S520 Problem Set 10

Solution Key

04/04/2023

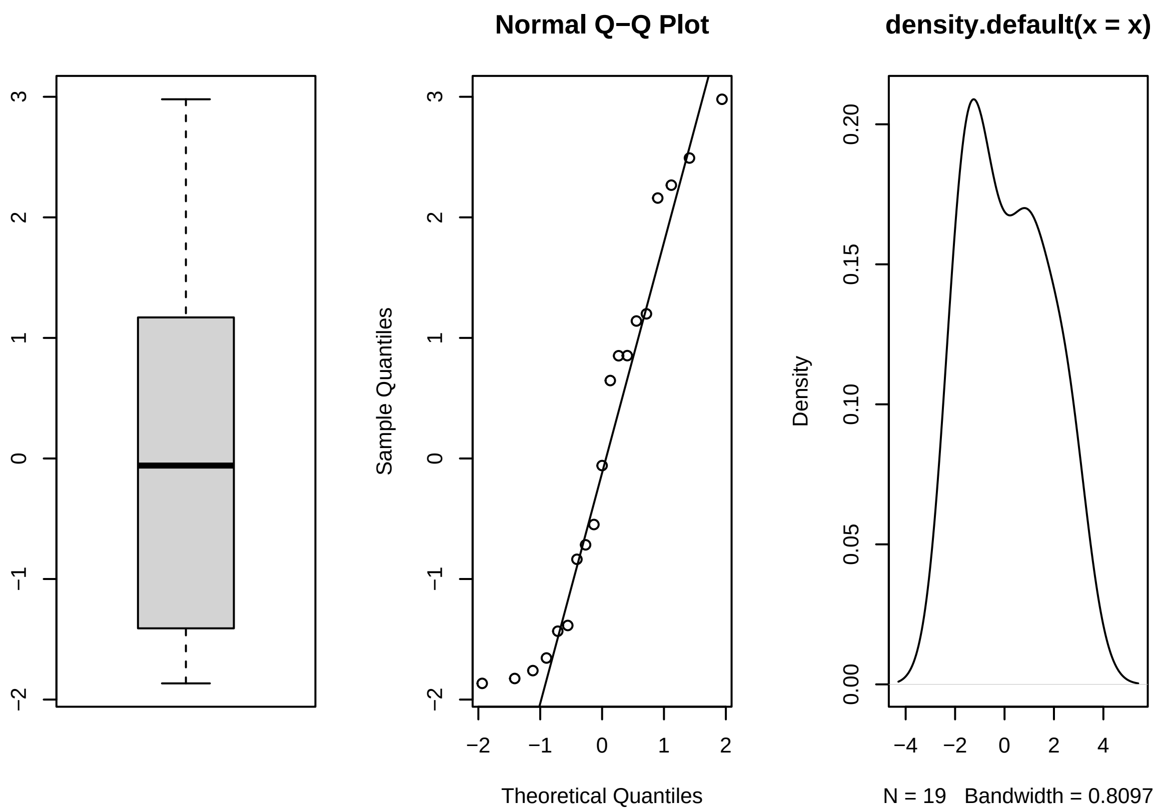
1.

a.

```
CPA = c(2.2041, 0.2744, 1.8050, 0.2822, 1.8062, 0.9600, 0.3175, 0.2953, 0.3704, 0.3828,
7.8867, 5.6250, 4.4694, 4.8133, 0.6840, 0.6086, 1.5651, 0.5600, 2.2969)
x = round(log2(CPA),4)
x
```

```
##  [1]  1.1402 -1.8656  0.8520 -1.8252  0.8530 -0.0589 -1.6552 -1.7597 -1.4328
## [10] -1.3853  2.9794  2.4919  2.1601  2.2670 -0.5479 -0.7164  0.6463 -0.8365
## [19]  1.1997
```

```
op1 = par(mfrow = c(1,3))
boxplot(x)
qqnorm(x)
qqline(x)
plot(density(x))
```



