HW1

Sam Mottahedi

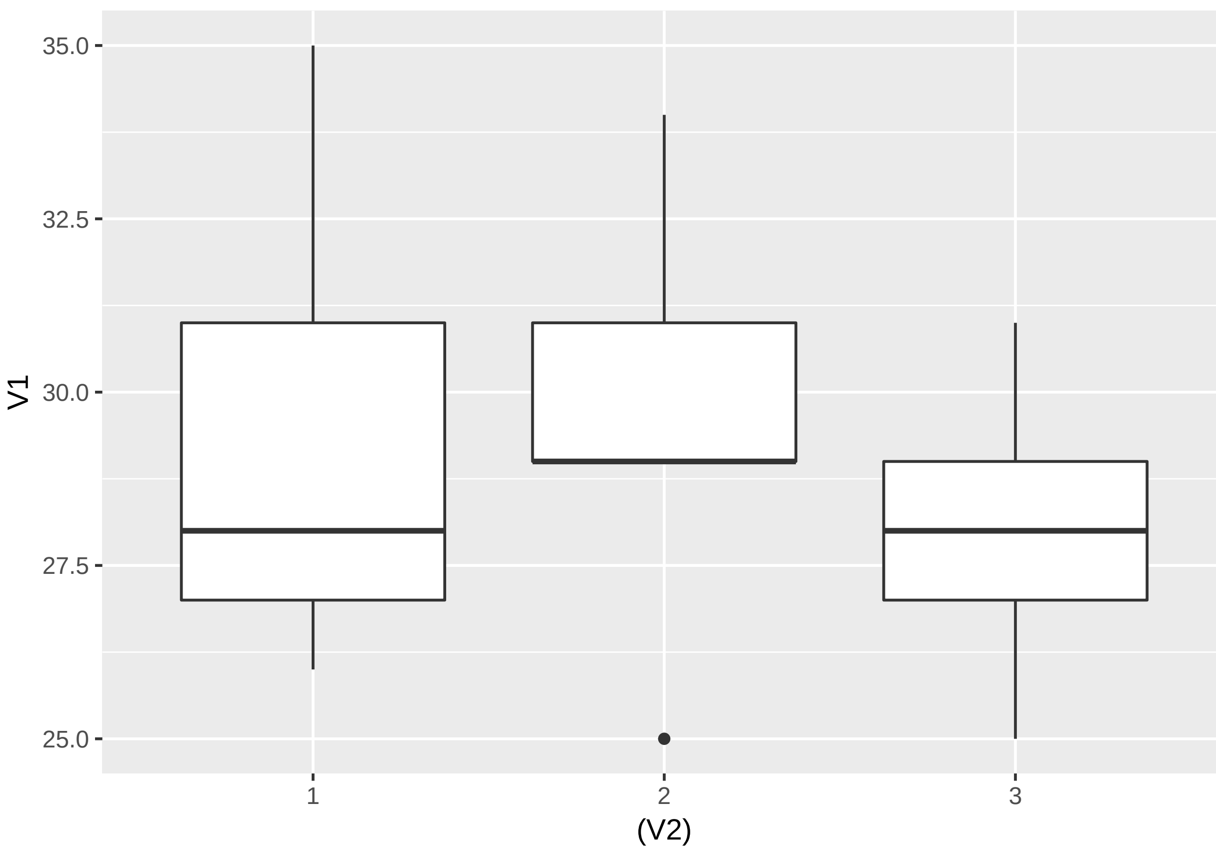
July 1, 2017

(adapted from KNNL 16.8) The data in "paper.txt" are from an experiment investigating the effect of paper color on response rates for a questionnaire.

(a) Create a box plot of the data separating the results by color. Does it appear there are differences in the response rates? Yes.

```
df <- read.table('./data/paper.txt', header = F)
df <- df[, c(1,2)]
df$V2 <- as.factor(df$V2)
p <- ggplot(df, aes(x= (V2), y=V1)) + geom_boxplot()

print(p)
```



(b) Calculate the mean and standard deviation of the response rate for each of the three colors. Also calculate the pooled sample standard deviation.

```
df.mean <- tapply(df$V1,df$V2,mean)
df.sd <- tapply(df$V1,df$V2,sd)
df.var <- tapply(df$V1,df$V2,var)

pooled.sd <- sqrt(sum(4 * tapply(df$V1,df$V2, var)) / (15 - 3))
pooled.sd
```

```
## [1] 3.114482
```

(c) Denoting the population means of the blue and green groups by 1 and 2, respectively, carry out the test of $H_0 : \mu_1 - \mu_2 = 0$ versus the two-sided alternative. Report the degrees of freedom, test statistic, and conclusion with $\alpha = 0.05$.

```
pooled.sd <- sqrt(sum((5-1) * df.var[1:2]) / (5 + 5 - 2))
t.value <- (df.mean[1] - df.mean[2])/ (pooled.sd * sqrt(1/5 + 1/5))
qt(0.025, 8)
```

```
## [1] -2.306004
```

$df = 8$

$t - value = -0.0911$

$p - value = 2 \times pt(0.0911) = `r2 * pt(-0.0911, 8)$

whe can't reject the null hypothesis.

