Capstone – Data Analysis

Hannah Wilson

3/11/2022

library(readxl)  
library(cfcdae)

## Registered S3 method overwritten by 'DoE.base':  
## method from   
## factorize.factor conf.design

library(car)

## Loading required package: carData

library(nlme)  
dataA <- read\_excel("C:/Users/hanna/OneDrive/Documents/Hannah1.xlsx", sheet="A")  
dataB <-read\_excel("C:/Users/hanna/OneDrive/Documents/Hannah1.xlsx", sheet="B")

## Set up Initial Factors

data1 <- na.exclude(dataA)  
block.f <- as.factor(data1$ExcavBlock)  
levels(block.f)

## [1] "Bld11" "Block1" "Block2Footing" "Block2N"   
## [5] "Block2S" "Block3" "M1" "M11"   
## [9] "M2" "M5" "M6" "M9"

geo.f <- as.factor(data1$Geostrat)  
levels(geo.f)

## [1] "A1" "A1a-2" "A4" "Aa" "B1" "B1a" "B1a-1" "B1A4-1"  
## [9] "B1c" "B1p" "B1x" "B1y" "B1z" "B1zk" "B2" "B2a"   
## [17] "B2ak" "B3" "Ba" "Bc" "II" "III" "IV" "Iva"   
## [25] "VI"

lev.f <- as.factor(data1$Level)  
levels(lev.f)

## [1] "1" "2" "2b" "2B" "3" "4" "4a"   
## [8] "5" "5a" "6" "A1" "A2" "A4-1" "A4b"   
## [15] "Aa" "B1" "B1a" "B1a-1" "B1a-2" "b1x" "B1x"   
## [22] "B1y" "B1yV" "B1z" "B2" "B2a" "B2ak" "B3"   
## [29] "Ba" "Bc" "I" "II" "IIa" "IIb" "IIB"   
## [36] "IIB2a" "III" "IIIB2ak" "IV" "V" "VB1y" "VI"

lev.f2 <- as.factor(data1$Level.G)  
levels(lev.f2)

## [1] "1" "2" "3" "4" "5" "6" "A" "A1" "A2" "A4" "B" "B1" "B2" "B3"

geo.f2 <- as.factor(data1$Geostrat.G)  
levels(geo.f2)

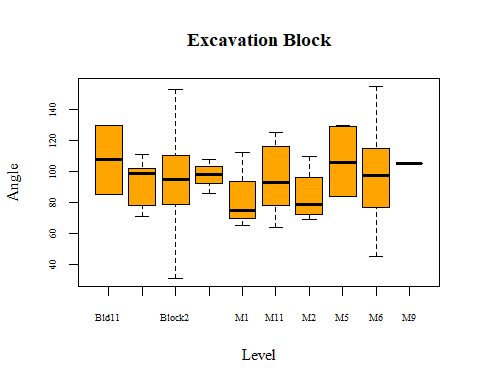
## [1] "2" "3" "4" "A" "A1" "A4" "B" "B1" "B2" "B3"

block.f2 <- as.factor(data1$Excav.G)  
levels(block.f2)

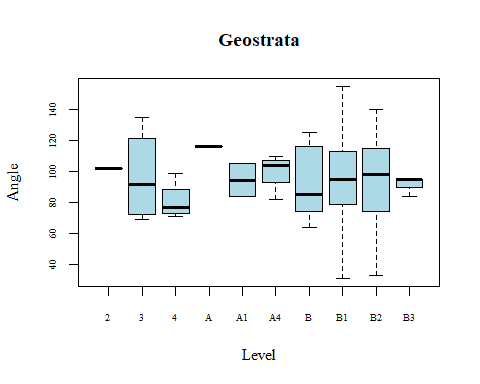
## [1] "Bld11" "Block1" "Block2" "Block3" "M1" "M11" "M2" "M5"   
## [9] "M6" "M9"

## Visual Representations of Difference

par(family = "serif")  
windowsFonts(A = windowsFont("Times New Roman"))  
boxplot(break1~block.f2, data=data1, cex.axis=0.6, main="Excavation Block", xlab="Level", ylab="Angle", col="orange")



par(family = "serif")  
boxplot(break1~geo.f2, data=data1, cex.axis=0.6, main="Geostrata", xlab="Level", ylab="Angle", col="light blue")



The box plots show that the means of each block and geostrata are at least slightly different: they also have very different ranges, which could signal unequal variance. The plot for the geostrata seems to have more variation present than the block factor.

