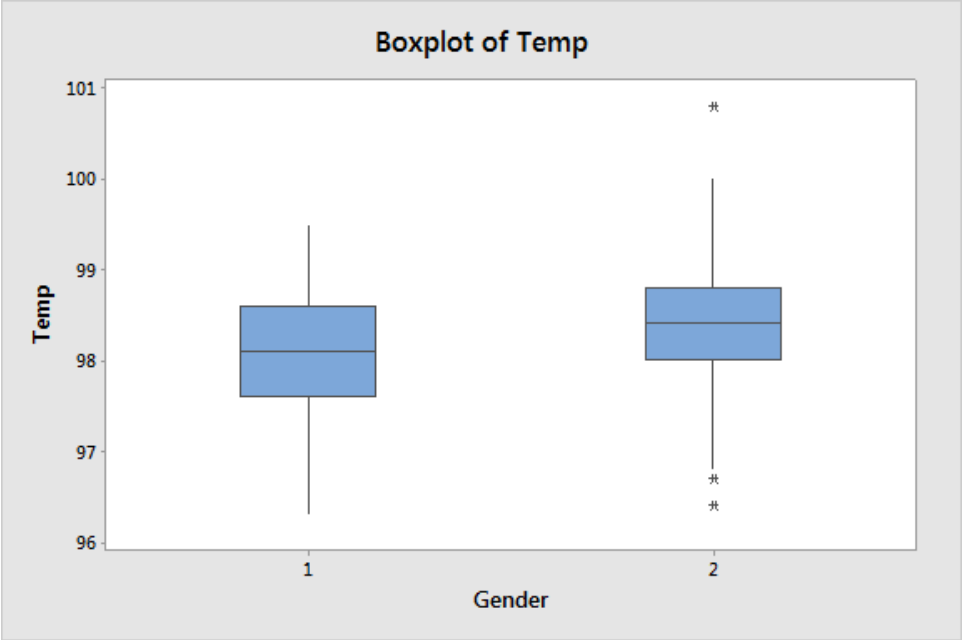
Homework #7 – Solutions
COR1-GB.1305 – Statistics and Data Analysis

Problem 1

Here, we consider the two-sample t-test, for the data set `NormTemp.CSV`*. The second column (*`Gender`*) is 1 for male, 2 for female, and the third column (*`HeartRate`*) is measured in beats per minute.*

(a) *Make side-by-side boxplots for the temperatures of males and females in the dataset. To do this, use* Graph ⇒ Boxplot *then select "One Y, With Groups". Select* `Temp` *in the "Graph variables" box, and select* `Gender` *in the "Categorical variables for grouping" box. Do there seem to be any differences?*



The temperature values for females appear to be slightly higher.

(b) *What are the two samples?*

The measured temperatures for the 65 females and 65 males in the study.

(c) *What are the two populations?*

The temperatures of all females and males.

(d) *What are the null and alternative hypotheses?*

Let μ_1 be the average temperature for all males, and let μ_2 be the average temperature for all females. The null and alternative hypotheses are $H_0 : \mu_1 = \mu_2$, and $H_a : \mu_1 \neq \mu_2$.

(e) *Get the descriptive statistics for the two samples. To do this, use* Stat ⇒ Basic Statistics ⇒ Display Descriptive Statistics. *Select* `Temp` *in the "Variables" box, and select* `Gender` *in the "By variables (optional)" box. Find* n_1, $\bar{x}_1$, s_1, n_2, $\bar{x}_2$, *and* s_2.

```
Variable  Gender   N  N*    Mean  SE Mean  StDev  Minimum      Q1  Median      Q3  Maximum
Temp      1       65   0  98.105   0.0867  0.699   96.300  97.600  98.100  98.600   99.500
          2       65   0  98.394   0.0922  0.743   96.400  98.000  98.400  98.800  100.800
```

From the output, we can see

$$\begin{aligned} n_1 &= 65, \\ \bar{x}_1 &= 98.105, \\ s_1 &= 0.699, \end{aligned}$$

$$\begin{aligned} n_2 &= 65, \\ \bar{x}_2 &= 98.394, \\ s_2 &= 0.743. \end{aligned}$$

(f) *Compute the test statistic.*

First, we compute

$$\begin{aligned} \bar{x}_1 - \bar{x}_2 &= 98.105 - 98.394 \\ &= -0.289, \\ \mathrm{se}(\bar{x}_1 - \bar{x}_2) &= \sqrt{\frac{s_1^1}{n_1} + \frac{s_2^2}{n_2}} \\ &= \sqrt{\frac{(0.699)^2}{65} + \frac{(0.743)^2}{65}} \\ &= 0.127. \end{aligned}$$

The test statistic is

$$\begin{aligned} t &= \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{se}(\bar{x}_1 - \bar{x}_2)} \\ &= \frac{-0.289}{0.127} \\ &= -2.27. \end{aligned}$$

(g) *Compute an approximate p-value.*

$$\begin{aligned} p &\approx P(|Z| > 2.27) \\ &\approx P(|Z| > 2.2) \\ &= 0.02781. \end{aligned}$$

Note: it is OK to use the approximation $p \approx P(|Z| > 2.3)$ instead.

(h) *Use the p-value to evaluate whether or not there appears to be a significant difference in average temperature between males and females.*

Since $p < 0.05$, there does appear to be a significant difference in average temperate between males and females.

