Homework 2 Problem Set

Statistics E-109

Due February 16, 2022 at 5:50 pm EST

Homework policies. Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving R, include the code in your solution, along with any plots.

Please submit your homework assignment via Canvas as a PDF file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Max points: 100

Name: Erin Lopez

PART 1 (30 points)

Use y <- rnorm(100) to generate a random sample of size 100 from a normal distribution.

```
set.seed(1311)
y <- rnorm(100)
```

1A

Calculate the mean and standard deviation of y.

```
mean(y)
```

```
## [1] -0.04795566
```

```
sd(y)
```

```
## [1] 1.006503
```

The mean of y is -0.048 with a standard deviation of 1.01.

1B

Use a loop to repeat the above calculation 30 times. Store the 30 means in a vector named AVG. Calculate the standard deviation of the values in AVG.

```
M <- 30
AVG <- numeric(M)
for (i in 1:M) {
  y <- rnorm(100)
  AVG[i] <- mean(y)
}
sd(AVG)
```

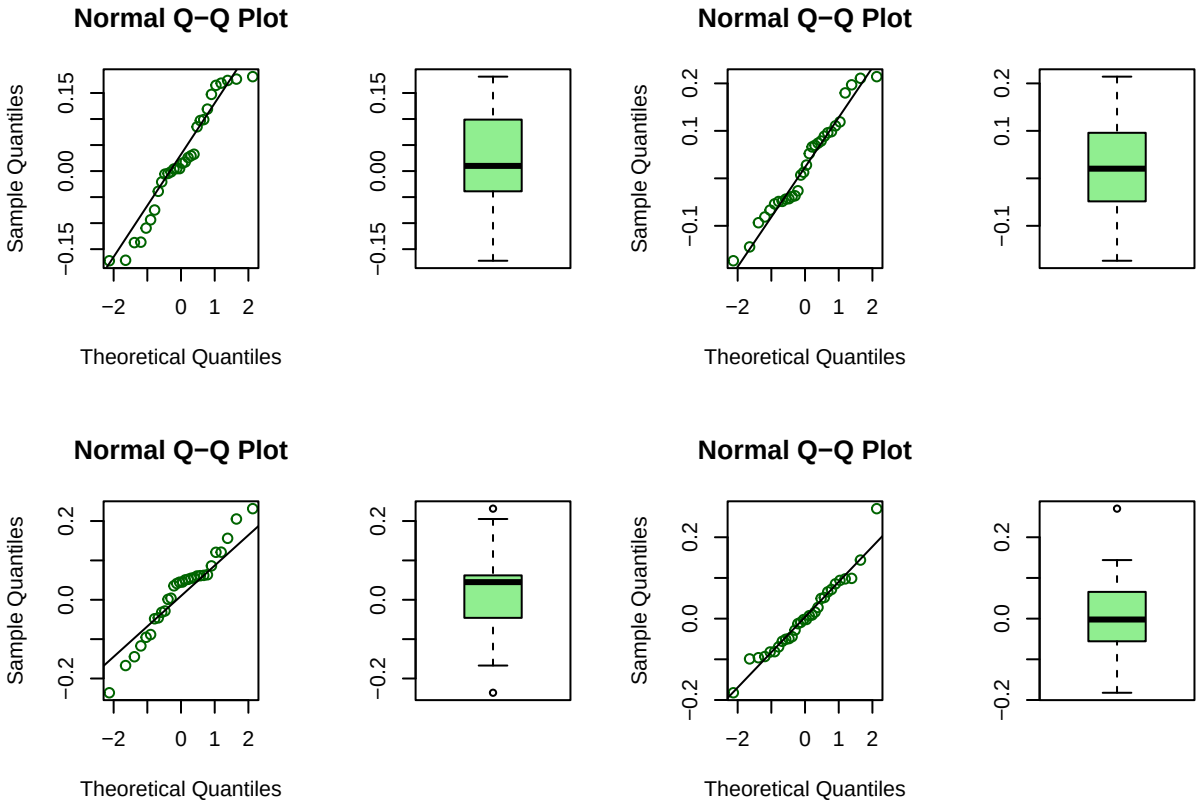
```
## [1] 0.1175322
```

The standard deviation of the values in AVG is 0.118.

1C

Run (b) 4 times, showing each of the distributions of 30 means in a normal probability plot and box plot.

```
par(mfrow = c(2,4))
for (i in 1:4){
  M <- 30
  AVG <- numeric(M)
  for (i in 1:M) {
    y <- rnorm(100)
    AVG[i] <- mean(y)
  }
  sd(AVG)
  qqnorm(AVG, col = 'dark green')
  qqline(AVG)
  boxplot(AVG, col = 'light green')
}
```



The boxplot of the second run shows the most normal distribution, with evenly weighted sides and the median in the middle. The first and fourth run show very slight skewing. The third run shows even boxplot whiskers; however, the median is skewed. The normal probability plots are the most linear for run two and four; however, there is some deviation from the fit line on the ends. The third run shows greater deviation and the line is not as good of a fit.

