p8130_hw5_ps3194

Pangsibo Shen

11/19/2020

```
#initial setup
library(tidyverse)
library(arsenal)
```

```
## Warning: package 'arsenal' was built under R version 4.0.3
```

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.0.3
```

```
library(broom)
theme_set(theme_minimal() + theme(legend.position = "bottom"))

options(
  ggplot2.continuous.colour = "viridis",
  ggplot2.continuous.fill = "viridis"
)

scale_colour_discrete = scale_color_viridis_d
scale_fill_discrete = scale_fill_viridis_d
```

Problem 1

Given the non-normal distributions, now you are asked to use an alternative, non-parametric test to assess and comment on the difference in Ig-M levels between the two groups (please ignore unanswered and missing values). You can use R to perform the calculations, but make sure you state test that you used, the hypotheses, test statistic, p-value and provide interpretation in the context of the problem.

```
#load antibodies dataset
antibodies = read_csv("./data/Antibodies.csv") %>%
  drop_na() #remove NA values

#extract IgM values for Normal group
Normal =
  antibodies %>%
  filter(Smell == "Normal") %>%
```

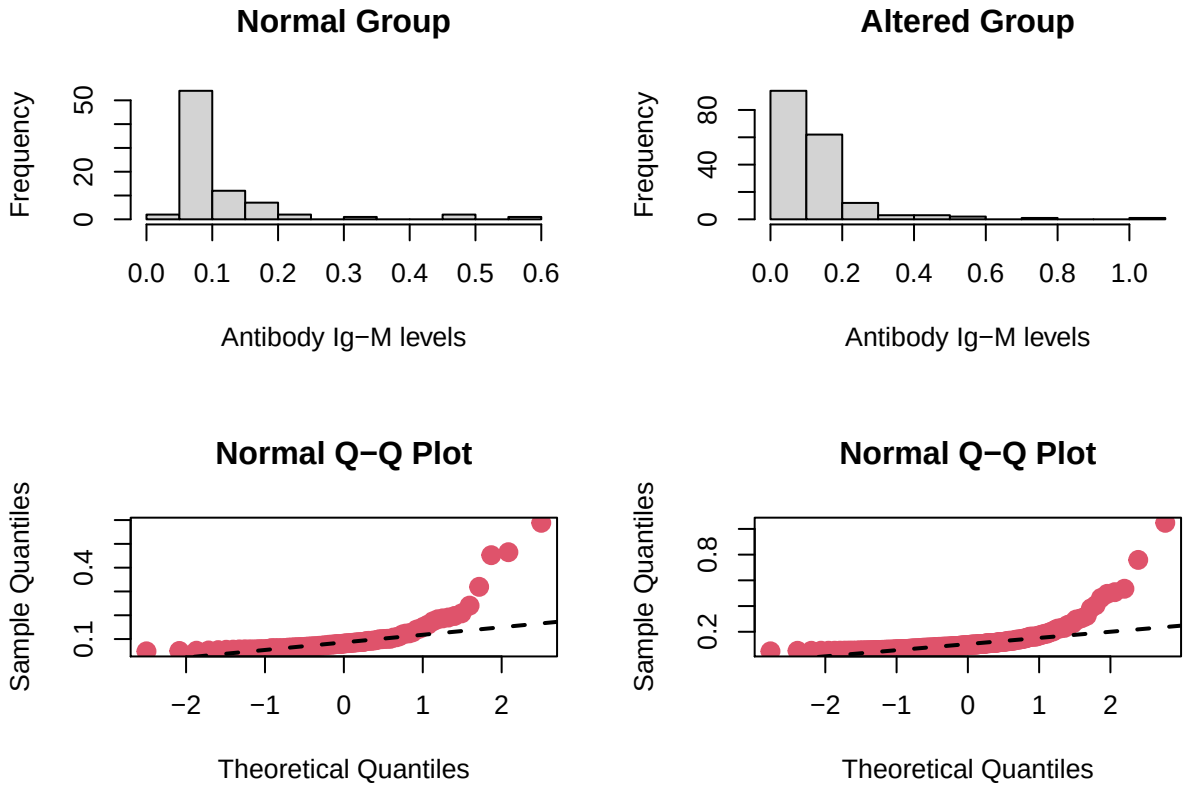
```
  pull(Antibody_IgM)

#extract IgM values for Altered group
Altered=
  antibodies %>%
  filter(Smell == "Altered") %>%
  pull(Antibody_IgM)
```

solution We first loaded the antibodies dataset and extract Ig-M levels for both Normal and Altered groups.Then we're going to do test for normality on both groups.

