Lab Homework Module:5
ST8114
mjh100

Mark Hill

10/30/2020

Contents

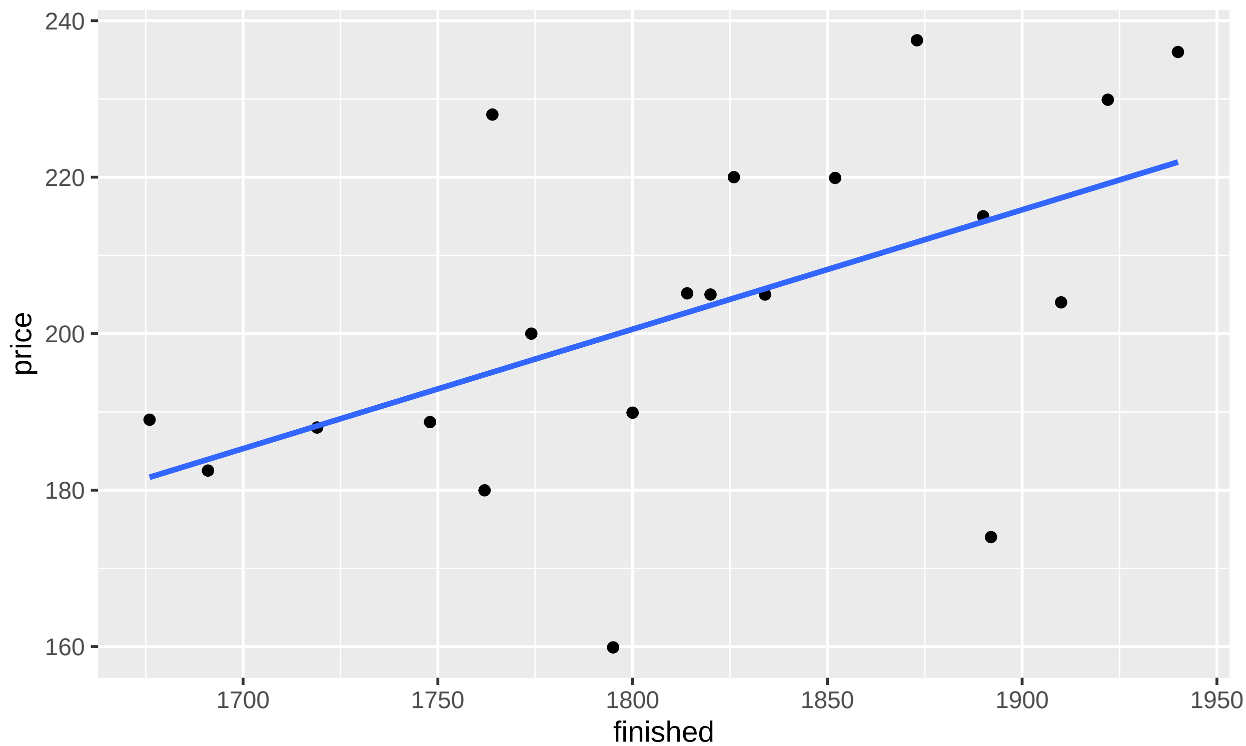
Practice 1

```
estate <- read.csv("~/ST 8114/lab/module 5/data/estate.txt", sep = " ", header = TRUE)
```

a

```
ggplot(estate, aes(x= finished, y= price))+
  geom_point()+
  geom_smooth(method = "lm", se=F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



b

```r
model <- lm(price ~ finished, estate)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ finished, data = estate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.608  -5.614   1.039  10.884  32.928
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -74.15350  102.73397  -0.722   0.4797
## finished      0.15262    0.05655   2.699   0.0147 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.67 on 18 degrees of freedom
## Multiple R-squared:  0.2881, Adjusted R-squared:  0.2485
## F-statistic: 7.283 on 1 and 18 DF,  p-value: 0.01469
```

