Stat 21 Test 1

Due: Oct 17, 2020 by noon ET

This test is due on to be submitted on Gradescope on **October 17** by **12:00pm ET**. Please use the `#test1_questions` channel on Slack to post any clairfication questions. Do not ask questions like "Is [this] the right answer?"

You must submit your solutions as a single **PDF** document uploaded to **Gradescope**. You may use R markdown to write up your solutions alone or you may use R markdown and hand-written solutions. You must show all of your work, including code input and output. Please make sure each problem is **clearly labeled** and that any handwritten components (such as pictures or equations) are easily readable in the PDF document. You may want to use a service like CamScanner (https://www.camscanner.com/) to help you upload handwritten pages and Small PDF (https://smallpdf.com/merge-pdf) to merge multiple PDFs into a single document.

You are permitted to reference all class material and use the internet (though I am not sure it will be very helpful). You are not permitted however, to get assistance from any person online or otherwise.

- Your file should contain the code to answer each question in its own code block. Your code should produce plots/output that will be automatically embedded in the output pdf file.
- Each answer must be supported by written statements and relevant plots.
- Each problem is worth 20 points for a total of 100 points possible.
- In order to knit this document, make sure you have installed the following packages in your version of RStudio: `ggplot2`, `tidyverse`, `gridExtra`, `gcookbook`, `knitr`

The table below is from an article titled "Class in America-2012" by Gregory Mantsios. This table shows the median combined SAT scores (ranging from 400 to 1600) and the household income (broken into 10 categories) of 1,647,123 SAT-takers in the year 2010.

Income	Median SAT Score
$<$ \$20,000	1323
\$20,000 - \$40,000	1398
\$40,000 - \$60,000	1461
\$60,000 - \$80,000	1503
\$80,000 - \$100,000	1545
\$100,000 - \$120,000	1580
\$120,000 - \$140,000	1594
\$140,000 - \$160,000	1619
\$160,000 - \$200,000	1636
$\geq$ \$200,000	1721

Based on this table, we may suspect that there is a relationship between SAT score and household income. Run the following lines of R code to import a data set consisting of a simple random sample of 100 students who took the SAT in 2010. (This sample was based on the data report by the College Board, feel free to talk to me about how I obtained this sample later![1]) Use this data set to answer Problems 1-2.

```
SAT_data <- read_table2(url("http://www.swarthmore.edu/NatSci/sthornt1/DataFiles/SAT_data.txt"))
```

Problem 1

State the null and alternative hypotheses for an ANOVA test of association between income and SAT scores. Comment on whether or not the necessary assumptions seem reasonable. Then perform the ANOVA test and interpret the results in the context of this data set.

Problem 2

Perform a Tukey HSD pairwise comparison to determine where the greatest disparities in SAT scores occur.

[1] https://secure-media.collegeboard.org/digitalServices/pdf/research/2010-total-group-profile-report-cbs.pdf

Problem 3

Suppose we decide to treat household income as a quantitative variable and interpolate the annual income in the following manner.

Income	Median SAT Score
10,000	1323
30,000	1398
50,000	1461
70,000	1503
90,000	1545
110,000	1580
130,000	1594
150,000	1619
180,000	1636
300,000	1721

Fit a SLR model to this data with the median SAT score as the response variable. Write out the estimated regression equation and interpret the slope in the context of the problem. Do you think a SLR is a reasonable model for this data? Justify your answer.

Hint: You may copy and paste the code below to create a data object in R based on the table above.

```
income_SAT2 <- tibble(income = c(10000, 30000, 50000, 70000, 90000, 110000,
                                 130000, 150000, 180000, 300000),
                      median_SAT = c(1323, 1398, 1461, 1503, 1545, 1580, 1594,
                                     1619, 1636, 1721))
```

Income Bracket	Number of Households
<15,000	11689.04
15,000-24,999	10276.08
25,000-34,999	10661.43
35,000-49,999	15028.77
50,000-74,999	21194.42
75,000-99,999	15799.47
100,000-149,999	19909.90
150,000-199,999	10661.43
>=200,000	13230.45

Problem 4

The data above was published by the US Census bureau Current Population Report in September 2020 on income and poverty in the United States.[2] In this data, income is broken into 9 different categories and the number of households (in thousands) within each income bracket is recorded for the year 2019. (E.g. In the table above, there were 10276.08×1000 households who reported an annual income between \$15,000 and \$24,999 in the 2019 census.) If income inequality is not a real issue in the US, then we would expect to see the total number of households is uniformly distributed across these different income brackets.

Determine which chi-square procedure to use to statistically determine if income inequality is present in the US. Explicitly write out your null and alternative hypotheses, comment on the validity of the required assumptions, and interpret the results of the test at an $\alpha = 0.05$ significance level. Show all of your work including R code and output.

Hint: You may copy and paste the code below to create a data object in R based on the table above.

```
household_income <- tibble(
  income_bracket = factor(c("<15,000", "15,000-24,999", "25,000-34,999",
                            "35,000-49,999", "50,000-74,999", "75,000-99,999",
                            "100,000-149,999", "150,000-199,999", ">=200,000")),
  households = c(11689.041, 10276.08, 10661.433, 15028.767, 21194.415, 15799.473,
                 19909.905, 10661.433, 13230.453))
```

[2] https://www.census.gov/content/dam/Census/library/publications/2020/demo/p60-270.pdf

	Asian	Black	White
<15,000	444.990	2933.288	6195.364
15,000-24,999	342.300	1961.210	6195.364
25,000-34,999	355.992	1944.156	6365.100
35,000-49,999	595.602	2336.398	9335.480
50,000-74,999	883.134	2865.072	13748.616
75,000-99,999	855.750	1671.292	10863.104
100,000-149,999	1225.434	1841.832	14257.824
150,000-199,999	855.750	716.268	7892.724
>=200,000	1293.894	784.484	10014.424

Problem 5

The data above is based on the same data from Problem 5. This data set however identifies not only the annual income of each household in the US in 2019 but also identifies the race of the household. Though the US census now allows households to select more than one racial identifier, we are going to consider only non-intersecting racial categories: white alone (not Hispanic), Black alone, and Asian alone. Based on the data above, perform a statistical test to determine if race and income bracket are independent variables. Explicitly write out your null and alternative hypotheses, comment on the validity of the required assumptions, and interpret the results of the test at an $\alpha = 0.05$ significance level. Show all of your work including R code and output.

Hint: You may copy and paste the code below to create a data object in R based on the table above.

```
household_income_race <- tibble(
  income_bracket = factor(c(rep(c("<15,000", "15,000-24,999", "25,000-34,999",
                                  "35,000-49,999", "50,000-74,999", "75,000-99,999",
                                  "100,000-149,999", "150,000-199,999", ">=200,000"),3)),
                          levels=c("<15,000", "15,000-24,999",
                                   "25,000-34,999","35,000-49,999",
                                   "50,000-74,999", "75,000-99,999",
                                   "100,000-149,999", "150,000-199,999", ">=200,000")),
  households = c(6195.364, 6195.364, 6365.1, 9335.48, 13748.616, 10863.104, 14257.824,
                 7892.724, 10014.424, 2933.288, 1961.21, 1944.156, 2336.398, 2865.072,
                 1671.292, 1841.832, 716.268, 784.484, 444.99, 342.3, 355.992, 595.602,
                 883.134, 855.75, 1225.434, 855.75, 1293.894),
  race = factor(c(rep("White",9),rep("Black",9),rep("Asian",9))))
```