```
#Graphic methods to test for normality
par(mfrow=c(2,2))
hist(Normal, xlab="Antibody Ig-M levels", freq=T, main="Normal Group")
hist(Altered, xlab="Antibody Ig-M levels", freq=T, main="Altered Group")
qqnorm(Normal, col=2, pch=19, cex=1.5)
qqline(Normal, col = 1,lwd=2,lty=2)

qqnorm(Altered, col=2, pch=19, cex=1.5)
qqline(Altered, col = 1,lwd=2,lty=2)
```



From both the histograms and Quantitle Quantitle plots, we found both groups failed their normality assumptions. Hence, we're going to use non-parametric Wilcoxon-Rank Sum Test to assess whether there is difference in Ig-M levels between the Normal and Altered groups since both groups have more than 10 observations and we're under normal-approximation. First, we establish our hypotheses:

$$H_0 : the\ medians\ of\ the\ two\ populations\ are\ equal$$
$$H_1 : the\ medians\ of\ the\ two\ populations\ are\ not\ equal$$

```
#check if there is ties in the data
duplicated.data.frame(antibodies$Antibody_IgM)
```

```
## [1] FALSE
```

Since there is no ties in the data, we're going to use test statistics with no ties

```
# Non-parametric Wilcoxon-Rank Sum test: two-independent groups
res = wilcox.test(Normal, Altered, mu=0)
# add the n1(n1+1)/2 term
res$statistic = res$statistic + length(Normal)*(length(Normal)+1)/2
res
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Normal and Altered
## W = 9157, p-value = 0.01406
## alternative hypothesis: true location shift is not equal to 0
```

```
Test_statistics = (abs(res$statistic-(length(Normal)*(length(Normal)+length(Altered)+1))/2)-0.5)/sqrt((
Test_statistics #T score
```

