Midterm Information

- In class on Tuesday (Oct 11): 1:30-2:45pm
- You are allowed a calculator (without internet access) and two pages of double-sided notes (4 sides total)
- The exam covers problem sets 0-4, lectures 1-10, and labs 1-5
- Practice exams are on Canvas
- Extra office hours are posted on Canvas (Monday and Tuesday)

Difficulty guide:

- Easy: 2a, 2d, 2f, 2g, 3a, 3c, 3e, 4a, 5a, 6d
- Medium: 2b, 2c, 2e, 3b, 3d, 3g, 4b, 5c, 6a, 6b, 7
- Difficult: 1, 3f (if using process of elimination), 4c, 5b, 6c,
- Very difficult: 3f (if done fully)

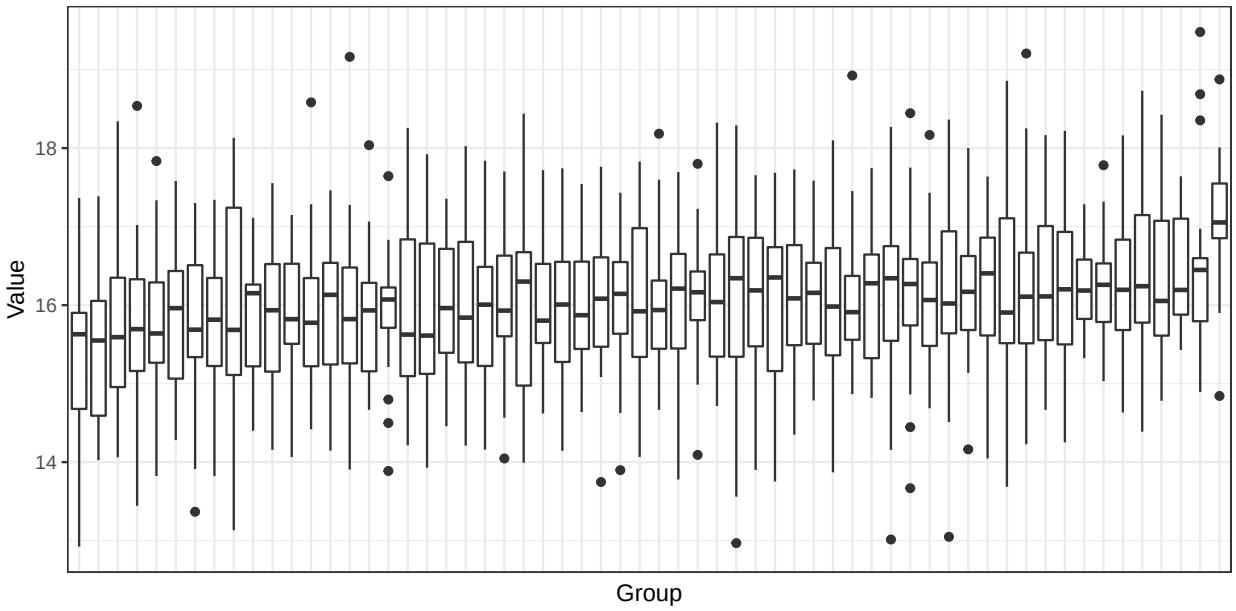
Problem 1: Pride goeth before a fall (Just for fun)

Alex Roygant (A. Roygant) is in a class with 99 other people, and the class's only assignments are 11 weekly quizzes. At the end of the semester, the quiz grades for each student are averaged to get a final score. Alex brags that he scored in the 90^{th} percentile on each quiz. What is the lowest his final percentile could be in the class? Note: There are no assumptions made about scores being independent or scores being unique.

Problem 2: Statistical testing (Parts are unrelated unless otherwise specified)

(a) In June 2022, the New York Times reported that in the last year (June 2021 to June 2022) the per capita death rate from COVID-19 had become higher for white Americans than for Black and Hispanic Americans. However, many public health officials and statisticians pointed out that in every age bracket, Black and Hispanic Americans were still more likely than white Americans to die of COVID-19. How is this possible?

