STAT 1293 - Quiz 3

Gordon Lu

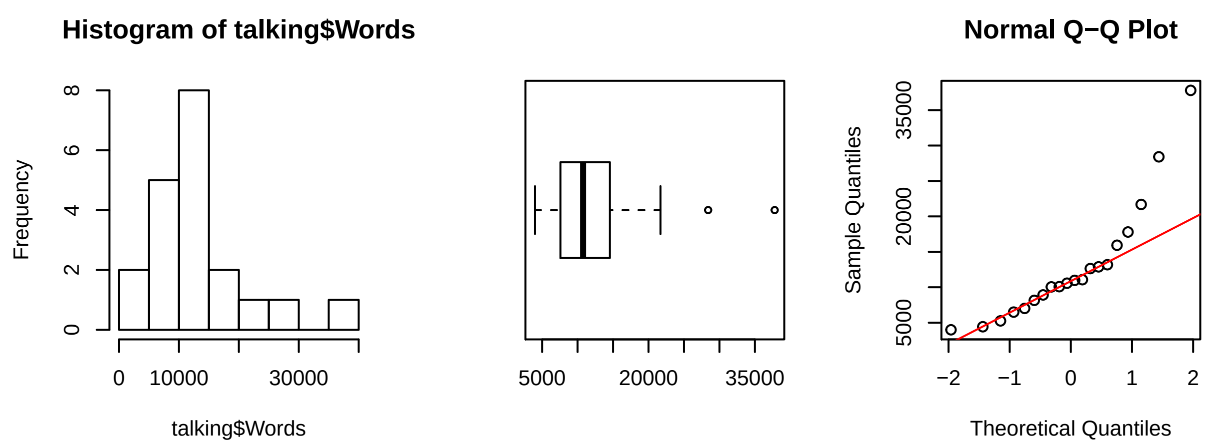
7/17/2020

Problem 1: Men of few words? (15 points)

1a) Create a histogram, boxplot, and a Q-Q plot of the data. Put them together using par() function. (6 points)

Solution:

```
talking <- read.table("C:/Users/gordo/Desktop/talking.txt", header = TRUE) #read in talking.txt
par(mfrow = c(1,3), pty = "s")
hist(talking$Words)
boxplot(talking$Words, horizontal = TRUE)
qqnorm(talking$Words)
qqline(talking$Words, col = 2)
```



1b) Examine the data. Are there any outliers? Is the distribution symmetric or skewed? (4 points)

Solution:

```r
summary(talking$Words)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3998    7868   10774   12867   13869   37786
```

By making use of the 5-number summary, and the generated plots above, it is apparent that there are 2 upper outliers from the box plot, and the Q-Q plot. By using the `1.5 x IQR` rule, any outliers above `13869 + (1.5 x (13869 - 7868))` are considered upper outliers. In particular, any value above `22870.5` are considered otuliers. Therefore, the two upper outliers are `28408` and `37786`. However, there do not seem to be any lower outliers. Additionally, by looking at the histogram, the boxplot and the Q-Q plot, it is apparent that the distribution is skewed. This can also be noticed since the `Mean > Median`, which typically indicates a positively/right-skewed distribution.

1c) Do the data give convincing evidence that the mean number of words per day of men at this university differs from 7,000? Conduct a one-sample t test. Show your R output and the 4 steps of hypothesis test. Use significance level $\alpha = 0.01$. (5 points)

Solution:

```r
t.test(talking$Words, mu = 7000, conf.level = 0.99)
```

```
##
##  One Sample t-test
##
## data:  talking$Words
## t = 3.145, df = 19, p-value = 0.005332
## alternative hypothesis: true mean is not equal to 7000
## 99 percent confidence interval:
##   7529.817 18203.583
## sample estimates:
## mean of x
##   12866.7
```

4-Step H.T.

Hypothesis: $H_0 : \mu = 7000$ vs. $H_a : \mu \neq 7000$

Test statistic: $t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} =$ `3.145`

P-value: $P(T < t_0) = P(T < t_0) =$ `0.005332`

Conclusion: Reject H_0, since p-value < 0.05.

Problem 2: Comparing two drugs. (15 points)

2a) Create a variable, diff, which is the difference between the two drugs (Reference-Generic). Show the vector. (3 points)

Solution:

