Homework 3, P8130

Emil Hafeez (eh2928)

10/15/2020

```r
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()
```

```r
library("animation")


theme_set(theme_minimal() + theme(legend.position = "bottom")) #setup and establish the colors schemes f
options(
  ggplot2.continuous.colour = "viridis",
  ggplot2.continuous.fill = "viridis"
)

scale_colour_discrete = scale_colour_viridis_d
scale_fill_discrete = scale_fill_viridis_d

exercise_df =
  read_csv(
      "./data/exercise.csv") #load in df
```

```
## 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.

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 changes in the Intervention group using raw data. This is explored in Problem 1.c.i and utilizes mean μ_{pre} and μ_{post} respectively. We consider the changes in systolic blood pressure 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 $t = \frac{\bar{d}-0}{s_d/\sqrt{n}}$ where $\bar{d}$ 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, Tukey's, or other correction, considering that we will be implementing multiple significance tests, but we say it is unnecessarily conservative for the case of this homework problem. Additionally, the standard deviation for the difference between Baseline and Endline is given and thus the variance is known.

Using $t = \frac{\bar{d}-0}{s_d/\sqrt{n}} = t = \frac{-8.58-0}{17.17/\sqrt{36}} = t = \frac{-8.58}{17.17/\sqrt{36}} \approx -2.99825$. 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 changes in the Control group 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 $t = \frac{\bar{d}-0}{s_d/\sqrt{n}}$ where $\bar{d}$ 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, 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. Additionally, the standard deviation for the difference between Baseline and Endline is given and thus the variance is known.

Using $t = \frac{\bar{d}-0}{s_d/\sqrt{n}} = t = \frac{-3.33-0}{14.81/\sqrt{36}} \approx -1.3491$. 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. Since we do not know the two population variances, we first check for equality of variances.