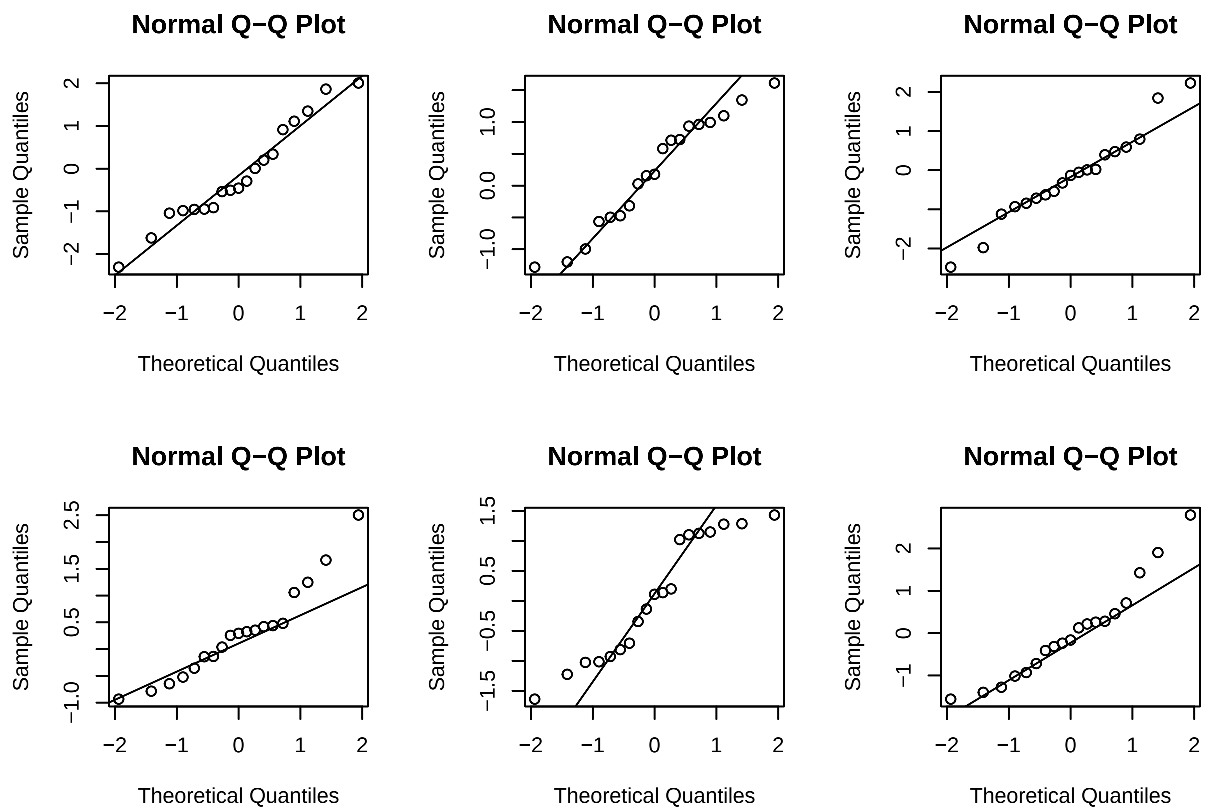

```
par(op1)
```

The sample doesn't seem to be drawn from the normal distribution, but the deviations are not major.

b.

```
# i.
y1 <- rnorm(19)
y2 <- rnorm(19)
y3 <- rnorm(19)
y4 <- rnorm(19)
y5 <- rnorm(19)
y6 <- rnorm(19)

# ii.
op2 = par(mfrow = c(2,3))
qqnorm(y1); qqline(y1)
qqnorm(y2); qqline(y2)
qqnorm(y3); qqline(y3)
qqnorm(y4); qqline(y4)
qqnorm(y5); qqline(y5)
qqnorm(y6); qqline(y6)
```




```
par(op2)

# iii

c(IQR(y1)/sqrt(mean(y1^2)-mean(y1)^2),
  IQR(y2)/sqrt(mean(y2^2)-mean(y2)^2),
  IQR(y3)/sqrt(mean(y3^2)-mean(y3)^2),
  IQR(y4)/sqrt(mean(y4^2)-mean(y4)^2),
  IQR(y5)/sqrt(mean(y5^2)-mean(y5)^2),
  IQR(y6)/sqrt(mean(y6^2)-mean(y6)^2))
```

