**RSCH 6120/8120: HW 3**

**Kristen Zhang | Oct 10**

**1. A study proposal indicates an expected power rate of 70% to detect an intervention effect. How should 70% power be interpreted? Identify at least one method that could be used to increase this power rate.**

Usually, statistical power means the probability of detecting an effect if one is present, and thus rejecting the null hypothesis correctly. The 70% power suggested that there is a 70% chance that the study will detect the intervention effect as statistically significant, but still has a 30% chance of failing to detect the effect that exists, which causes Type Ⅱ error.

To increase the power rate, the most direct way is to increase the sample size. A larger sample size allows for greater statistical power and larger sensitivity to detect the effects that exist.

**The following questions use several data sets: eclsk.csv (eclsk); hayden\_2005.sav (hayden2005); baguley\_payne\_2000.sav (bp2000). See wk7 R code for data import help.**

**2. Get descriptive statistics for the bp2000 data set**

**a) Include code and results**

> summary(bp2000)

group match trials percent\_accuracy

Min. :0.0 Min. : 1.00 Min. : 8 Min. : 12.50

1st Qu.:0.0 1st Qu.: 6.75 1st Qu.: 8 1st Qu.: 75.00

Median :0.5 Median :12.00 Median :16 Median : 87.50

Mean :0.5 Mean :13.43 Mean :16 Mean : 80.65

3rd Qu.:1.0 3rd Qu.:21.00 3rd Qu.:24 3rd Qu.: 91.67

Max. :1.0 Max. :24.00 Max. :24 Max. :100.00

> describe(bp2000)

vars n mean sd median trimmed mad min max range

group\* 1 56 1.50 0.50 1.5 1.50 0.74 1.0 2 1.0

match 2 56 13.43 7.82 12.0 13.48 11.86 1.0 24 23.0

trials 3 56 16.00 8.07 16.0 16.00 11.86 8.0 24 16.0

percent\_accuracy 4 56 80.65 19.04 87.5 83.51 12.36 12.5 100 87.5

skew kurtosis se

group\* 0.00 -2.04 0.07

match 0.00 -1.79 1.04

trials 0.00 -2.04 1.08

percent\_accuracy -1.59 2.43 2.54

**b) Note any strange results. Are typical descriptive statistics (measures of central tendency and dispersion) appropriate for all of the variables?**

I highlighted the group variable’s results. The group variable is a categorical variable, not interval/ratio. Therefore, measures of central tendency and dispersion would be meaningless for these kinds of variables.

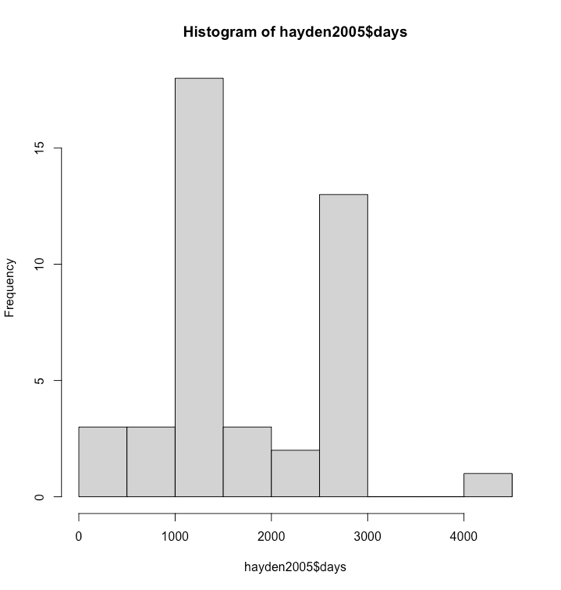
**Graphical assessment of statistical model assumptions**

**3.-Create a histogram using ‘days’ in the hayden\_2005.sav data**

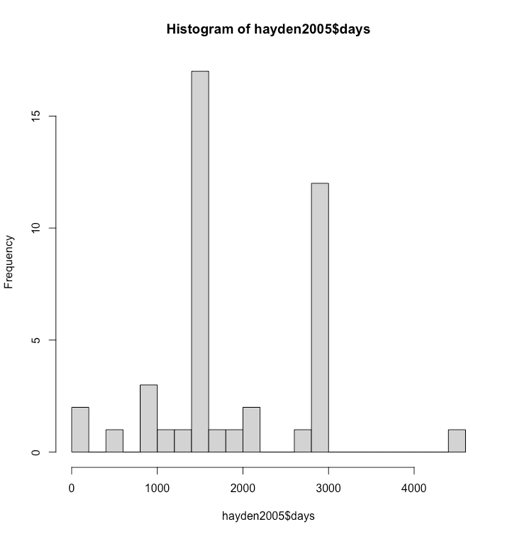
**-Create another histogram using ‘days’ in the hayden\_2005.sav data with a different number of breaks or bins**

**a) Include code and plots**

> hist(hayden2005$days)



> hist(hayden2005$days, breaks = 16)



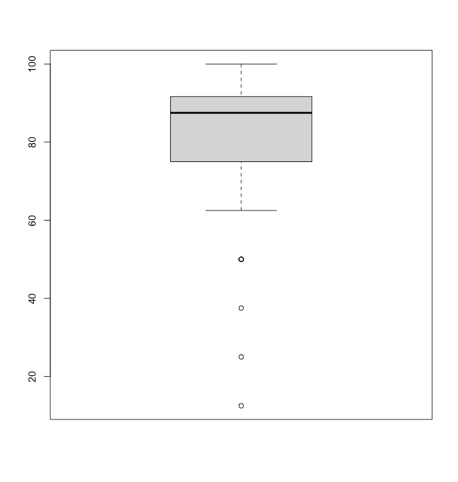
**b) Describe the distribution of ‘days’**

The distribution of ‘days’ has a mode between 1000 and 1500, and then the second most number comes from 2500-3000. It is not a normal distribution, but a robust distribution. Kind of like bimodal.

