Assignment 4: Statistical inference

Akshat Valse

Due Saturday, July 22 at 8am

Please upload the `PDF` that you obtain by knitting the `Rmd` file that contains your `R` code and your text answering other questions. So this uploaded file will also show any output that `R` produces in addition to your code.

You should not touch the starter code as it will print out the necessary data frames and results for grading purposes.

```
set.seed(123)
```

Part 1. Bootstrap: true or false

For this problem, indicate whether the statement is true or false, and **briefly explain or justify your answer**.

Exercise 1.1

Let $X_1, \ldots, X_{50}$ be independent observations generated from $\mathcal{N}(\mu, \sigma)$. This means, we generate 50 observations from a Normal distribution with parameters mean μ and standard deviation σ. We denote the sample average of our 50 generated samples $X_1, \ldots, X_{50}$ by $\bar{X}$.

True or false: $\bar{X}$ is likely to be off from μ by approximately $\sigma/\sqrt{50}$, just due to random error. True as per the central limit theorem the random error can be approximated by dividing the standard deviation by the number of samples used to create the sample mean. Increasing this number not only makes the sample mean distribution more normal, it also reduces the standard error.

Exercise 1.2 - Exercise 1.4

For Exercises 1.2 to 1.4, we will be using the following set up: Let X_{ik} be independent observations sampled from $\mathcal{N}(\mu, \sigma)$, for $i = 1, 2, \ldots, 100$ and $k = 1, 2, \ldots, 50$. You can interpret this as: we have 100 trials, each with sample size = 50. The index i represents the trial repetition we are on right now, and the index k represents the number of observation we are on in the ith trial.

Let

$$\bar{X}_{(i)} = \frac{1}{50} \sum_{k=1}^{50} X_{ik}$$

$$s^2_{(i)} = \frac{1}{50} \sum_{k=1}^{50} \left(X_{ik} - \bar{X}_{(i)}\right)^2$$

$$\bar{X}_{\text{ave}} = \frac{1}{100}\sum_{i=1}^{100} \bar{X}_{(i)}$$

$$V = \frac{1}{100}\sum_{i=1}^{100} \left(\bar{X}_{(i)} - \bar{X}_{\text{ave}}\right)^2$$

Exercise 1.2

True or false and briefly explain: $\{\bar{X}_{(i)} : i = 1, \ldots, 100\}$ is a sample of size 100 from $\mathrm{N}(\mu, \sigma/\sqrt{50})$. True as it is a sample of size 100 as we can interpret it as having 100 trials each with sample size of 50. The reason these samples are from $\mathrm{N}(\mu, \sigma/\sqrt{50})$ is due to the fact that the central limit theorem dictates that the sample mean is equivalent to the population mean however the standard deviation of the sample means is equal to the population standard deviation dived by the number of samples which in this case is 50.

Exercise 1.3

True or false and briefly explain: $|\bar{X}_{(i)} - \bar{X}_{\text{ave}}| < 2\sqrt{V}$ for about 95% of the i's. True as the absolute difference between each Xi and Xave will be less than 2xsqrt(V) for 95% of Xi's because sqrt(V) is the computation for the standard deviation and multiplying it by two is equal to two standard deviations which encompasses about 95% of the normal distribution.

Exercise 1.4

True or false and briefly explain: $\bar{X}_{\text{ave}}$ follows $\mathrm{N}(\mu, \sigma/\sqrt{5000})$. True. Given the setup and the definition of Xave and V from the provided expressions, we can determine the distribution of X¯ave. Since Xave is the mean of 100 independent sample means, each with a normal distribution N(μ, σ/√50), the Central Limit Theorem applies. According to the Central Limit Theorem, the distribution of the sample mean of a sufficiently large number of independent and identically distributed random variables will be approximately normally distributed, regardless of the original distribution. As the sample size is 100 and the observations are from a normal distribution N(μ, σ), the distribution of Xave will follow N(μ, σ/√5000). The standard deviation of the sample mean Xave is the population standard deviation σ divided by the square root of the total number of observations (100 trials * 50 observations per trial), which is √5000. Hence, Xave follows N(μ, σ/√5000).

Part 2. Assumptions in bootstrapped inference

In this question, we will further investigate the (implicit) assumptions that we make about our data in the bootstrap procedure.