(b) Consider the following plot comparing 60 groups. Look carefully for groups that might be different from the rest.



i. Which (if any) of the ANOVA assumptions are violated?

ii. Do you expect the F-test to be significant? Why or why not?

iii. If you ran pairwise t-tests for all of the groups, what proportion of the tests would you expect to be significant?

A) 0
B) 0.033
C) 0.050
D) 0.082

(c) A company is testing a new drug intended to reduce plaques of a misfolded protein in a rare disease. The company's biostatistics team has determined that for a preliminary study where all participants will be given the drug, they need a sample size n to achieve a Type 1 error rate of 0.05 and a Type 2 error rate of 0.2 for a t-test of H_0 : the drug has 0 effect on reducing plaques vs H_a : the drug has some effect on reducing plaques. However, when it comes time to enroll participants in the trial, the doctors can only find one large family where n people are affected by the disease. If the trial consists of just the n people from this family, provide plausible estimates of the Type 1 and Type 2 error rates and a brief explanation of why they will or will not change.

(d) Lotsa Cash is comparing incomes between Massachusetts and New Hampshire and decides to use a $\log_2$ transformation on the data because of its right skew. Lotsa obtains a confidence interval of (0.23, 0.66) for the difference in means (Massachusetts - New Hampshire) on the log scale. What should she conclude on the original scale? What assumption is required for this conclusion?

(e) If 10 i.i.d. observations are generated from a Normal distribution with mean 0, which of the following have the t_9 distribution?

A) $\frac{\bar{X}}{\sqrt{\frac{S^2_{\bar{X}}}{10}}}$

B) $\frac{\bar{X}}{\text{SE}(\bar{X})}$

C) $\frac{\bar{X}}{\frac{\text{UB}-\bar{X}}{t^*_9}}$ where UB is the upper bound in a 95% confidence interval for μ

D) $F^{-1}_{t_9}(U)$ where $F^{-1}_{t_9}$ is the t_9 quantile function and U is the p-value of a one-sample 2-sided t-test when the null is true.

(f) For each variable, choose the transformation most likely to make it normal from the following list: log, exponential, square root, logit, reciprocal.

i. Proportion of people below the poverty line in US counties:
ii. Population in US cities:
iii. Area of individual napkins produced by a factory:

(g) For each of the following scenarios, choose the 2-sample comparison from the following list most likely to fix any issues with the original data: unpooled t-test, paired t-test, log-transformed t-test, rank-sum test, permutation test.

i. The values in each group are bounded above by 5 and are left skewed with many points less than 0. You do not care about any particular statistic but rather about comparing the whole distributions.

ii. The values in each group are bounded above by 5 and are left skewed with many points less than 0. You want to compare the means of the groups.

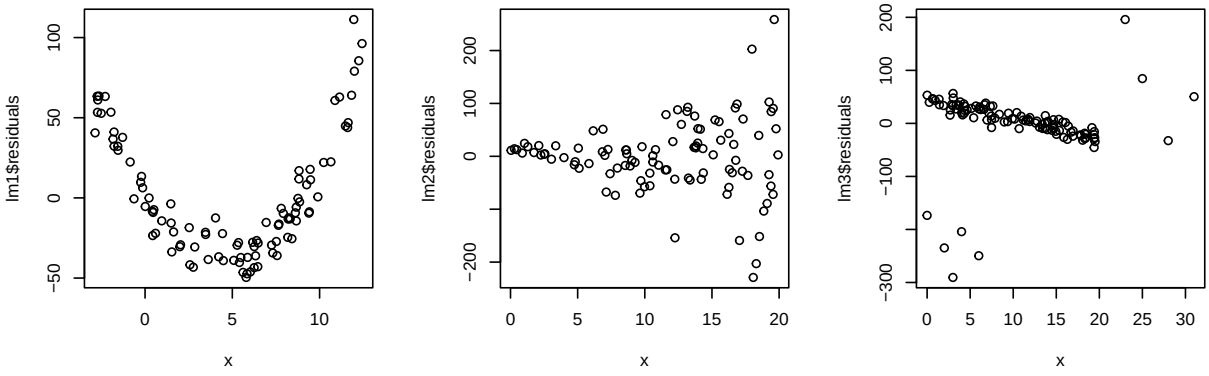
iii. The values in each group are measurements of the same thing at different times (one time in each group). You want to compare the means of the groups.