```
##        W
## 2.455635
```

```
qnorm(0.975) #critical value for 0.05 significance level
```

```
## [1] 1.959964
```

$$T = \frac{|T_1 - \frac{n_1(n_1+n_2+1)}{2}| - \frac{1}{2}}{\sqrt{(\frac{n_1 n_2}{12})(n_1+n_2+1)}}$$

$$T = \frac{|9157 - \frac{81(81+178+1)}{2}| - \frac{1}{2}}{\sqrt{(\frac{81*178}{12})(81+178+1)}} = 2.45563$$

$$Z_{1-\alpha/2} = 1.959964$$

$$T > Z_{1-\alpha/2}$$

$$Under\ Normal - Approximation : n_1\ and\ n_2 > 10$$

We got T = 2.4556349, p-value = 0.01406 and T is greater than the critical value($Z_{1-\alpha/2} = 1.959964$).Since the test statistic is greater than the critical value and p-value is less than 0.05, we reject the Null hypothesis and conclude that at significance level of 0.05, the medians of Ig-M levels between the control and altered groups are different.

Problem 2

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

1) $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, $i = 1, 2, \dots, n$

The likelihood of a normal distribution:

$$L(\mu, \sigma^2 | Y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(Y_i - \mu)^2}{2\sigma^2}}$$

The likelihood of the linear model:

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}}$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \cdot e^{\frac{-\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}}$$

Maximize the log-likelihood function:

$$\ln L(\beta_0, \beta_1, \sigma^2) = \log\left[(2\pi\sigma^2)^{-n/2} \cdot e^{\frac{-\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}} \right]$$

$$= -\frac{n}{2} \log(2\pi\sigma^2) + \frac{-\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}$$

$$= -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}$$

Take the first derivative respect to β_0 and set to 0.

$$\frac{\partial \ln L}{\partial \beta_0} = -\frac{2}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-1) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 X_i = 0 \quad \Rightarrow \quad n\beta_0 = \sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i$$

$$\Rightarrow \hat{\beta_0} = \frac{\sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i}{n} = \bar{Y} - \hat{\beta_1} \bar{X}$$

$$\frac{\partial \ln L}{\partial \beta_1} = -\frac{2}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-X_i) = 0$$

$$\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} \beta_0 X_i - \sum_{i=1}^{n} \beta_1 X_i^2 \right) = 0$$

$$\sum_{i=1}^{n} Y_i X_i - \beta_0 \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \beta_1 X_i^2 = 0 \quad \text{substitute } \beta_0 \text{ with } \hat{\beta}_0$$

$$\sum_{i=1}^{n} Y_i X_i - (\bar{Y} - \beta_1 \bar{X}) \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \beta_1 X_i^2 = 0$$

$$\sum_{i=1}^{n} Y_i X_i = (\bar{Y} - \beta_1 \bar{X}) \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \beta_1 X_i^2$$

$$= \bar{Y} \sum_{i=1}^{n} X_i - \beta_1 \bar{X} \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \beta_1 X_i^2$$

$$\sum_{i=1}^{n} X_i \hookrightarrow n \cdot \bar{X}$$

$$= \bar{Y} \sum_{i=1}^{n} X_i - n \beta_1 \bar{X}^2 + \beta_1 \sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n} X_i \hookrightarrow n \cdot \bar{X}$$

$$= n \cdot \bar{X} \cdot \bar{Y} - \beta_1 \left(\sum_{i=1}^{n} X_i^2 - n \bar{X}^2 \right)$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - n \bar{X} \cdot \bar{Y}}{\sum_{i=1}^{n} X_i^2 - n \bar{X}^2} = \text{Corr}(Y_i, X_i) \cdot \frac{SD(Y_i)}{SD(X_i)}$$

---

2) 'estimated errors' / residuals denoted by $e_i = Y_i - \hat{Y}_i$

Show that the sum of the residual is zero. $\sum_{i=1}^{n} e_i = 0$.

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{Y}_i$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$= n \cdot \bar{Y} - \sum_{i=1}^{n} (\beta_0 + \beta_1 X_i) \qquad = n \cdot \bar{Y} - \sum_{i=1}^{n} (\bar{Y} - \beta_1 \bar{X} + \beta_1 X_i)$$

$$= n \cdot \bar{Y} - \sum_{i=1}^{n} \bar{Y} + \beta_1 \sum_{i=1}^{n} \bar{X} - \beta_1 \sum_{i=1}^{n} X_i$$

$$= n \cdot \bar{Y} - n \bar{Y} + n \cdot \bar{X} \cdot \beta_1 - n \cdot \bar{X} \cdot \beta_1 = 0$$

□

Problem 3

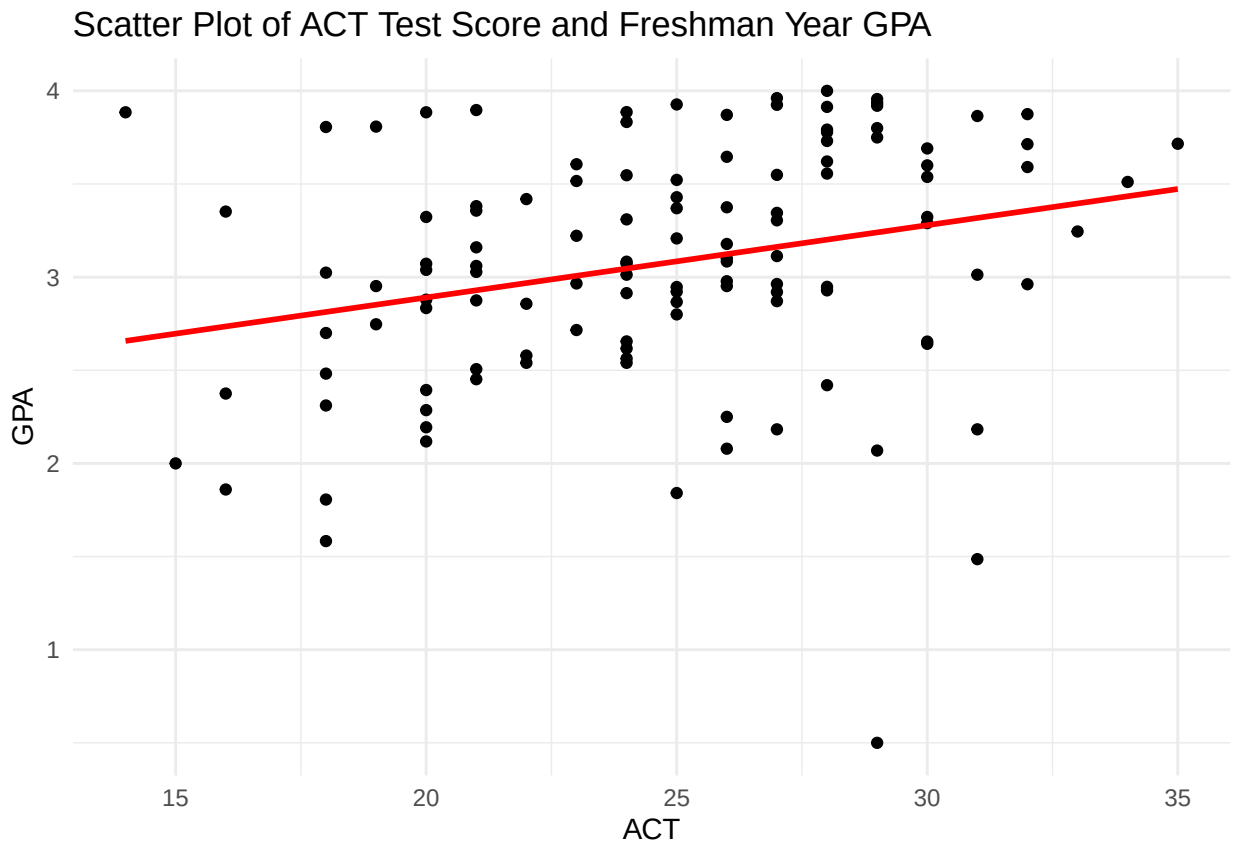
```
#load the gpa data
gpa = read_csv("./data/GPA.csv")
```

solution

1)

