# STAT511: Hw3

*Amy Fox*

## Question 1

### Suppose the mean oxygen level of a certain lake was of interest. A total of n=10 samples were taken (from randomly selected locations) and oxygen level was measured in ppm. The sample mean oxygen level was found to be ($\bar{y}$) 5.3 and the sample standard deviation was found to be (s) 0.5. Use $\alpha = 0.05$.

### A. Calculate the SE (standard error) and the 95% ME (margin of error).

Standard error of the mean = standard deviation/sqrt(n) Margin of error = t(alpha/2, n-1)*SE = t(0.025, 9)

```
# standard error
0.5/sqrt(10)
```

```
## [1] 0.1581139
```

```
# margin of error
qt(1-0.025, df = 9) * 0.5/sqrt(10)
```

```
## [1] 0.3576785
```

### B. Use your ME from above to construct a 95% confidence interval for $\mu$ (population mean).

(sample mean - ME, sample mean + ME)

(5.3 - 0.358, 5.3 + 0.358)

(4.94, 5.66)

### C. Use your CI from above to test H0: $\mu$=5 vs HA: $\mu \neq 5$. Make a conclusion about the test. Justify your conclusion based on the CI.

Because the null hypothesis ($\mu = 5$) is within the confidence interval (4.94, 5.66), we fail to reject H0. This means that the true population mean could be 5ppm.

### D. Now test H0: $\mu$=5 vs HA: $\mu \neq 5$ using a formal hypothesis test. Be sure to define the rejection region, calculate the test statistic and state your conclusion. (4 pts)

A two-sided t test should be performed.

Test statistic: $\bar{y}$ - $\mu$ / (s/sqrt(n))

Rejection region: t($\alpha$/2, n-1) = 2.26

```
# test statistic  aka t
(5.3 - 5)/(0.5/sqrt(10))
```

```
## [1] 1.897367
```

```
# rejection region aka t alpha
qt(1-0.025, df = 9)
```

```
## [1] 2.262157
```

Reject H0 if |t| > t$\alpha$

1.90 > 2.26

Therefore, fail to reject H0; the true population mean for the oxygen could be 5ppm.

### E. Now test H0: $\mu \leq 5$ vs HA: $\mu$>5 using a formal hypothesis test. Be sure to define the rejection region, calculate the test statistic and state your conclusion. (4 pts)

One sided test must be used now

```r
# rejection region - different from two-sided because don't divide alpha by 2
qt(1-0.05, df = 9)
```

```
## [1] 1.833113
```

```r
# test statistic
(5.3 - 5)/(0.5/sqrt(10))
```

```
## [1] 1.897367
```

t$\alpha$ = 1.83

t = 1.90

Reject H0 if t > t$\alpha$

1.90 > 1.83

Therefore, reject H0. The true population mean for the oxygen is greater than 5ppm.

### F. Now suppose that the summary statistics were based on a sample of size n=51. Rerun the hypothesis test from part D (H0: $\mu$=5 vs HA: $\mu \neq 5$) based on this larger sample size. Be sure to define the rejection region, calculate the test statistic, and state your conclusion. (4 pts)

Reject H0 if |t| > t$\alpha$

```r
# rejection region aka t alpha
qt(1-0.025, df = 50)
```

```
## [1] 2.008559
```

```r
# test statistic  aka t
(5.3 - 5)/(0.5/sqrt(51))
```

```
## [1] 4.284857
```

|4.28| > 2.00

Therefore, reject H0. The true population mean for the oxygen is different from 5ppm.

