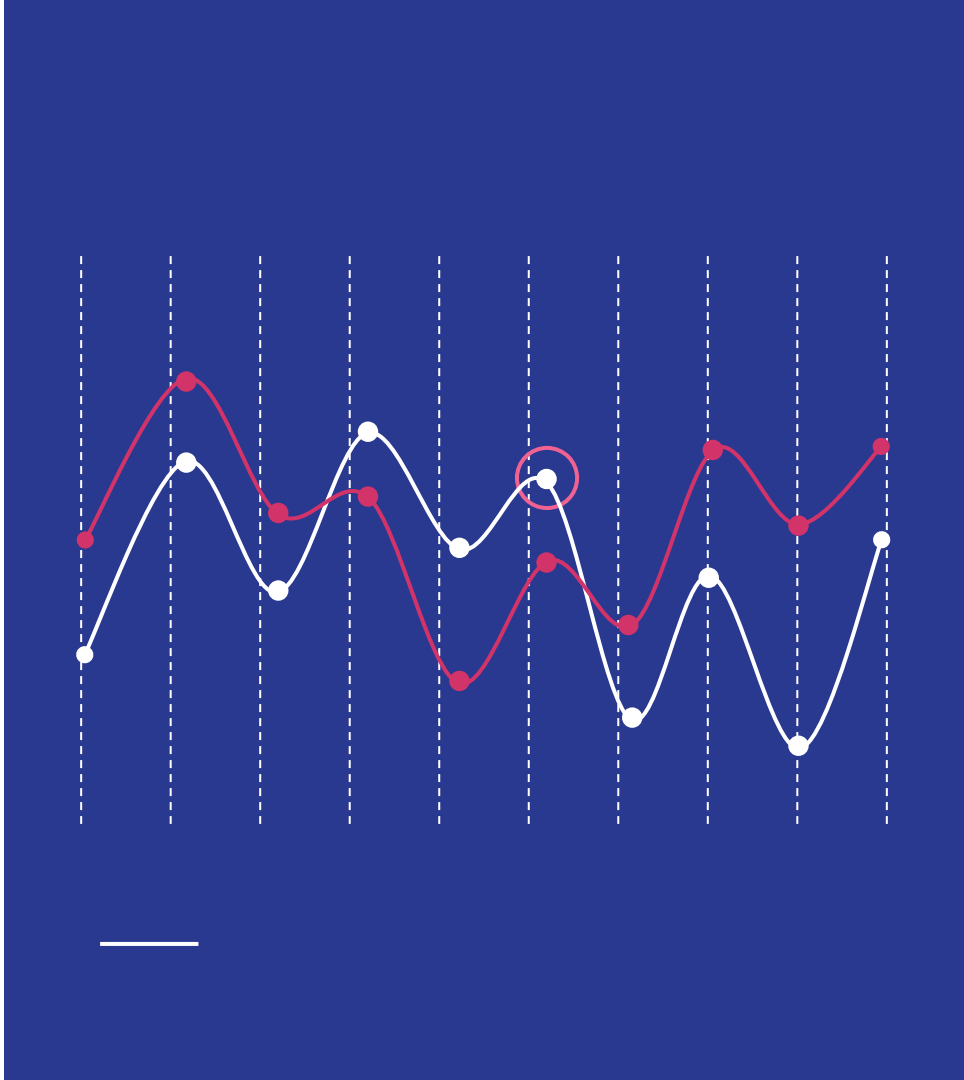
Non-Exam Strategies: PSYC 300 Statistics in Psychology

Steven Zhou
Innovations in Teaching & Learning
September 22, 2021



Existing Challenges due to Online Classes

Testing Capabilities

Current testing setup focuses on memorizing key formulas and applying them to “math” problems - which isn’t what statistics is about!

Access to SPSS

Current curriculum relies on SPSS - which isn’t what the vast majority of employers and researchers use!

