MVE155, Statistical Inference
Assignment 3

March 4, 2020

Ella Guiladi
930509-0822
guiladi@student.chalmers.se

Exercises

The code to the following solutions and data can be found in the source code files attached in the mail. Since the code is written in cells, you need to run through all the cells in order for the code to work.

a.

The source code for this subsection can be found in the attached file, by opening the script-file "Exercise3A".

(i) By using normal theory and forming a 95% confidence interval for the difference of mean body temperatures between males and females, the resulting confidence interval was: [0.0412,0.5373]. To calculate the confidence interval it was assumed that the two samples were independent and since the sample sizes were large a normal approximation was used as

$$\frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_{\bar{X}}^2+S_{\bar{Y}}^2} \approx \mathrm{N}(0,1). \tag{1}$$

The confidence interval was calculated with the following formula,

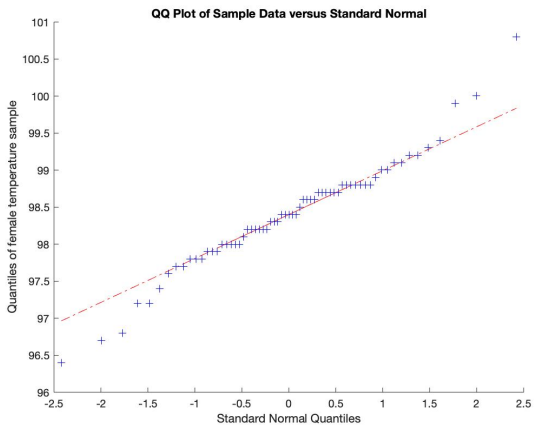
$$I_{\mu_1-\mu_2} = \bar{x}-\bar{y} \pm z_{\alpha/2}\cdot\sqrt{s_{\bar{x}}^2+s_{\bar{y}}^2} \tag{2}$$

Where $\bar{x}$ and $\bar{y}$, is the mean of the female respectively the male temperature, and were calculated by the built in matlab function "mean" that returns the mean of the elements of the input.

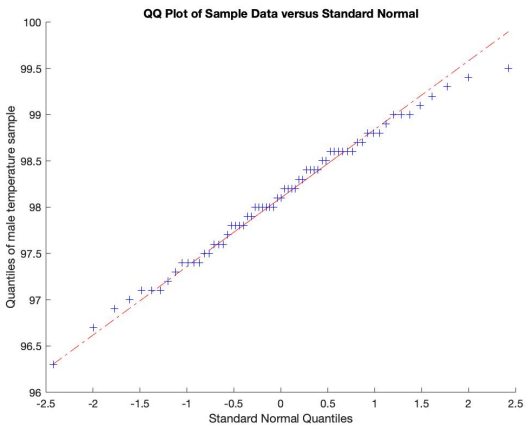
The variance of each sample is calculated by the built in function "var", which returns the variance of the elements of the input. The variance of each sample is thereby divided by each sample size in order to get $s_{\bar{x}}^2, s_{\bar{y}}^2$. By taking the square root of $s_{\bar{x}}^2, s_{\bar{y}}^2$, we get the estimated standard errors.

As the interval doesn't include zero, it is clear that we can reject the null hypothesis $H_0 : \mu_1 = \mu_2$, since if the mean were the same, the difference should be zero and thereby zero should be included in the confidence interval.

Since the sample sizes are large, a normal approximation is reasonable but since the null hypothesis was rejected, the distribution of the samples should be slightly different then from the normal distribution. By investigating the QQ-plots of the samples one can analyse if the quantiles come from the same distribution, i.e if we do a statistical analysis that assumes normal distribution, we use the normal QQ-plot to check that assumption. If the distribution of the sample is normal, then the data plot appears linear. As one can see in figure [1], the QQ-plot for the female sample shows some deviations, i.e outliers, which might be the reason for the deviation from the normal approximation.



(a) QQ-plot of the quantiles of female temperature sample against the theoretical quantile values from a normal distribution



(b) QQ-plot of the quantiles of male temperature sample against the theoretical quantile values from a normal distribution

Figure 1: QQ-plots of the quantiles of respectively sample data versus the theoretical quantile values from a normal distribution

(ii) The chosen parametric test to compare the body temperatures between males and females was the two-sample t-test, which assumes that two normal population distribution have equal variances. In our case the variances for the two samples are approximately equal, the variance for the male sample = 0.4883 and for the female sample = 0.5528. Given $\sigma_1^2 = \sigma_2^2 = \sigma^2$, the pooled sample variance is given by,

$$s_{\mathrm{p}}^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{n+m-2} = \frac{n-1}{n+m-2} \cdot s_1^2 + \frac{m-1}{n+m-2} \cdot s_2^2, \tag{3}$$

where m=n=the sample size, which is equal to 65. The exact confidence interval formula is given by,

$$I_{\mu_1-\mu_2} = \bar{x} - \bar{y} \pm t_{n+m-2}(\alpha/2) \cdot s_{\mathrm{p}} \cdot \sqrt{\frac{n+m}{nm}}, \tag{4}$$

with a 95% confidence level.

