Homework 12 Solution

STAT 509 Statistics for Engineers
Summer 2017 Section 001
Instructor: Tahmidul Islam

1. A textile fiber manufacturer is investigating a new drapery yarn, which the company claims has a mean thread elongation of 12 kilograms with a standard deviation of 0.5 kilograms. The company wishes to test the hypothesis

$$H_0 : \mu = 12$$
$$H_a : \mu \neq 12$$

using a random sample of four specimens. Suppose the random sample is from a normal population. *(Hint: notice in this question, the population variance is assumed to be known with $\sigma = 0.5$)*

(a) Given the sample mean $\overline{y}$ is 11.3 and the confidence level is 95%, follow the 4-step procedure to conduct a hypothesis test. What is your conclusion?

- Step 1: State H_0 and H_a.

$$H_0 : \mu = 12$$
$$H_a : \mu \neq 12$$

- Step 2: Calculate the test statistic.

$$z_0 = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} = \frac{11.3 - 12}{0.5/\sqrt{4}} = -2.8$$

- Step 3: Calculate p-value.

$$\text{p-value} = 2P(Z < -|-2.8|) = 2P(Z < -2.8) = 2 \times 0.0026 = 0.0052$$

- Step 4: Make Conclusion.
Because p-value$= 0.0052 < 0.05 = \alpha$, we reject H_0. With 95% confidence, we conclude that the population mean thread elongation is not 12 kilograms.

(b) Using the confidence interval approach to calculate a 95% two-sided confidence interval for μ. Does the confidence interval cover 12? Is the results of confidence interval consistent to the testing conclusion?

$$\begin{aligned} \text{C.I.} &= \bar{Y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &= 11.3 \pm 1.96 \times \frac{0.5}{\sqrt{4}} \\ &= 11.3 \pm 0.49 \\ &= (10.81, 11.79) \end{aligned}$$

The 95% confidence interval does not cover 12, which means we should conclude that the population mean thread elongation is not 12 kilograms by the confidence interval. It is consistent to the testing result.

(c) What is margin of error and the length of interval in (b)? If we want to control the length of the confidence interval to be 0.6, how many observations do we need in the sample?

$$\text{Margin of Error} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{0.5}{\sqrt{4}} = 0.49$$
$$\text{Length of CI} = 2 \times \text{Margin of Error} = 2 \times 0.49 = 0.98$$

If we want to control the length of the confidence interval to be 0.6, the margin of error must be $\frac{0.6}{2} = 0.3$. Therefore, we need to solve the following equation to find the number observations needed.

$$0.3 = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{0.5}{\sqrt{n}}$$

which gives

$$n = \left(\frac{1.96 \times 0.5}{0.3}\right)^2 = 10.67 \approx 11$$

2. A manufacturing firm is interested in the mean batteries hours used in their electronic games. To investigate mean batteries life in hours, say μ. The following data are collected

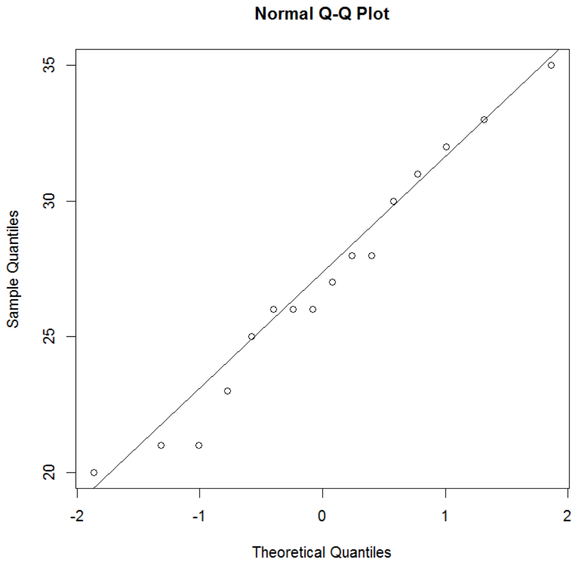
```
20,25,21,28,21,30,23,27,26,26,28,31,26,32,33,35
```