Need to Practice

Statistics builds up week after week, and the current setup of one-time-use homework does not emphasize practice enough

Proposed Solutions

Project-Based

Use of real-world projects on real datasets, to emphasize actual mastery of technique over memorizing formulas

Using R

Both industry and academia are moving towards use of R, a free programming language, for statistical analysis; students are taught to run all their analyses in R

Weekly Practice Sets

In the weekly problem sets, students could turn them in early and get a grade within 24 hours so they could make corrections, maintaining high expectations on keeping up with the material while giving the chance to improve

Practice Set 5

Please download and complete the following questions. You may show your work by typing it out and submitting the document. or printing it out to complete by hand (then scan it and submit the scanned document). Either is fine!

If you turn in the assignment early (before Friday afternoon, September 24), I will grade it and get it back to you within 24 hours. You can then resubmit the assignment as many times as you want to earn full credit (15 points). All submissions and resubmissions are due on Blackboard by Sunday September 26 at 11:59 PM.

Questions

For each of the following scenarios, calculate the 95% Confidence Interval for the population mean. (4 points each)

1. You take a sample of 36 students (out of the entire population of George Mason students) and calculate that the average GPA for this sample of students is 3.32. The population standard deviation is 0.46.

2. You want to know how much money, on average, the population of all US citizens spends eating out for lunch (you know the population SD = $1.80). You survey the lunch habits of 9 employees in your company and get the following results on how much each employee spends for lunch:
   $15.10, $9.30, $10.80, $10.90, $8.70, $12.20, $14.90, $10.10, and $11.40

Test/Method	Use	R Code
mean	Represents the “weighted middle” of the dataset	mean()
standard deviation	Represents the “spread” of the dataset around the mean	sd()
Welch’s two-sample t-test	Compare two sample means	t.test(sample1, sample2)
paired sample t-test	Compare pre- and post- test using the same sample	t.test(pre, post, paired = T)
one-way ANOVA	Compare three or more sample means	aov(DV ~ IV, data = data)
factorial ANOVA	Used with 2+ categorical IVs and one numerical CV	aov(DV ~ IV1*IV2, data = data)
correlation	Relationship between two numerical variables	cor(data$x, data$y)
regression	Used to predict a numerical DV based on one or more IVs	lm(DV ~ IV, data = data)

Project #1

The attached dataset is a sample of 238 employees from a large (3000+ employee) technology company in the United States. The dataset includes five variables:

(a) Number of hours worked per week
(b) Gender
(c) Conscientiousness score, obtained by adding together five self-report questions on a scale of 1 to 5. Conscientiousness is a personality trait measuring one's diligence, carefulness, and organization. Sample question items include "I do things according to a plan" and "I follow a schedule"
(d) Task leadership score, obtained by adding together four self-report questions on a scale of 1 to 5. Task leadership is a set of behaviors focused on leading by directing and managing tasks. Sample question items include "I schedule the work to be done" and "I let group members know what is expected of them"

Your job is to produce a full set of descriptive statistics for these five variables, including the frequencies, means, and standard deviations. Your final deliverable should include:

(1) A table with the *frequencies* and *percentages* by gender. In other words, how many employees of each gender were there? What percent of the overall dataset was this?
(2) A table with the *means* and *standard deviations* for each of the four continuous variables (hours worked, conscientiousness, life satisfaction, and task leadership).
(3) A table with the *means only* of each of the four continuous variables, *separated by gender*. In other words, what was the mean hours worked for males? For females?

“Testimonials”

I just finished this weeks lectures and I wanted to thank you for taking the time to thoroughly explain everything. It is something you have done from the start but currently I am taking another math course and I am having a hard time because my professor does not take the time to walk through everything step by step like you. I was dreading stats because I almost failed ap stats in HS, but you have made it painless and understandable. So thank you for taking the time to walk through everything, and then summarizing and reviewing the next class, it means a lot and it helps a lot.

“Testimonials”

Hi professor I just wanted to say how great of a teacher you’ve been, even in an online setting. I usually have trouble watching lectures but the way you teach with energy and examples as well as clear instructions really helps me listen and take notes. Even that first video with the Pokémon stuff was pretty fun and I find it hard to find fun teachers. When I was in high school I took an AP Statistics class and pretty much bombed it and I hated statistics afterwards. Now while I’m taking your class I find myself considering a career in statistics because it’s fun to actually know how to do all the formulas.