iv. Each group has 80 observations, and the distribution of each group's values is slightly right skewed. You want to compare the means of the groups.

Problem 3: Regression (Parts are unrelated unless otherwise specified)

(a) In *Thinking Fast and Slow*, Daniel Kahneman recalls teaching Israeli fighter pilots the evidence-based idea that rewarding good performance is more effective than punishing poor performance. One of the experienced instructors disagreed, noting that when he praised a pilot for good performance the pilot rarely performed better on the next flight and when he criticized a pilot for poor performance the pilot often performed better on the next flight. Both Kahneman's teaching and the instructor's experience were likely true. How can this be?

(b) Estimate the correlation of the residuals and the predictors in the following plots from simple linear regressions:



(c) Circle which of the following is not equivalent to the rest:

A) $\sum_{i=1}^{n}(X_i - \bar{X})\bar{X}$
B) $(\sum_{i=1}^{n} X_i^2) - n\bar{X}^2$
C) $\sum_{i=1}^{n}(X_i - \bar{X})X_i$
D) $\sum_{i=1}^{n}(X_i - \bar{X})^2$

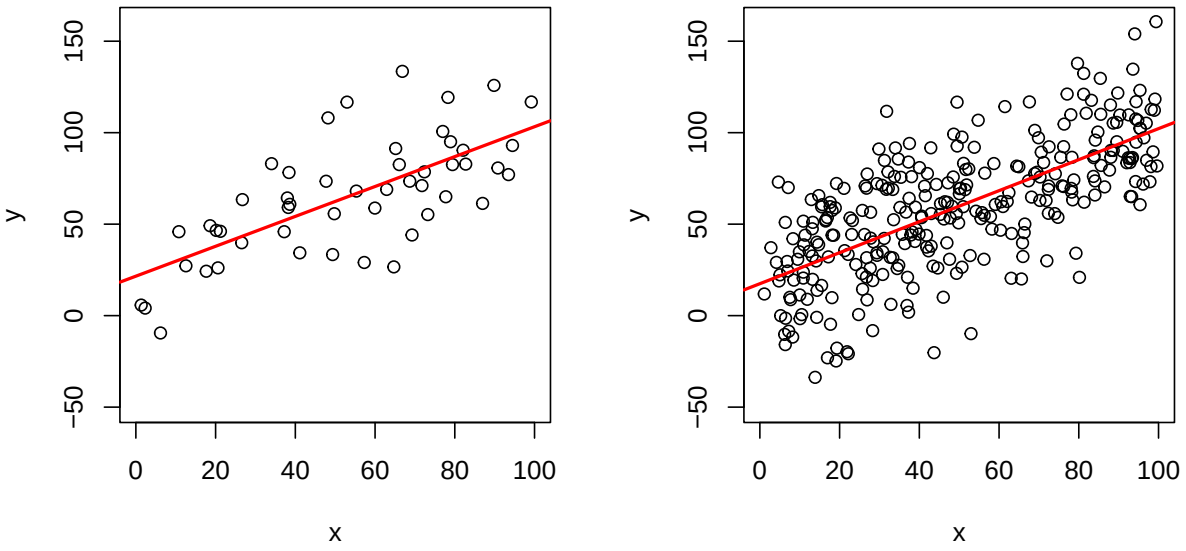
(d) The following summary is from a data set with the following columns:

- `epi_eh`: Environmental health policy score measuring how well a country is protecting its citizens from environmental health risks.
- `epi_cch`: Climate change score measuring progress to reduce pollutants including carbon dioxide, methane, and fluorinated gases.
- `mad_gdppc`: GDP per capita

Interpret the coefficients and significances of each predictor in the model.

```
##
## Call:
## lm(formula = epi_eh ~ epi_cch + log(mad_gdppc, 2), data = countries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.7246  -6.8701  -0.2615   8.3251  29.2407
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -94.15448    8.60541 -10.941  < 2e-16 ***
## epi_cch             0.43548    0.07876   5.529 1.32e-07 ***
## log(mad_gdppc, 2)   8.86217    0.79336  11.170  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.94 on 156 degrees of freedom
##   (35 observations deleted due to missingness)
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 207.9 on 2 and 156 DF,  p-value: < 2.2e-16
```

(e) Two scatterplots with regression lines are shown below. Trace a 95% confidence and 95% prediction interval for each.



(f) Consider a data set with a continuous variable Y and a categorical variable X with equal proportions of 0s and 1s (n each). Exactly one of the following tests gives a different p-value. Which one? Assume the sample variances of each group are not the same.

A) Unpooled t-test
B) t-test for $\beta_1 = 0$ in the linear model `lm(y ~ x)`
C) Overall regression f-test for the resulting model of `lm(y ~ x)` having any predictive ability
D) Contrast test with $C^T = [0 \;\; 1]$ and $\vec{\beta}^T = [\beta_0 \;\; \beta_1]$ for the linear model `lm(y ~ x)`

(g) In a last ditch effort to draw conclusions from her data, Auda Fidese has decided to test every combination of her five β coefficients for significance in contrast tests. Specifically, she is going to test all $\vec{C}^T$ of the form $[0 \;\; I_1 \;\; I_2 \;\; I_3 \;\; I_4 \;\; I_5]$ where I_j is an indicator of β_j being in the test. Assume the overall model F-test null is true and she is rejecting tests where the p-value is below α.

i. Find the expected number of Type 1 errors in this setup.

ii. Explain why the probability of making any Type 1 error is not $1 - (1 - \alpha)^{31}$.

Problem 4: Ben O'Meal's bootstraps

Ben O'Meal is trying to create a 95% confidence interval for the proportion of nights he stays up past midnight. He has recorded $n = 30$ night's worth of data and found that he stayed up past midnight 18 times. In his sleep deprivation, Ben cannot remember the proper way to build a confidence interval from this data, so he decides to use a studentized bootstrap.

(a) Show the correct way to build such a confidence interval, commenting on any assumptions necessary.

(b) Find the confidence interval Ben will produce, and show that it's wider than the proper confidence interval.

(c) Find the exact probability that a confidence interval created from Ben's method will capture the true p. Your answer may be left in terms of indicator functions and one sum.

Problem 5: Recruiting the right Baller

A small liberal arts college outside of Boston believes that their secret to basketball success is a statistician under the pseudonym Coach t. Coach t knows of two identical twins in the area who are high school juniors, Alice Baller (A. Baller) and Nota Baller. One of the twins is very good at basketball, but the other is not very good. Coach t just received 20 games of film from the Baller family, but the Baller family neglected to label whether the film was of Alice or Nota. Coach t is too embarrassed to ask which twin is in the film, but she knows that Alice scores Pois(18) points a game while Nota scores Pois(9). Coach t only has time to watch one game, and she figures the family would only send film of Alice, so she decides that unless the twin makes 12 or fewer points in the first game she'll add all the games to Alice's history. (Otherwise she'll add it to Nota's.)

(a) Find the probability of making a Type 1 and a Type 2 error with this plan. You may leave your answer as a single sum.

