BEE552 Biometry Week 5

Maria Feiler

2/23/2022

My Learning Journey

Over the last week, I participated in Biometry in the following ways:

- I asked / answered **6** questions posed in class.
- I asked **2** questions in Slack.
- I answered **0** questions posed by other students on Slack.
- I came to Heather's office hours: **Yes**
- I came to Jose's office hours: **No**
- I met with Heather or Jose separately from office hours: **No**

Anything not falling into one of the above categories?

> **No**

On a scale of 1 (no knowledge) to 10 (complete expert), how would I rate my comfort with R programming after this week?

> **7**

Any topics from last week that you are still confused about?

> **Doing fine for now**

Problem Set

Part I

One concern about applying metal tags to penguins is that doing so may increase drag in the water. In a study of the impact of foraging tags on penguin foraging efficiency, the mean length of 344 foraging trips for penguins with a metal tag was 12.70 days with a standard deviation of 3.71 days. For those with an electronic tag (which presumably would not increase drag), the mean foraging trip length was 11.60 days with standard deviation of 4.53 days over 512 trips. Calculate the confidence interval for the difference in foraging trip length and calculate the p-value for the null hypothesis that there is no difference between tag types. (Do not use the R function 't.test'. I'm asking you to calculate these yourself.) Do these data provide evidence that mean foraging trips are longer for penguins with a metal tag?

An unpaired two sample t-test will be used to determine if there is a difference in foraging lengths of penguins with different tag lengths.

$$\begin{aligned} H_{null} &: \overline{x}_{metal} = \overline{x}_{electronic} \\ H_{alt} &: \overline{x}_{metal} \neq \overline{x}_{electronic} \end{aligned} \qquad \begin{aligned} \overline{x}_{metal} &= 12.70 \quad & \overline{x}_{electronic} &= 11.60 \\ s_{metal} &= 3.71 \quad & s_{electronic} &= 4.53 \\ n_{metal} &= 344 \quad & n_{electronic} &= 512 \end{aligned}$$

Assuming the distributions of penguin foraging lengths are normally distributed, then the distributions of foraging lengths are represented by $X_{metal} \sim N\left(\overline{x}_{metal}, s_{metal}\right)$ and $X_{metal} \sim N\left(\overline{x}_{electronic}, s_{electronic}\right)$ for metal and electronic tags, respectively.

Assuming that penguin foraging lengths between those with metal versus electronic tag are independent, then...

$$\overline{X}_{metal} - \overline{X}_{electronic} \sim N(\overline{x}_{metal} - \overline{x}_{electronic}\ , \frac{s^2_{metal}}{n_{metal}} + \frac{s^2_{electronic}}{n_{electronic}})$$

Therefore, the standard error of the difference of means is given by...

$$SE = \sqrt{\frac{s^2_{metal}}{n_{metal}} + \frac{s^2_{electronic}}{n_{electronic}}}\ ,$$

the test statistic is...

$$T^* = \frac{\overline{x}_A - \overline{x}_B}{\sqrt{\frac{s^2_A}{n_A} + \frac{s^2_B}{n_B}}}\ ,$$

and the degrees of freedom are calculated using...

$$dof = \frac{\left(\frac{s^2_{metal}}{n_{metal}} + \frac{s^2_{electronic}}{n_{electronic}}\right)^2}{\frac{\left[\frac{s^2_{metal}}{n_{metal}}\right]^2}{n_{metal}-1} + \frac{\left[\frac{s^2_{electronic}}{n_{electronic}}\right]^2}{n_{electronic}-1}}\ .$$

Finally, the 95 percentile confidence interval of the difference in means is given by...

$$\overline{x}_{metal} + \overline{x}_{electronic}\ \pm\ z\sqrt{\frac{s^2_{metal}}{n_{metal}} + \frac{s^2_{electronic}}{n_{electronic}}}$$