Quick Example:
How to Predict the Future

THE DATA

YEAR	TUITION (IN-STATE, GRADUATE)
2020	15,648
2019	15,648
2018	15,138
2017	14,480
2016	13,724
2015	13,304
2014	12,614
2013	12,038
2012	11,690
2011	11,264
2010	10,556

source: https://www.collegetuitioncompare.com/trends/george-mason-university/cost-of-attendance

Learning Outcomes

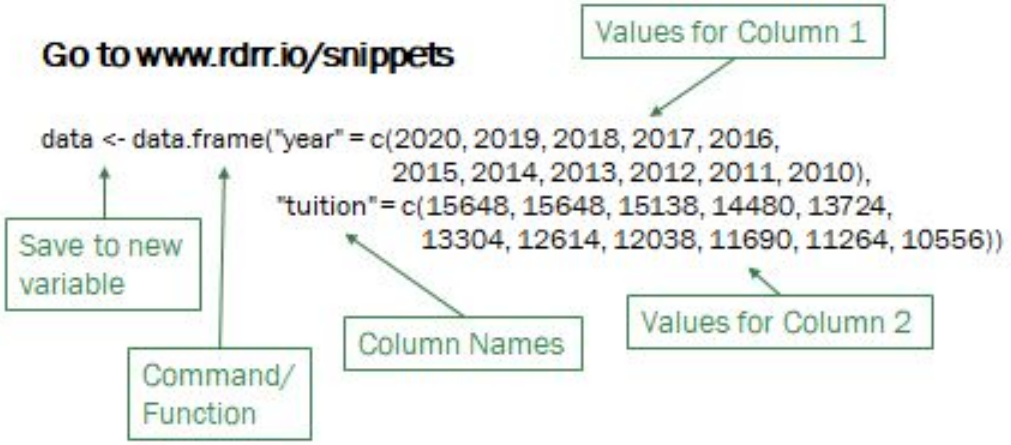
1. Use R
2. Build a regression model
3. Use it to predict next year's tuition (2021)
4. Add a covariate to the model
5. Use the new model to predict next year's tuition

THE DATA

	year	tuition
1	2020	15648
2	2019	15648
3	2018	15138
4	2017	14480
5	2016	13724
6	2015	13304
7	2014	12614
8	2013	12038
9	2012	11690
10	2011	11264
11	2010	10556

source: https://www.collegetuitioncompare.com/trends/george-mason-university/cost-of-attendance

Go to www.rdrr.io/snippets




```
data
```

THE MODEL

	year	tuition
1	2020	15648
2	2019	15648
3	2018	15138
4	2017	14480
5	2016	13724
6	2015	13304
7	2014	12614
8	2013	12038
9	2012	11690
10	2011	11264
11	2010	10556

```
model1 <- lm(tuition ~ year, data = data)
```

Linear model

Use *year* column to predict *tuition* column

Based on data in the *data* variable you just created

```
summary(model1)
```

```
Call:
lm(formula = tuition ~ year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-331.18 -113.18   21.82  129.22  237.62

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.074e+06  3.499e+04  -30.68 2.03e-10 ***
year         5.394e+02  1.737e+01   31.06 1.82e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 182.1 on 9 degrees of freedom
Multiple R-squared:  0.9908,    Adjusted R-squared:  0.9897
F-statistic: 964.8 on 1 and 9 DF,  p-value: 1.822e-10
```

a = -1074000

B = 539.4

significant!

THE PREDICTION

```
Call:
lm(formula = tuition ~ year, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-331.18 -113.18   21.82  129.22  237.62

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.074e+06  3.499e+04  -30.68 2.03e-10 ***
year         5.394e+02  1.737e+01   31.06 1.82e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 182.1 on 9 degrees of freedom
Multiple R-squared:  0.9908,    Adjusted R-squared:  0.9897
F-statistic: 964.8 on 1 and 9 DF,  p-value: 1.822e-10
```

```
newdata <- data.frame("year" = c(2021))

predict(model1, newdata, interval = "confidence")
```

Use *model1* to predict *newdata*

```
       fit      lwr      upr
1 16518.58 16252.14 16785.02
```

In-state tuition for graduate students next year will be $16,518.58, with a 95% confidence interval of [$16252.14, $16785.02]