### G. Comparing the tests in part D (n = 10) vs part F (n = 51), which has higher power? No need to justify.

n = 51 has a higher power

## Question 2

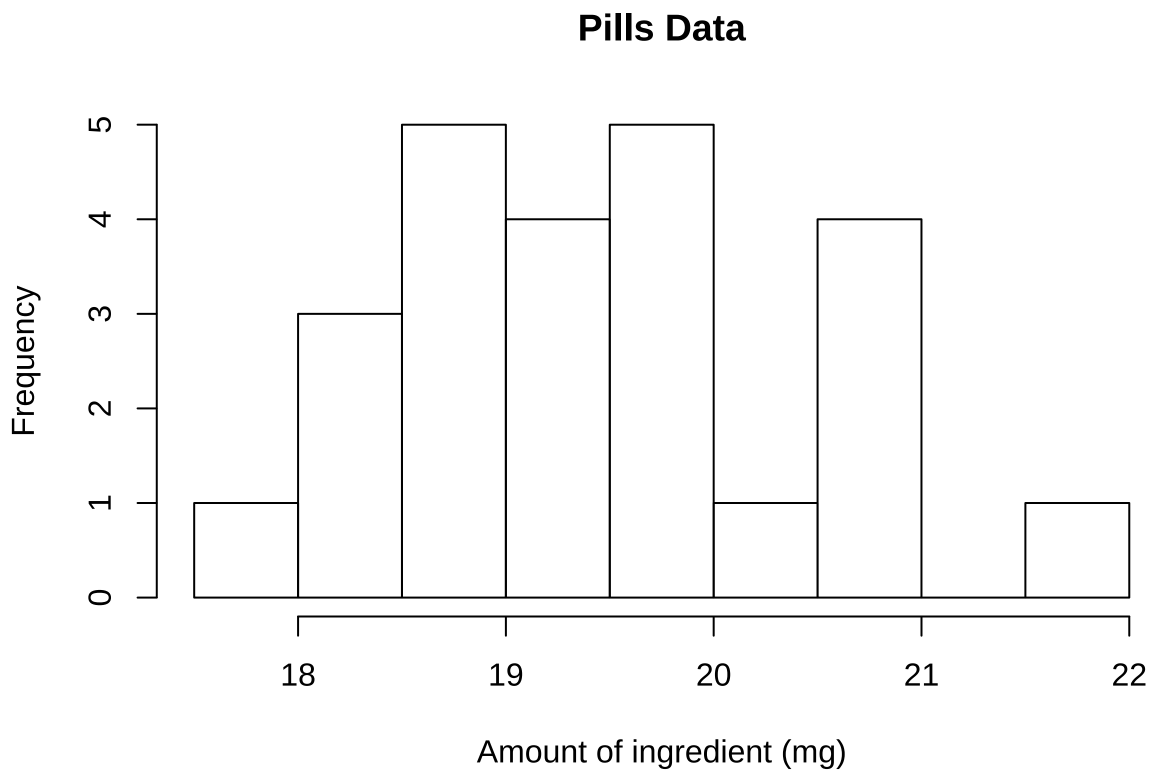
Manufacturers must test the amount of the active ingredient in medications before releasing the batch of pills. The data Pills.csv (available from Canvas) represents the amount (in mg) of the active ingredient in n = 24 pills (from a random sample of the same large batch). Use $\alpha = 0.05$.

A. Create a histogram and qqplot and test for normality of the data (using Shapiro-Wilk test). Based on this evidence, does the data appear to be normally distributed? Justify your response. Be sure to include the plots in your assignment. (4 pts)

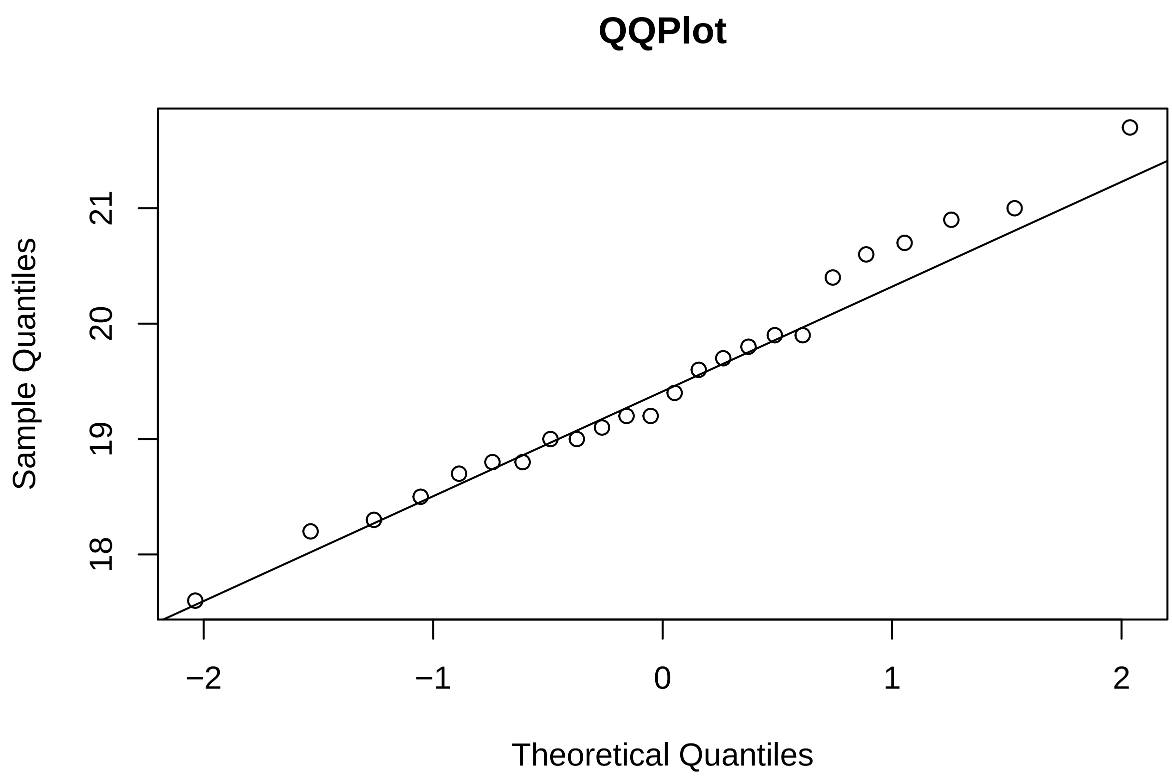
```r
# read in data
library(readr)
pills_data <- read_csv("../Data/Pills.csv")
```

```
## Parsed with column specification:
## cols(
##   y = col_double()
## )
```

```r
# plot histogram
hist(pills_data$y, main = "Pills Data", xlab = "Amount of ingredient (mg)")
```




```r
qqnorm(pills_data$y, main = "QQPlot")
qqline(pills_data$y)
```




```
# normality test
shapiro.test(pills_data$y)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  pills_data$y
## W = 0.97988, p-value = 0.8936
```

