STAT 509 Homework 7 Solution

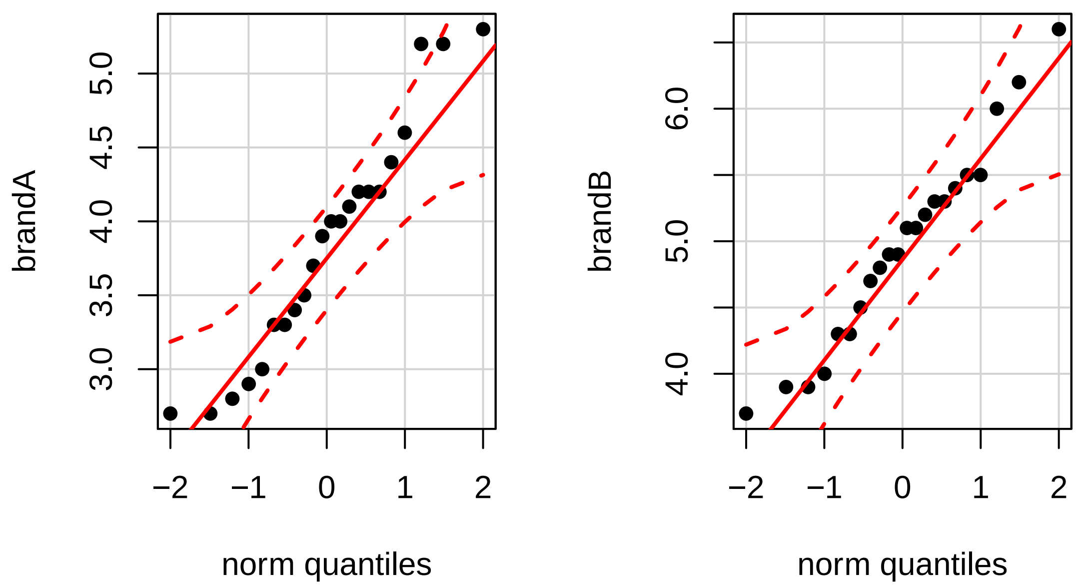
Chong Ma

April 3, 2017

Problem 1

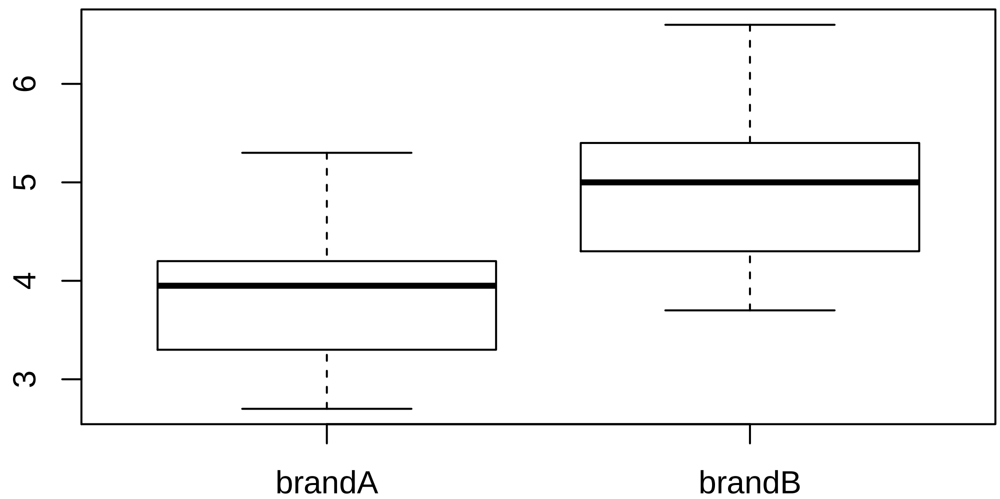
```
brandA=c(3.5,2.7,3.9,4.2,4.6,2.7,3.3,5.2,4.2,2.9,4.4,
         5.2,4.0,4.1,3.4,3.3,4.2,5.3,3.7,3.0,4.0,2.8)
brandB=c(4.7,3.9,4.5,5.5,4.0,5.3,4.3,6.0,5.3,3.7,5.5,
         6.2,5.1,5.4,4.8,4.9,6.6,4.3,4.9,5.1,3.9,5.2)

##check the normal assumption for both of data
op<-par(mfrow=c(1,2))
qqPlot(brandA,distribution = "norm",pch=16)
qqPlot(brandB,distribution = "norm",pch=16)
```




```
par(op)

##check if the two variances are equal or not
boxplot(cbind(brandA,brandB))
```



