Name: Samuel L. Peoples

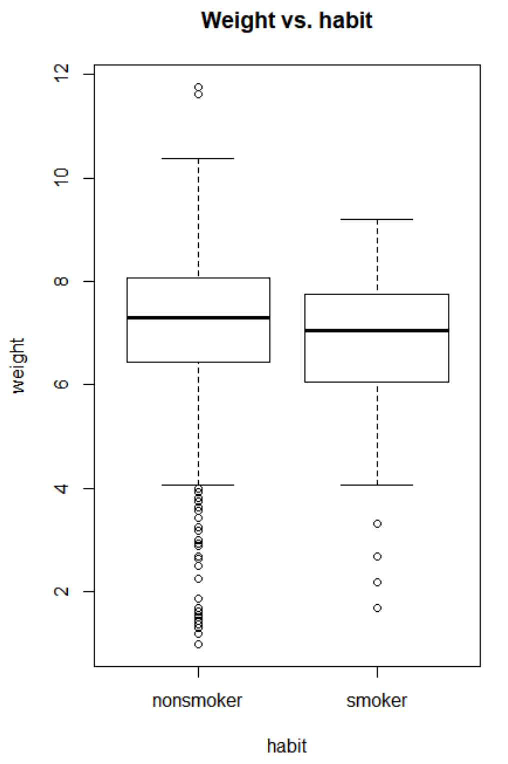
R Lab 6: Inference for Numerical Data

Exercise 1: What are the cases in this data set? How many cases are there in our sample?

There are 1000 cases of pregnancy information over 13 variables.

Exercise 2: Make a side-by-side boxplot of habit and weight. What does the plot highlight about the relationship between these two variables?

```
> boxplot(nc$weight[nc$habit =="nonsmoker"],nc$weight[nc$habit ==
"smoker"],xlab="habit",ylab="weight",main="Weight vs. habit", names =
c("nonsmoker", "smoker"))
```



In this side-by-side histogram, we can gather that there may be some relationship between smoking and low birth rates.

Exercise 3: Check if the conditions necessary for inference are satisfied. Note that you will need to obtain sample sizes to check the conditions. You can compute the group size using the same by command above but replacing mean with length.

```
> by(nc$weight, nc$habit, mean)
nc$habit: nonsmoker
                                               7.144273
nc$habit: smoker
                                                6.82873
> by(nc$weight, nc$habit, length)
nc$habit: nonsmoker
                                                    873
nc$habit: smoker
                                                    126
> s=subset(nc, habit=="smoker")
> n=subset(nc, habit=="nonsmoker")
> qqnorm(s$weight)
> qqline(s$weight)
> qqnorm(n$weight)
> qqline(n$weight)
> sSamp = sample(s$weight, 50)
> nSamp = sample(n$weight, 50)
> qqnorm(nSamp)
> qqline(nSamp)
> qqnorm(sSamp)
> qqline(sSamp)
```

Our sample size is 873 for nonsmoker and 126 from smoker, which will be enough to obtain normally distributed samples, given that they are normally distributed. The QQ plots for smokers and nonsmokers have quite a few entries near zero along the line. Taking samples of fifty yields normally distributed plots as well.

Exercise 4: Write the hypotheses for testing if the average weights of babies born to smoking and non-smoking mothers are different.

H_0: `mean(s$weight) = mean(n$weight)` : Smoking does not affect birth weight on average.

H_a: `mean(s$weight) < mean(n$weight)` : On average, babies born to women who smoke weight less when compared to women who do not smoke.

Exercise 5: Change the type argument to "ci" to construct and record a confidence interval for the difference between the weights of babies born to smoking and non-smoking mothers.

```
inference(y = nc$weight, x = nc$habit, est = "mean", type = "ht", null = 0,

          alternative = "twosided", method = "theoretical")

Summary statistics:
n_nonsmoker = 873, mean_nonsmoker = 7.1443, sd_nonsmoker = 1.5187
n_smoker = 126, mean_smoker = 6.8287, sd_smoker = 1.3862
Observed difference between means (nonsmoker-smoker) = 0.3155

H0: mu_nonsmoker - mu_smoker = 0
HA: mu_nonsmoker - mu_smoker != 0
Standard error = 0.134
Test statistic: Z =  2.359
p-value =  0.0184


inference(y = nc$weight, x = nc$habit, est = "mean", type = "ci", null = 0,

          alternative = "twosided", method = "theoretical")

Response variable: numerical, Explanatory variable: categorical

Difference between two means

Summary statistics:

n_nonsmoker = 873, mean_nonsmoker = 7.1443, sd_nonsmoker = 1.5187

n_smoker = 126, mean_smoker = 6.8287, sd_smoker = 1.3862

Observed difference between means (nonsmoker-smoker) = 0.3155


Standard error = 0.1338

95 % Confidence interval = ( 0.0534 , 0.5777 )
```