The statistic is given by F= $\frac{{s_1}^2}{{s_s}^2}$, 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: $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)} = s^2 = \frac{(35)s_1^2+(35)s_2^2}{70} = \frac{(35)(14.81^2)+(35)(17.17^2)}{70} = 257.0725$ 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 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 t = $\frac{-3.33-(-8.58)}{s\sqrt{\frac{1}{36}+\frac{1}{36}}} = \frac{5.25}{s\sqrt{0.055}} = 1.3892$. The critical value is given by $t_{n_1+n_2-2,0.975}$ = `qt(0.975,70)` = 1.994437. Therefore, we fail to reject the null hypothesis and conclude that the difference in mean systolic blood pressure change from Baseline to Endline in the Control group is not significantly different from the mean systolic blood pressure change from Baseline to Endline in the Intervention group.

For the confidence interval,

$$\left(\overline{X_1} - \overline{X_2} - t_{n_1+n_2-2,1-\alpha/2} \cdot s \cdot \sqrt{1/n_1 + 1/n_2}, \overline{X_1} - \overline{X_2} + t_{n_1+n_2-2,1-\alpha/2} \cdot s \cdot \sqrt{1/n_1 + 1/n_2}\right) \tag{1}$$

This is to say, the lower limit of the 95% CI is 5.25 - $t_{n_1+n_2-2,1-\alpha}$ * $s\sqrt{\frac{1}{36}+\frac{1}{36}} = -2.2872$ and the upper limit of the confidence interval is (-3.33 - (-8.58)) $+t_{n_1+n_2-2,1-\alpha} * s\sqrt{\frac{1}{36}+\frac{1}{36}} = 12.7872$, thus the 95% CI is approximately (-2.2872, 12.7872). This provides further evidence for the decision to fail to reject the null hypothesis, because the CI contains the null hypothesized value of 0.

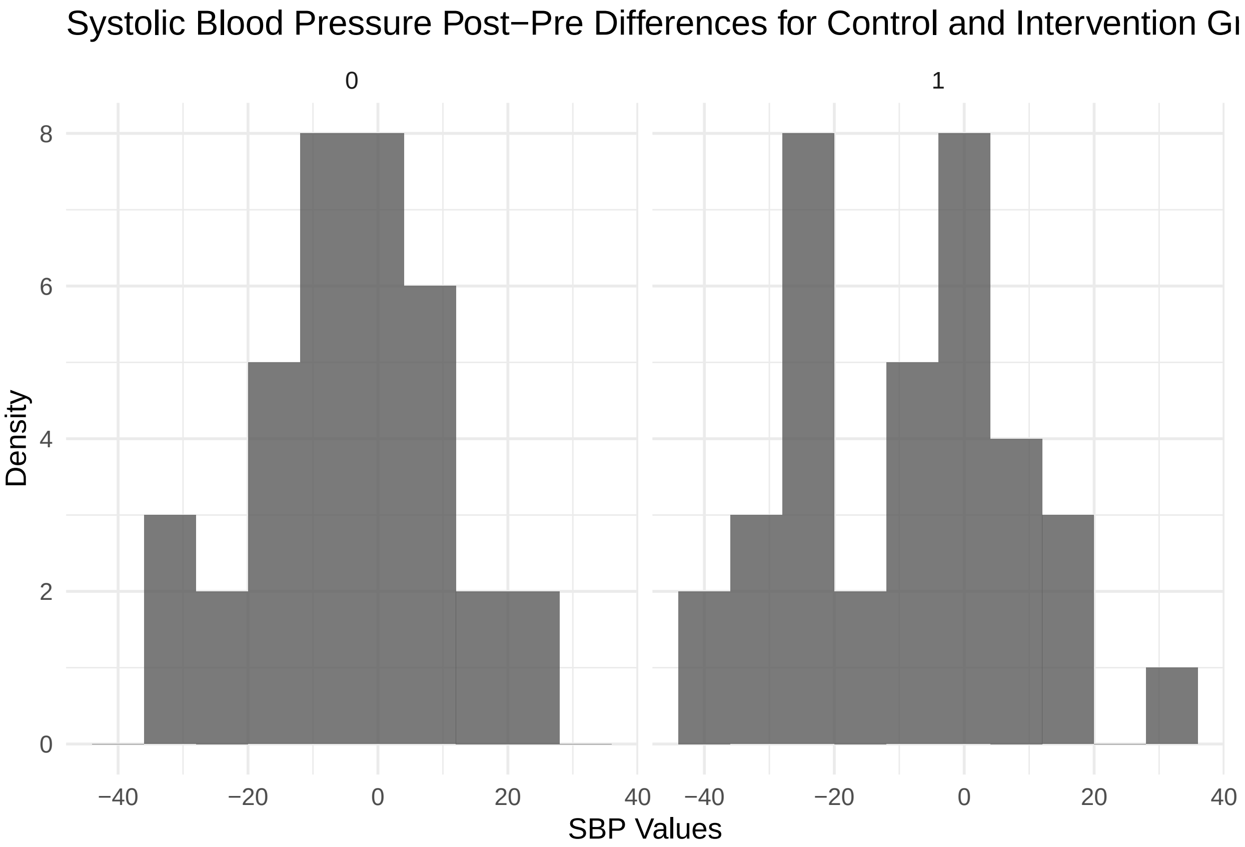
Problem 1.c

Problem 1.c.i The main assumptions for part a) the two-sided paired t-test, are that we assume the systolic blood pressure measurements are normally distributed with mean $\mu_{baseline}$ $\mu_{endline}$, and that these differences also follow a normal distribution. We therefore also assume that the observed SBP differences constitute a random sample from this normally distributed population of differences. We're assuming the sampled patients have been drawn randomly, and also that each of their sampling is independent from another. We also assume that the variance is known.

First, for the Paired two-sided t-test: we examine the distribution of blood pressure differences between baseline and endline (post-pre) within the treatment and control groups (separately).

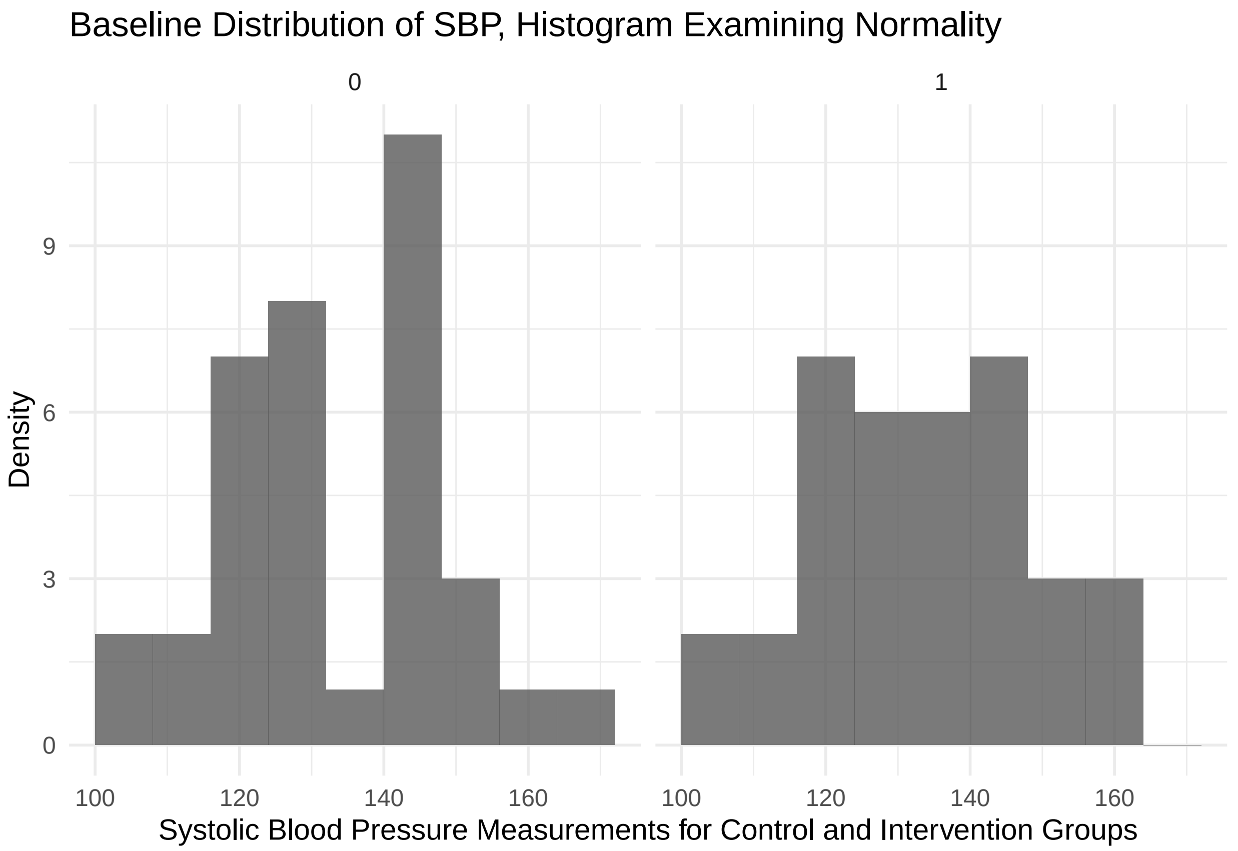
We can see that, for the Paired test of SBP change in the control group, it appears the distribution of SBP differences in the Control group appear approximately normally distributed, though not well-centered on zero. The intervention group does not appear to be normally distributed.