Part 2 (30 points)

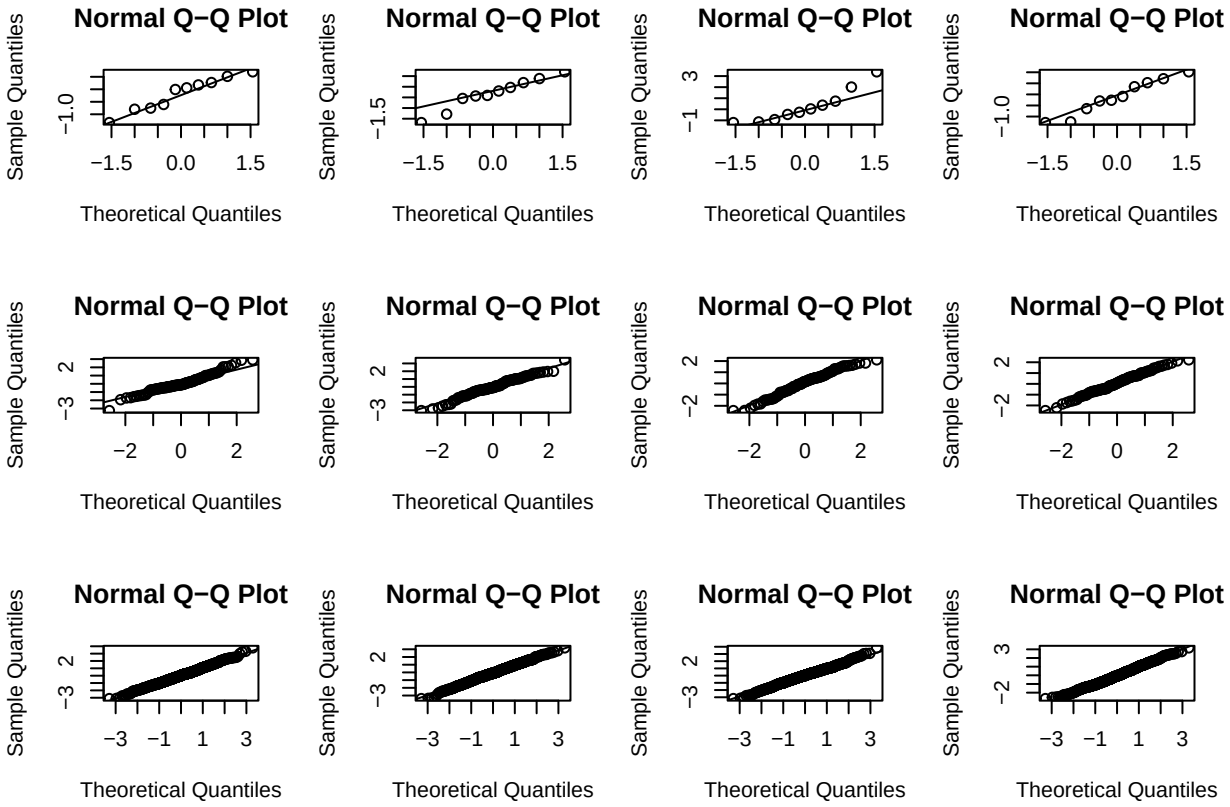
Use mfrow to set up the layout for a 3 by 4 array of plots. In the top 4 panels, show normal probability plots for 4 separate "random" samples of size 10, all from a normal distribution. In the middle 4 panels, display plots for samples of size 100. In the bottom 4 panels, display plots for samples of size 1000. How does the appearance of plots change as the sample size changes?

```
par(mfrow = c(3, 4))

#In the top 4 panels, show normal probability plots for 4 separate "random" samples of size 10, all fro
for (i in 1:4) {
  r <- rnorm(10)
  qqnorm(r)
  qqline(r)
}

#In the middle 4 panels, display plots for samples of size 100.
for (i in 1:4) {
  r <- rnorm(100)
  qqnorm(r)
  qqline(r)
}

#In the bottom 4 panels, display plots for samples of size 1000.
for (i in 1:4) {
  r <- rnorm(1000)
  qqnorm(r)
  qqline(r)
}
```