Now days, it is not uncommon for people to apply data analytic techniques to any dataset (without thinking carefully about what assumptions are made and whether these assumptions are satisfied in reality). There are many websites that investigate interesting correlations (http://tylervigen.com/spurious-correlations) between seemingly unrelated things. For example, For example, the site Spurious Correlations (http://tylervigen.com/view_correlation?id=1703) shows a high correlation between the divorce rate in Maine and per capita consumption of margarine in the U.S.

```
# from http://tylervigen.com/view_correlation?id=1703
divorce_data = c(5, 4.7, 4.6, 4.4, 4.3, 4.1, 4.2, 4.2, 4.2, 4.1)
margarine_data = c(8.2, 7, 6.5, 5.3, 5.2, 4, 4.6, 4.5, 4.2, 3.7)
```

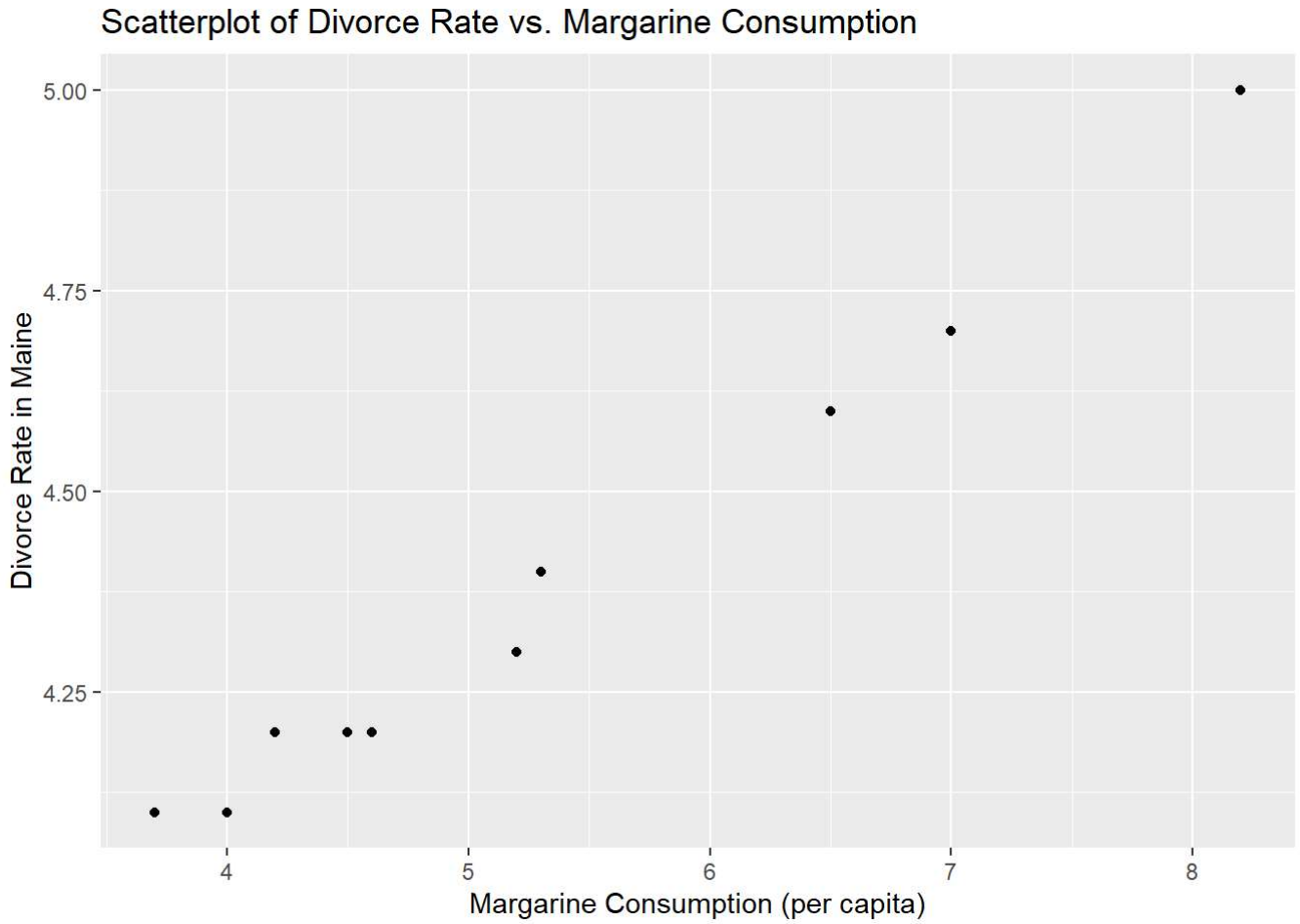
Exercise 2.1

Make a scatterplot of `divorce_data` vs. `margarine_data` . Using the plot, briefly describe whether the two variables appear related. Your plot should be adequately titled and labeled.

```
### YOUR CODE HERE
library(ggplot2)

# Create a data frame
df <- data.frame(divorce = divorce_data, margarine = margarine_data)

ggplot(df, aes(x = margarine, y = divorce)) +
  geom_point() +
  labs(title = "Scatterplot of Divorce Rate vs. Margarine Consumption",
       x = "Margarine Consumption (per capita)",
       y = "Divorce Rate in Maine")
```




```
### END OF YOUR CODE
```

YOUR EXPLANATION HERE. There seems to be a strong positive linear correlation between these two variables.

Exercise 2.2

Compute the sample correlation between `divorce_data` and `margarine_data` using `cor` . Then, use the bootstrap procedure to provide an estimate of the standard error of your estimated sample correlation coefficient.

```
### YOUR CODE HERE
# Compute the sample correlation
correlation <- cor(divorce_data, margarine_data)

# Number of bootstrap samples
n_bootstrap <- 100000

# Empty vector to store bootstrap estimates
bootstrap_correlations <- numeric(n_bootstrap)

# Bootstrap procedure
for (i in 1:n_bootstrap) {
  # Randomly sample with replacement
  bootstrap_sample <- sample(1:length(divorce_data), replace = TRUE)
  bootstrap_correlations[i] <- cor(divorce_data[bootstrap_sample], margarine_data[bootstrap_samp
le])
}

# Estimate the standard error of the sample correlation
se_estimate <- sd(bootstrap_correlations)