(d) Estimate 1 2 with 95% confidence.

```
df.mean[1] - df.mean[2] - abs(qt(0.025, 8)) * pooled.sd * sqrt(1/5 + 1/5)
```

```
##         1
## -5.262716
```

```
df.mean[1] - df.mean[2] + abs(qt(0.025, 8)) * pooled.sd * sqrt(1/5 + 1/5)
```

```
##        1
## 4.862716
```

$lb = -5.262716$

$ub = 4.862716$

(e) What assumptions are you making for parts (c) and (d)?

Two the samples have approximately the same variance.

Problem 2

```
data=read.table("./data/CH01PR19.txt")
y=data[,1]
x=data[,2]
fit=lm(y~x)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## x            0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

a)

$y = 2.11405 + 0.03883 \times x$

b)

The estimated standard deviation of the coefficient.

c)

$\beta_{lb} = \beta - t_{\alpha/2,n-2} \times Std.Error = 0.017659$

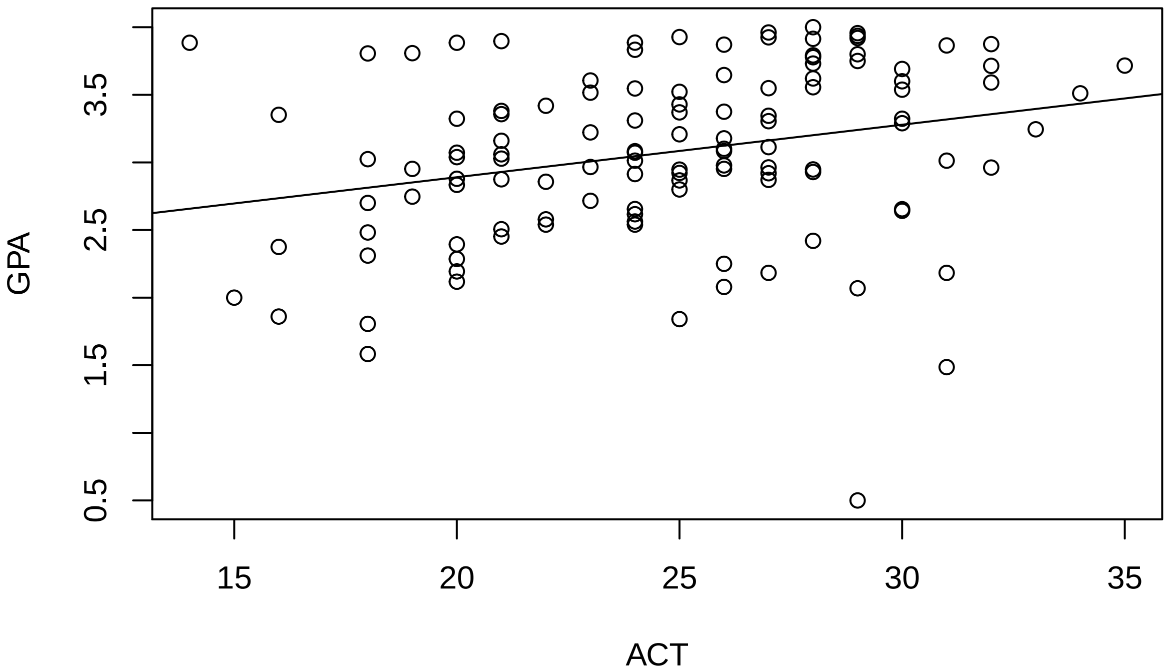
$\beta_{up} = \beta + t_{\alpha/2,n-2} \times Std.Error = 0.060001$

d)

$y = 2.11405 + 0.03883 \times x = 3.27895$

e)

```r
plot(x, y, xlab='ACT', ylab='GPA')
abline(fit)
```



Problem 3

a)

```
df1 <- data[data$V2 <= 25,]
df2 <- data[data$V2 > 25,]
fit1 <- lm(V1 ~ V2, data=df1)
fit2 <- lm(V1 ~ V2, data=df2)
```

$mean(GPA_low) = 2.9495077$

$mean(GPA_high) = 3.2212364$

$meandifference = 0.2717287$

b)

```
x1 = 1 * (x>25)
```

The $x > 25$ returs True and False. When multiplied by 1 the returned value is 0 and 1.

c)

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.72124 -0.30905  0.07163  0.47220  0.97749
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.94951    0.07845  37.597   <2e-16 ***
## x1           0.27173    0.11588   2.345   0.0207 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6325 on 118 degrees of freedom
## Multiple R-squared:  0.04452,    Adjusted R-squared:  0.03643
## F-statistic: 5.499 on 1 and 118 DF,  p-value: 0.0207
```

$differenceexpectedGPA = 2.94951$

$Stderror = 0.07845$

$df = 120 - 2 = 118$

d)

$t = \frac{\beta}{Std.Error} = 2.3449258$

$t^* = 1.9802722 < 2.345$

$P - value = 0.0206973$

e)

Yes.

```
low=y[x<=25];high=y[x>25]
```

Returns the value of y if the conidtion in bracket is true.

f)

$mean_{high} = 3.2212364$

$mean_{low} = 2.9495077$

$differece_{mean} = 0.2717287$

$df = 118$ is equal to df in part c.

g)

Yes, because both d, and g tests the difference in mean in two groups.

```
t.test(low, high, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  low and high
## t = -2.3449, df = 118, p-value = 0.0207
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.50120290 -0.04225445
## sample estimates:
## mean of x mean of y
##  2.949508  3.221236
```