BY SUBMITTING THIS TEST, I HEREBY PLEDGE ON MY HONOR THAT I HAVE TAKEN IT IN PERSON AND WITHOUT ASSISTANCE FROM ANY OTHER PERSON. I ACKNOWLEDGE THAT THE PENALTY FOR VIOLATING ACADEMIC INTEGRITY IS MOST SEVERE.

- This is an online test given during the Corona Virus outbreak.
- For questions that have only 1 correct response, the system will not allow multiple answers; but for questions that may have multiple responses, the system will allow you to select multiple responses.
- It is open book/web/hw/solutions/past_tests/calculator/etc., but closed collaboration.
- The questions are presented to students in random order.
- You have the option of changing your answer to an already answered question.
- The list of questions and the remaining time for the whole test appear at the bottom of each page.
- Questions 1-10 are worth 1 point and do not require much calculation or writing.
- Questions 11-13 are worth 2 points each, and require a bit more work.
- Questions 14-15 are worth about 3 points and require varying levels of calculation, all of which are to be done on paper (or Tablet), and saved/scanned/photographed, and uploaded to canvas before 3:20. For these non-multiple-choice questions, SHOW WORK. NO CREDIT FOR CORRECT ANSWER WITHOUT EXPLANATION/WORK.

1. The adjacent figure shows the histogram of 2 variables. Which one is (more) likely to have come from a Poisson dist?
   a) The black one
   b) The red one
   c) Neither one

2. Here is the class evaluation Table again!
   Let 5, 4, ..., 1, 0 denote “Excellent,” “Very Good,” etc., respectively. A meaningful notion of the mean rating of “The Course as a whole” is
   a) \( \frac{1}{6}(0.27 + 0.33 + \cdots + 0.03) \)
   b) \( \frac{1}{6}(5(0.27) + 4(0.33) + \cdots + 0(0.03)) \)
   c) \( 5(0.27) + 4(0.33) + \cdots + 0(0.03) \)
   d) none of the above.

3. Circle the correct statement(s) about qq-plots, in general (not just normal qq plots).
   a) A qq-plot can display scatter (in the same sense used in scatterplots).
   b) It is perfectly appropriate to fit an OLS curve to a qq-plot.
   c) A qq-plot may look like a parabola (as in U-shaped).
   d) If the qq-plot of x is curved (i.e., not linear), some transformation of x may yield a linear qq-plot.
   e) None of the above.

4. It can be shown that if there is no linear association between x and y, then equation for the OLS fit is \( \hat{y}(x) = \bar{y} \). Then,
   a) \( \text{SST} = 0 \)
   b) \( \text{SST} = \text{SSexplained} \)
   c) \( \text{SST} = \text{SSunexplained} \)
   d) none of the above.
5. Which step (if any) is incorrect?
   We skipped this question, because the LaTeX code didn’t work in Canvas.

6. Suppose we have transformed our data on $x$ and $y$ and have found that the model $\frac{1}{y} = \alpha + \beta \sqrt{x} + \epsilon$ is appropriate for fitting to data. Then, (circle all TRUE statements)
   a) $\hat{\beta}$ measures the average change in $y$ for 1 unit change in $x$.
   b) $\hat{\beta}$ measures the average change in $\frac{1}{y}$ for 1 unit change in $\sqrt{x}$.
   c) $R^2$ measures the proportion of the variability in $y$ that is explained by $x$.
   d) $\frac{1}{R^2}$ measures the proportion of the variability in $\frac{1}{y}$ that is explained by $\sqrt{x}$.
   e) $s_y$ measures the typical error in the prediction of $y$.
   f) $\frac{1}{y}$ measures the typical error in the prediction of $\frac{1}{y}$.