(Hint: the population variance is not given, therefore we assume it is not known)

(a) Is it reasonable to assume that the sample data has come from a normal distribution? The R code is given below (*Hint: use fat pencil test in R.*)

```
battery<-c(20,25,21,28,21,30,23,27,26,26,28,31,26,32,33,35)
qqnorm(battery)
qqline(battery)
```

By the QQ plot, we see that most of the points are close to the 45 degree line, which means we can use a "fat pencil" to cover those points. Thus, it is reasonable to assume that the sample data has come from a normal distribution.



(b) Suppose it is reasonable to assume the data has come from a normal distribution, construct a 99% two-sided confidence interval for μ. The quantile can be found via R or t-table. The sample mean and the sample standard deviation can be computed via the following command:

```
mean(battery)
sd(battery)
```

By R

```
> mean(battery)
[1] 27
> sd(battery)
[1] 4.442222
```

Therefore,

$$\begin{aligned}
\text{C.I.} &= \bar{Y} \pm t_{n-1,\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \\
&= 27 \pm t_{16-1,\frac{0.01}{2}} \frac{4.44}{\sqrt{16}} \\
&= 27 \pm 2.947 \left(\frac{4.44}{4}\right) \\
&= 27 \pm 3.27 \\
&= (23.73, 30.27)
\end{aligned}$$

(c) **Using R** to test the following hypothesis with the level of significance $\alpha = 0.01$:

$$\begin{aligned}
H_0 &: \mu = 24 \\
H_a &: \mu \neq 24
\end{aligned}$$

You need to print out both your R code and testing results.

Based on the R results, p-value is 0.01641, which is greater than $\alpha = 0.01$. Therefore, we fail to reject the H_0, and say with 99% confidence, the mean battery life is 24 hours. Since 24 is included in the 99% confidence interval, we get the same conclusion comparing to the result in (b).

```
> t.test(battery, alternative="two.sided", mu=24)

One Sample t-test

data:  battery
t = 2.7014, df = 15, p-value = 0.01641
alternative hypothesis: true mean is not equal to 24
95 percent confidence interval:
24.63291 29.36709
sample estimates:
mean of x
27
```

3. Inexperienced data analysts often erroneously place too much faith in qq plots when assessing whether a distribution adequately represents a data set (especially when the

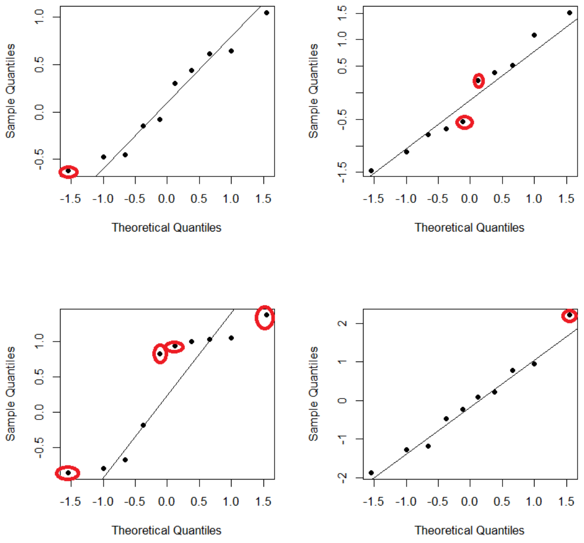
sample size is small). The purpose of this problem is to illustrate to you the dangers that can arise. In this problem, you will use R to simulate the process of drawing repeated random samples from a given population distribution and then creating normal probability plots (Q-Q plots). Follow the code provided

(a) Generate your own data and create a qq plot for each sample using this R code:

```
# create 2 by 2 figure
par(mfrow = c(2,2))
B = 4
n = 10
# create matrix to hold all data
data = matrix(round(rnorm(n*B,0,1),4), nrow = B, ncol = n)
# this creates a qq plot for each sample of data
for (i in 1:B){
qqnorm(data[i,],pch=16,main="")
qqline(data[i,])
}
```

