STAT 2300 Exam 2

Name: _____

Friday, April 4, 2025

Write directly on this exam. You may "show work" by handing in an R script, .Rmd file, or knit Markdown.

You may use R, the internet, and any reference material. You are not allowed to communicate with anyone - no email, messaging, internet forums, AI, etc.

Honor Pledge (10 points)

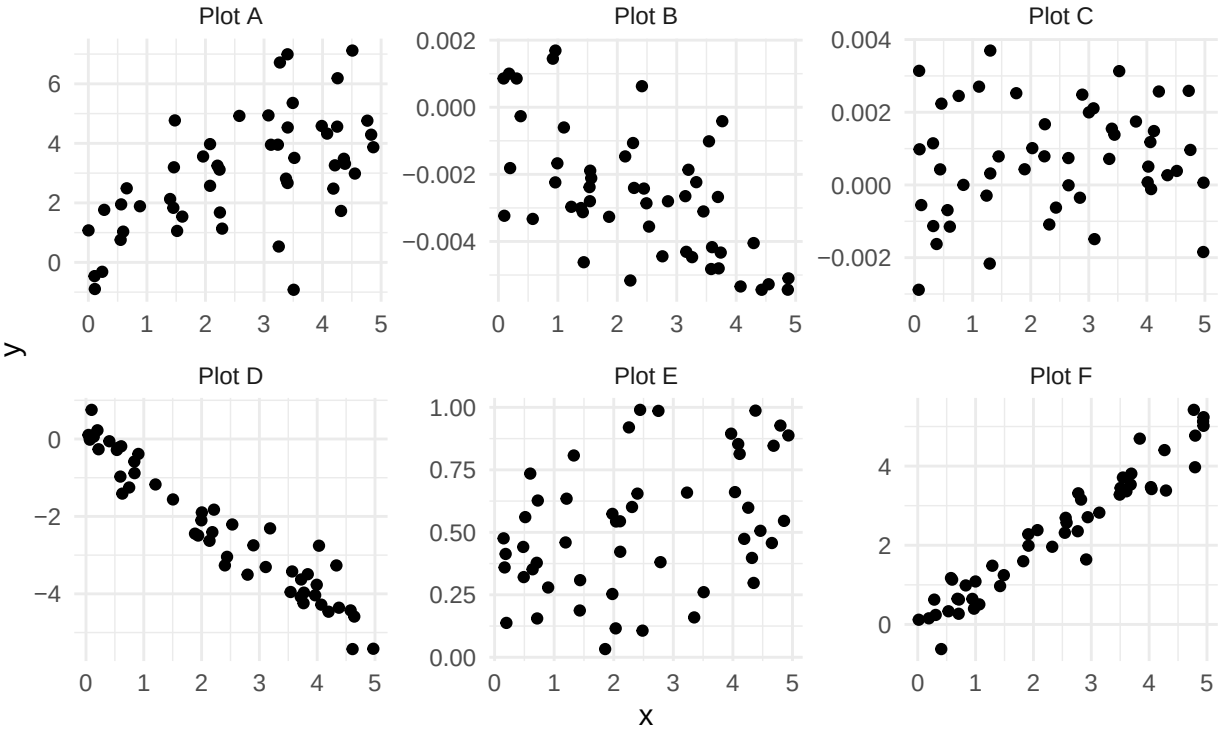
The work I have submitted represents my own effort. While working on this exam, I did not use generative AI or communicate in any form with individuals other than the instructor.

Signed:

_____

Problem 1 (10 points)

For each plot, make your best estimate of the correlation coefficient r. Write your answer on the plot itself.



Reading across then down, something like 0.5, -0.5, 0, -0.9, 0.3, 0.9.

COMPAS is a rating system that US courts use to help decide the likelihood that an arrested person will commit another crime. It is used to decide whether to release prisoners on bail.

On our webpage, the data `compas2014.csv` contains COMPAS information for all violent criminals in 2014 in Broward County, Florida. The COMPAS score is the variable `score`, ranging from low risk 1 to high risk 10.

```
compas2014 <- read.csv("https://turtlegraphics.org/stat2300/data/compas2014.csv")
```

Problem 2 (10 points)

COMPAS scores are ordinal, so it is reasonable to use rank-based methods.

Perform a Wilcoxon rank sum test to determine if there is a difference in COMPAS scores between Male and Female prisoners. Report your results with a p-value.

Solution

```
wilcox.test(score ~ sex, data=compas2014)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  score by sex
## W = 88974, p-value = 0.2906
## alternative hypothesis: true location shift is not equal to 0
```

There is not significant evidence of a difference in scores between Male and Female prisoners ($p = 0.29$)

Problem 3 (10 points)

Does COMPAS score depend on age? Fit a linear model for score in terms of age.

What is the coefficient of `age` in your model? How do you interpret this value?