```
#plot of group data differences to examine normality
exercise_df %>%
  ggplot(aes(x = Systolic_POST - Systolic_PRE)) +
  geom_histogram(binwidth = 8, alpha = 0.8) +
  facet_grid(. ~ Group) +
  labs(
    x = "SBP Values",
    y = "Density",
    title = "Systolic Blood Pressure Post-Pre Differences for Control and Intervention Groups") +
  scale_fill_viridis_d("") +
  theme(legend.position = "none")
```

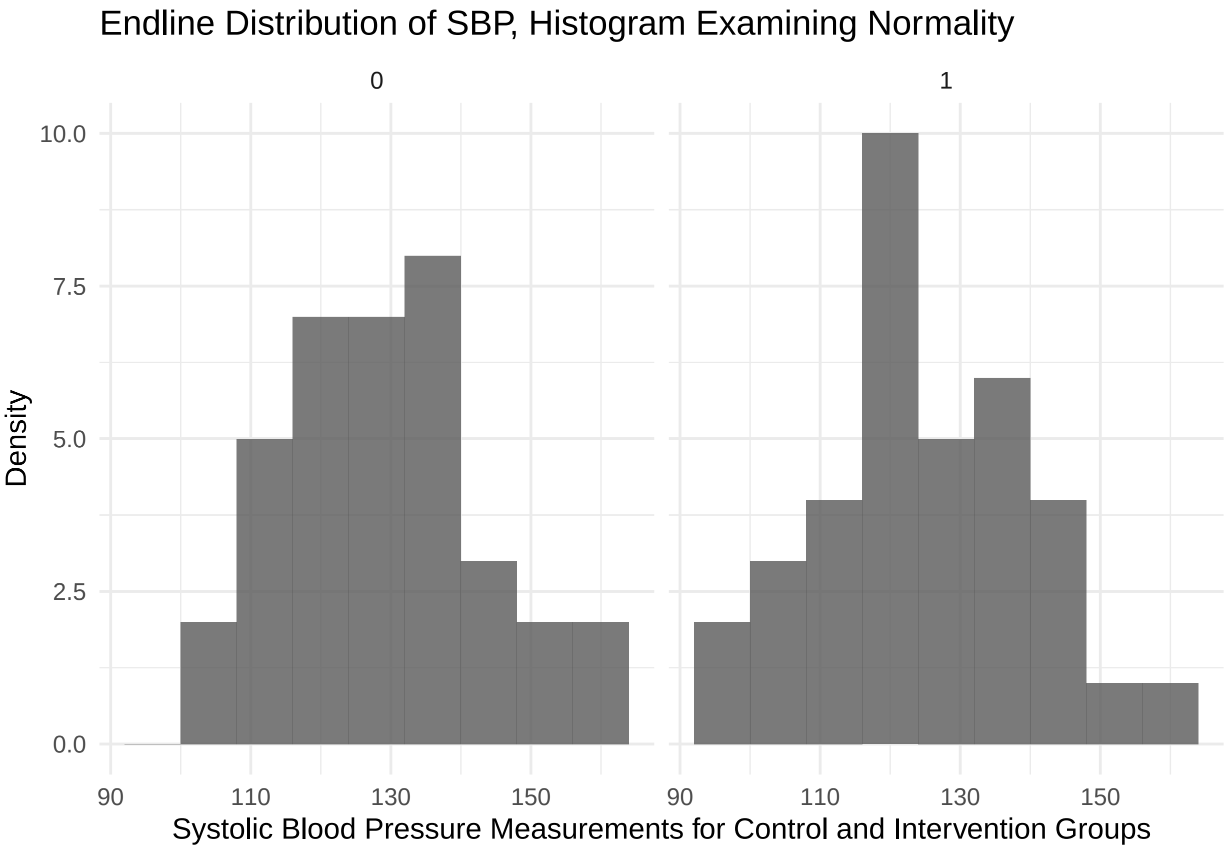


For the two-sample independent two-sided t-test, the main assumptions for part b) assumes similarly as part a) regarding independence and normality, and then regarding variance instead assumes the homogeneity of variance (which we tested