... where z is 1.96, or the z-value corresponding with 95% confidence.

```r
# Penguins with metal tags
metal <- list(12.70, 3.71, 344, 3.71^2)
names(metal) <- c("mean", "sd", "n", "var")

# Penguins with electronic tags
electronic <- list(11.60, 4.53, 512, 4.53^2)
names(electronic) <- c("mean", "sd", "n", "var")

# Create list for hypothesis testing results
penguinHypTest <- vector("list", length = 4)
names(penguinHypTest) <- c("stErr", "t", "dof", "p")

# Calculate standard error of the means
penguinHypTest$stErr <- sqrt(((metal$var)/metal$n)+((electronic$var)/electronic$n))

# Calculate the test statistic
penguinHypTest$t <- (metal$mean-electronic$mean)/penguinHypTest$stErr

# Calculate the degrees of freedom
top <- ((metal$var/metal$n)+(electronic$var/electronic$n))^2
bottom1 <- ((metal$var/metal$n)^2)/(metal$n-1)
bottom2 <- ((electronic$var/electronic$n)^2)/(electronic$n-1)

penguinHypTest$dof <- top/(bottom1 + bottom2)

# Find the p value
penguinHypTest$p <- 2*pt(penguinHypTest$t,
                         penguinHypTest$dof,
                         lower.tail = FALSE
                         )

# Calculate the 95% confidence interval
penguinsCI <- vector("list", length = 2)
names(penguinsCI) <- c("LL", "UL")

penguinsCI$LL <- metal$mean-electronic$mean-1.96*penguinHypTest$stErr
penguinsCI$UL <- metal$mean-electronic$mean+1.96*penguinHypTest$stErr
```

Table 1: Two-sample t-test results.

SE	T*	dof	p
0.28	3.89	821.22	0

The 95% confidence interval of the difference of mean foraging length between tag types is (0.545, 1.655). The difference in foraging lengths between penguins with different tag types is statistically significant (p = 0). Therefore, the tag type has an appreciable influence on penguin foraging efficiency.

Part II

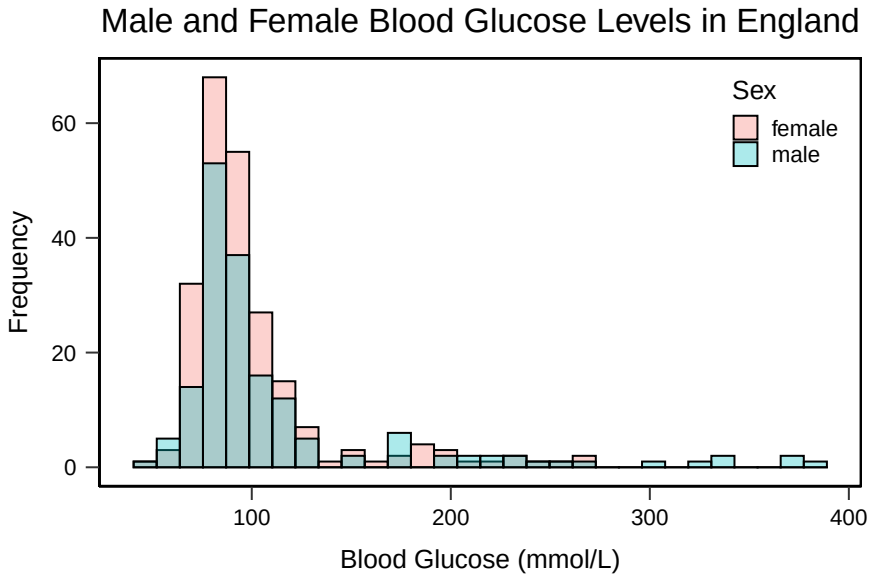
Download the dataset "diabetes.csv". This dataset describes a diabetes study by Willems et al. 1997 in which blood glucose (stab.glu) was recorded for males and females (gender) in each of two locations (location).

```
link <- "https://raw.githubusercontent.com/hlynch/Biometry2022/master/_data/diabetes.csv"
diabetes <- read_csv(link,
                     show_col_types = FALSE
                     )

# Rename "gender" to "sex" for biological accuracy, as "male" and "female"
# are not genders
diabetes <- rename(diabetes, sex = gender)
```

Table 2: A selection of data from Willems et al. 1997.