Solution The coefficient of age is -0.09, which means that for every year of age, the average COMPAS score drops by 0.09. Or, for every decade older, the average COMPAS score drops by almost a point.

```
lm(score ~ age, data=compas2014) |> summary()
```

```
##
## Call:
## lm(formula = score ~ age, data = compas2014)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8560 -2.1219 -0.8441  1.8702  7.9733
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.229017   0.253751   28.49   <2e-16 ***
## age         -0.091269   0.006881  -13.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.624 on 1014 degrees of freedom
## Multiple R-squared:  0.1479, Adjusted R-squared:  0.147
## F-statistic: 175.9 on 1 and 1014 DF,  p-value: < 2.2e-16
```

Problem 4 (10 points)

Using your model of score in terms of age, test if the relationship between age and score is significant.

Solution The relationship is significant ($p < 2 \times 10^{-16}$)

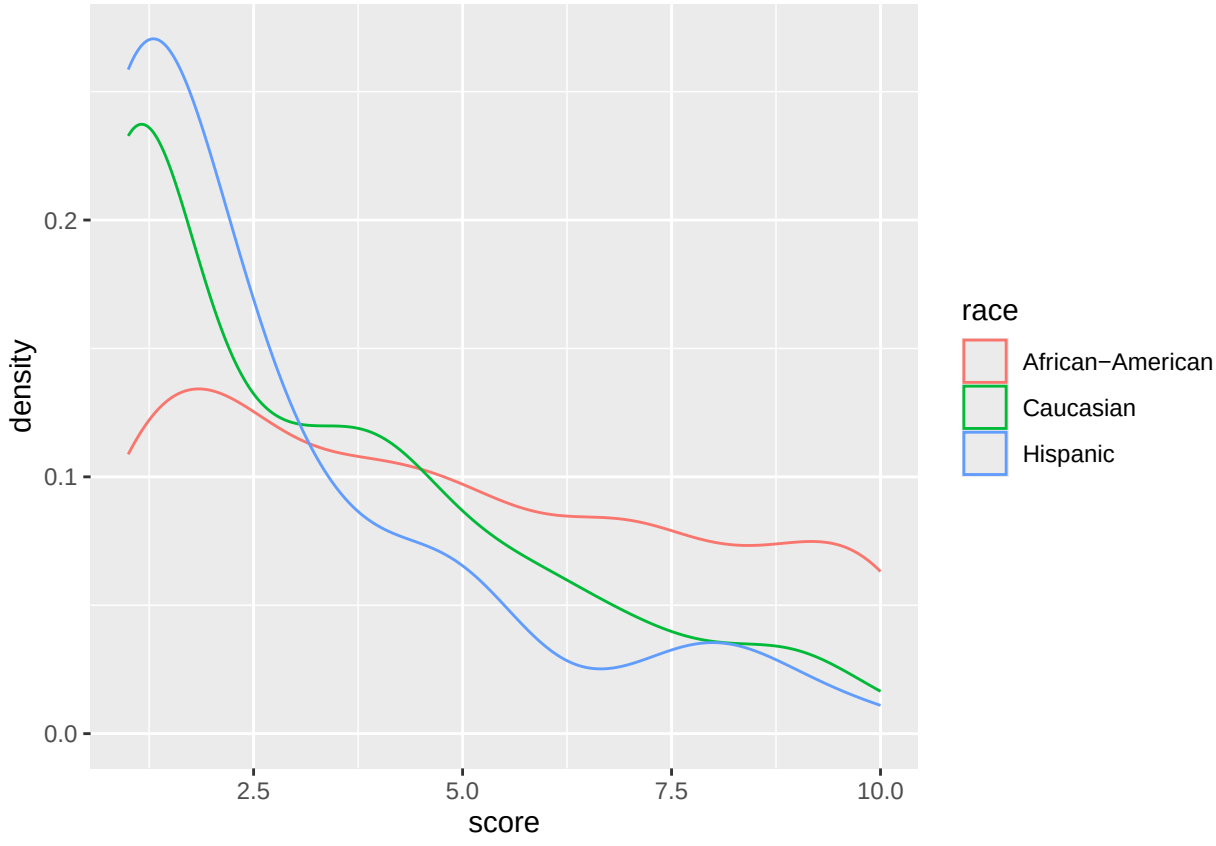
Problem 5 (10 points)

Is the distribution of COMPAS scores different across racial lines? Make a plot using `geom_histogram` or `geom_density` so that you can compare the distributions across the different race categories. Limit your picture to only African-American, Caucasian, and Hispanic (since the other race categories are relatively rare).

Hand in the plot as part of a knit markdown or else simply a screenshot.

Solution

```
compas_race <- compas2014 |> filter(race %in% c("African-American", "Caucasian", "Hispanic"))
compas_race |> ggplot(aes(x=score, color=race)) + geom_density()
```



Problem 6 (10 points)

Perform an analysis of variance to test if COMPAS scores are significantly different for the three race categories African-American, Caucasian, and Hispanic.