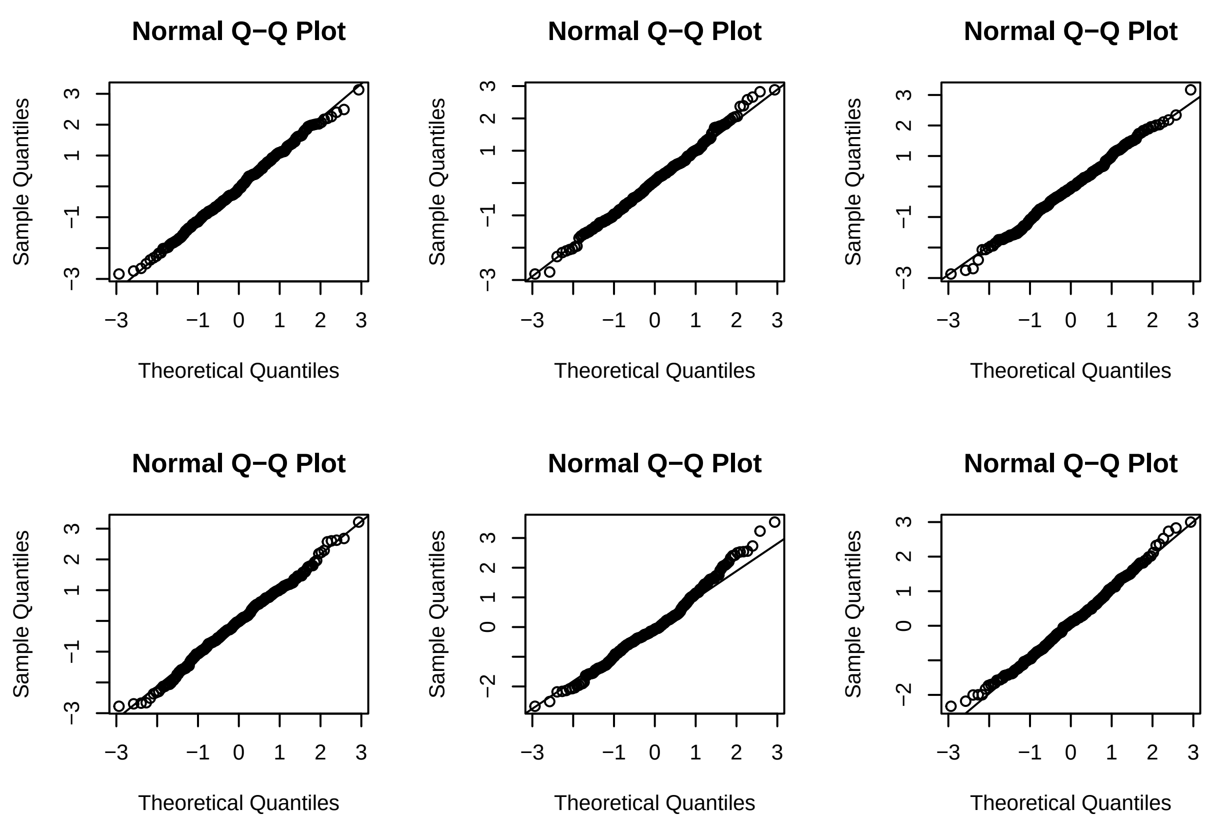
```
## [1] 1.3856120 1.6736461 1.0952629 0.8445421 2.0045032 1.0825824
```

iv. The six simulated samples don't seem to be drawn from a normal distribution, even though they are. This result is not uncommon with small samples. Thus, we may still use the methods learned when the sample doesn't seem to be drawn exactly from a normal distribution, as long as the departures are not extreme and there are no obvious outliers present. That said, always be more cautious with your conclusions when the sample is small.