z <- qnorm(0.975, lower.tail = T)

# Compute the confidence interval
lower_bound <- correlation - z * se_estimate
upper_bound <- correlation + z * se_estimate

cat("95% Confidence Interval: [", lower_bound, ", ", upper_bound, "]\n")
```

```
## 95% Confidence Interval: [ 0.9726875 ,  1.012429 ]
```

```
### END OF YOUR CODE
```

Exercise 2.3

Report the $\pm\mathbf{qnorm}(0.975, \text{lower. tail} = \text{T}) \times \text{SE}$ confidence interval. Does your confidence interval contain 0? Hence, what can you conclude about whether the divorce rate and margarine consumption are related, based on your answer to whether the confidence interval contains 0?

YOUR EXPLANATION HERE. The confidence interval does not contain 0 and we can say that the divorce rate to margarine consumption are well correlated however we cannot infer causation based on this.

Exercise 2.4

In this case, we have good reasons to believe that the seeming correlation is actually *spurious*. Think of a mechanism by which such spurious correlations might easily arise and describe it briefly below.

YOUR EXPLANATION HERE. The observed correlation between divorce rate and margarine consumption is not meaningful, and any apparent relationship is due to chance. Spurious correlations can easily arise due to coincidental patterns in data. In this case, the high correlation between the divorce rate in Maine and per capita

consumption of margarine in the U.S. is likely spurious because there is no plausible causal relationship between these two variables. It is an example of correlation without causation, where both variables might be influenced by some third variable or factors unrelated to each other.

Exercise 2.5

In order to trust your answer in Exercise 2.3, what assumption(s) did you make (implicitly) when 1) you used the bootstrap to estimate the standard error, and 2) computed the confidence interval as specified in 2.3? Do you think these assumption(s) hold here?

YOUR EXPLANATION HERE. Assumption for bootstrap: The bootstrap assumes that the observed data is a representative sample of the population. It assumes that the data points are independently and identically distributed. It is also assumed that the underlying data generating process is stable, meaning that the relationship between the variables does not change over time or across different conditions.