Solution

```
lm(score ~ race, data=compas_race) |> anova()
```

```
## Analysis of Variance Table
##
```

```
## Response: score
##             Df Sum Sq Mean Sq F value    Pr(>F)
## race         2  686.9  343.43  46.962 < 2.2e-16 ***
## Residuals 956 6991.3    7.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problem 7 (10 points)

The data `case0601` from Sleuth3 is from an experiment that rated employment interviews for people with various handicaps.

Is there a significant difference in Score between any actual Handicap and the "None" Handicap? Use a `pairwise.t.test` to check, and report the four relevant p-values.

Solution None of them are significant. The p-values are 1, 0.719, 0.866, 1.

```
pairwise.t.test(case0601$Score, case0601$Handicap)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  case0601$Score and case0601$Handicap
##
##            Amputee Crutches Hearing None
## Crutches   0.165   -        -       -
## Hearing    1.000   0.035    -       -
## None       1.000   0.719    0.866   -
## Wheelchair 0.860   1.000    0.321   1.000
##
## P value adjustment method: holm
```

Problem 8 (10 points)

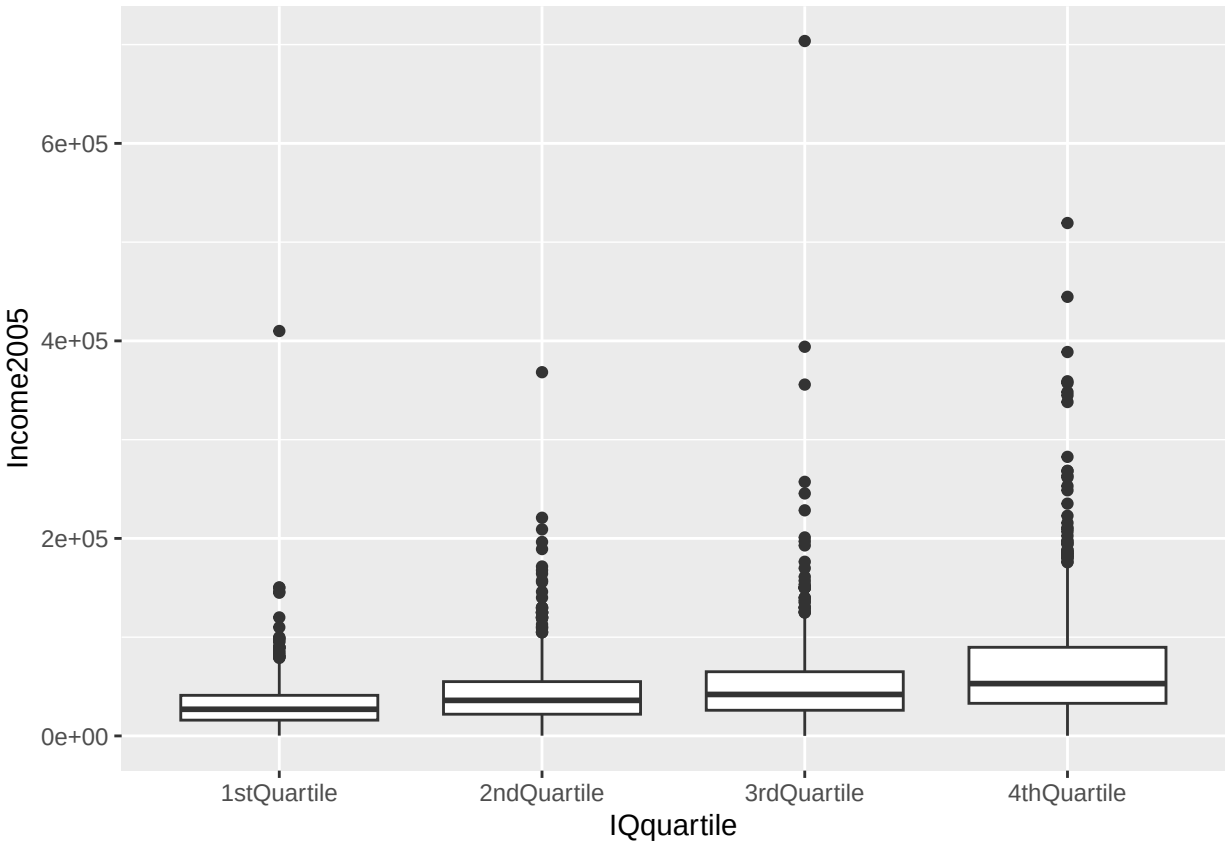
A two-sample t-test of score between Amputee and Crutches gives a significant result of $p = 0.0164$. However, the result from `pairwise.t.test` is no longer significant. Explain what happened here.

Solution Pairwise t test applies a correction for multiple tests, increasing all the p values. Here, they are multiplied by 10 since we are doing 10 tests.

Problem 9 (10 points)

Here is a boxplot showing income in 2005 for Americans who took IQ tests in 1981:

```
ex0524 |> ggplot(aes(x=IQquartile, y=Income2005)) + geom_boxplot()
```



a. Why is this data unsuitable for an ANOVA to test the relationship between IQ quartile and Income?

b. What could you do to fix the problem?

Solution This data is skew, and the variances aren't equal. Taking logarithms of the income fixes the problem.