7. In a regression problem involving variables $x_1, x_2, y$, the scatterplot of $y$ vs. $x_1$ exhibits an extremely strong linear association. Suppose, we know that there exists no collinearity at all. Then, ___ linear association between $y$ and $x_2$. Hint: Think about the 3D geometry of the data.
   a) there exists an extremely strong
   b) there cannot exist an extremely strong
   c) nothing can be said about any

8. A measure of spread based on $\sum_n (x_i^2 - \overline{x}^2)^2$ should be divide by
   a) $n$
   b) $n-1$
   c) $n-2$
   d) $(n-1)^2$

9. Which of the following histograms is a good candidate for the (empirical) sampling distribution of the product of elements in a sample of size 2 taken from a Unif(0,2)?
   a) The black one
   b) The red one
   c) The blue one
   d) The green one
   e) none of the above

FYI: the peak is at zero because anything times zero is zero.

10. You want to compute the numerical value for the probability that a random sample mean will fall within an observed confidence interval. Circle the item(s) that is/are necessary and sufficient.
    a) The confidence level.
    b) The sample size.
    c) A random confidence interval.
    d) None of the above.

Important practical Moral: Even in the simplest situation where you expect/believe a linear plane to be adequate for your problem, when you're choosing what predictors to include in your study, do NOT pick the ones that have high corr. with $y$!
11. An observed 95% confidence interval for the population proportion is \((0.1, 0.3)\). Then, (circle true statement(s))
   a) there is evidence that the population proportion is 0.2.
   b) there is evidence that the population proportion is not 0.2.
   c) there is no evidence that the population proportion is 0.2.
   d) there is no evidence that the population proportion is not 0.2.

12. The \(t\)-based confidence interval is ___ (narrower/wider) than the \(z\)-based confidence interval, because there are ___ (fewer/more) sources of variability in \(t\) than in \(z\).

13. A given confidence level, if data are truly paired, then an un-paired confidence interval will cover the true ___ (meansifference between means) ___ (less/more) frequently than the paired confidence interval.

14. If you want an observed confidence interval for a population mean to cover zero, then the sample size must be ___
   a) \(\left(z^*\sigma_x/\bar{x}_{\text{obs}}\right)^2\)
   b) \(\left(z^*\sigma_x/\bar{y}_{\text{obs}}\right)^2\)
   c) \(\left(z^*\sigma_x/B\right)^2\)
   d) \(\left(z^*\sigma_x/B\right)^2\)

15. A 50% confidence interval (CI) for the population mean is designed to cover the mean 50% of the time. So, out of 100 CIs, the typical number of times a CI will cover the mean is 50. The typical deviation in that number is ___.
   a) 5
   b) 10
   c) 25
   d) 50
   e) 100

16. Starting from the expression for SSE, derive the Normal equations of regression that must be satisfied by \(\hat{\alpha}, \hat{\beta}\) for the model \(y_i = \alpha + \beta x_i + \epsilon_i\). Write the final equations in “bar” notation. Show work, but do NOT waste time to solve the equations for \(\hat{\alpha}, \hat{\beta}\).

17. It is known that \(s \sim N(\sigma_x, \sigma_x/\sqrt{2n})\), where \(s\) is the sample standard deviation based on a sample of size \(n\), and \(\sigma_x\) is the population standard deviation. Starting from a “self-evident fact” build the CI for \(\sigma_x\)

\[
S \sim N(\sigma_x, \sigma_x/\sqrt{2n}) \Rightarrow 2^* = \frac{S-\sigma_x}{\sigma_x/\sqrt{2n}} \sim N(0,1)
\]

\[
\Pr(-z^* < z < z^*) = \text{conf. level} \quad (1 \text{ point})
\]

\[
-\frac{z^*}{\sigma_x} < \frac{S-\sigma_x}{\sigma_x/\sqrt{2n}} < \frac{z^*}{\sigma_x} \quad \Rightarrow \quad \frac{1}{1-\frac{z^*}{\sigma_x}} < \sigma_x < \frac{1}{1+\frac{z^*}{\sigma_x}}
\]

\[
\text{C.I for } \sigma_x = \left[\frac{S}{1+\frac{z^*}{\sigma_x}}\right] \left[\frac{S}{1-\frac{z^*}{\sigma_x}}\right]
\]

FYI: If \(n\) is large, then \(\frac{S}{\sqrt{2n}} \approx (1 + \frac{z^*}{\sqrt{2n}})
\]

\[
\text{CI for } \sigma_x = \frac{S}{1+\frac{z^*}{\sqrt{2n}}}
\]