(a)

Step 1. Assume that

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{i.i.d}{\sim} N(\mu_1, \sigma_1^2)$$
$$Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{i.i.d}{\sim} N(\mu_2, \sigma_2^2)$$

The normal qq plots for both of Brand A and Brand B data demonstrate no apparent departure from the normal assumption.

Step 2. The boxplots indicate that the two sample variances are not obviously different from each other, therefore we can consider $\sigma_1 = \sigma_2$ for constructing the 95% confidence interval (I use a default confidence level 95%). Note that $\bar{y}_1 = 3.845$, $\bar{y}_2 = 4.959$ and $\sigma_1 = 0.799, \sigma_2 = 0.767$, and $s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} = 0.171$, thus substitute the values into the followling formula and we obtain

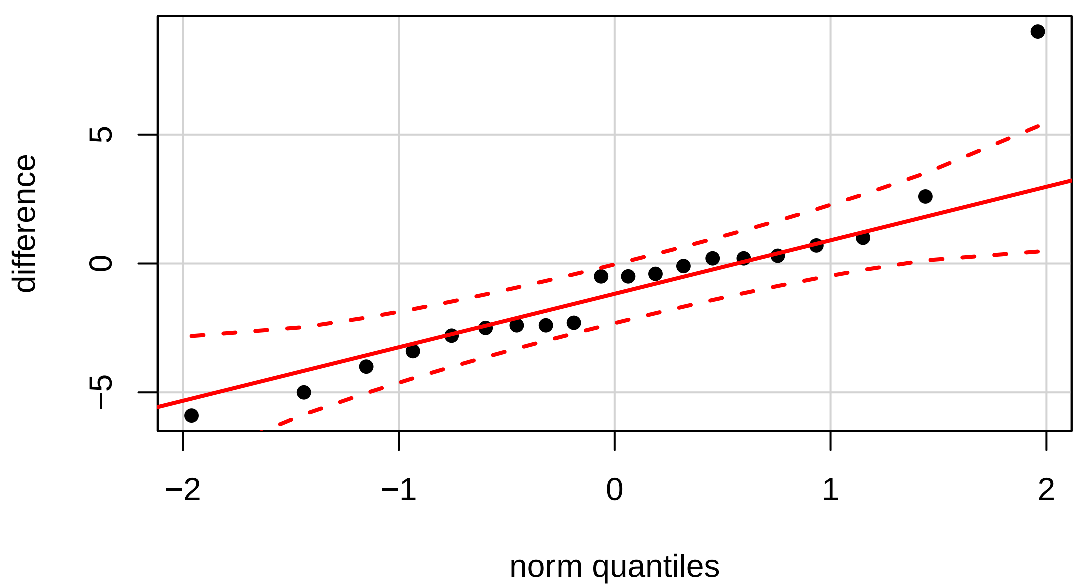
$$[\bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2,\alpha/2}\sqrt{S_P^2(\frac{1}{n_1} + \frac{1}{n_2})}] = [-1.2176, -1.0096]$$

Step 3. We are 95% confident that the difference of population mean time it takes the paint to dry for Brand A and Brand B is between -1.22 and -1.01. Because 0 is not contained in the according confidence interval, we have strong evidence to conclude that the population mean time it takes the paint to dry for Brand A and Brand B differ each other.