```r
cat("yhat=", model$coefficients[1], "+", model$coefficients[2],"x")
```

```
## yhat= -74.1535 + 0.1526223 x
```

We would expect the cost of houses to increase by 0.1526 * thousand dollars for each additional square foot in size. ie. $152 more for each additional square foot.

Practice 2

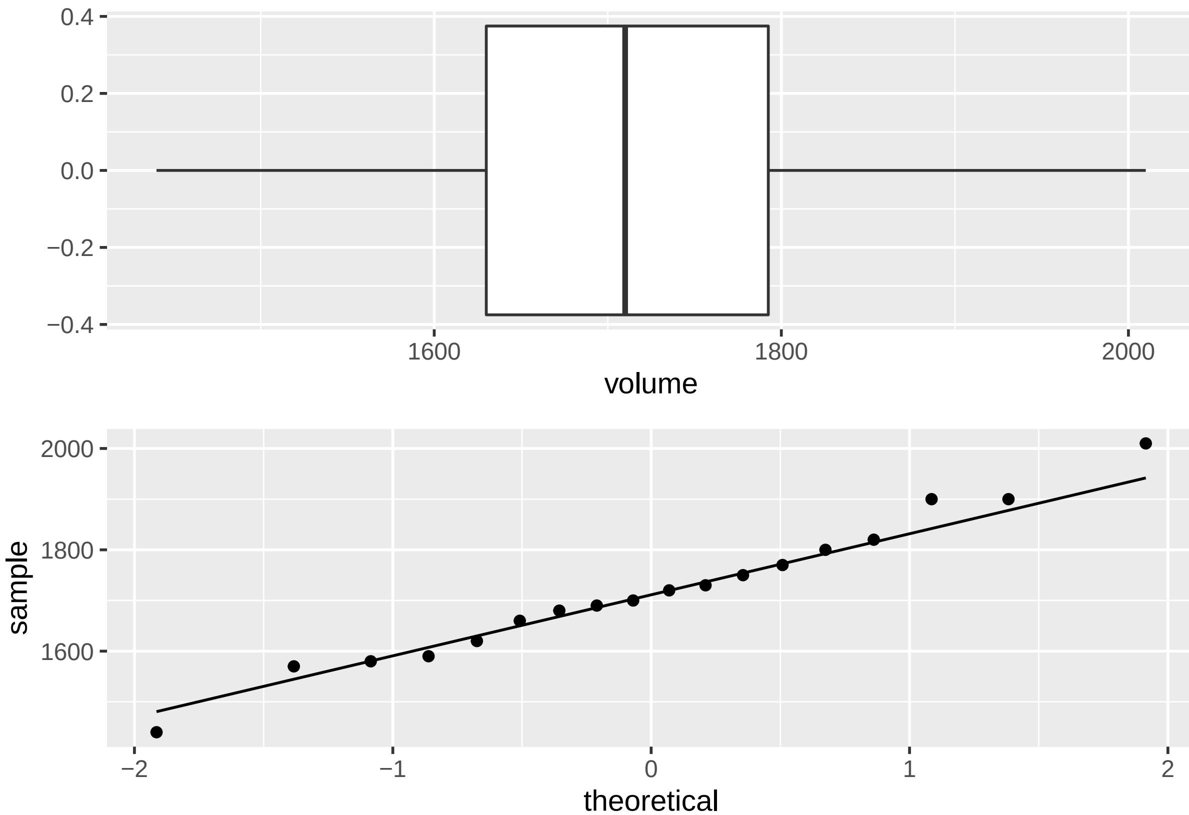
```
garbage <- read.csv("~/ST 8114/lab/module 5/data/garbage.txt", header = TRUE)
```

a

```
a <- ggplot(garbage, aes(x= volume))+
      geom_boxplot()

b <- ggplot(garbage, aes(sample= volume))+
  stat_qq()+
  stat_qq_line()

a / b
```



These appear to be normally distributed values

b

```
t1 <- shapiro.test(garbage$volume)
t2 <- pearson.test(garbage$volume)
t3 <- sf.test(garbage$volume)
t4 <- ad.test(garbage$volume)
tab <- map_df(list(t1,t2,t3,t4),tidy)
tab
```

```
## # A tibble: 4 x 3
##   statistic p.value method
##       <dbl>   <dbl> <chr>
## 1     0.986   0.990 Shapiro-Wilk normality test
## 2     0.667   0.955 Pearson chi-square normality test
## 3     0.980   0.894 Shapiro-Francia normality test
## 4     0.169   0.921 Anderson-Darling normality test
```

No rejection of H_0 in any of the tests.

c

