Math 150 - Methods in Biostatistics - Homework 1

your name here

Due: Wednesday, January 30, 2019, in class

```
knitr::opts_chunk$set(message=FALSE, warning=FALSE, fig.height=3, fig.width=5,
                      fig.align = "center")
library(tidyverse)
library(broom)
```

1. Chapter 2, A10: Use statistical software to calculate a two-sample test statistic (assuming equal variances) and find the p-value corresponding to this statistic. In addition, use software to calculate a 95% confidence interval for the difference between the two means $(\mu_1 - \mu_2)$. The end of chapter exercises will provide details on conducting this calculation by hand. If $H_0 : \mu_1 = \mu_2$ is true, the p-value states how likely that just random sampling variability would create a difference between two sample means $(\overline{y}_1 - \overline{y}_2)$ at least as large as we observed. Based on the p-value, what can you conclude about these two types of games?

```
t.test(Time ~ Type, data=games1)

t.test(Time ~ Type, data=games1) %>%
    tidy()
```

2. Chapter 2, A11: Use the software instructions to create dummy variables for game type and develop a linear regression model of the `Games1` data. Use `X = 1` to represent the color distracter game and `X = 0` represents the standard game. Develop a regression model using `Time` as the response and the indicator as the explanatory variable.

Assign the linear model (lm) to a new variable and then `tidy` the model.

```
games1 <- games1 %>%
  mutate(Type2 = ifelse(Type=="Color", 1, 0))

lm(Time ~ Type2, data=games1)

lm(Time ~ Type2, data=games1) %>%
  tidy()
```

3. Chapter 2, A12: Use statistical software to calculate the t-statistic and p-value for the hypothesis test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$. In addition, construct at 95% confidence interval for β_1. Based on these statistics, can you conclude that the coefficient β_1 is significantly different from zero?

The argument `conf.int = TRUE` inside `tidy` on the linear model will find confidence intervals for the coefficients.

4. Chapter 2, E1: Assume you are conducting a t-test to determine if there is a difference between two means. You have the following summary statistics: $\overline{x}_1 = 10, \overline{x}_2 = 20$ and $s_1 = s_2 = 10$. Without completing the hypothesis test, explain why $n_1 = n_2 = 100$ would result in a smaller p-value than $n_1 = n_2 = 16$.

5. Chapter 2, E2: If the hypothesis test $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$ results in a small p-value, can we be confident that the regression model provides a good estimate of the response value for a given value of x_i? Provide an explanation for your answer.

6. Chapter 2, E3: What model technical conditions (if any) need to be satisfied in order to calculate b_0 and b_1 in a simple linear regression model?

7. Chapter 3, E4: Explain why the model $y_i = \beta_0 + \beta_1 x_i$ is not appropriate, but $\hat{y}_i = \beta_0 + \beta_1 x_i$ is appropriate.

General notes on homework assignments (also see syllabus for policies and suggestions): - please be neat and organized, this will help me, the grader, and you (in the future) to follow your work.

- be sure to include your name on the assignment, and staple the pages together *prior* to class
- please include at least the number of the problem, or a summary of this question (this will also be helpful to you in the future to prepare for exams).
- it is strongly recommended that you write out the questions as soon as you get the assignment. This will help you to start thinking how to solve them!
- for R problems, it is required to use R Markdown
- please do not print errors, messages, warnings, or anything else that makes your homework unwieldy. You will be graded down for superfluous printouts.
- in case of questions, or if you get stuck please don't hesitate to email me (though I'm much less sympathetic to such questions if I receive emails within 24 hours of the due date for the assignment).

Homework assignments will be graded out of 5 points, which are based on a combination of accuracy and effort. Below are rough guidelines for grading.

Score & Description
5 points: All problems completed with detailed solutions provided and 75% or more of the problems are fully correct.

4 points: All problems completed with detailed solutions and 50-75% correct; OR close to all problems completed and 75%-100% correct

3 points: Close to all problems completed with less than 75% correct

2 points: More than half but fewer than all problems completed and $> 75\%$ correct

1 point: More than half but fewer than all problems completed and $< 75\%$ correct; OR less than half of problems completed

0 points: No work submitted, OR half or less than half of the problems submitted and without any detail/work shown to explain the solutions.