number of successes and failures be for us to apply the CLT to approximate the sample proportion distribution under the null?

For a fixed sample and a specific parameter of interest, is the 95% CI or 99% CI wider?

Chapter 5 - example questions

How large should the expected number of successes and failures be for us to apply the CLT to approximate the sample proportion distribution under the null?

For a fixed sample and a specific parameter of interest, is the 95% CI or 99% CI wider?

If $\alpha = 0.01$ and p-value $= 0.005$, should we reject H_0? Based on this decision, what type of error could we have made?

Chapter 6

- single proportions: samples proportion as the point estimate, SE, success-failure conditions
- two proportions: sample difference as the estimate for difference, SE

Chapter 6

- single proportions: samples proportion as the point estimate, SE, success-failure conditions
- two proportions: sample difference as the estimate for difference, SE
- For both single and two proportions inference, critical value $z^\star$ w.r.t normal distribution
    - if confidence level = 0.95, then $z^\star = 1.96$

Chapter 6

- single proportions: samples proportion as the point estimate, SE, success-failure conditions
- two proportions: sample difference as the estimate for difference, SE
- For both single and two proportions inference, critical value $z^\star$ w.r.t normal distribution
  - if confidence level = 0.95, then $z^\star = 1.96$
- chi-square test for goodness-of-fit: when to use (one-way table), formula for the χ^2 statistic, df, compute p-value, conditions
- chi-square test for independence: when to use (two-way table), formula for the χ^2 statistic, df, compute p-value, conditions

Chapter 6 - example questions

You surveyed 100 random students from Duke and UNC, respectively, and found that 80 Duke students are fully vaccinated while 60 UNC students are. Conduct a hypothesis test for whether or not the proportions of fully vaccinated students are different between Duke and UNC at significance level $\alpha = 0.05$.

Chapter 6 - example questions

You surveyed 100 random students from Duke and UNC, respectively, and found that 80 Duke students are fully vaccinated while 60 UNC students are. Conduct a hypothesis test for whether or not the proportions of fully vaccinated students are different between Duke and UNC at significance level $\alpha = 0.05$.

If we are to conduct a χ^2 test on the following two-way table for independence between gender and employment status (using a random sample among adults in a developing country), what is the χ^2 statistic and what is the degrees of freedom?

	unemployed	**part-time employed**	**full-time employed**	**total**
Male	30	55	65	150
Female	70	48	32	150
total	100	103	97	300

Chapter 7

- Basics of the t distribution and F distribution: shape, degrees of freedom, how to calculate percentiles and quantiles

Chapter 7

- Basics of the t distribution and F distribution: shape, degrees of freedom, how to calculate percentiles and quantiles
- One sample t-test: conditions (independence + normality), formula of SE, using t-distribution critical values for CIs
- t-test for paired data (similar to one-sample)
- two sample t-test: setup of the hypotheses, conditions, formula of SE

Chapter 7

- Basics of the t distribution and F distribution: shape, degrees of freedom, how to calculate percentiles and quantiles
- One sample t-test: conditions (independence + normality), formula of SE, using t-distribution critical values for CIs
- t-test for paired data (similar to one-sample)
- two sample t-test: setup of the hypotheses, conditions, formula of SE
- power calculation: what increases power, calculate sample size to reach a power level (extra credit)
- ANOVA: different sources of variability, formula (and interpretation) of the F statistic, calculation of p-value

Chapter 7 - example questions

50 students are recruited to rate two laptops (labelled by A and B) on a scale from 0 to 100 (each student would rate both A and B). What kind of test should we perform if we want to test for difference between the rating scores of laptops A and B?

Chapter 7 - example questions

50 students are recruited to rate two laptops (labelled by A and B) on a scale from 0 to 100 (each student would rate both A and B). What kind of test should we perform if we want to test for difference between the rating scores of laptops A and B?

Suppose the score differences have sample mean 2.6, sd = 3.5. Then what is the t-statistic value, and what is the p-value?

Chapter 7 - example questions

Suppose we want to compare the average weight of 578 chickens on 4 different diets.