```
t5 <- t.test(garbage$volume, conf.level = .9)
cat("90% confidence interval for mean of garbage$volume",t5$conf.int)
```

```
## 90% confidence interval for mean of garbage$volume 1661.845 1774.821
```

d

```
t6 <- t.test(garbage$volume, conf.level = .90, mu= 1600, alternative = "greater")
t6
```

```
##
##  One Sample t-test
##
## data:  garbage$volume
## t = 3.6442, df = 17, p-value = 0.001003
## alternative hypothesis: true mean is greater than 1600
## 90 percent confidence interval:
##  1675.036      Inf
## sample estimates:
## mean of x
##  1718.333
```

There is evidence, at $\alpha = 0.1$, to reject $H_0 : \mu \leq 1600$. Thus the two week schedule is desirable since they would usually pick up more than 1600 cubic feet.

Practice 3

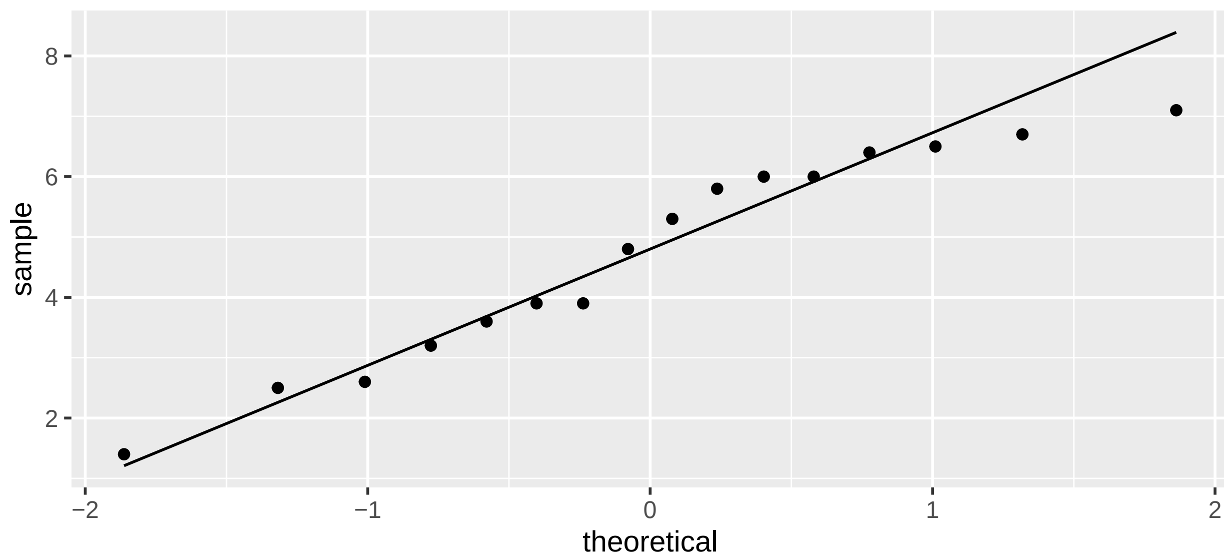
```
wgtgain <- read.csv("~/ST 8114/lab/module 5/data/wgtgain.txt", header = TRUE, sep = " ")
```

a

```
wgtgain$diff <- wgtgain$after - wgtgain$before
```

b

```
ggplot(wgtgain, aes(sample= diff))+
  stat_qq()+
  stat_qq_line()
```



c

```
t1 <- tidy(shapiro.test(wgtgain$diff))
t1
```

```
## # A tibble: 1 x 3
##   statistic p.value method
##       <dbl>   <dbl> <chr>
## 1     0.938   0.325 Shapiro-Wilk normality test
```

d

```
t2 <- t.test(x= wgtgain$after, y= wgtgain$before, paired = TRUE)
cat("95% confidence interval for weight gain after treatment is", "(",t2$conf.int,") kg.")
```

```
## 95% confidence interval for weight gain after treatment is ( 3.801008 5.661492 ) kg.
```

e