Location	Sex	Blood Glucose
Louisa	female	101.29
Buckingham	female	203.17
Louisa	female	101.14
Louisa	female	103.27
Buckingham	female	233.01
Buckingham	male	113.15
Louisa	male	81.17
Buckingham	female	76.31
Louisa	female	120.30
Louisa	female	67.24



Question 1 *Aggregating across locations, calculate the test statistic for the t-test to test the null hypothesis that males and females do not differ in their average blood glucose levels and compare it to its distribution under the null hypothesis.*

An unpaired two sample t-test will be used to determine if there is a difference between male and female blood glucose levels in England.

$$H_{null} : \overline{x}_{male} = \overline{x}_{female} \quad H_{alt} : \overline{x}_{male} \neq \overline{x}_{female}$$

```
# Males
males <- list(mean(diabetes$stab.glu[diabetes$sex == "male"]),
              sd(diabetes$stab.glu[diabetes$sex == "male"]),
              sum(diabetes$sex == "male"),
              var(diabetes$stab.glu[diabetes$sex == "male"])
              )
names(males) <- c("mean", "sd", "n", "var")

# Females
females <- list(mean(diabetes$stab.glu[diabetes$sex == "female"]),
                sd(diabetes$stab.glu[diabetes$sex == "female"]),
                sum(diabetes$sex == "female"),
                var(diabetes$stab.glu[diabetes$sex == "female"])
                )
names(females) <- c("mean", "sd", "n", "var")

# Create list for hypothesis testing results
diabetesHypTest <- vector("list", length = 4)
names(diabetesHypTest) <- c("stErr", "t", "dof", "p")

# Calculate standard error of the means
diabetesHypTest$stErr <- sqrt(((males$var)/males$n)+((females$var)/females$n))

# Calculate the test statistic
diabetesHypTest$t <- (males$mean-females$mean)/diabetesHypTest$stErr

# Calculate the degrees of freedom
top <- ((males$var/males$n)+(females$var/females$n))^2
bottom1 <- ((males$var/males$n)^2)/(males$n-1)
bottom2 <- ((females$var/females$n)^2)/(females$n-1)

diabetesHypTest$dof <- top/(bottom1 + bottom2)

# Find the p value
diabetesHypTest$p <- 2*pt(diabetesHypTest$t,
                          diabetesHypTest$dof,
                          lower = FALSE
                          )

# Calculate the 95% confidence interval
diabetesCI <- vector("list", length = 2)
names(diabetesCI) <- c("LL", "UL")

diabetesCI$LL <- males$mean-females$mean-1.96*diabetesHypTest$stErr
diabetesCI$UL <- males$mean-females$mean+1.96*diabetesHypTest$stErr
```

Table 3: Male and female blood glucose two-sample t-test results.

SE	T*	dof	p	95CI Lower	95CI Upper
5.471	2.318	251.076	0.021	1.959	23.406

Question 2 *Compare your results with the R function "t.test". (Copy and paste results below.) Explain what every single number reported in the output of t.test means.*

```
diabetesttest <- t.test(diabetes$stab.glu[diabetes$sex == "male"],
                        diabetes$stab.glu[diabetes$sex == "female"]
                        )
```

```
    Welch Two Sample t-test

data:  Males and Females
t = 2.318, df = 251.08, p-value = 0.02125
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.906998 23.458207
sample estimates:
  Mean Male Glucose Mean Female Glucose
          112.49538            99.81278
```

The p-values produced by the t.test() calculations are the same as the p-values obtained from manual test statistic calculations.

Question 3 *Interpret your results in words as you would in the results section of a manuscript (1-2 sentences)*

An unpaired two sample t-test was used to test if there was a significant difference between the blood glucose levels of males (112.5± 63.61) and females (99.81± 37.12). At the critical value of $\alpha_C = 0.05$, the difference between male and female blood glucose levels is statistically significant (p = 0.021, t* = 2.318, df = 251.08), and the null hypothesis is rejected. The 95^{th} percentile confidence interval for the difference in mean blood glucose levels between males and females is (1.96, 23.41), suggesting that the females had significantly lower blood glucose levels then their male counterparts.