with an F-test). For this test, we examine that the SBP is approximately normally distributed within each group and timepoint, and we examine the distributions of Post-Pre score differences within each group.

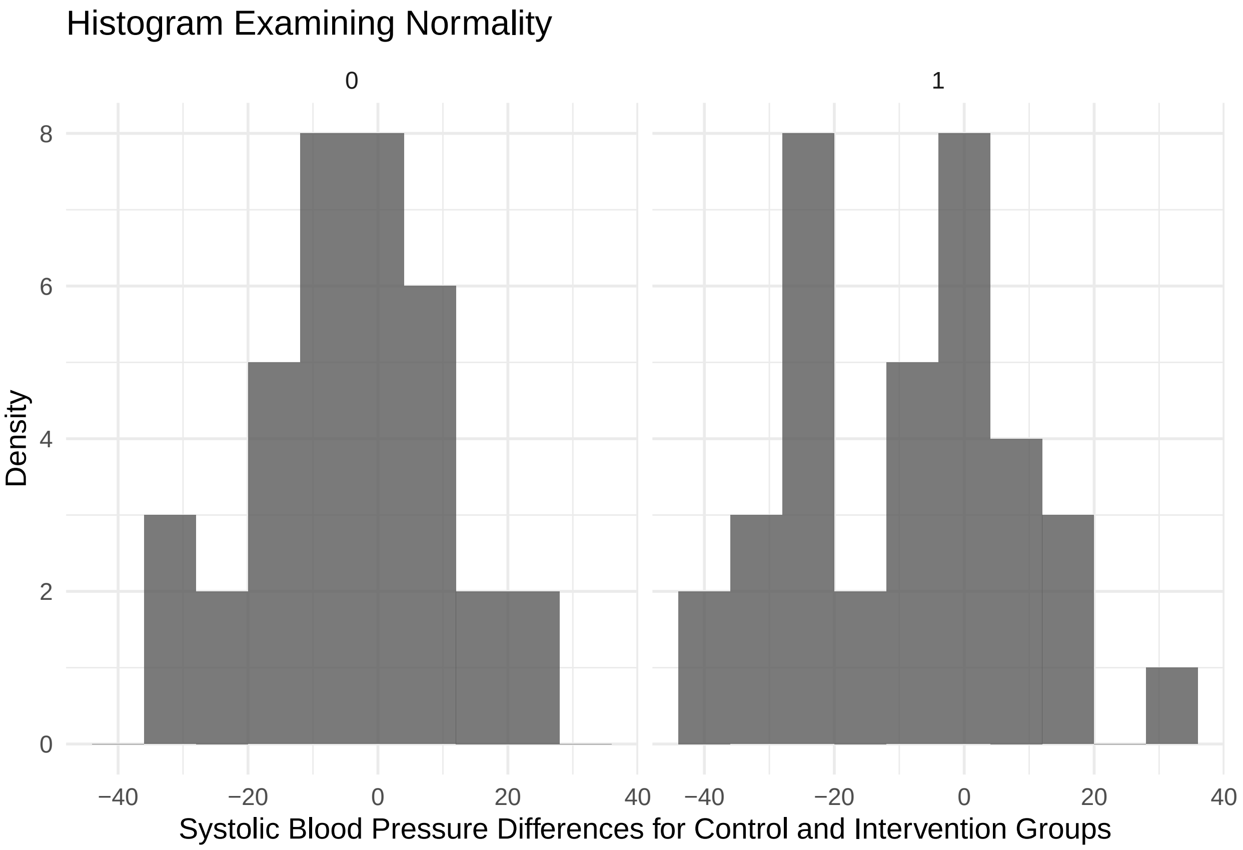
```
#plot of group data differences to examine normality
exercise_df %>%
  ggplot(aes(x = Systolic_PRE)) +
  geom_histogram(binwidth = 8, alpha = 0.8) +
  facet_grid(. ~ Group) +
  labs(
    x = "Systolic Blood Pressure Measurements for Control and Intervention Groups",
    y = "Density",
    title = "Baseline Distribution of SBP, Histogram Examining Normality") +
  scale_fill_viridis_d("") +
  theme(legend.position = "none")
```