```
t3 <- t.test(wgtgain$diff, alternative = "two.sided", mu= 7)
```

f

```
t4 <- t.test(wgtgain$diff, alternative = "greater", mu= 7)
```

g

```
t5 <- t.test(wgtgain$diff, alternative = "less", mu= 7)

tab <- map_df(list(t3,t4,t5), tidy)
tab
```

```
## # A tibble: 3 x 8
##   estimate statistic  p.value parameter conf.low conf.high method    alternative
##      <dbl>     <dbl>    <dbl>     <dbl>    <dbl>     <dbl> <chr>     <chr>
## 1     4.73     -5.20  1.08e-4        15     3.80      5.66 One Samp~ two.sided
## 2     4.73     -5.20  1.00e+0        15     3.97    Inf     One Samp~ greater
## 3     4.73     -5.20  5.40e-5        15  -Inf         5.50 One Samp~ less
```

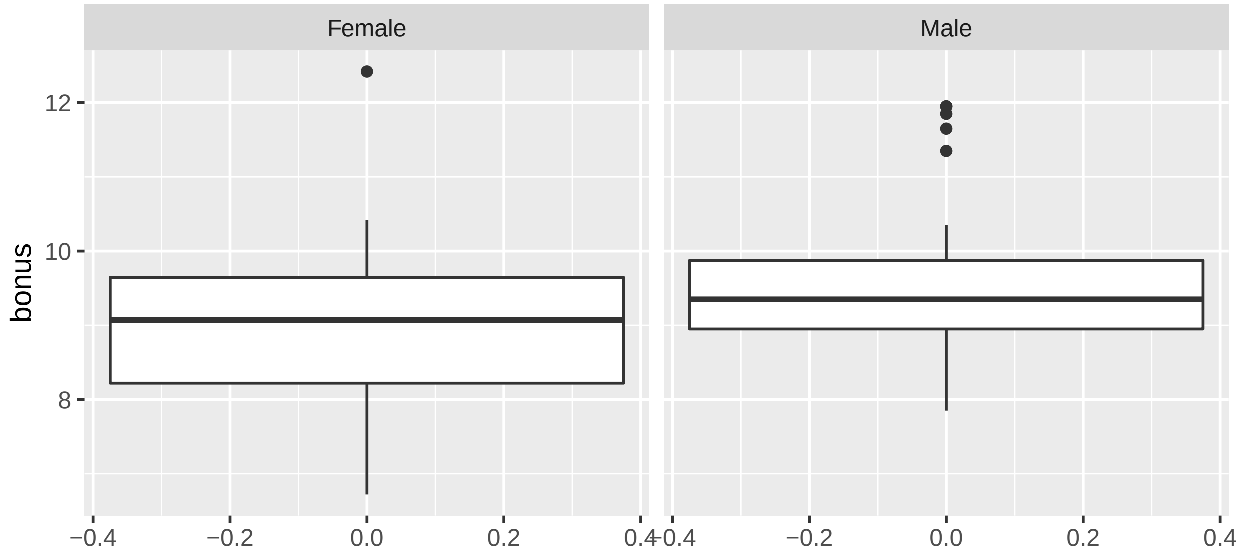
1. p-value $< \alpha$ =0.05, accept H_a that differences are not equal to 7 Kg.
2. p-value $\approx$ 1, fail to reject H_0 that weight gain after the treatment is greater than 7 Kg.
3. p-value $< \alpha$ =0.05, accept H_a that differences are less than 7 Kg.

Practice 4

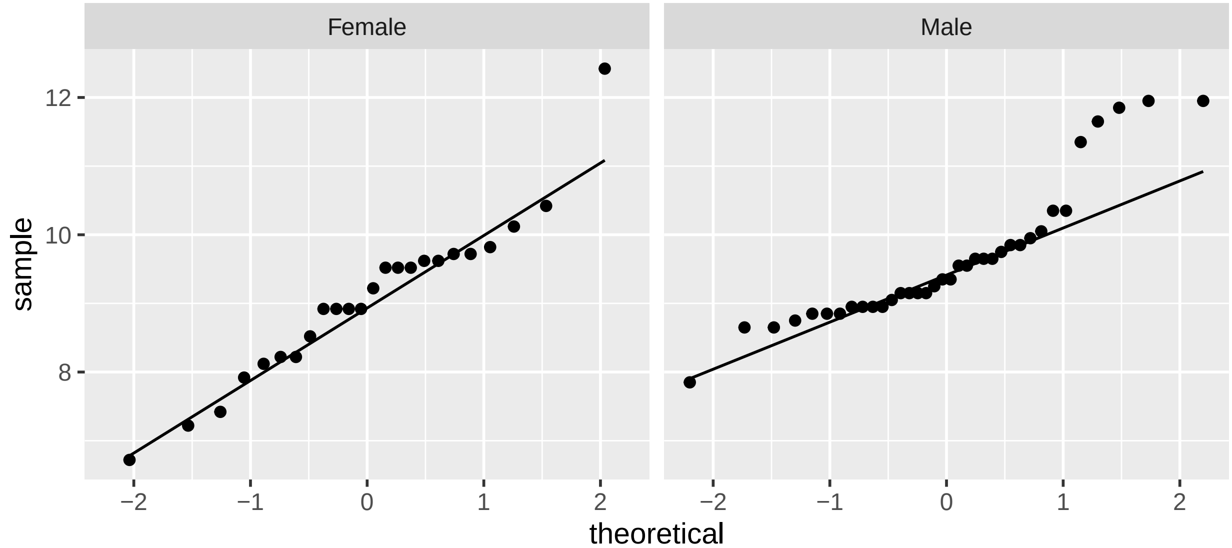
```
bonuses <- read.csv("~/ST 8114/lab/module 5/data/bonuses.txt", header = TRUE, sep = " ",
                    stringsAsFactors = TRUE)
```

a