Question 4 *How would the results differ if you used the median rather than the mean as the test statistic? Write a script to test the null hypothesis $H_0 : median_{male} = median_{female}$. Keep in mind that you will have to determine the distribution of the test statistic under the null hypothesis.*

```
# This code was sourced in part from the original poster's updated code:
# https://stats.stackexchange.com/questions/176691/running-a-permutation-test-with-
# different-sample-sizes-in-r/176714

# Calculate thesample difference in medians
theta_hat <- median(diabetes$stab.glu[diabetes$sex == "male"]) -
        median(diabetes$stab.glu[diabetes$sex == "female"])

# Create vector capture the permutation test results
dif <- c()
```

```r
# Create vector of the differences between the sample permutation medians
for (i in 1:1000) {
        samp <- sample(x = c(diabetes$stab.glu[diabetes$sex == "male"],
                             diabetes$stab.glu[diabetes$sex == "female"]),
                       replace = FALSE)

        samp_male <- samp[1:sum(diabetes$sex == "male")]
        samp_female <- samp[sum(diabetes$sex == "male")+1:sum(diabetes$sex == "female")]

        dif[i] <- median(samp_male) - median(samp_female)
}

p_value <- sum(abs(dif) > abs(theta_hat))/length(dif)
```

The difference between the male and female permutation medians was 0.07 ± 2.23. The quartiles of the permutations were -6.73(0%), -1.8(25%), 0.02(50%), 1.9(75%), 5.89(100%). The p-value of the original sample is 0.868. At $\alpha_c = 0.05$, the null hypothesis is not rejectd.

Question 5 *If these were your data, under what circumstances would you feel comfortable lumping the data for each location together? Assuming you could lump the data together, what would be one advantage of doing so?*

In the case of this data, it would depend on what question I was asking. For something like blood glucose, which is dependent on so many socioeconomic and biological factors, my first instinct is to not lump the locations together to preserve potential differences in food quality, stress, health care availability, etc. If I were interested in the whole of England, I would probably lump them together, though I would question why I only sampled two cities to characterize a whole country. In the event I can lump the data together, I am gaining more power in my statistical tests due to the higher sample size and the assumption that the variance/mean is similar across the samples.

Part III

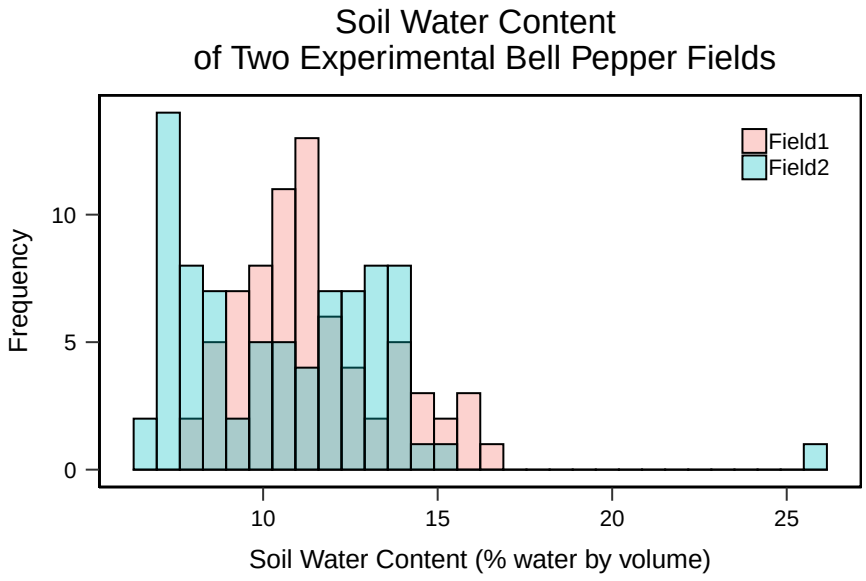
Download the dataset "WaterSamples.csv", which represent soil water content (% water by volume) for independent random samples of soil from two experimental fields growing bell peppers. (Reference: Journal of Agricultural, Biological, and Environmental Statistics, Vol. 2, No. 2, p 149-155)

Here we are interested in the variance within each field, with the null hypothesis: $H_0 : \sigma^2_{Field1} = \sigma^2_{Field2}$

```
link <- "https://raw.githubusercontent.com/hlynch/Biometry2022/master/_data/WaterData.csv"
water <- read_csv(link,
                  show_col_types = FALSE
                  )
```

Table 4: A selection of soil water content data (% water by volume).

