Homework 3, P8130

Emil Hafeez (eh2928)

10/15/2020

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
exercise_df =
  read_csv(
      "./data/exercise.csv")
```

```
## Parsed with column specification:
## cols(
##   Group = col_double(),
##   Age = col_double(),
##   Gender = col_double(),
##   Race = col_double(),
##   HTN = col_double(),
##   T2DM = col_double(),
##   Depression = col_double(),
##   Smokes = col_double(),
##   Systolic_PRE = col_double(),
##   Systolic_POST = col_double()
## )
```

Problem 1

For each question, make sure to state the formulae for hypotheses, test-statistics, decision rules/p-values, and provide interpretations in the context of the problem. Use a type I error of 0.05 for all tests.

Problem 1.a

Perform appropriate tests to assess if the Systolic BP at 6 months is significantly different from the baseline values for the intervention group.

********Check for normality, consider a bonferroni adjustment, state that it's a a paired two-tailed test using an estimate. Then provide the formulas for the test statistic needed, critical value needed, the value, and then get these. and then interpret in context.********

H0, HA, mu1, mu2, test statistic formula, critical value needed, interpret

Problem 1.a.i In order to determine an appropriate test, we check for normality using visual inspection of the plot of systolic blood pressure in the Intervention group at both baseline and endline, using raw data. This is explored in Problem 1.c.i and utilizes mean μ_{pre} and μ_{post} respectively.

We first consider the changes in the intervention group between Baseline and Endline first. We determine, given that these are the same patients with data collected at two different timepoints and we do not have reason to test a specific directionality, to use a two-sided Paired t-test.

The H_0 is that $\mu_{pre} - \mu_{post} = 0$ or $\Delta = 0$. The H_A is $\mu_{pre} - \mu_{post} \neq 0$ or $\Delta \neq 0$.

The test statistic is $t = \frac{\overline{d}-0}{s_d/\sqrt{n}}$ where $\overline{d}$ is the point estimate of the mean difference, $s_d/\sqrt{n}$ is the estimated standard error of the differences, and we use the critical value of $t_{n-1,1-\alpha/2}$. We could use a Bonferroni correction or Tukey's, considering that we will be implementing multiple significance tests, but we say it is not necessary for the case of this homework problem.