```
ggplot(bonuses)+
  geom_boxplot(aes(y= bonus))+
  facet_wrap(~gender)
```




```
ggplot(bonuses,aes(sample= bonus))+
  stat_qq()+
  stat_qq_line()+
  facet_wrap(~gender)
```



b

```r
do.call("rbind",
        with(bonuses, tapply(bonus, gender,
              function(x)
                unlist(shapiro.test(x)
                       [c("statistic", "p.value")])
                     )
             )
        )
```

```
##        statistic.W      p.value
## Female   0.9468766 0.2316268301
## Male     0.8655588 0.0004434335
```

Based on the Q-Q plot, boxplot, and the small p-value, the male sample follows a non-normal distribution. It exhibits a pattern of right skewness.

c

```r
t.test(bonuses$bonus ~ bonuses$gender)
```

```
##
##  Welch Two Sample t-test
##
## data:  bonuses$bonus by bonuses$gender
## t = -1.9676, df = 43.587, p-value = 0.0555
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.17424165  0.01424165
## sample estimates:
## mean in group Female   mean in group Male
##             9.053333             9.633333
```

Here, at the assumed $\alpha = 0.05$ we would fail to reject $H_0 : \mu_1 = \mu_2$

But, our assumption of normality has been violated from one of the groups. Lets see if the findings will be different utilizing a nonparametric test.

```r
wilcox.test(bonuses$bonus ~ bonuses$gender, correct= FALSE)
```

```
## Warning in wilcox.test.default(x = c(9.72, 8.22, 12.42, 6.72, 9.52, 8.92, :
## cannot compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test
##
## data:  bonuses$bonus by bonuses$gender
## W = 317, p-value = 0.08248
## alternative hypothesis: true location shift is not equal to 0
```

Again, we don't reject the null in this `two.sided` test. This less "efficient" test leaves us even farther away despite the aforementioned normality violation.

d

```r
var.test(bonuses$bonus~ bonuses$gender)
```

```
##
##  F test to compare two variances
##
## data:  bonuses$bonus by bonuses$gender
## F = 1.4028, num df = 23, denom df = 35, p-value = 0.3584
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6760627 3.0910792
## sample estimates:
## ratio of variances
##          1.402802
```

e

```r
t.test(bonuses$bonus ~ bonuses$gender, var.equal= TRUE, alternative= "less")
```

```
##
##  Two Sample t-test
##
## data:  bonuses$bonus by bonuses$gender
## t = -2.0359, df = 58, p-value = 0.02317
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -0.1038034
## sample estimates:
## mean in group Female   mean in group Male
##             9.053333             9.633333
```

Assuming the groups have equal variance and specifying the alternative presents a different result in our `t.test`. The last test showed a difference in means hence the direction of the alternative. Even a two tailed test results in p-value $< \alpha = 0.05$. Conclusion: accept H_a : female bonuses are less than male bonuses.