Field 1	Field 2
9.1	8.1
14.1	11.3
12.0	9.7
11.9	13.2
14.0	7.6



Question 1 *What is the equation for and numerical value of the test statistic for testing this null hypothesis?*

```
field1 <- list("n" = length(na.omit(water$Field1)),
               "var" = var(water$Field1, na.rm = TRUE)
               )
field2 <- list("n" = length(na.omit(water$Field2)),
               "var" = var(water$Field2, na.rm = TRUE)
               )

# Determine which is the larger variance
field1$var > field2$var
```

```
 FALSE
```

Therefore, the test statistic will be calculated using the following:

$$F_s = \frac{s^2_{Field2}}{s^2_{Field1}}$$

```
# Rename lower sample size as m
names(field1)[1] <- "m"

# Calculate test statistic F*
Fs <- field2$var/field1$var
```

The value of the test statistic is 2.116.

Question 2 *If we assume a two-sided test, what is the p-value for testing this null hypothesis? (Use the test statistic from 1 and the appropriate distribution under the null hypothesis.)*

For a two tailed test, the p-value will be calculated using

$$\frac{s^2_{Field2}}{s^2_{Field1}} > F_{[1-\frac{\alpha}{2}](n-1,\, m-1)} \quad .$$

Under the null hypothesis,

$$(F_s = \frac{s^2_{Field2}}{s^2_{Field1}} | H_0) \sim F_{n-1,\, m-1} \quad .$$

```
# Calculate one-sided p_value
p_value <- 1-pf(Fs,
                df1 = field2$n-1,
                df2 = field1$m-1
                )

# Multiply by two to get the two-tailed p_value
p_value <- 2*p_value
```

The p_value for the difference in sample variances between the two samples is 0.00153, and we reject the null hypothesis.

Question 3 *What is the p-value if we are interesting only in* $H_A : \sigma^2_{Field1} > \sigma^2_{Field2}$*?*

$$\frac{s^2_{Field1}}{s^2_{Field2}} > F_{[1-\alpha](n-1,\, m-1)} \ .$$

```
# Calculate test statistic F*
Fs <- field1$var/field2$var

# Calculate one-sided p_value
p_value <- 1 - pf(Fs,
                  df1 = field1$m-1,
                  df2 = field2$n-1
                  )
```

The p_value if the variance of Field 1 is greater than the variance of Field 2 is 0.999. Therefore, the variance of Field 1 is not greater than Field 2.

Question 4 *What is the p-value if we are interesting only in $H_A : \sigma^2_{Field1} < \sigma^2_{Field2}$?

```
# Calculate test statistic F*
Fs <- field2$var/field1$var

# Calculate one-sided p_value
p_value <- 1 - pf(Fs,
                  df1 = field2$n-1,
                  df2 = field1$m-1
                  )
```

The p_value if the variance of Field 1 is greater than the variance of Field 2 is 0.000766. Therefore, the variance of Field 2 is greater than Field 1.

Question 5 *Confirm that you can reproduce the answers for 2-4 above using the R function 'var.test'.*

```
vartestresults <- list("TwoTailed" = var.test(water$Field1,
                                              water$Field2,
                                              alternative = "two.sided"),
                       "Field1Greater" = var.test(water$Field1,
                                                  water$Field2,
                                                  alternative = "greater"),
                       "Field2Greater" = var.test(water$Field1,
                                                  water$Field2,
                                                  alternative = "less")
                       )
```

Table 5: var.test() results for Part 3 Questions 2-4.