On Your Own:

1. Calculate a 95% confidence interval for the average length of pregnancies (weeks) and interpret it in context. Note that since you're doing inference on a single population parameter, there is no explanatory variable, so you can omit the x variable from the function.

```
> inference(y = nc$weeks, est = "mean", type = "ci", null = 0,
+           alternative = "twosided", method = "theoretical")
Single mean
Summary statistics: mean = 38.3347 ;  sd = 2.9316 ;  n = 998
Standard error = 0.0928
95 % Confidence interval = ( 38.1528 , 38.5165 )
```

We are 95% confident that the average length of pregnancies falls between 38.15 and 38.52 weeks.

2. Calculate a new confidence interval for the same parameter at the 90% confidence level. You can change the confidence level by adding a new argument to the function: conflevel = 0.90.

```
> inference(y = nc$weeks, conflevel = 90, est = "mean", type = "ci", null= 0,
+           alternative = "twosided", method = "theoretical")
Single mean
Summary statistics: mean = 38.3347 ;  sd = 2.9316 ;  n = 998
Standard error = 0.0928
90 % Confidence interval = ( 38.182 , 38.4873 )
```

3. Conduct a hypothesis test evaluating whether the average weight gained by younger mothers is different than the average weight gained by mature mothers.

```
> inference(y = nc$gained, x = nc$mature, est = "mean", type = "ht", null = 0
,
+           alternative = "twosided", method = "theoretical")
Response variable: numerical, Explanatory variable: categorical
Difference between two means
Summary statistics:
n_mature mom = 129, mean_mature mom = 28.7907, sd_mature mom = 13.4824
n_younger mom = 844, mean_younger mom = 30.5604, sd_younger mom = 14.3469
Observed difference between means (mature mom-younger mom) = -1.7697

H0: mu_mature mom - mu_younger mom = 0
HA: mu_mature mom - mu_younger mom != 0
Standard error = 1.286
Test statistic: Z =  -1.376
p-value =  0.1686
```

4. Now, a non-inference task: Determine the age cutoff for younger and mature mothers. Use a method of your choice, and explain how your method works.

```
> max(subset(nc,mature == "younger mom")$mage)
                                                                             34
> min(subset(nc,mature == "mature mom")$mage)
                                                                             35
```

I calculated the maximum of the subset of nc, where the variable “mature” is equal to “younger mom”, with the parameter of the mother’s age. I did the same to find the minimum age of mature mothers, and because the two subsets do not intersect, we know that the cutoff is at the age of 34.

5. Pick a pair of numerical and categorical variables and come up with a research question evaluating the relationship between these variables. Formulate the question in a way that it can be answered using a hypothesis test and/or a confidence interval. Answer your question using the inference function, report the statistical results, and also provide an explanation in plain language.

Does the gender of a child have any impact on how much weight the mother gains? We will conduct a hypothesis test evaluating whether the average weight gained by mothers of females is different from the average weight gained by mothers of males.

```
> inference(y = nc$gained, x = nc$gender, est = "mean", type = "ht", null = 0
,
+           alternative = "twosided", method = "theoretical")
Response variable: numerical, Explanatory variable: categorical
Difference between two means
Summary statistics:
n_female = 488, mean_female = 29.8135, sd_female = 14.2506
n_male = 485, mean_male = 30.8412, sd_male = 14.228
Observed difference between means (female-male) = -1.0277

H0: mu_female - mu_male = 0
HA: mu_female - mu_male != 0
Standard error = 0.913
Test statistic: Z =  -1.126
p-value =  0.2604
```

Because the pvalue is greater than any population parameter we could assign, we fail to reject the null, which suggests that the gender of the child has no bearing on how much weight will be gained.