(b) It is an open question. It would be possible to conduct a matched-pairs design for comparing the two brands of latex paint. Because matched-pairs design requires dependent samples, each of which produces corresponding measurements for the two brands of latex paint. When it comes to the measurement of time it takes the paint to dry, it is possible that paint a sample of construction objects on both sides, each side for a construction object with a brand of latex paint and then record the time it takes both of latex paint to dry. In this case, we could consider it as a matched-pairs design. Of course, there are probably a lot of different ways to conduct matched-pairs design by using different measurements of the quality of the two brands of latex paint.

Problem 2

```r
plasma=c(2.5,3.1,2.1,3.5,3.1,1.8,6.0,3.0,36.0,4.7,
         6.9,3.9,4.2,1.6,7.2,1.8,20.0,2.0,2.5,4.1)
fat=c(4.9,5.9,4.4,6.9,9.0,4.2,10.0,5.5,41.0,4.4,
      7.0,2.9,4.6,1.4,7.7,1.1,11.0,2.5,2.3,1.5)
##get the difference of TCDD levels in the two tissues
difference=plasma-fat
##check the normal assumption for the difference
qqPlot(difference,distribution = "norm",pch=16)
```



(a) It is likely that the TCDD levels in different organisms might triger or result from different disease. So it makes sense to research how the TCDD levels distribute in organisms in order to further study how different levels of TCDD cause or lead to a certain disease.

(b) Denote that μ_p is the population mean TCDD levels in Plasma and μ_f is the population mean TCDD levels in Fat Tissues. Because this is a matched-pairs design, the inference study for $\mu_D = \mu_p - \mu_f$ reduces to a one-sample inference. A 95% confidence interval for μ_D is $[-2.397, 0.577]$ calculated as follows.

$$[\bar{x}_D \pm t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}] = [-2.397, 0.577]$$

We are 95% confident that the difference of population TCDD levels in Plasma and Fat Tissue are between -2.397 and 0.577. Becasue 0 is contained in the according confidence interval, we don't have sufficient evidence to conclude that the population TCDD levels are different compared in plasma to in Fat Tissue.

(c) The matched-pairs design can remove variations among the veterans, which makes compare the TCDD levels for the plasma and Fat Tissue under more homogeneous conditions. In this sense, the inference about the population mean difference is more precise compared the matched-pairs design to the two independent samples design.

Problem 3

(a) Note that the sample proportion of exceedence for airline SAS is $\hat{p}_1 = \frac{8}{86} \approx 0.093$ and the sample proportion of exceedence for airline Lufthanza is $\hat{p}_2 = \frac{10}{142} \approx 0.07$. The 95% confidence interval is $(-0.052, 0.097)$ calculated as follows.

$$[(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}] = [-0.0518, 0.097]$$

Because in the sample each airline has at least 5 exceedences and 5 normals, that is,

$$n_1\hat{p}_1 = 8 \geq 5, n_1(1-\hat{p}_1) = 78 \geq 5$$
$$n_2\hat{p}_2 = 10 \geq 5, n_2(1-\hat{p}_2) = 132 \geq 5$$

there is no concern for us to derive the confidence interval by using the above formula.

(b) Recall the optimal sample size for one sample inference that is

$$n = (\frac{Z_{\alpha/2}}{B})^2 p_0(1-p_0)$$

where B is the required margin of error and α relates to the required confidence level and p_0 is the best guess for the population proportion. In this specific problem, the optimal sample size can be calculated by

$$B = Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n}}$$

After algebric manipulations, we have

$$n = (\frac{Z_{\alpha/2}}{B})^2(p_1(1-p_1) + p_2(1-p_2))$$

Note that $B = 0.03$, $\alpha = 1 - 0.99 = 0.01$, therefore $Z_{0.01/2} = 2.58$. The best guesses for p_1 and p_2 can be achieved by preivous study, which is $p_1 = 0.093$ and $p_2 = 0.07$. Hence, the optimal sample size is 1105 that is calculated as follows.

$$n = (\frac{Z_{\alpha/2}}{B})^2(p_1(1-p_1) + p_2(1-p_2)) = (\frac{2.58}{0.03})^2(0.093(1-0.093) + 0.07(1-0.07)) = 1105.3396$$