(i) *Find a 95% confidence for the difference in average temperatures in the populations.*

An approximate 95% confidence interval for $\mu_1 - \mu_2$ is

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2) \pm 2\text{se}(\bar{x}_1 - \bar{x}_2) &= (-0.289) \pm (2)(0.127) \\ &= -0.280 \pm 0.254 \\ &= (-0.534, -0.026).\end{aligned}$$

(j) *Now, use Minitab to perform the test and construct the confidence interval. Do do so, use* Stat ⇒ Basic Statistics ⇒ 2-Sample t. *Choose the option "Both samples are in one column". Set "Samples" to* `Temp` *and set "Sample IDs" to* `Gender`*. The p-value and confidence interval that Minitab computes will be slightly more accurate than the one you compute in part (g), because Minitab uses a t distribution instead of a z distribution to compute the probability.*

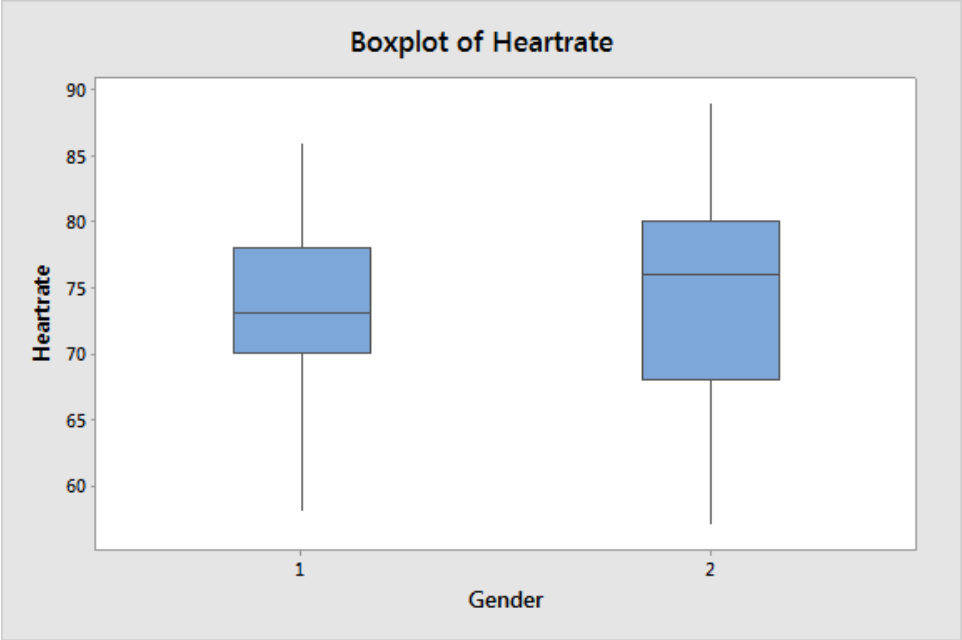
```
Two-sample T for Temp

Gender   N    Mean  StDev  SE Mean
1       65  98.105  0.699    0.087
2       65  98.394  0.743    0.092


Difference = mu (1) - mu (2)
Estimate for difference:  -0.289
95% CI for difference:  (-0.540, -0.039)
T-Test of difference = 0 (vs ): T-Value = -2.29  P-Value = 0.024  DF = 127
```

.....

Problem 2

We will again use `NormTemp.CSV` *data, but now we will investigate* `HeartRate`*.*

(a) *Make side-by-side boxplots for the heart rates of males and females in the dataset.*



(b) *Test whether or not there is a significant difference in average heart rate between all males and females. You can either compute the p-value by hand, or you can compute it using Minitab.*

```
Two-sample T for Heartrate

Gender   N   Mean  StDev  SE Mean
1       65  73.37   5.88     0.73
2       65  74.15   8.11      1.0


Difference = mu (1) - mu (2)
Estimate for difference:  -0.78
95% CI for difference:  (-3.24, 1.67)
T-Test of difference = 0 (vs ): T-Value = -0.63  P-Value = 0.529  DF = 116
```

There does not appear to be a significant difference ($p \geq 0.05$).

(c) *Find a 95% confidence interval for the difference in average heart rates between all males and all females. Again, you can either compute the confidence interval by hand, or you can compute it using Minitab.*

From the Minitab output: $(-3.24, 1.67)$.

(d) *What assumptions do you need for the p-value and the confidence interval to be valid?*

We need the observed samples to be simple random samples from the populations. We do not need to assume that the populations are normal, because the sample sizes are both above 30.

. . . . . .

Problem 3

(Adapted from Stine and Foster, 17.32) The dataset `retail_sales.csv` *gives the sales volume (in dollars per square foot) for 37 retail outlets specializing in women's clothing in 2006 and 2007. Note that these samples are paired: the same stores are measured in both years. Did sales change by a statistically significant amount from 2006 to 2007? To answer the question, use a paired t-test; that is, subtract the sales in the two years, then analyze the store-specific differences. Answer the following:*

(a) *What is the sample?*

The 2006 and 2007 sales volumes for the 37 retail outlets specializing in women's clothing.

(b) *What is the population?*

The 2006 and 2007 sales volumes for all retail outlets specializing in women's clothing.

(c) *If there were no difference in expected sales between the two years, what would be the chance of getting data like that observed?*

We use a paired t-test to answer this question. Here is the Minitab output:

```
Paired T for Sales, 2007 - Sales, 2006

               N    Mean  StDev  SE Mean
Sales, 2007   37  332.97  48.98     8.05
Sales, 2006   37  328.28  49.44     8.13
Difference    37    4.69  10.06     1.65


95% CI for mean difference: (1.33, 8.04)
T-Test of mean difference = 0 (vs not = 0): T-Value = 2.83  P-Value = 0.007
```

If there were no difference in expected sales, then the chance of getting data like that observed would be 0.7% (the observed data would be very unlikely).

(d) *Find a 95% confidence interval for the difference in expected sales between 2006 and 2007.*

From the Minitab output, a 95% confidence interval for the difference (2007 − 2006) is $(1.33, 8.04)$.

.....