Math 150 - Methods in Biostatistics - Homework 1

Your Name Here

Due: Friday, February 5, 2021

Assignment Summary (Goals)

- Practice using R to run t-tests and linear models
- Providing details about what the models mean

Dataset load

I downloaded the dataset from Sakai, and read the file into R using `read_delim` with a tab-delimiter.

```
games1 <- read_delim("C2 Games1.txt", delim="\t")
games1 %>% head()
```

```
## # A tibble: 6 x 3
##   studentID Type      Time
##       <dbl> <chr>    <dbl>
## 1         1 Standard    38
## 2         2 Color       36
## 3         3 Color       42
## 4         4 Standard    35
## 5         5 Standard    32
## 6         6 Color       37
```

Looks good!

Q0. PodQ Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

This is the story of Lian and the donkey: Lian was SO excited to visit a donkey in her neighborhood. Sometimes he stands beside the road and Lian could pet him. Today, Lian wanted to meet the donkey, so she brought some carrots and tried to attract him, but alas the donkey ignored her!

Q1. Chapter 2, A10. Use R to calculate a two-sample test statistic (assuming equal variances) and find the p-value corresponding to this statistic. In addition, 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?

```r
#two-sample t test
games1 %>%
    t.test(Time ~ Type, data=., var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Time by Type
## t = 2.2862, df = 38, p-value = 0.02791
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2920254 4.8079746
## sample estimates:
##    mean in group Color mean in group Standard
##                  38.10                  35.55
```

```r
#tidy version
games1 %>%
    t.test(Time ~ Type, data=., var.equal = TRUE) %>%
    tidy()
```

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low conf.high
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl>
## 1     2.55      38.1      35.6      2.29  0.0279        38    0.292      4.81
## # ... with 2 more variables: method <chr>, alternative <chr>
```

Since the p-value is < 0.05, we can reject the null hypothesis ($H_0 : \mu_1 = \mu_2$) at the $\alpha = 0.05$ level, meaning that the **true** mean time to play the color game is significantly different from the **true** mean time to play the standard game.

Q2. Chapter 2, A11 To fit a linear model, the `Type` variable will need to be binary. Fit a linear model in R using `lm()` and notice which level of `Type` gets set to 0 and which gets set to 1. How can you tell?

Develop a regression model using `Time` as the response and the indicator on `Type` as the explanatory variable.

Create a linear model (`lm()`) and then `tidy()` the model. The following example code might help.

```r
model<-games1 %>%
  lm(Time ~ Type, data = .)

model %>%
  tidy()
```

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic  p.value
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     38.1      0.789     48.3  1.01e-35
## 2 TypeStandard    -2.55     1.12      -2.29 2.79e- 2
```

The level of `Type` that gets set to 1 is Standard, and the level of `Type` that gets set to 0 is Color. I know this because the mean time of Color (38.1) is greater than the mean time of Standard (35.55), and the regression

coefficient is negative, meaning that the smaller level of `Type` must be associated with the level 1. Also printing the model summary in tidy format indicates that `TypeStandard` is a term in the model, illustrating that `TypeStandard` is associated with level 1 (and `TypeColor` is the baseline, or level 0 variable).

Q3. Chapter 2, A12 Use R 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?

```
model<-games1 %>%
  lm(Time ~ Type, data = .)

model %>% tidy(conf.int=TRUE)
```

```
## # A tibble: 2 x 7
##   term         estimate std.error statistic  p.value conf.low conf.high
##   <chr>           <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)     38.1      0.789     48.3  1.01e-35    36.5     39.7
## 2 TypeStandard    -2.55     1.12      -2.29 2.79e- 2    -4.81    -0.292
```

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

Based on the t-test (t-stat=-2.29 and p-value=0.0279) and confidence interval ($CI_{95\%}(\beta_1) = [-4.81, -0.292]$), we can reject the null hypothesis that $\beta_1 = 0$. Thus, we can conclude that β_1 is significantly different from 0.

Q4. 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 $n_1 = n_2 = 16$.

Qualitatively, a difference in sample means=10 (under the null hypothesis of equality of means) is much less likely if the samples are of size 100 than size 16 because larger samples reduce the effect of random noise on the sample means (per the law of large numbers). In other words, drawing samples of size 100 that differ by more than 10 from populations with equal means is far less likely than if you drew samples of size 16. This yields a smaller p-value in the $n_1 = n_2 = 100$ case. Quantitatively, the t-statistic is calculated according to the following formula: $t = \frac{\overline{x}_1 - \overline{x}_2 - 0}{\sqrt{s^2\left(\frac{1}{n_1} \cdot \frac{1}{n_2}\right)}}$. Thus, for all other terms constant, supplying larger n_1, n_2 will boost the value of the t-statistic, and large magnitude t-statistics correspond to significant deviations from the null hypothesis, i.e. lower p-values.

Q5. 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.

No, the hypothesis test merely tells us whether β_1 is significantly different from 0 or not. This doesn't yield any guarantees on the accuracy of our predictions! Imagine a case with a noisy response variable; even if your linear regression model produces a β_1 parameter significantly different from 0, your predictions will still be relatively poor estimates of the response variables due to the inherent noise in the data.

In fact, a small p-value doesn't even guarantee a good estimate of β_1. To wit, in case of Question 3, we obtained a significant p-value (0.027) for the β_1 coefficient associated with game type, but a fairly wide confidence interval ($CI_{95\%}(\beta_1) = [-4.81, -0.292]$), indicating that while significantly different from 0, our estimate may not be precise.

Q6. 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?

In order to calculate meaningful b_0 and b_1 values (without violating linear model assumptions): 1. The average of the response variable $\mathbb{E}[Y]$ must be a linear function of the predictors. 2. The error terms $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, meaning the noise about the linear signal are independent, normally distributed, centered at zero, and have constant variance.

Q7. Chapter 2, 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.

$y_i = \beta_0 + \beta_1 x_i$ implies that the true response is **exactly** a linear combination of the predictors (i.e. the underlying model is deterministic). In nearly every real-world problem, there exists some noise/randomness in the response variable, which is not accounted for in the first model. Thus, the first model is not appropriate for the vast majority of real world problems. $\hat{y}_i = \beta_0 + \beta_1 x_i$ is an appropriate model, because this model admits that the linear combination of the predictors **is an estimate** of the response variable (as indicated by $\hat{y}_i$).