```
#Scatter plot with regression line overlaid
gpa %>%
  ggplot(aes(ACT,GPA)) +
  geom_point() +
  ggtitle("Scatter Plot of ACT Test Score and Freshman Year GPA") +
  geom_smooth(method='lm', se=FALSE, color='red')
```

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



```r
GPA = gpa$GPA
ACT = gpa$ACT
lm1 = lm(GPA~ACT)
# Summarize regression
summary(lm1)
```

```
##
## Call:
## lm(formula = GPA ~ ACT)
##
## 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 ***
## ACT          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
```

```r
# Get the critical t value for alpha=0.05 and n-2 df
qt(0.975, 118)
```

```
## [1] 1.980272
```

```r
# calulate test statistics
Test_statistics_t = (summary(lm1)$coefficients[2,1])/(summary(lm1)$coefficients[2,2])
```

Since we're going to test whether there is a linear association exists between student's ACT score and GPA, our null hypothesis will be $\beta_1 = 0$. The hypotheses, test statistics and critical value are written below:

$$H_0 : \beta_1 = \beta_{10} = 0 H_1 : \beta_1 \neq \beta_{10} t = \frac{\hat{\beta}_1 - \beta_{10}}{se(\hat{\beta}_1)} t = \frac{0.03883 - 0}{0.01277} = 3.039777 t_{n-2,1-\alpha/2} = 1.980272 |t| > t_{n-2,1-\alpha/2}$$

We got test statistics 3.0397768 which is larger than the critical value 1.9802722. Hence we reject the null hypothesis and conclude that at significance level of 0.05, there is a significant linear association between s between student's ACT score (X) and GPA at the end of the freshman year (Y).

2)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 XGPA = 2.11405 + 0.03883 * ACT$$

3)

```r
#calculate the 95% CI for the slope

coef<-summary(lm1)$coefficients[2,1]
err<-summary(lm1)$coefficients[2,2]
slope_int<-coef + c(-1,1)*err*qt(0.975, 118)
slope_int
```

```
## [1] 0.01353307 0.06412118
```

The 95% confidence interval for $\beta 1$ is (0.0135331, 0.0641212) which does not include zero. We are 95% confidence that the true slope(the expected increase in GPA for one point increase in ACT score) will between those two values. If the confidence interval includes zero, then we conclude that there is no evidence of a relationship between the predictor X(ACT score) and the response Y(freshman year GPA) in the population or we fail to reject the null hypothesis.

4)

```r
new <- data.frame(ACT = 28)
predict.lm(lm1, new, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 3.201209 3.061384 3.341033
```

The 95% interval estimate of the mean freshman GPA for students whose ACT test score is 28 is (3.061384,3.341033.). We are 95% confidence that the true freshman year GPA will lay between 3.061384 and 3.341033 for students whose ACT test score is 28.

5)

```r
#Prediction interval
predict.lm(lm1, new, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 3.201209 1.959355 4.443063
```

The 95% prediction estimate of the freshman GPA for Anne who obtained a score of 28 on the entrance test is (1.959355,4.443063). We are 95% confidence that the true freshman year GPA for Anne who obtained a score of 28 on the entrance test will lay between 1.959355 and 4.443063.

6)

The prediction interval in part 5) is wider than the confidence interval in part 4). Because we're taking the error term into account in prediction interval, the standard error is augmented by 1. As a result, the prediction interval is wider than the confidence interval.