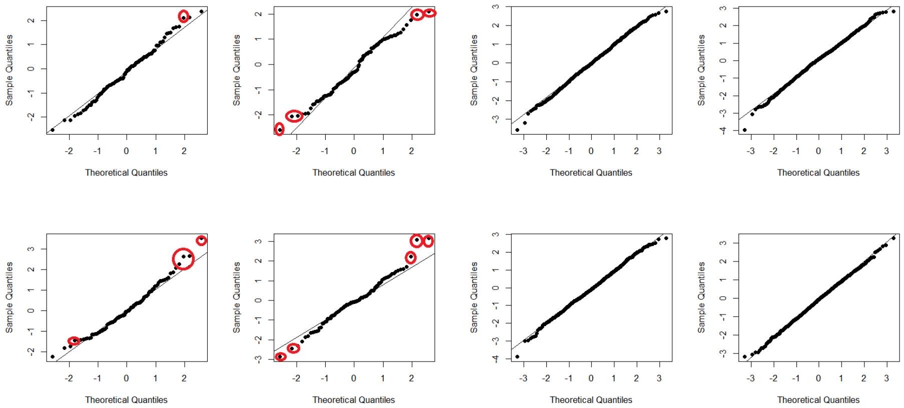
mark the qq plot that appears to violate the normal assumption the most. Note: In theory, all of these plots should display perfect linearity! Why? Because we are generating the data from a normal distribution! **Therefore, even when we create normal qq plots with normally distributed data, we can get plots that don't look perfectly linear.** This is a byproduct of sampling variability. This is why you don't want to rush to discount a distribution as being plausible based on a single plot, especially when the sample size n is small (e.g. $n = 10$).

Each student should generate different 4 pieces of qq plots. Based on my data, we can find that there are several observation points which are pretty far away from the 45 degree line.



(b) Increase your sample size to $n = 100$ and repeat. What happens? What if $n = 1000$? Just change n in the R code on the last page and re-run.

The left four qq plots are generated with $n = 100$ and the right ones are generated with $n = 1000$. When $n = 100$ we still can find a couple of points deviating from the 45 degree line, however, the proportion of deviated points are decreased. Once we increase the sample size to 1000, we get almost perfet lines. After all, we sample observations directly from the normal distribution! This problem also tells us when we have more information (observations), we can reveal the truth more precise.



(c) Take $n = 100$, replace

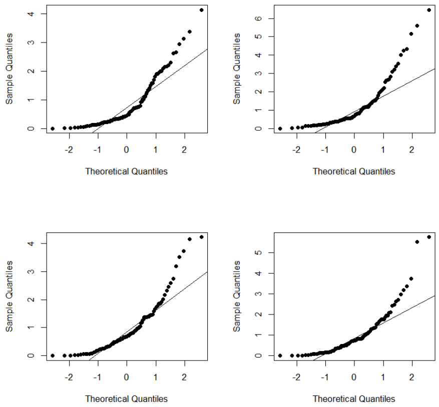
```
data = matrix(round(rnorm(n*B,0,1),4), nrow = B, ncol = n)
```

with

```
data = matrix(round(rexp(n*B,1),4), nrow = B, ncol = n)
```

and re-run. By doing this, you are changing the underlying population distribution from $\mathcal{N}(0, 1)$ to exponential(1). What do these normal qq plots look like? Are you surprised?

The shape of the exponential pdf is right skewed. We are supposed to get qq plots, which are highly deviated from the 45 degree line. By my generated data, we get what we expected. No surprise at all!