Complete the following ANOVA table:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Diet			51954.219		
Residuals	574	2758693	4806.086		
Total		2914556			

Chapter 8

- Residuals, correlation, line-fitting
- Least squares: what is being minimized, conditions, calculating the parameter estimates
- Outliers: leverage points, influential points
- Inference about the slope: t-test for nonzero-ness, constructing CIs, interpretation of the slope

Chapter 8 - example questions

We use simple linear regression to predict Gift Aid (y) with Family Income (x). Calculate the estimates for intercept β_0 and slope β_1 using the summary statistics below.

	Family Income (x)	Gift Aid (y)
mean	$\bar{x} = \$101{,}780$	$\bar{y} = \$19{,}940$
sd	$s_x = \$63{,}200$	$s_y = \$5{,}460$
		$R = -0.499$

Figure 8.13: Summary statistics for family income and gift aid.

Also, what is the R^2 of this simple linear regression model?

Chapter 8 - example questions

Below is the output of a simple linear regression model to predict the trunk circumference (in mm) of orange trees with the age (in days) of trees.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.400	8.623	2.018	0.052
age	0.107	0.008	12.900	1.93e-14

- Interpret the estimated slope for “age”.
- Is “age” a significant predictor for trunk circumference? (At $\alpha = 0.05$ level)
- Would a 99% confidence interval for the slope (of “age”) cover 0?

Chapter 9

- How the levels of categorical variables are handled (1 reference, $k - 1$ dummy variables)
- Interpret slopes in MLR ("all else held constant/fixed...")
- Model selection: backward, forward, using adjusted R^2 or p-values
- Checking model conditions via plots

Chapter 9 - example questions

We use MLR to predict petal length of iris flowers of 3 species (setosa, versicolor, virginica) using petal width, sepal length, and sepal width. Model output is captured as below:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.11099    0.26987  -4.117 6.45e-05 ***
Petal.Width        0.60222    0.12144   4.959 1.97e-06 ***
Sepal.Length       0.60801    0.05024  12.101  < 2e-16 ***
Sepal.Width       -0.18052    0.08036  -2.246   0.0262 *
Speciesversicolor  1.46337    0.17345   8.437 3.14e-14 ***
Speciesvirginica   1.97422    0.24480   8.065 2.60e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Which is the reference level for “Species”? How to interpret the slope estimates for “Species”?

