Practice Class Test2 Template

2416963E

Question 1

Large samples of iron ore were mined from quarries "x" and "y", and each of the two samples were broken down into 10 smaller sub-samples for analysis. Each of these 20 sub-samples were sent to a chemical laboratory, and the percentage of iron in each sub-sample was measured. These data are stored in `practice1.csv`. Using bootstrapping, is there evidence in these data that the population mean iron percentage in each quarry is 35%, and are the population mean percentages different between the two quarries?

```
q1data <- read_csv("practice1.csv")
data1 <- gather(data = q1data,
                key = ore,    ## select column to collapse
                value = percent, ## column name for values
                1:2)
```

Part (a) asks if there is evidence in these data that the population mean iron percentage in each quarry, μ_x and μ_y, is 35% using bootstramp sampling. We can test each of these with a 95% confidence interval on the bootstrap sample means, $\bar{x}_x$ and $\bar{x}_y$ respectively.

```
samples_x <- data1 %>%
  filter(ore == "x") %>%
  specify(formula = percent ~ NULL) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

samples_y <- data1 %>%
  filter(ore == "y") %>%
  specify(formula = percent ~ NULL) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")

x_bar <- data1 %>%
  filter(ore == "x") %>%
  summarize(stat = mean(percent))
y_bar <- data1 %>%
  filter(ore == "y") %>%
  summarize(stat = mean(percent))

px <- samples_x %>%
  visualize(obs_stat = x_bar)
py <- samples_y %>%
  visualize(obs_stat = y_bar)

ci_x <- samples_x %>%
  get_ci(level = 0.95, type = "percentile")
ci_y <- samples_y %>%
  get_ci(level = 0.95, type = "percentile")

#px <- samples_x %>%
#  visualize(endpoints = ci_x, direction = "between")
#py <- samples_y %>%
```

```r
#  visualize(endpoints = ci_y, direction = "between")

# Bootstrap sample is close to symmetric and bell-shaped (Normal) so we can use the Standard Error meth
ci_x <- samples_x %>%
  get_ci(level = 0.95, type = "se", point_estimate = x_bar)
ci_y <- samples_y %>%
  get_ci(level = 0.95, type = "se", point_estimate = y_bar)

px <- samples_x %>%
  visualize(endpoints = ci_x, direction = "between")
py <- samples_y %>%
  visualize(endpoints = ci_y, direction = "between")
```