**4.-Create a box plot with ‘percent\_accuracy’ in the bp2000 data**

**a) Include code and plots**

> boxplot(bp2000$percent\_accuracy)



**b) Describe the shape/skew of percent\_accuracy**

From the boxplot, it’s quite leaning to the top side. It could be a negative skew if we use the histogram. Also, the long tail below the box proves that very small values form the tail to the left.

**c) Are there any possible outliers?**

Yes, four outliers are quite below the main body of the box, making it with a long tail. The minimum is 12.5, which is the lowest point on the boxplot.

These outliers represent natural variations in the population (true outliers), rather than some outliers that are problematic and should be moved (caused by measurement errors, data entry, or processing errors). These outliers match the trials. So we don’t need to remove them.

**5.-Create a QQ Plot from the following simple regression:**

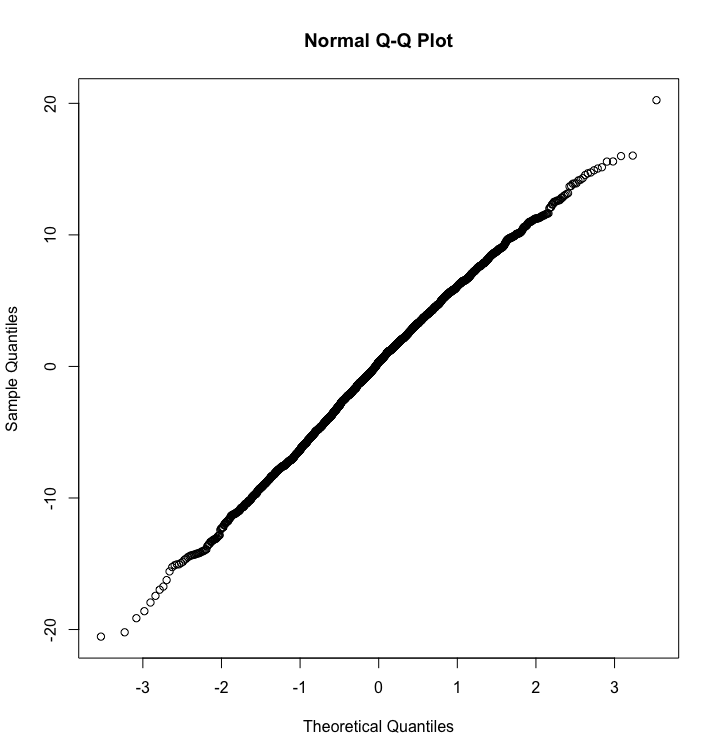
**lin.model <- lm (gen~math,eclsk)**

1. **Include code and plot**

> lin.model <- lm (gen~math,eclsk)

> lin.model.residuals<-lin.model$residuals

> qqnorm(lin.model.residuals)



**b) Interpret the results**

From the results, we can tell that it’s almost a linear line, so the dataset could be a normal distribution. The sample distribution is similar to the theory distribution.

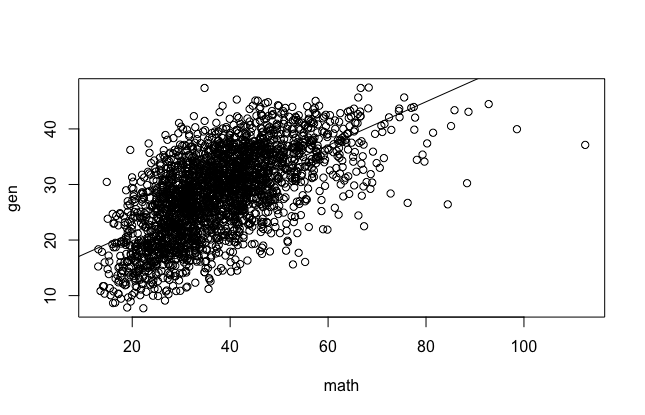
**6.-Create a scatter plot of eclsk$math and eclsk$gen with a fitted regression line**

**a) Include code and plot**

> plot(eclsk$gen,eclsk$math)

> with(eclsk, plot(math,gen)) #(y=x)

> abline(lm (gen~math,eclsk))



1. **What does this plot tell us about linearity and homogeneity of variance?**

From what I can tell in the picture, the line is quite straight and the points look evenly above and below the line, so it could meet the requirement of linearity. It’s a weakly positive relationship between math and gen. However, the variance of residue can’t be the same (math scores could be higher as the scatter plot fans out). It doesn’t meet the homogeneity of variance.