Aside: Let's assume we use larger samples, say n=300 (as shown below). Then the samples are much closer to the expected behavior and the ratios of the sample interquartile range to the sample standard deviation are also close to what we would expect.

```
Y1 = rnorm(300)
Y2 = rnorm(300)
Y3 = rnorm(300)
Y4 = rnorm(300)
Y5 = rnorm(300)
Y6 = rnorm(300)


op3 = par(mfrow =c(2,3))
qqnorm(Y1); qqline(Y1)
qqnorm(Y2); qqline(Y2)
qqnorm(Y3); qqline(Y3)
qqnorm(Y4); qqline(Y4)
qqnorm(Y5); qqline(Y5)
qqnorm(Y6); qqline(Y6)
```




```
par(op3)


c(IQR(Y1)/sqrt(mean(Y1^2)-mean(Y1)^2),
  IQR(Y2)/sqrt(mean(Y2^2)-mean(Y2)^2),
  IQR(Y3)/sqrt(mean(Y3^2)-mean(Y3)^2),
  IQR(Y4)/sqrt(mean(Y4^2)-mean(Y4)^2),
  IQR(Y5)/sqrt(mean(Y5^2)-mean(Y5)^2),
  IQR(Y6)/sqrt(mean(Y6^2)-mean(Y6)^2))
```

```
## [1] 1.381741 1.290395 1.251632 1.336670 1.168801 1.364156
```

c.

- $H_0 : \mu = 0$
- $H_1 : \mu \neq 0$

```
n <- length(x)
mu0 <- 0
xbar <- mean(x)
se <- sqrt(var(x)/n)
(t <- xbar - mu0 /se) # test statistic
```

```
## [1] 0.1319
```

```
2*(1-pt(abs(t), n-1)) # p value
```

```
## [1] 0.8965266
```

We fail to reject the null hypothesis. We haven't found evidence that μ is different than 0.

d.

If the true population mean for our problem is $\mu = 0.20$, then we have made a Type II error by failing to reject that $\mu = 0$, when in fact it is.

2.

This is a right-tailed test.

2a.

Here is the code in R:

```
xbar = 15.2
n = 36
mu0 = 14
sigma = 6
z = (xbar - mu0)/(sigma/sqrt(n))
z
```

```
## [1] 1.2
```

```
1 - pnorm(z)
```

```
## [1] 0.1150697
```

Using $\alpha = 0.1$, we fail to reject the null hypothesis, so μ may be less than or equal 14 as there is no evidence to conclude the opposite.

2b.

The quantile under the normal would be

```
q = qnorm(0.9)
q
```

```
## [1] 1.281552
```

So, the corresponding mean would be about 1.28 standard deviation above the hypothesized mean:

```
xbar_alpha = 14  + q*sigma/sqrt(n)
xbar_alpha
```

```
## [1] 15.28155
```

the sample mean that corresponds to the critical value is about 15.58.

2c.

In part a we failed to reject the null ($\mu \leq 14$), which now we learned it was false (as $\mu = 15 > 14$), so we have committed a type II error.

2d.

So, the probability of committing a type II error would be failing to reject the null given that the null is false. Since this is a right-tailed test, we fail to reject the null whenever we get a sample that is smaller than `xbar_alpha`=15.2815516. The probability of that happening on our true distribution of $\bar{X}_n$ is

```
true_mu = 15
beta = pnorm(xbar_alpha, true_mu, sigma/sqrt(n))
beta
```

```
## [1] 0.6108563
```

The probability of committing a type II error is $\beta = 0.61$.

2e.

The power of the test is $1 - \beta = 0.39$.

3.

3a.

```
xbar = mean(r_sample)
s = sd(r_sample)
n = length(r_sample)
alpha = 1- 0.92
q = qt(1 - alpha/2, n-1)
c(xbar - q*(s/sqrt(n)), xbar + q*(s/sqrt(n)))
```

```
## [1] -2.724155  7.884155
```

We are 92% confident that the arrival delay is between -2.72 and 7.88 (observe that negative values represent a flight arriving before the scheduled time).

3b.

```
df3 <- data.frame(r_sample)
boot_dist <- df3 %>%
  specify(response = r_sample) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

percentile_ci <- get_ci(boot_dist, level = .92)
percentile_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    -2.46     7.91
```

The result is fairly close to the theory-based approach.

3c.

```
n = length(r_sample)
phat = mean(r_sample <= 0)
se = sqrt(phat*(1-phat)/n)
alpha = 1 - 0.96
q = qnorm(1 - alpha/2)
c(phat - q*se, phat + q*se)
```

```
## [1] 0.4889898 0.6910102
```

We are 96% confident that somewhere between 48.8% and 69% of the flights will arrive without delays.