Assumption for confidence interval computation: The confidence interval is based on the assumption that the bootstrap distribution of the sample correlation is approximately normally distributed. This is a reasonable assumption when the sample size is large enough, as the Central Limit Theorem comes into play.

In this case, the bootstrap assumption of random and independent sampling is likely reasonable since the data is sourced from a known website. However, the assumption that the underlying relationship between divorce rate and margarine consumption is stable and meaningful may not hold. As discussed earlier, the correlation between these two variables is most likely spurious and not indicative of any real causal relationship. Therefore, the confidence interval computed based on the bootstrap procedure may not be meaningful in drawing conclusions about the relationship between divorce rate and margarine consumption.

Part 3. Permuation test

For this exercise, we will analyze whether or not there is a **statistically significant** difference in the weight of babies of those individuals who smoker vs. those who do not smoke. Instead of using the entire dataset, we will use a subsample of the dataset for analysis.

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
url = 'https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/extdata/babies.txt'
baby_data = read.table(url(url), header=T)
## bwt is baby weight
bwt_nonsmoke = filter(baby_data, smoke==0) %>% select(bwt) %>% unlist
bwt_smoke = filter(baby_data, smoke==1) %>% select(bwt) %>% unlist
```

We will sample 10 observations at random from each of the smoking and nonsmoking group.

```
set.seed(123)
n = 10
nonsmokers_wt = sample(bwt_nonsmoke, n)
smokers_wt = sample(bwt_smoke, n)
```

Exercise 3.1

You will first compute the absolute difference between the average baby weight of the smokers and the nonsmokers and store this in a variable named `abs_diff_mean` for observed difference in means.

```
### YOUR CODE HERE
abs_diff_mean <- abs(mean(smokers_wt) - mean(nonsmokers_wt))
abs_diff_mean
```

```
## [1] 8
```

```
### END OF YOUR CODE
```

Exercise 3.2

Recall that we are interested in whether this observed difference is statistically significant. We do not want to reply on the assumptions needed for the normal or t-distribution approximations to hold. So instead, we will use the permutation test. In short, we will reshuffle the data many times and recompute the mean each time. Then, using the mean of the reshuffled samples, we can generate the null distribution, and, subsequently, compute the p-value.

Let us do this in steps.

First, you will create a vector of length 10,000 prepopulated with `NA` s. Call this vector `perm_mean_vec` .

```
### YOUR CODE HERE
perm_mean_vec <- rep(NA, 10000)

### END OF YOUR CODE
```

Then, you will concatenate `smokers_wt` and `nonsmokers_wt` into one vector; call this concatenated vector `data` . You will then shuffle `data` 10,000 times using the `sample` function.

For each reshuffled data vector, we will create a pseudo-smoker-baby group using the **first** 10 observations, and a pseudo-nonsmoker-baby group using the **latter** 10 observations. Compute the absolute difference between the mean weight of babies from the pseudo-smoker-baby group and the mean weight of babies from the pseudo-

nonsmoker-baby group. Store each absolute difference between the means in the `perm_mean_vec` vector that you just created.