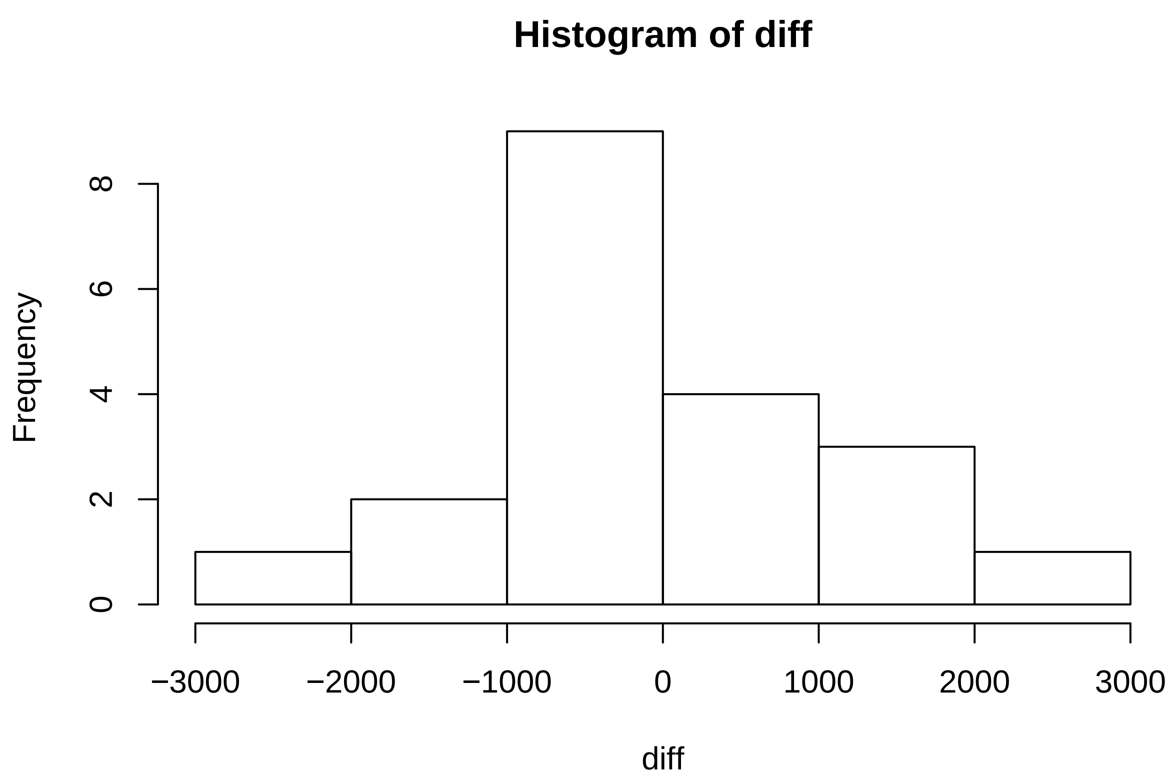
```
drugs <- read.table("C:/Users/gordo/Desktop/drugs.txt", header = TRUE) #read in drugs.txt
diff <- drugs$Ref - drugs$Generic
diff
```

```
##  [1]  2353  1388  1166  1598   523   343   208  -169   147   -19  -394  -438
## [13]  -262  -796  -349  -591  -702 -1111 -1605 -2030
```

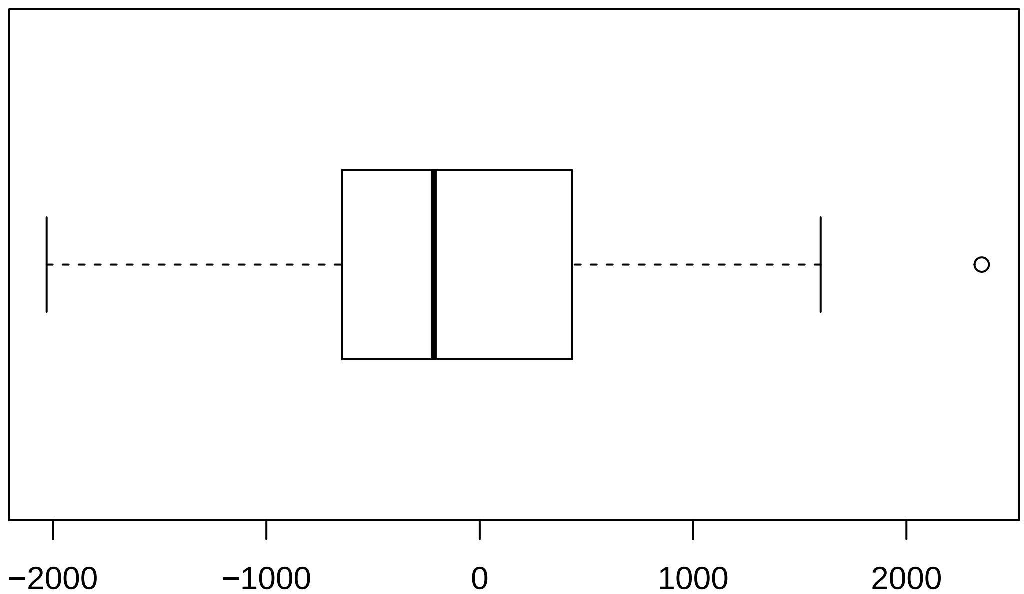
2b) Create a histogram and a boxplot of diff. (5 points)

Solution:

```
hist(diff)
```




```
boxplot(diff, horizontal = TRUE)
```



Yes, the distribution is positively skewed, this is apparent through the shape of the histogram, and the upper outlier that is 2353.

Yes, there is a single outlier, 2353, this is apparent by looking at the boxplot.