Using $t = \frac{\overline{d}-0}{s_d/\sqrt{n}} = t = \frac{-8.58-0}{17.17/\sqrt{36}} = t = \frac{-8.58}{17.17/\sqrt{36}} \approx -2.99825$. The critical value is given the the percentile of the t distribution with (n-1) degrees of freedom, `qt(0.975,35)` ≈ 2.03, such that we find evidence to reject the null hypothesis and conclude that in the intervention group, the mean systolic blood pressure at Endline is significantly different than the mean systolic blood pressure at Baseline.

Problem 1.a.ii Similarly, in order to determine an appropriate test, we check for normality using visual inspection of the plot of systolic blood pressure in the Control group at both Baseline and Endline, using raw data. This is explored in Problem 1.c.i and utilizes mean μ_{pre} and μ_{post} respectively.

We determine, given that these are the same patients with data collected at two different timepoints and we do not have reason to test a specific directionality, to use a two-sided Paired t-test.

The H_0 is that $\mu_{pre} - \mu_{post} = 0$ or $\Delta = 0$. The H_A is $\mu_{pre} - \mu_{post} \neq 0$ or $\Delta \neq 0$.

The test statistic is $t = \frac{\overline{d}-0}{s_d/\sqrt{n}}$ where $\overline{d}$ is the point estimate of the mean difference, $s_d/\sqrt{n}$ is the estiamted standard error of the differences, and we use the critical value of $t_{n-1,1-\alpha/2}$. We could use a Bonferroni correction, Tukey's or others, considering that we will be implementing multiple significance tests, but we say it is not necessary for the case of this homework problem.

Using $t = \frac{\overline{d}-0}{s_d/\sqrt{n}} = t = \frac{-3.33-0}{14.81/\sqrt{36}} \approx -1.3491$. The critical value is given the the percentile of the t distribution with (n-1) degrees of freedom, `qt(0.975,35)` ≈ 2.03, such that we fail to reject the null hypothesis and conclude that in the Control group, the mean systolic blood pressure at Endline is not significantly different than the mean systolic blood pressure at Baseline.

Problem 1.b Now, we assess the systolic blood pressure absolute changes between the two groups using an independent, two-sampled t-test. We assume independence based on the given information on the sampling mechanism. Since we do not know the two population variances, we first check for equality of variances.

