Stat509 Fall 2014 Homework 6

Instructor: Peijie Hou

October 26, 2014

Instruction: Please finish this homework before the class on 11/6. You will have a quiz based on this homework during the class.

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 $H_0 : \mu = 12$ against $H_a : \mu \neq 12$ 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) Follow the 4 steps of conducting a hypothesis test, what is your conclusion if the sample mean $\overline{y} = 11.3$ and we use $\alpha = 0.05$.?

   - Step 1: $H_0 : \mu = 12$ vs. $H_a : \mu \neq 12$
   - Step 2:
   $$z_0 = \frac{11.3 - 12}{0.5/\sqrt{4}} = -2.8$$
   - Step 3: p-value$= 2P(Z < -|-2.8|) = 0.005$
   - Step 4: Since p-value$< \alpha$, we reject H_0. We have sufficient evidence to conclude that the mean thread elongation of a new drapery yarn 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?

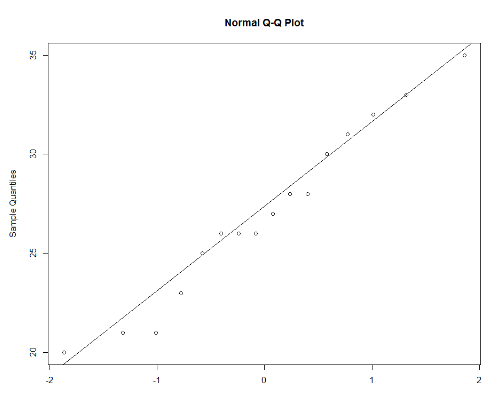
   $$11.3 \pm 1.96 * \frac{0.5}{\sqrt{4}} = (10.81, 11.79)$$

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.*)



The points are approximately on the straight line, there is no gross departure from the normal assumption. Therefore, it is reasonable to assume the data comes 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 standard deviation can be computed via the following command:

The CI formula is

$$\overline{Y} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$$

where $\overline{y} = 27$, $s = 4.44$ and $t_{15,0.005} = 2.947$. Therefore, the 95% CI is:

$$27 \pm 2.947 \times \frac{4.44}{\sqrt{16}} = (23.73, 30.27)$$

Note, t quantile can be found using t-table, or via the following command:

```
> qt(0.995,15)
[1] 2.946713
```

(c) Construct a hypothesis testing question (4 steps) to test the following hypothesis: $H_0 : \mu = 24$ vs. $H_a : \mu \neq 24$. The p-value can be found through R. The significance level $\alpha = 0.01$.

- Step 1: $H_0 : \mu = 24$ vs. $H_a : \mu \neq 24$
- Step 2:

$$t_0 = \frac{27 - 24}{4.44/\sqrt{16}} = 2.70$$

- Step 3: p-value$= 2P(t < -|2.70|) = 0.016$
- Step 4: Since p-value $> \alpha$, we do not reject H_0. We do not have sufficient evidence to conclude that the mean battery life is not 24 hours.

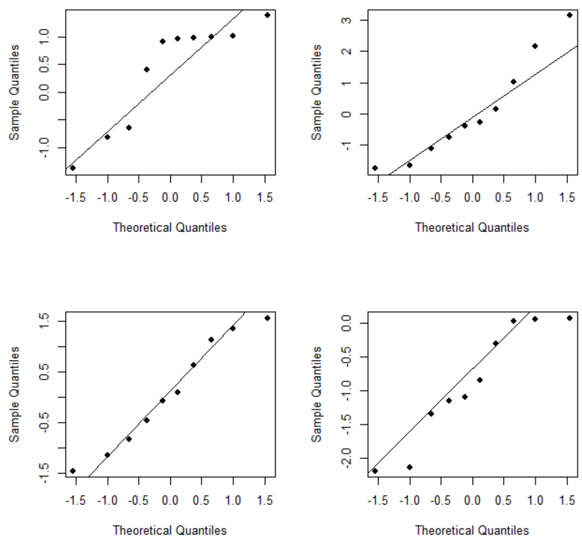
3. (For fun) Inexperienced data analysts often erroneously place too much faith in qq plots when assessing whether a distribution adequately represents a data set (especially when the sample size is small).The purpose of this problem is to illustrate to you the dangers that can arise. In this problem, 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:

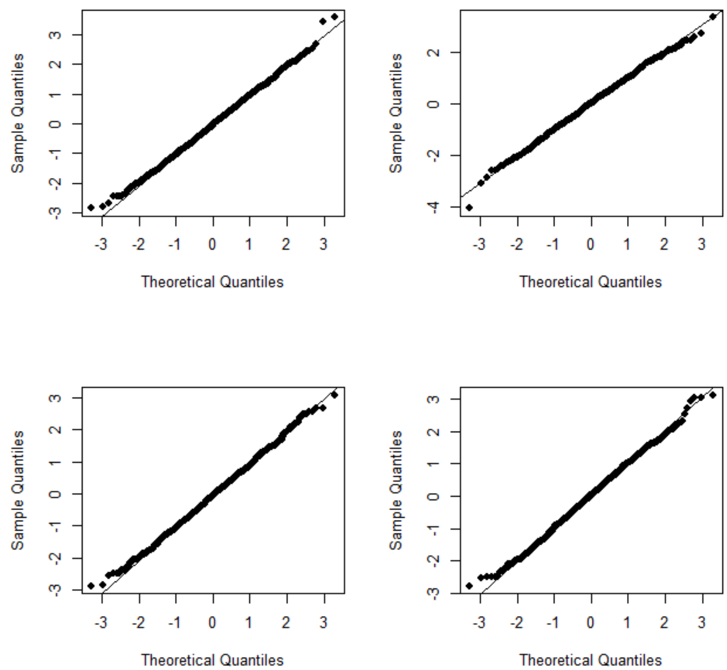
```
# create 2 by 2 figure
par(mfrow = c(2,2))
B = 4
n = 10
# create matrix to hold all data
data = matrix(round(rnorm(n*B,0,1),4), nrow = B, ncol = n)
# this creates a qq plot for each sample of data
for (i in 1:B){
   qqnorm(data[i,],pch=16,main="")
   qqline(data[i,])
}
```

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 (like $n = 10$).
The plot one the top left corner apparently violates the normal assumption. Note: your plot should different from mine plot due to the nature of random number generation.



(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.
Let us look at the case when $n = 1000$. Now, all the Q-Q plots works almost perfectly! It suggests that we can trust Q-Q plot if our sample size is large.



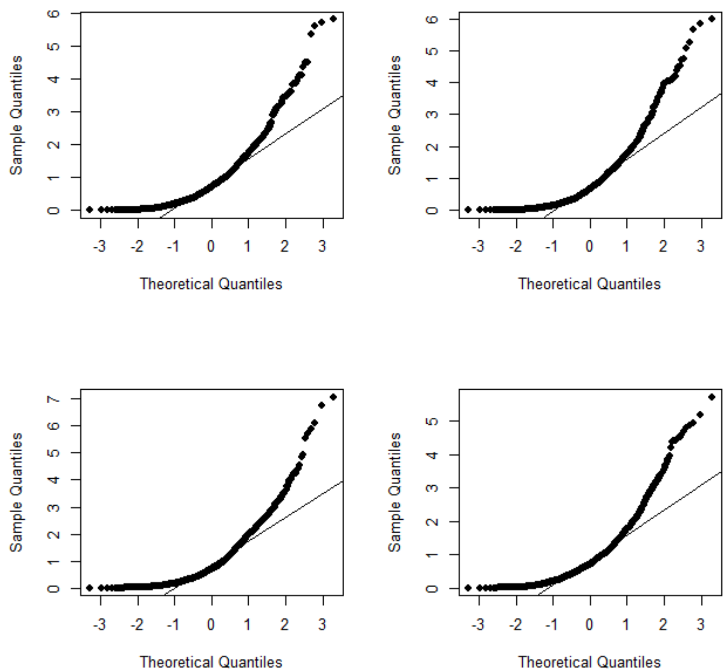
(c) Take $n = 100$, replace

```
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? No Surprise, we expect to see curvature in the Q-Q plots, since we generate our data from exponential distribution instead of normal.


Sample Quantiles
Theoretical Quantiles
Sample Quantiles
Theoretical Quantiles
Sample Quantiles
Theoretical Quantiles
Sample Quantiles
Theoretical Quantiles