```
exercise_df %>%
  ggplot(aes(x = Systolic_POST)) +
  geom_histogram(binwidth = 8, alpha = 0.8) +
  facet_grid(. ~ Group) +
  labs(
    x = "Systolic Blood Pressure Measurements for Control and Intervention Groups",
    y = "Density",
    title = "Endline Distribution of SBP, Histogram Examining Normality") +
  scale_fill_viridis_d("") +
  theme(legend.position = "none")
```



```
exercise_df %>%
  ggplot(aes(x = Systolic_POST - Systolic_PRE)) +
  geom_histogram(binwidth = 8, alpha = 0.8) +
  facet_grid(. ~ Group) +
  labs(
    x = "Systolic Blood Pressure Differences for Control and Intervention Groups",
    y = "Density",
    title = "Histogram Examining Normality") +
  scale_fill_viridis_d("") +
  theme(legend.position = "none")
```



Problem 1.c.ii Unfortunately, since normality of the SBP and its differences between Post and Pre measurements is questionable, this brings the tests' validity into question, as well as brings doubt to the additional assumption that the underlying distribution of the difference between Baseline and Endline scores is normal. This needs to be remedied. One way would be beginning with the Central Limit Theorem's recommendation and conduct the experiment with a larger sample such that the distribution of the sample means will be approximately normally distributed (which holds true even if the population SBP distribution is not normal). While 36 measurements per group is not small, it was not sufficient. Additionally, we could examine if there are measurement errors or outliers in the data that may be uncharacteristically skewing the distribution (though this doesn't seem to be the case). We could additionally attempt transformation of the data (for example, by taking the logarithm or square root), though this may not be effective. Finally, I would suggest the alternative is to use an Exact method rather than assume normality, and calculate the confidence interval.

Problem 2

In this exercise you learn how to create your own 'true' scenario, simulate corresponding data, and quantify the type I error over many repetitions.

Problem 2.a

Given scenario is that the average IQ score of Ivy League Colleges is 120 and assume this to be the null hypothesis $H_0 : \mu = \mu_0 = 120$. The alternate hypothesis is that the true mean is less than 120, $H_A : \mu < 120$.

The standard deviation and alpha are given as $\sigma = 15$ and $\alpha = 0.05$.

$$X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2)$$

Firstly, we generate one random sample of size n=20 using `rnorm(20, mean = 120, sd = 15)`.