Because the p-value for the Sharpiro-Wilk Normality test $> 0.05$, we can assume that it is normally distributed. Further, because the data on the qqplot looks linear (and the histogram looks normal), we can also assume it is normally distributed.

### B. Give an estimate of the mean and corresponding 95% confidence interval.

```
# mean and 95% CI
t.test(pills_data$y)
```

```
##
##  One Sample t-test
##
## data:  pills_data$y
## t = 95.158, df = 23, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.07609 19.92391
## sample estimates:
## mean of x
##      19.5
```

The mean is 19.5mg and the confidence interval is (19.08, 19.92).

### C. For this question, suppose that if there is evidence that the mean is different from 20mg, the batch of pills will be destroyed. Is there evidence that the batch of pills has a mean amount different from 20mg? State your hypotheses, test statistic, p-value and make a conclusion. (4 pts)

**Hypothesis** H0: $\mu = 20$ HA: $\mu$ not equal 20

```
library(dplyr)
# this gives us n
n<- nrow(pills_data)

# test statistic  aka t
t <- (mean(pills_data$y) - 20)/(sd(pills_data$y)/sqrt(n)) %>%
  print()
```

```
## [1] 0.2049213
```

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

```
## [1] 0.02281102
```

```
# alternative way to calculate test statistic and p-value
t.test(pills_data$y, mu = 20)
```

```
##
##  One Sample t-test
##
## data:  pills_data$y
## t = -2.44, df = 23, p-value = 0.02281
## alternative hypothesis: true mean is not equal to 20
## 95 percent confidence interval:
##  19.07609 19.92391
## sample estimates:
## mean of x
##      19.5
```

**Conclusion** Since p-value $= 0.02 < \alpha = 0.05$, we reject H0 $= 20$. Therefore, the true populaiton mean amount is different than 20mg, so the pills should be destroyed.

### D. For this question, suppose that if there is evidence that the mean is less than 20mg, the batch of pills will be destroyed. Is there evidence that the batch of pills has a mean amount less than 20mg? State your hypotheses, test statistic, p-value and make a conclusion. (4 pts)

**Hypothesis** H0: $\mu >= 20$ HA: $\mu < 20$

```
# test statistic  aka t
t <- (mean(pills_data$y) - 20)/(sd(pills_data$y)/sqrt(n)) %>%
  print()
```

```
## [1] 0.2049213
```

```
# p-value - took out the 2X because it's one-sided
1-pt(abs(t), n-1)
```

```
## [1] 0.01140551
```

```r
# another way to look at the p-value & test-statistic
t.test(pills_data$y, mu = 20, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  pills_data$y
## t = -2.44, df = 23, p-value = 0.01141
## alternative hypothesis: true mean is less than 20
## 95 percent confidence interval:
##      -Inf 19.85121
## sample estimates:
## mean of x
##      19.5
```

**Conclusion** Since p-value $= 0.01 < \alpha = 0.05$, we reject H0 $>= 20$. Therefore, the true populaiton mean amount of drug is less than 20mg so the pills should be destroyed.