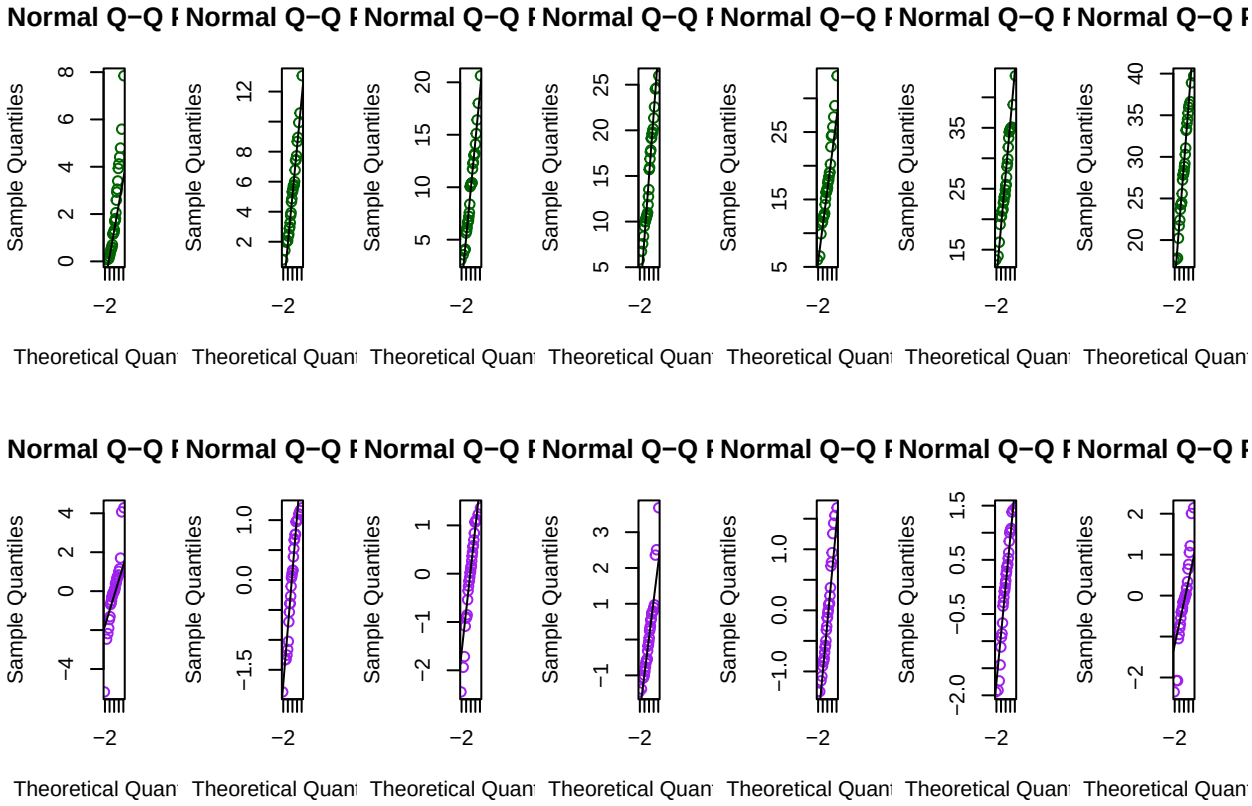


Comments **Greater sample size results in more linear plots. With a sample size of 10, deviation from the qqline is very apparent. A sample size of 100 improves the linearity; however, there is still some deviation from the line at the ends. With 1000 samples, the plot looks very linear. While there is still deviation at the ends, it is much less than with the sample size of 100.**

Part 3 (30 points)

The statement x <- rchisq(n, 1) generates n random values from a chi-squared distribution with one degree of freedom. The statement x <- rt(n, 1) generates n random values from a t-distribution with one degree of freedom. Make normal probability plots using 30 random values for various degrees of freedom from each of these distributions. Approximately how large degrees of freedom is necessary, in each instance, to obtain a consistent normal distribution shape?

```
par(mfrow = c(2, 7))
qqnorm(rchisq(30, 2), col = 'dark green')
qqline(rchisq(30, 2))
for (i in 1:6){
  x <- rchisq(30, i*5) #df: 5, 10, 15, 20, 25, 30
  qqnorm(x, col = 'dark green')
  qqline(x)
}
qqnorm(rt(30, 2), col = 'purple')
qqline(rt(30, 2))
for (i in 1:6){
  t <- rt(30, i*5)
  qqnorm(t, col = 'purple')
  qqline(t)
}
```



Comments **Testing degrees of freedom of 2, 5, 10 15, 20, 25, and 30, rchisq consistently showed a linear normal distribution shape starting at 10 degrees of freedom. This consistent normal distribution shape was found from 10 to 30 degrees of freedom. The rt probability plots showed consistent normal distribution shape at 15 degrees of freedom. This remained true for 15 to 25 degrees of freedom. In this run specifically, 5-30 degrees of freedom showed consistent normal distribution for chi-squared, and 5-25 degrees of freedom showed consistent normal distribution for rt.**

Part 4 (10 points)

A bank has kept records of the checking balances of its customers and determined that the average daily balance of its customers is $350 with a standard deviation of $58. A random sample of 144 checking accounts is selected.

4A

What is the probability that the sample mean will be less than $356.60?

```
sd = 58/sqrt(144)
mean = 350
pnorm(356.6, mean, sd)
```

```
## [1] 0.9139547
```

The probability that the sample mean will be less than $356.60 in 0.91 or 91.4%.

4B

What is the probability that the sample mean will lie between $340 and $350?

```
pnorm(350, mean, sd) - pnorm(340, mean, sd)
```

```
## [1] 0.4807253
```

The probability that the sample mean will lie between $340 and $350 is 0.48 or 48.07%.