The statistic is given by F= $\frac{s_1^2}{s_2^2}$, where the null hypothesis is $\sigma_1^2 = \sigma_2^2$, and the alternate hypothesis is $\sigma_1^2 \neq \sigma_2^2$. In our case, we use F= $\frac{14.81^2}{17.17^2} = 0.74399$ and the critical value is `qf(0.975, 35, 35)`$= 1.961$, such that we fail to reject the null hypothesis and conclude that the variance of sample 1 is not significantly different from the variance of sample 2.

We use the pooled estimate of the variance (and associated standard deviation) from two independent samples given by: $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)} = s^2 = \frac{(35)s_1^2+(35)s_2^2}{68} = \frac{(35)(14.81^2)+(35)(17.17^2)}{70} = 257.0725$ and $s = \sqrt{257.0725} = 16.0335$

So, we can proceed with the independent two sampled t-test assuming equal variances.

Given the null hypothesis H_0: $\mu_{control} = \mu_{intervention}$ and the alternate, H_A: $\mu_{control} \neq \mu_{intervention}$, we use $\text{t} = \frac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$ under H_0

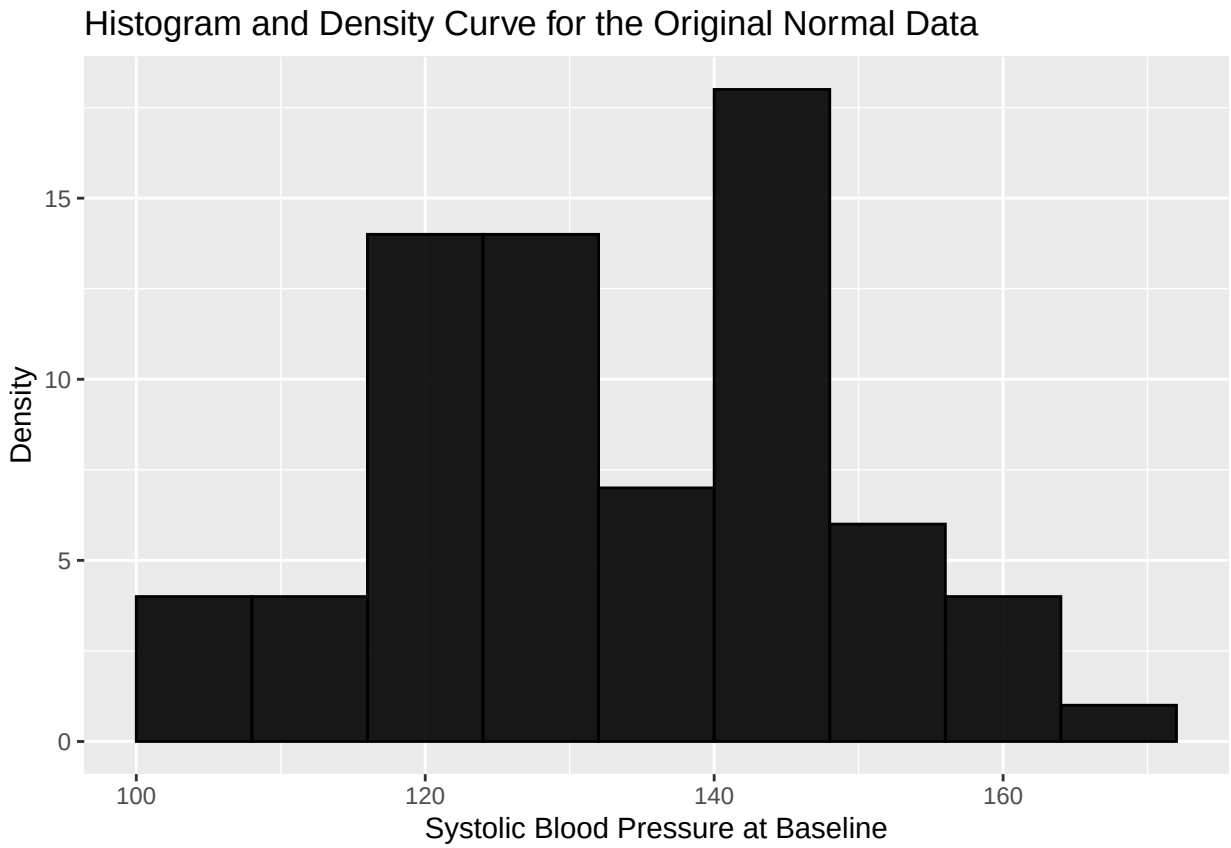
Computing this, we have $\text{t} = \frac{-3.33-(-8.58)}{s\sqrt{\frac{1}{36}+\frac{1}{36}}} = \frac{5.25}{s\sqrt{0.055}} = 1.3892$.

We can also use `std_pooled<-sqrt(((14.81^2*35)+(17.17^2*35))/70)` and `t_stats<-(-3.33-(-8.58))/(std_pooled*sq`

The critical value is given by $t_{n_1+n_2-2,0.975}$ = `qt(0.975,70)` = 1.994437.

Problem 1.c

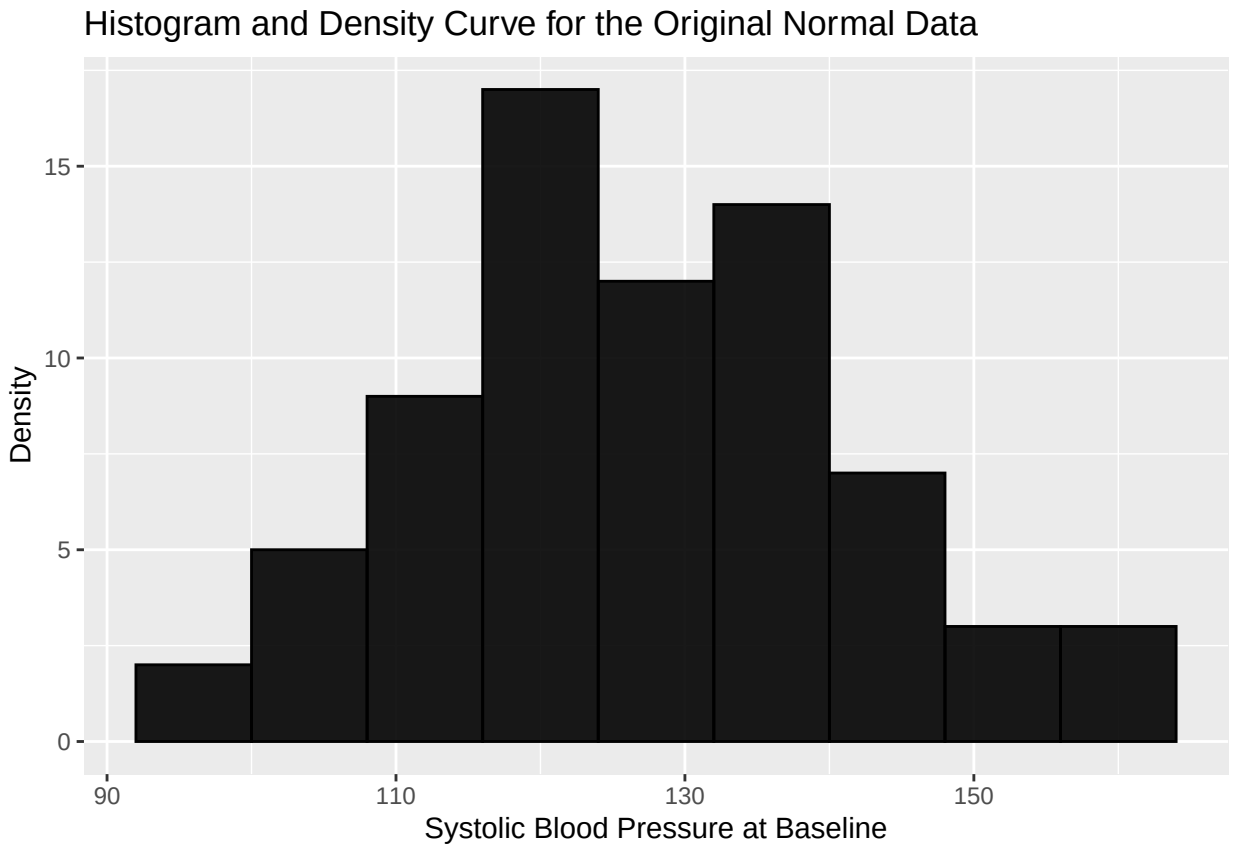
```
#plot of intervention at baseline to examine normality
exercise_df %>%
  ggplot(aes(x = Systolic_PRE)) +
  geom_histogram(binwidth = 8, fill = "black", colour = "black", alpha = 0.9) +
  labs(
    x = "Systolic Blood Pressure at Baseline",
    y = "Density",
    title = "Histogram and Density Curve for the Original Normal Data") +
  scale_fill_viridis_d("") +
  theme(legend.position = "none")
```



Problem 1.c.i

```
#plot of intervention at endline to examine normality
exercise_df %>%
  ggplot(aes(x = Systolic_POST)) +
  geom_histogram(binwidth = 8, fill = "black", colour = "black", alpha = 0.9) +
  labs(
```

```
    x = "Systolic Blood Pressure at Baseline",
    y = "Density",
    title = "Histogram and Density Curve for the Original Normal Data") +
  scale_fill_viridis_d("") +
  theme(legend.position = "none")
```



DISCUSS #### Problem 1.c.ii