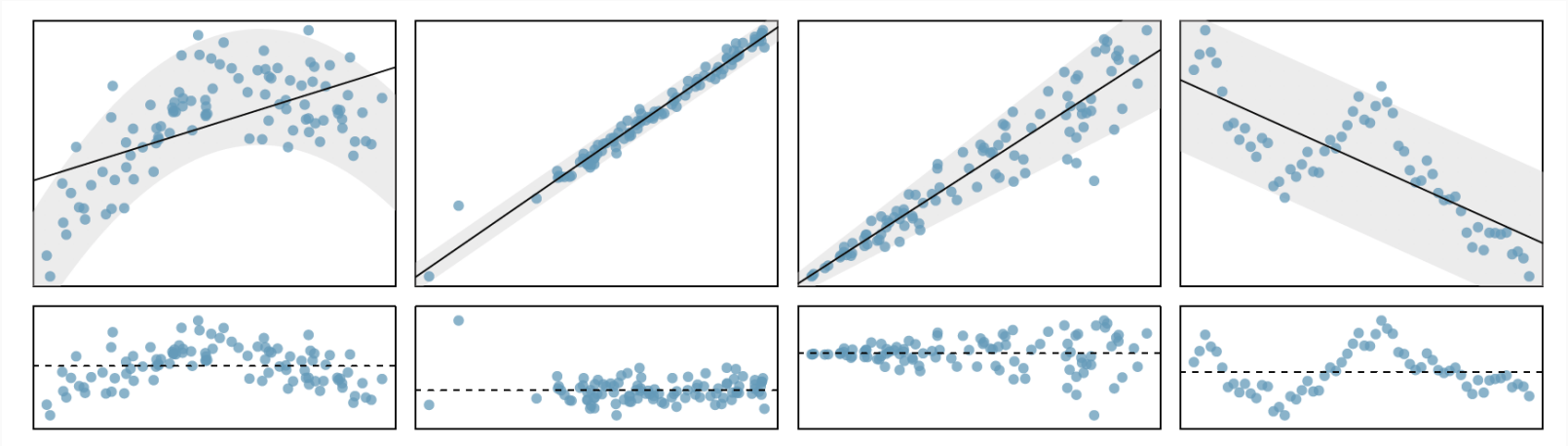
Chapter 9 - example questions

We use MLR to predict petal length of iris flowers of 3 species (setosa, versicolor, virginica) using petal width, sepal length, and sepal width. Model output is captured as below:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.11099    0.26987  -4.117 6.45e-05 ***
Petal.Width        0.60222    0.12144   4.959 1.97e-06 ***
Sepal.Length       0.60801    0.05024  12.101  < 2e-16 ***
Sepal.Width       -0.18052    0.08036  -2.246   0.0262 *
Speciesversicolor  1.46337    0.17345   8.437 3.14e-14 ***
Speciesvirginica   1.97422    0.24480   8.065 2.60e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

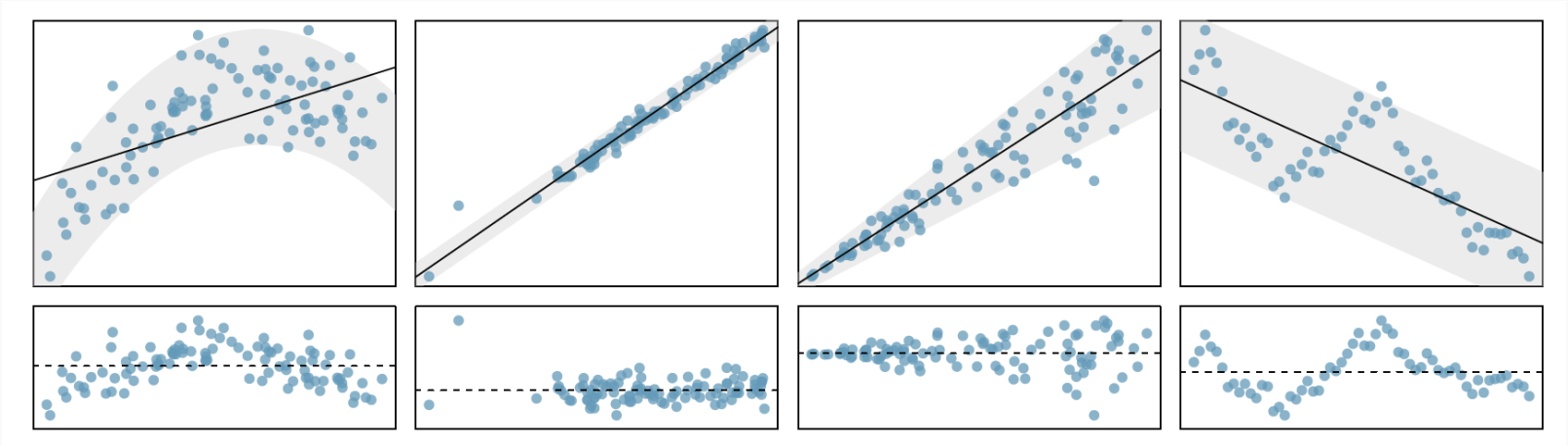
- If we want to do backward elimination based on p-values (at significance level $\alpha = 0.01$), which variable should be removed first?

Chapter 9 - example questions



Which condition for using linear regression is violated for each dataset? (top: y vs. $\hat{y}$; bottom: residuals vs. fitted value/order of observation)

Chapter 9 - example questions



(1) nonlinearity

(2) there is a clear outlier (a potential issue, not necessarily break everything)

(3) unconstant variance

(4) dependent observations (this example is time series data)

Some R stuff

- To get percentiles & quantiles:
  - Normal distribution: `pnorm` and `qnorm`
  - χ^2 distribution: `pchisq` and `qchisq`
  - t-distribution: `pt` and `qt` - don't forget to specify the df!
  - F-distribution: `pf` and `qf` - there are two dfs you have to specify!
- E.g., find the cuttoff value with tail probability 0.05 for $F(2, 5)$:

```
> qf(0.05, df1 = 2, df2 = 5, lower.tail = FALSE)

[1] 5.786135
```

- Check out the help document: "?" before a function name
- Have your `RStudio` container (or local environment) ready for the exam

Course evaluations!

- Please fill out the course evaluations survey
- Link in the email (sent out on June 15 or 16)
- Or log on to *duke.evaluationkit.com*
- *1 bonus point* if you complete the surveys and email me a screenshot of the submit/confirmation page (before 12pm June 24)

Thank you!