Yes, it seems reasonable to use the `t` procedure, since based on the histogram, the data looks fairly normal. Additionally, the boxplot implies the shape, disregarding the upper outlier, that the distribution is rather symmetric.

3c) Conduct a one-sample t test. Use significance level $\alpha = 0.05$. (7 points)

Solution:

```
t.test(diff, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  diff
## t = -0.15455, df = 19, p-value = 0.8788
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -538.0663  464.0663
## sample estimates:
## mean of x
##       -37
```

4-Step H.T.

Let d = (Reference - Generic)

Hypothesis: $H_0 : \mu_d = 0$ vs. $H_a : \mu_d \neq 0$

Test statistic: $t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} =$ -0.15455

P-value: $P(T < t_0) = P(T < t_0) =$ 0.8788

Conclusion: Fail to reject H_0, since p-value > 0.05.

Problem 3: Student drinking (20 points)

3a) Compare the summary statistics of the female sample and male sample (use `summary()`). Which group has a higher mean? (4 points)

Solution:

```
drinks <- read.table("C:/Users/gordo/Desktop/drinks.txt", header = TRUE) #read in drinks.txt
drinks.male <- drinks$Drinks[drinks$Sex == "M"]
drinks.female <- drinks$Drinks[drinks$Sex == "F"]

summary(drinks.male)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   6.000   6.519   8.000  16.000
```

```
summary(drinks.female)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   4.274   5.500  10.000
```

It appears that the `male` group has a higher mean.

3b) Create histograms for female students and male students. Make sure that they are on the same scale (same `xlim`). (4 points)

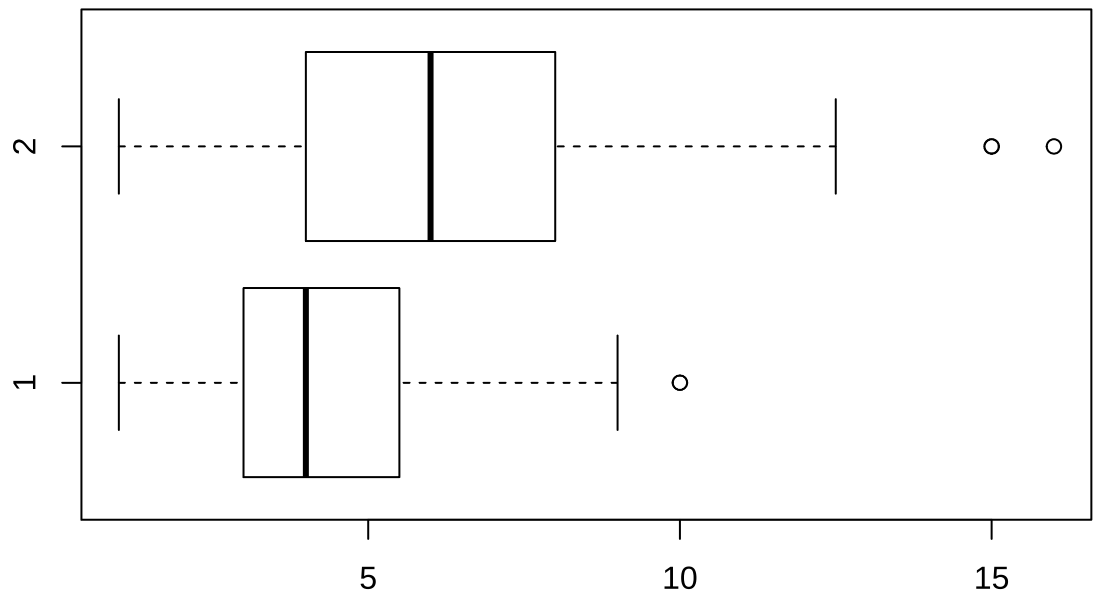
Solution:

```
par(mfrow = c(1,2), pty = "s")
hist(drinks.male, xlim = c(0, 20), ylim = c(0, 30))
hist(drinks.female, xlim = c(0, 20), ylim = c(0, 30))
```

3c) Create a side-by-side boxplot for female and male students. (4 points)

Solution:

```
boxplot(drinks.female, drinks.male, horizontal = TRUE)
```



3d) Conduct a general two-sample t test (without equal-variance assumption). Do male students drink more than female students, on average? (8 points)

Solution:

```r
t.test(drinks.male, drinks.female, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  drinks.male and drinks.female
## t = 5.1934, df = 132.15, p-value = 3.807e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.528832      Inf
## sample estimates:
## mean of x mean of y
##  6.518519  4.273684
```

4-Step H.T.

Hypothesis: $H_0 : (\mu_{Male} - \mu_{Female}) = 0$ vs. $H_a : \mu_{Male} > \mu_{Female}$

Test statistic: $t_0 = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} = 5.1934$

P-value: $P(T < t_0) = P(T < t_0) =$ `3.807e-07`

Conclusion: Reject H_0, since p-value < 0.05.