4. Two machines are used for filling plastic bottles with a net volume of 16.0 ounces. The fill volume can be assumed to be normal with standard deviation $\sigma_1 = 0.020$ and $\sigma_2 = 0.025$ ounces. A member of the quality engineering staff suspects that both machines fill to the same mean net volume, whether or not this volume is 16.0 ounces. A random sample of 10 bottles is taken from the output of each machine.

```
Machine 1:  16.03, 16.04, 16.05, 16.05, 16.02, 16.01, 15.96, 15.98, 16.02,
15.99
Machine 2:  16.02, 15.97, 15.96, 16.01, 15.99, 16.03, 16.04, 16.02, 16.01,
16.00
```

(a) Do you think the engineer is correct? Conduct a formal 4-step procedure with $\alpha = 0.05$. What is your conclusion? *(Hint: Sample means can be computed using R.)*

By R, we have $\bar{y}_1 = 16.015$ and $\bar{y}_2 = 16.005$.
Step 1: The null and alternative hypothesis

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0$$

Step 2: The test statistic is

$$z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{16.015 - 16.005}{\sqrt{\frac{0.02^2}{10} + \frac{0.025^2}{10}}} = 0.99$$

Step 3: Calculate p-value

$$\text{p-value} = 2P(Z < -|z_0|) = 2P(Z < -0.99) = 2 \times 0.1611 = 0.3222$$

Step 4: Because p-value $= 0.3222 > 0.05 = \alpha$, we fail to reject the null, which means we are 95% confident that both machines fill to the same mean net volume.

(b) Calculate a 95% confidence interval on the difference in population means. Provide a practical interpretation of this interval.

$$\begin{aligned}\text{C.I.} &= \bar{y}_1 - \bar{y}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= 16.015 - 16.005 \pm 1.96\sqrt{\frac{0.02^2}{10} + \frac{0.025^2}{10}} \\ &= (-0.010, 0.030)\end{aligned}$$

The 95% confidence interval on $\mu_1 - \mu_2$ is $(-0.010, 0.030)$, in which 0 is included. We have sufficient evidence that both macines fill to the same mean net volume.

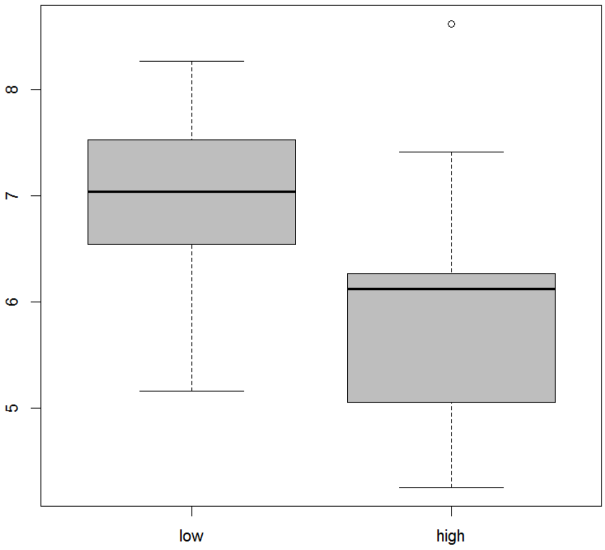
5. Data on pH for 16 random batches of low and high volt electrolyte were collected. The data are given by

```
Low volt:  7.78, 5.77, 7.08, 6.75, 7.09, 8.27, 6.5, 5.16, 6.81, 7.28, 7.88,
7.87, 7.2, 5.95, 6.58, 6.99
high volt:  4.54, 5.04, 5.07, 6.18, 8.62, 6.28, 7.41, 6.17, 6.25, 4.25, 6.08,
7.23, 4.68, 6.19, 5.85, 5.83
```

(a) Population variances σ_1^2 and σ_2^2 are unknown. Do you think it is reasonable to assume $\sigma_1^2 = \sigma_2^2$? Let's figure it out! First, draw a side-by-side boxplot in R. Based on the boxplot, do you believe $\sigma_1^2 = \sigma_2^2$?

```
low <- c(7.78,5.77,7.08,6.75,7.09,8.27,6.5,5.16,6.81,7.28,7.88,7.87,7.2,
5.95,6.58,6.99)
high <- c(4.54,5.04,5.07,6.18,8.62,6.28,7.41,6.17,6.25,4.25,6.08,7.23,
4.68,6.19,5.85,5.83)
boxplot(low,high,names=c("low","high"),col="grey")
```

The width of two boxes are similar, which indicates that their variances might be the same.