(b) Coach t gets a text from her friend Student t saying he's going to miss their scheduled lunch because he's too hungover on account of his "job" that he never speaks about. Coach t decides to use this extra time to watch another game and now decides that she'll add the games to Alice's history if the twin makes more than 12 points averaged across the two games. Assume that the points made in each game are independent. What are the probabilities of Type 1 and Type 2 errors now?

(c) Using a Normal approximation to the Poisson, show that every additional game Coach t watches will reduce both the chance of a Type 1 and the chance of a Type 2 error. (Note that a Normal approximation to a Poisson does not hold generally, but here we are essentially using a scaled Poisson.)

Problem 6: Imka Fused and her mixed up data

Imka Fused is scrambling to finish her final project in Stat 931. She has a data frame with two columns, and she has run a simple linear regression to predict one from the other. Unfortunately, she mixed up the column names and can't remember which is the predictor and which is the response. The deadline is minutes away and she doesn't have time to figure out which is which, but she still wants whatever she writes to be correct. Help Imka make some true claims about her data regardless of whether she used the model `response ~ predictor` or the model `predictor ~ response`.

(a) Show that the R^2s of both models are the same.

(b) Show that the overall model F-statistics are the same and that the F-test gives the same p-value in both models.

(c) Show that the t-statistic for a test of $\beta_1 = 0$ is the same in both models and that the p-values of the tests are the same.

(d) Briefly show that the $\hat{\beta}_1$ and $\text{SE}(\hat{\beta}_1)$ of each model need not be the same.

Problem 7: To add or not to add (predictors)

Read through the code for the following simulation and summarize what it is showing.

```
library(ggpubr)
set.seed(139)

nsims = 1000
n = 20
beta_1 = 2
beta_2 = 2
sigma = 2

run_simulation <- function(Sigma) {
  # Model with predictors for only X_1
  pvals_single = vector(length = nsims)
  for (i in 1:nsims) {
    x = mvrnorm(n = n, rep(0, 2), Sigma)
    y = x %*% c(beta_1, beta_2) + rnorm(n, 0, sigma)
    pvals_single[i] = summary(lm(y ~ x[,1]))$coefficients[2, 4]
  }

  # Model with predictors for X_1 and X_2
  pvals_double = vector(length = nsims)
  for (i in 1:nsims) {
    x = mvrnorm(n = n, rep(0, 2), Sigma)
    y = x %*% c(beta_1, beta_2) + rnorm(n, 0, sigma)
    pvals_double[i] = summary(lm(y ~ x))$coefficients[2, 4]
  }

  out_data = data.frame(Predictors = as.factor(c(rep(1, nsims), rep(2, nsims))),
                        pvalue = c(pvals_single, pvals_double))

  return(out_data)
}

# Version 1
Sigma = cbind(c(1, 0.5), c(0.5, 1))

out_data <- run_simulation(Sigma)

plot1 <- ggplot(out_data, aes(x = log(pvalue), fill = Predictors)) +
  geom_histogram(alpha=0.5, position="identity", bins=30) +
  theme_bw() +
  ggtitle("Version 1")

# Version 2
Sigma = cbind(c(1, 0), c(0, 1))

out_data <- run_simulation(Sigma)

plot2 <- ggplot(out_data, aes(x = log(pvalue), fill = Predictors)) +
  geom_histogram(alpha=0.5, position="identity", bins=30) +
  theme_bw() +
```

```
  ggtitle("Version 2")

# Version 3
Sigma = cbind(c(1, 0), c(0, 10))

out_data <- run_simulation(Sigma)

plot3 <- ggplot(out_data, aes(x = log(pvalue), fill = Predictors)) +
  geom_histogram(alpha=0.5, position="identity", bins=30) +
  theme_bw() +
  ggtitle("Version 3")

ggarrange(plot1, plot2, plot3, ncol = 2, nrow = 2)
```