The resulting confidence interval was: [0.0412,0.5373], i.e we get the same result as in sub-exercise ai), that the null hypothesis is rejected.

(iii) The nonparametric test that was chosen to compare the body temperatures between males and females, was the rank sum test. The rank sum test is a nonparametric test for two independent samples, which does not assume normality of population distributions. We now assume continuous population distributions F1 and F2, and consider the null hypothesis:

$$H_0 : F_1 = F_2 \text{ against } H_1 : F_1 \neq F_2 \tag{5}$$

The rank sum test procedure starts with pooling of the samples and replacing the data values by their ranks, i.e 1, 2, . . . , n + m, starting from the smallest sample value to the largest. Then the two statistics r_1 = sum of the ranks of x-observations, and r_2 = sum of y-ranks are computed.

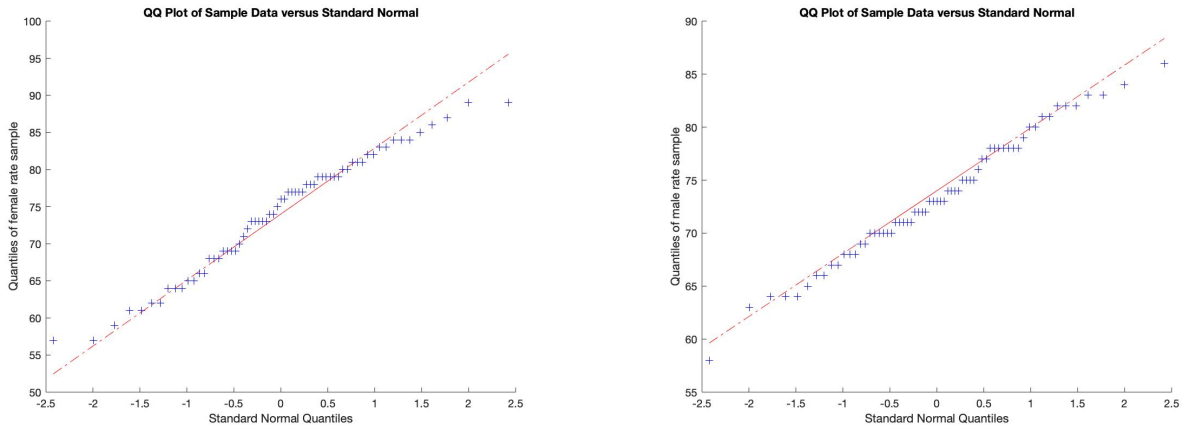
To perform the rank sum test, the built in function "ranksum" in matlab, where [p,h] = ranksum(x,y) returns a logical value indicating the test decision. When h = 1 the null hypothesis is rejected and when h = 0 the null hypothesis is not rejected at a 5% significance level.

The resulting p-value and the returned value of h from the "ranksum" function was $p = 0.0268$ and $h = 1$, i.e rejection of the null hypothesis, meaning that the samples aren't from the same distribution.

b.

The source code for this subsection can be found in the attached file, by opening the script-file "Exercise3B".

(i) By using normal theory and forming a 95% confidence interval for the difference of mean heart rates between males and females, the resulting confidence interval was: [-1.6490 3.2183]. The same assumption and calculations was done here as in exercise ai) for the difference in temperature in this report. In this case, the confidence interval actually includes zero, and thereby the null hypothesis: $H_0 : \mu_1 = \mu_2$ is not rejected. From figure [2] it is also clear that the distribution is normal, since the data plot seems linear and since the sample sizes are large, a normal approximation is reasonable here as well.



(a) QQ-plot of the quantiles of female heart rate sample against the theoretical quantile values from a normal distribution

(b) QQ-plot of the quantiles of male heart rate sample against the theoretical quantile values from a normal distribution

Figure 2: QQ-plots of the quantiles of respectively sample data versus the theoretical quantile values from a normal distribution

(ii) The chosen parametric test to compare the heart rates between males and females was the two-sample t-test, where the same assumtions and calculations were done as in subexercise aii) in this report. The resulted confidence interval was [-1.6490 3.2183], i.e we get the same result as in subexercise bi), that the null hypothesis is not rejected.

(iii) The nonparametric test that was chosen to compare the heart rates between males and females, was the rank sum test. The assumptions and calculations that was made, was the same as in exercise aiii) in this report.

The resulting p-value and the returned value of h from the "ranksum" function was $p = 0.3898$ and $h = 0$, i.e we cannot reject the null hypothesis, meaning that the samples are from the same distribution. This conclusion also matches the result in the two sample t-test from subexercise bii).