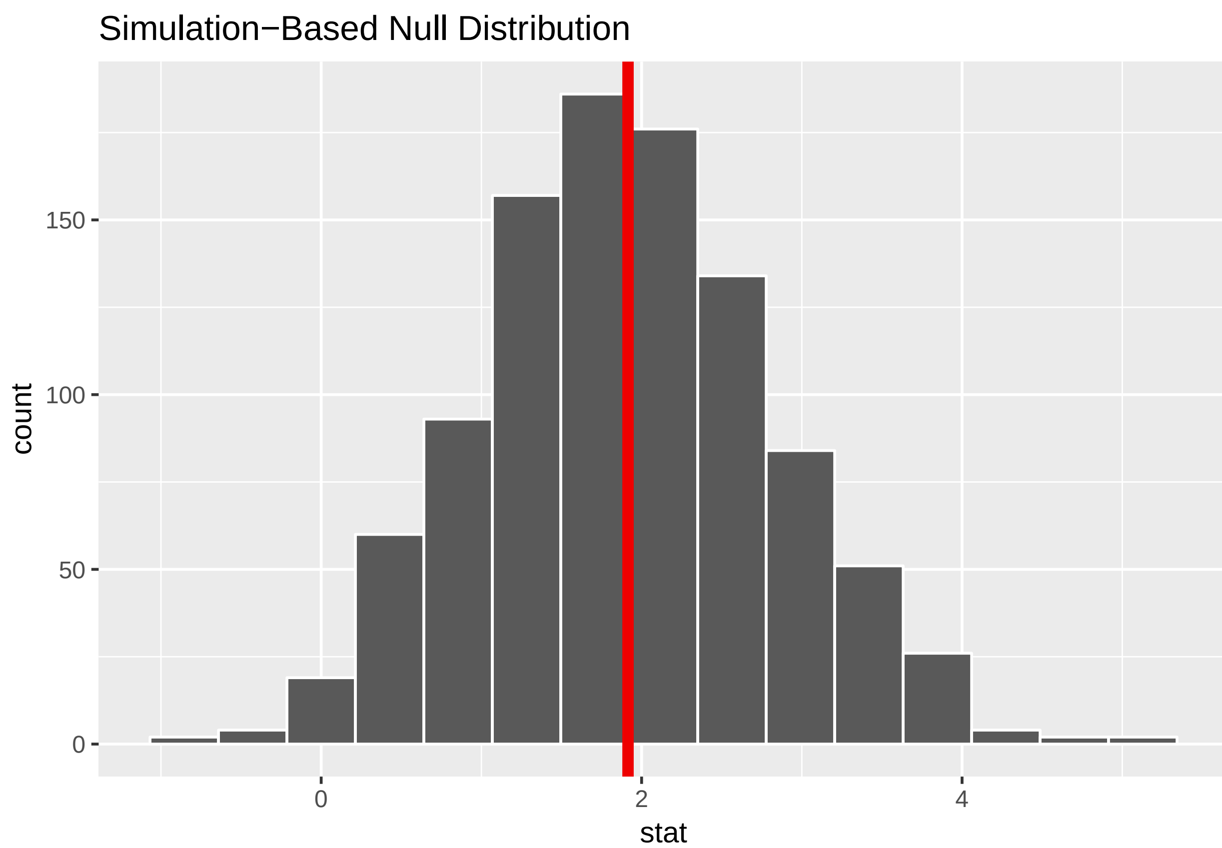
Since the distribution of bootstrap samples of x and y are both close to symmetric and bell-shaped a 95% confidence interval was created using the standard error method. The 95% confidence interval for x is [36.5, 38.58]. The confidence interval does not include 0.35 (35%) and we conclude that there is not evidence in these data that the population mean iron percentage in quarry x is 35%. For quarry y the 95% confidence interval is [34.17, 37.08]. The confidence interval includes 0.35 and we conclude there is evidence in these data that the population mean iron percentage in quarry y is 35%.

Part (b) asks if there is evidence in these data that the population mean is difference between the two quarries.

```r
set.seed(222)

samples_xy <- data1 %>%
  specify(formula = percent ~ ore) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("x", "y"))


samples_xy %>%
  visualize(obs_stat = x_bar - y_bar)
```




```
ci_xy <- samples_xy %>%
  get_ci(level = 0.95, type = "percentile")
```

The bootstrap samples are created by subtracting y from x. The 95% confidence interval is made using the percentiles method and is found to be [0.24, 3.75], which contains zero. Since the confidence interval contains zero we conclude that there is no evidence that the two quarries are different.

Question 2

(a) Using the data contained in `practice2.csv` build a regression model that adequately describes the `response` in terms of the potential explantory variables `X1`, `X2`, `X3` and `X4`. Your chosen model will therefore be the one that you believe best represents the `response`. Use only the theoretical confidence intervals generated under standard assumptions (which you should check) to identify the correct model. Note you **do not** need to consider interaction terms.

First, a pairs plot, Figure **??**, is made to explore the data visually and see if there are any patterns or correlations in the data. The data consists of

(b) Construct a table of **all** the possible linear models (without interactions or transformations) that could be fitted to the `response` variable in `practice2.csv`. In the table include the R^2_{adj} and AIC values for model comparisons. Do these measures lead you to the same conclusion about the model that best represents the data as in part (a)? Note: you are **not** required to check assumptions for each of these models in this task.

Table 1: Parameter estimates obtained from the model response X1 + X2 + X3 + X4

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	30.992	1.049	29.533	0.000	28.931	33.054
X1	0.050	0.459	0.108	0.914	-0.852	0.951
X22	-3.505	1.215	-2.885	0.004	-5.891	-1.118
X23	-2.352	1.482	-1.587	0.113	-5.264	0.561
X3	1.218	0.456	2.671	0.008	0.322	2.113
X4	0.952	0.466	2.045	0.041	0.038	1.867