(b) Now, let's conduct a formal test of

$$H_0 : \sigma_1^2/\sigma_2^2 = 1$$
$$H_a : \sigma_1^2/\sigma_2^2 \neq 1$$

using the following R code.

```
var.test(low, high)
```

What is the **p-value** of the testing result? With significance level 0.05, do you reject H_0 or fail to reject H_0? Is the result consist to the one you get from (a)?

Here is the R output:

```
F test to compare two variances

data:  low and high
F = 0.5338, num df = 15, denom df = 15, p-value = 0.2356
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.1865103 1.5278126
sample estimates:
ratio of variances
0.5338097
```

p-value is 0.2356, which is greater than 0.05. This means that we fail to reject the H_0, therefore, we have sufficient evidence to say two variances are the same. The result is consist to (a).

(c) Assuming the two samples are independent. The engineer want to test that the low volt average pH is greater than the high volt average pH. Let μ_L be the average pH of low volt electrolyte and μ_H be the average pH of high volt electrolyte. State the null and alternative hypotheses.

$$\begin{aligned} H_0 &: \mu_L - \mu_H = 0 \\ H_a &: \mu_L - \mu_H > 0 \end{aligned}$$

(d) Calculate the appropriate test statistic for the test. The sample means and sample variances can be computed using R.

```
> mean(low)
[1] 6.935
> var(low)
[1] 0.6896
> mean(high)
[1] 5.979375
> var(high)
[1] 1.291846
```

R output tells us that $\bar{y}_1 = 6.9350$, $\bar{y}_2 = 5.9793$, $S_1^2 = 0.6896$, $S_2^2 = 1.2918$. We need to calculate the pooled variance S_p^2 first, because we assume two variances are the same.

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{15 \times 0.6896 + 15 \times 1.2918}{16+16-2} = 0.9907$$

The test statistic is

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{6.9350 - 5.9793}{\sqrt{0.9907} \times \sqrt{\frac{1}{16}+\frac{1}{16}}} = 2.716$$

(e) Use R to calculate the p-value of the test. *(Hint: $P(T_{n_1+n_2-2} > t_0)$ can be calculated using R with command* `1 - pt(`t_0`, `$n_1 + n_2 - 2$`)`*).*

```
> 1 - pt(2.716, 16+16-2)
[1] 0.005428769
```

By R result, p-value is approx. 0.005.

(f) Make decision and state your conclusion at a 0.05 level of significance.

Since $0.005 < 0.05$, we reject the H_0, and conclude with 95% confidence that low volt average pH is greater than the high volt average pH.

(g) Use `t.test` in R to check your work.

```
t.test(low,high,alternative="greater",paired = FALSE, var.equal = TRUE)
```

We achieve the same p-value as the R output.

```
Two Sample t-test

data:  low and high
t = 2.7155, df = 30, p-value = 0.005435
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
0.3583425       Inf
sample estimates:
mean of x mean of y
6.935000  5.979375
```

(h) Construct a two-sided 95% confidence interval of $\mu_L - \mu_H$ by hand. Provide a practical interpretation of this interval.

$$\begin{aligned}\text{C.I.} &= (\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2,n_1+n_2-2}\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})} \\ &= (6.9350 - 5.9793) \pm 2.042\sqrt{0.9907(\frac{1}{16} + \frac{1}{16})} \\ &= (0.237, 1.674)\end{aligned}$$

The 95% confidence interval of $\mu_L - \mu_H$ is $(0.237, 1.674)$, in which 0 is not included, we claim that we have sufficient evidence to say the low volt average pH is greater than the high volt average pH.