**RSCH 6120/8120: HW 1**

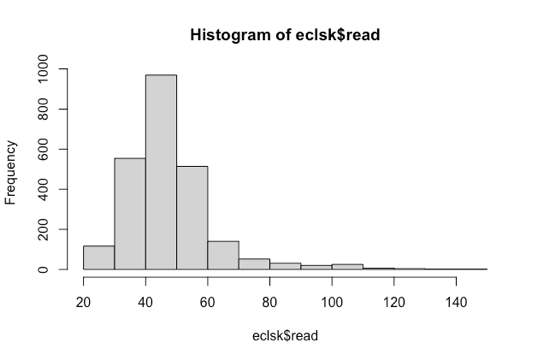
**Kristen Zhang | Sept 5**

**Answer the following questions using the eclsk.csv data. Provide code and results for each question.**

**1. a) Histogram of eclsk$read**

eclsk <- read.csv("eclsk.csv")

hist(eclsk$read)

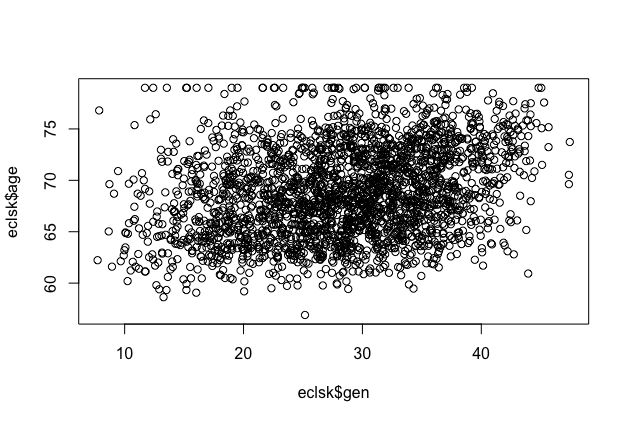


**b) Describe the shape of the distribution of reading scores (normal, positive skew, negative skew, bimodal, uniform)?**

Positive skew: as the mode of the reading scores located at (40, 50), which is leaning to the left side of the x-scale. Also, we can tell that in the eclsk$read data, the mean > median > mode.

**2. a) Scatter plot of eclsk$gen and eclsk$age.**

plot(eclsk$gen,eclsk$age)



**b) Describe the relationship (strength and direction) of the two variables in the scatter plot.**

It’s really hard to tell the strength and direction here. It looks scattered. If I have to pick one, I would say that it’s a little leaning to the right and top side.

**c) Determine the correlation of eclsk$gen and eclsk$age**

> cor(eclsk$gen,eclsk$age)

[1] 0.2782645

**d) Interpret the correlation coefficient. Does it match the scatter plot in part a)?**

The correlation here is quite low. Lower than 0.3. Therefore we can say that the relationship between eclsk$gen and eclsk$age is very weakly positive. I think it matches the plot in part a).

**3. a) Conduct a t-test to examine differences between male and female students (eclsk$female) on reading scores (eclsk$read).**

> t.test(read~female, data=eclsk)

Welch Two Sample t-test

data: read by female

t = 4.0632, df = 2426.5, p-value = 4.994e-05

alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0

95 percent confidence interval:

1.218705 3.492236

sample estimates:

mean in group 0 mean in group 1

48.83038 46.47491

**b) Interpret the results (T value and p value).**

From the results, we can tell that the p value (4.994e-05) is pretty small, therefore the difference between male and female on the reading scores is quite significant; For the T value, t = 4.0632 is very high, which means that the difference of mean between the two groups are significant. Therefore, we can reject the null hypothesis that male and female students don’t have different reading scores, and we can say that there is significant difference between male and female students’ reading scores.

**4. a) Conduct an ANOVA to determine if reading scores are significantly different across age categories (see ANOVA code for eclsk$age.cat set-up).**

> # Create a group variable: low, med, or high age: Age ranges from 57-79:

> eclsk$age.cat <- "b - average"

> eclsk$age.cat[which(eclsk$age >= 72)] <- "c - older" #72 months or older

> eclsk$age.cat[which(eclsk$age <= 66)] <- "a - younger" #66 months or younger

> eclsk$age.cat <- as.factor(eclsk$age.cat)

> # ANOVA analysis

> anova(lm(read~age.cat, data=eclsk))

Analysis of Variance Table

Response: read

Df Sum Sq Mean Sq F value Pr(>F)

age.cat 2 12724 6362.0 31.708 2.545e-14 \*\*\*

Residuals 2432 487961 200.6

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

**b) Interpret results (F value and p value).**

From the result (F value =31.708, p value= 2.545e-14), F value is very high, to indicate that the the differences between age groups are more significant relative to differences within the model. P-value is very close to 0 (<0.001), showing that the difference between age groups are statistically significant. We can reject the null hypothesis that different age groups don’t have significant difference on reading scores.