```
set.seed(4)
rnorm_20 = rnorm(20, mean = 120, sd = 15)
mean_rnorm20 = mean(rnorm_20)
```

We are assuming the normal distribution and compute the test statistic as given by $Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$, such that

$$Z = \frac{\overline{X} - 120}{15/\sqrt{20}}$$

$= 0.08414241$. The critical value is given by $t_{n-1,1-\alpha}$ = `qnorm(0.95,lower.tail = TRUE)` ≈ 1.644854 such that we fail to reject the null hypothesis and conclude that the true mean is not less than 120 units.

Problem 2.b

Using qnorm(0.05), we run a simulation of 100 random samples of size n=20 from the underlying null distribution, where for each sample we run a one-sided t-test, and record the result.

```
sampled = rep(NA, 100) #creates a vector for the samples, filling with NAs, so that these can be replac
decision = rep(NA,100) #similar, for the 0 or 1 for the decision to reject/fail to reject the null

for (i in 1:100) {
  set.seed(i) #this is flexible and I think can be anything (including any constant number?)
  sampled = rnorm(20, mean = 120, sd = 15) #take a random sample, each time the loop repeats (made cons
  t_stats = (mean(sampled) - 120) / (sd(sampled)/sqrt(20)) #generate the tstat by using the mean and sd
  decision[i] = ifelse(t_stats < qnorm(0.05), 1, 0) # replace each of the 100 NAs in the vector made ea
 }
mean(decision)
```

```
## [1] 0.05
```

There is a 95% rate of 1's and 5% rate of 0's produced by this simulation, suggesting that in 5% of the simulations we reject the null hypothesis which, since we know the true population is 120 IQ points, this is a Type I error rate.

Problem 2.c

Similarly, using qnorm(0.05), we run a simulation of 1000 random samples of size n=20 from the underlying null distribution, where for each sample we run a one-sided t-test, and record the result.

```
sampled = rep(NA, 1000)
decision = rep(NA,1000)

for (i in 1:1000) {
  set.seed(i)
  sampled = rnorm(20, mean = 120, sd = 15)
  t_stats = (mean(sampled) - 120) / (sd(sampled)/sqrt(20))
  decision[i] = ifelse(t_stats < qnorm(0.05), 1, 0)
 }
mean(decision)
```

```
## [1] 0.056
```

There is a 5.6% rate of 1's and 94.4% rate of 0's produced by this simulation, suggesting that in 5.6% of the simulations we reject the null hypothesis which, since we know the true population is 120 IQ points, this is a Type I error rate.

Problem 2.d

The type I error rate in part b) was 5% and in part c) was 5.6%. While we would normally anticipate that the type I error rate should be closer and closer to the theoretically imposed value of 0.05 as the sampling size increases, this is not always the result that we observe, due to chance.