Two-Tailed	Field 1 Greater	Field 2 Greater
0.00153	0.999	0.000766

Extra Work

Part II Questions 2-4, Locations Separated

Question 1 *Aggregating across locations, calculate the test statistic for the t-test to test the null hypothesis that males and females do not differ in their average blood glucose levels and compare it to its distribution under the null hypothesis.*

```
# Get unique location names
uniqueLoc <- unique(diabetes$location)
```

Louisa Buckingham

```
# Divide into two datasets based on location
Louisa <- diabetes[which(diabetes$location == uniqueLoc[1]), c("sex", "stab.glu")]

Bucky <- diabetes[which(diabetes$location == uniqueLoc[2]), c("sex", "stab.glu")]
```

Louisa

An unpaired two sample t-test will be used to determine if there is a difference between male and female blood glucose levels in Louisa, England.

$$H_{null} : \overline{x}_{male} = \overline{x}_{female} \quad H_{alt} : \overline{x}_{male} \neq \overline{x}_{female}$$

```
# Males
LMales <- list(mean(Louisa$stab.glu[Louisa$sex == "male"]),
               sd(Louisa$stab.glu[Louisa$sex == "male"]),
               sum(Louisa$sex == "male"),
               var(Louisa$stab.glu[Louisa$sex == "male"])
               )
names(LMales) <- c("mean", "sd", "n", "var")

# females
Lfemales <- list(mean(Louisa$stab.glu[Louisa$sex == "female"]),
                 sd(Louisa$stab.glu[Louisa$sex == "female"]),
                 sum(Louisa$sex == "female"),
                 var(Louisa$stab.glu[Louisa$sex == "female"])
                 )
names(Lfemales) <- c("mean", "sd", "n", "var")

# Create list for hypothesis testing results
LouisaHypTest <- vector("list", length = 4)
names(LouisaHypTest) <- c("stErr", "t", "dof", "p")

# Calculate standard error of the means
LouisaHypTest$stErr <- sqrt(((LMales$var)/LMales$n)+((Lfemales$var)/Lfemales$n))

# Calculate the test statistic
LouisaHypTest$t <- (LMales$mean-Lfemales$mean)/LouisaHypTest$stErr

# Calculate the degrees of freedom
top <- ((LMales$var/LMales$n)+(Lfemales$var/Lfemales$n))^2
```

```
bottom1 <- ((LMales$var/LMales$n)^2)/(LMales$n-1)
bottom2 <- ((Lfemales$var/Lfemales$n)^2)/(Lfemales$n-1)

LouisaHypTest$dof <- top/(bottom1 + bottom2)

# Find the p value
LouisaHypTest$p <- 2*pt(LouisaHypTest$t,
                        LouisaHypTest$dof
                        )

# Calculate the 95% confidence interval
LouisaCI <- vector("list", length = 2)
names(LouisaCI) <- c("LL", "UL")

LouisaCI$LL <- LMales$mean-Lfemales$mean-1.96*LouisaHypTest$stErr
LouisaCI$UL <- LMales$mean-Lfemales$mean+1.96*LouisaHypTest$stErr
```

Table 6: Louisa male and female blood glucose two-sample t-test results.

SE	T*	dof	p	95CI Lower	95CI Upper
8.298	1.803	106.714	1.926	-1.302	31.226

Buckingham

An unpaired two sample t-test will be used to determine if there is a difference between male and female blood glucose levels in Buckingham, England.