3d.

```r
on_time <- as.factor(r_sample<=0)
df3d <- data.frame(on_time)
boot_dist <- df3d %>%
  specify(response = on_time, success = "TRUE") %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "prop")

percentile_ci <- get_ci(boot_dist, level = .96)
percentile_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     0.49     0.69
```

The results are again, fairly close to the ones obtained using the theory-based approach.

4.

Observe that we can obtain the estimated variance, based on the results obtained earlier, namely

$$\hat{\sigma}^2 = 0.05 \times (1 - 0.05)$$

and the estimated standard deviation would be the square root of this value. We can use this value in our formula, using also the other values given:

```r
L = 0.002
sigmahat = sqrt(0.05 * (1 - 0.05))
alpha = 1 - 0.99
q = qnorm(1 - alpha/2)
n = (2 * q * sigmahat / L) ^ 2
n
```

```
## [1] 315157.6
```

We need around 315k samples to be 99% confident that the interval estimate would be at most 0.002 in length.

5. ISI 11.4 Problem set B.

3.

- Experimental unit: A student
- 1-population, 1-sample problem. 20 units drawn.
- 2 measurement per e.u. Number of watts expended for each protocol (S and D).
- $X_i = S_i - D_i$ for $i = 1, \ldots, 20$ so $X_1, X_2, \ldots X_{20} \sim \mathbb{P}$, $E\bar{X} = \mu$, and μ is the parameter of interest.
- $H_0 : \mu \geq 0$ versus $H_1 : \mu < 0$

6.

- Experimental unit: A runner
- 1 population:
  1. 120 units drawn from the population
  2. This is a 1-sample problem.
- 2 measurements taken from each experimental unit; Race time of each runner is measured when a runner does not wear the race flats on the first race and then does wear them on the second race.
- We define $X_i = F_i - S_i$, the difference in time for the first race (F) minus the the second (S) of the ith runner with $i : 1, \ldots, 120$; The parameter of interest is $\mu = \mathrm{E}\bar{X}_{120}$, positive differences would imply the second race took less time. Since the company's claim is that races will improve by more than 30 seconds, if we want to show that the company's claim are overstating the improvement using the racing flats, then the hypotheses are:
  1. $H_0 : \mu \leq 30$
  2. $H_1 : \mu < 30$

7.

- Experimental unit: A wood block.
- 2-populations (wood blocks with IGR vs wood block with solvent only), 2-sample problem. 120 units drawn in total, 60 from each population.
- 2 measurement per e.u. The weight before(A) and after the experiment(B).
- Let X_i be the ith woodblock with IGR and Y_j the jth wood block with only solvent. One possible representation of the relevant result would be $X_i = (B_i - A_i)/B_i$ (Population 1) where we divide the difference by the weight before termites to measure the relative difference in weights (using just the absolute difference would have been acceptable as well). Similarly, we let $Y_j = (B_j - A_j)/B_j$ (Population 2). Then, we have $X_1, X_2, \ldots X_{60} \sim \mathbb{P}_1$ and $Y_1, Y_2, \ldots, Y_{60} \sim \mathbb{P}_2$. Then, $\Delta = \mu_1 - \mu_2$ and Δ is the parameter of interest.
- $H_0 : \Delta \geq 0$ versus $H_1 : \Delta < 0$

8.

- Experimental unit: A couple
- 1 population: Couples who enrolled in an introductory swing dance class
  - 20 units drawn from the population
  - This is a 1-sample problem.
- 4 measurements taken from each experimental unit; Each participant's resting pulse is measured at the beginning and at the end of the ten-week class.
- One alternative is to find the average of the difference of scores for each couple. For example,

$$X_i = \frac{(B_{1i} - E_{1i}) + (B_{2i} - E_{2i})}{2}$$

where $B_{1i} - E_{1i}$ is the the difference of resting pulse at the beginning (B) minus resting pulse at the end (E) for the first member and $B_{2i} - E_{2i}$ the difference for the second member of the ith couple, with $i = 1, \ldots, 20$. Then μ is the population mean of the average of these differences. If the swing dancing classes work, we would expect the difference to be positive (lower resting pulse at the end), so the average of the differences should be also positive. The hypotheses are:

1. $H_0 : \mu \leq 0$
2. $H_1 : \mu > 0$