```
### YOUR CODE HERE
data <- c(smokers_wt, nonsmokers_wt)
for (i in 1:10000) {
  shuffled_data <- sample(data)
  perm_smokers_wt <- shuffled_data[1:10]
  perm_nonsmokers_wt <- shuffled_data[11:20]
  perm_mean_vec[i] <- abs(mean(perm_smokers_wt) - mean(perm_nonsmokers_wt))
}
### END OF YOUR CODE
```

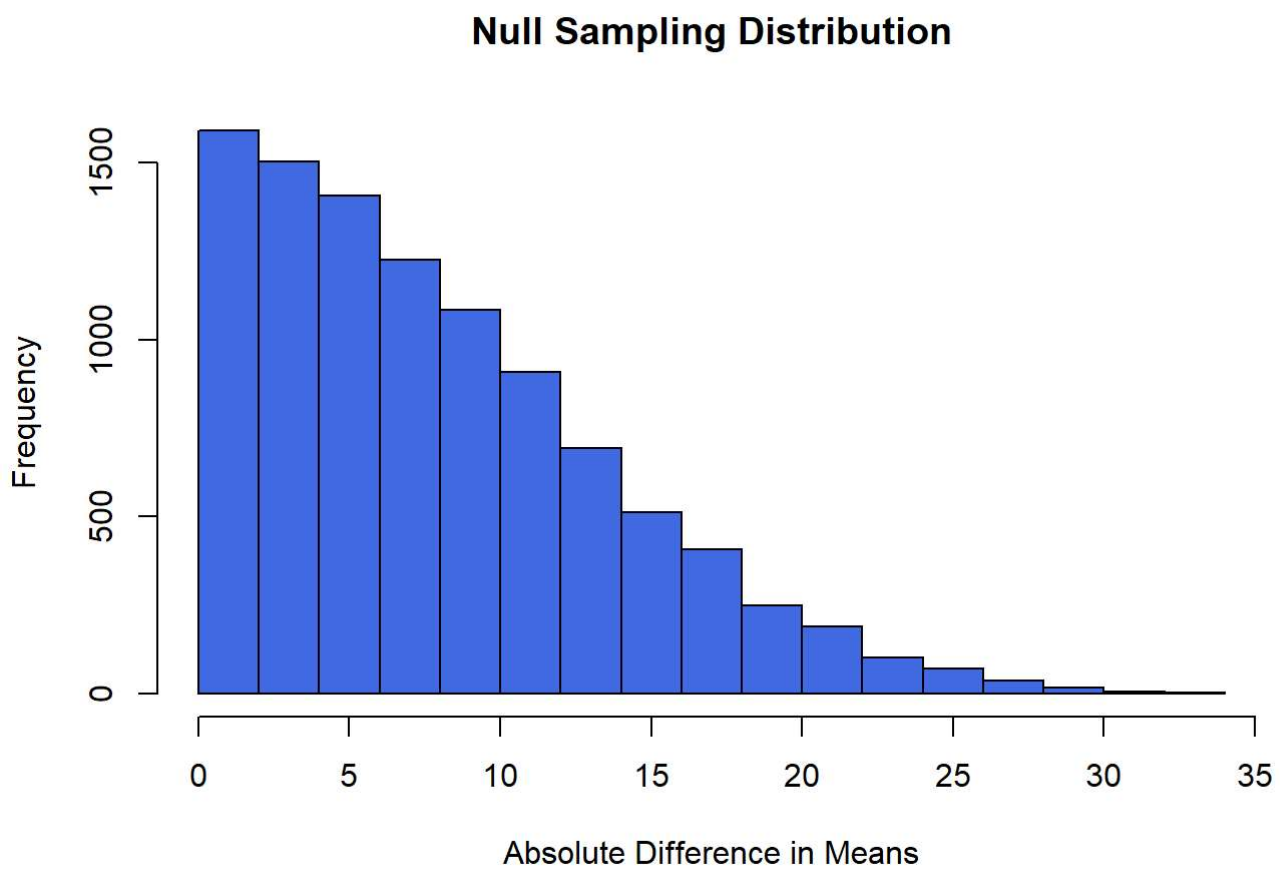
Exercise 3.3

Now, using the null sampling distribution of the test statistic generated by the permutation procedure in Exercise 3.2, compute the p-value of observing an absolute difference in means that is at least as large as the one we observed (recall that the observe value is stored in `abs_diff_mean`).

First, plot a histogram of `perm_mean_vec`. This is the null sampling distribution of the absolute difference in means. Title and label the axes accordingly.

State the null and alternative hypotheses and whether or not you can reject the null hypothesis at 95% significance. Interpret your conclusion in terms of the question in which we are interested.

```
### YOUR CODE HERE
# Plotting the null sampling distribution
hist(perm_mean_vec, main="Null Sampling Distribution",
     xlab="Absolute Difference in Means", ylab="Frequency", col = "royalblue")
```




```
# Computing the p-value
p_value <- sum(perm_mean_vec >= abs_diff_mean) / length(perm_mean_vec)
p_value
```

```
## [1] 0.4396
```

```
### END OF YOUR CODE
```

YOUR EXPLANATION HERE. As the p-value is greater than 0.05 we fail to reject the null hypothesis. The null hypothesis is that there is no difference between the weights of smoker babies and non smoker babies. Ha is that there is a difference in weights between smoker and non smoker babies.

(This exercise is adapted from Rafael Irizarry and Michael Love's PH525x course)