$$H_{null} : \overline{x}_{male} = \overline{x}_{female} \quad H_{alt} : \overline{x}_{male} \neq \overline{x}_{female}$$

```
# Males
BMales <- list(mean(Bucky$stab.glu[Bucky$sex == "male"]),
               sd(Bucky$stab.glu[Bucky$sex == "male"]),
               sum(Bucky$sex == "male"),
               var(Bucky$stab.glu[Bucky$sex == "male"])
               )
names(BMales) <- c("mean", "sd", "n", "var")

# Females
Bfemales <- list(mean(Bucky$stab.glu[Bucky$sex == "female"]),
                 sd(Bucky$stab.glu[Bucky$sex == "female"]),
                 sum(Bucky$sex == "female"),
                 var(Bucky$stab.glu[Bucky$sex == "female"])
                 )
names(Bfemales) <- c("mean", "sd", "n", "var")

# Create list for hypothesis testing results
BuckyHypTest <- vector("list", length = 4)
names(BuckyHypTest) <- c("stErr", "t", "dof", "p")

# Calculate standard error of the means
BuckyHypTest$stErr <- sqrt(((BMales$var)/BMales$n)+((Bfemales$var)/Bfemales$n))

# Calculate the test statistic
BuckyHypTest$t <- (BMales$mean-Bfemales$mean)/BuckyHypTest$stErr

# Calculate the degrees of freedom
top <- ((BMales$var/BMales$n)+(Bfemales$var/Bfemales$n))^2
bottom1 <- ((BMales$var/BMales$n)^2)/(BMales$n-1)
bottom2 <- ((Bfemales$var/Bfemales$n)^2)/(Bfemales$n-1)

BuckyHypTest$dof <- top/(bottom1 + bottom2)

# Find the p value
BuckyHypTest$p <- 2*pt(BuckyHypTest$t,
                       BuckyHypTest$dof
                       )

# Calculate the 95% confidence interval
BuckyCI <- vector("list", length = 2)
names(BuckyCI) <- c("LL", "UL")

BuckyCI$LL <- BMales$mean-Bfemales$mean-1.96*BuckyHypTest$stErr
BuckyCI$UL <- BMales$mean-Bfemales$mean+1.96*BuckyHypTest$stErr
```

Table 7: Buckingham male and female blood glucose two-sample t-test results.

SE	T*	dof	p	95CI Lower	95CI Upper
7.246	1.434	149.953	1.846	-3.814	24.591

Question 2 *Compare your results with the R function "t.test". (Copy and paste results below.) Explain what every single number reported in the output of t.test means.*

```
ttestLouisa <- t.test(Louisa$stab.glu[Louisa$sex == "male"],
                      Louisa$stab.glu[Louisa$sex == "female"]
                      )

ttestBucky <- t.test(Bucky$stab.glu[Louisa$sex == "male"],
                     Bucky$stab.glu[Louisa$sex == "female"]
                     )
```

```
    Welch Two Sample t-test

data:  Louisa Males and females
t = 1.8031, df = 106.71, p-value = 0.07419
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.48783 31.41240
sample estimates:
  Mean Male Glucose Mean Female Glucose
           113.4417             98.4794

    Welch Two Sample t-test

data:  Buckingham Males and females
t = 1.0965, df = 134.02, p-value = 0.2748
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.580805 22.957418
sample estimates:
  Mean Male Glucose Mean Female Glucose
           110.4972            102.3089
```

The p-values produced by the t.test() calculations are the same as the p-values obtained from manual test statistic calculations.

Question 3 *Interpret your results in words as you would in the results section of a manuscript (1-2 sentences)*

An unpaired two sample t-test was used to test if there was a significant difference between the blood glucose levels of Louisa males (113.44± 70.5) and females (98.48± 32.4). At the critical value of $\alpha_C = 0.05$, the difference between Louisa male and female blood glucose levels is not statistically significant (p = 1.926, t* = 1.803, df = 106.71), and the null hypothesis is not rejected. The 95^{th} percentile confidence interval for the difference in mean blood glucose levels between Louisa males and females is (-1.3, 31.23), suggesting that the Louisa females had significantly lower blood glucose levels then their male counterparts.

An unpaired two sample t-test was used to test if there was a significant difference between the blood glucose levels of Buckingham (111.58± 56.58) and females (101.19± 41.56). At the critical value of $\alpha_C = 0.05$, the difference between Buckingham male and female blood glucose levels is not statistically significant (p = 1.846, t* = 1.434, df = 149.95), and the null hypothesis is not rejected. The 95^{th} percentile confidence interval for the difference in mean blood glucose levels between Buckingham males and females is (-3.81, 24.59), suggesting that the Buckingham females had significantly lower blood glucose levels then their male counterparts.