## Tests for Assumptions

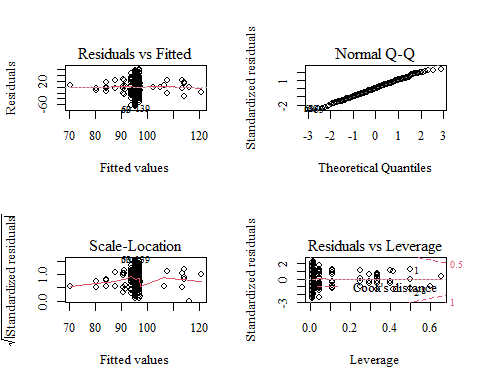
To have an accurate test, the assumptions of independence, normality, and equal variance must be met or otherwise accounted for.

The independence assumption is met, because we assume that the angle measurements of each bone don’t depend on the other bones.

To test for normality and equal variance, we look at residual plots.

bone.mod1 <- lm(break1~block.f+geo.f, data=data1)  
bonegroup.mod1 <- lm(break1~block.f2+geo.f2, data=data1)  
par(mfrow=c(2,2))  
par(family="serif")  
plot(bonegroup.mod1)

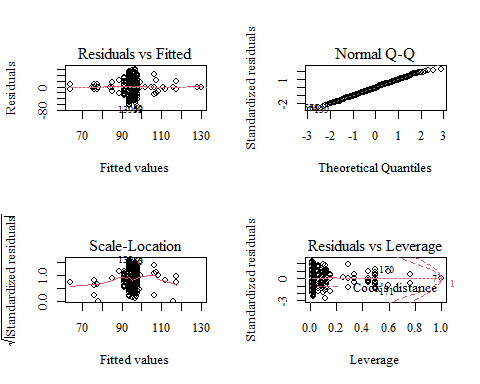
## Warning: not plotting observations with leverage one:  
## 4, 192, 264



par(mfrow=c(2,2))  
plot(bone.mod1)

## Warning: not plotting observations with leverage one:  
## 3, 4, 172, 178, 192, 193, 194, 195, 264

## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced  
  
## Warning in sqrt(crit \* p \* (1 - hh)/hh): NaNs produced



outlierTest(bone.mod1)

## No Studentized residuals with Bonferroni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferroni p  
## 139 -2.6681 0.0081709 NA

outlierTest(bonegroup.mod1)

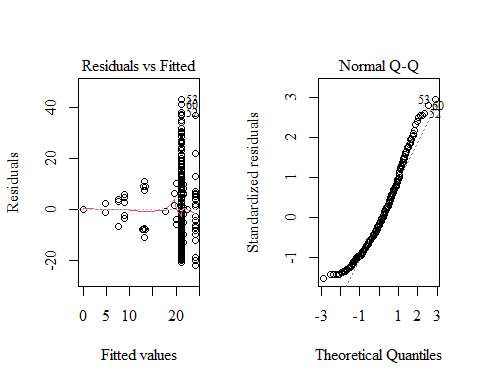
## No Studentized residuals with Bonferroni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferroni p  
## 53 -2.518704 0.012419 NA

Unfortunately, it looks like we do have some problems. The grouping of data near the center of the first plot is not that big of a concern (it just means we have more data near the middle), but the non-flat line in the scale-location plot indicates that we do have unequal variance problems. This could be due to an outlier, but the outlier test shows no evidence for this. The normal plot is pretty good, but we have slight evidence of long tails.

To further test the possibility of unequal variance, we need to fit the factors against the residuals. Then, the ANOVA test will be looking at the variances between factor levels versus the means between factor levels. If the test is significant, then we know we have unequal variance.

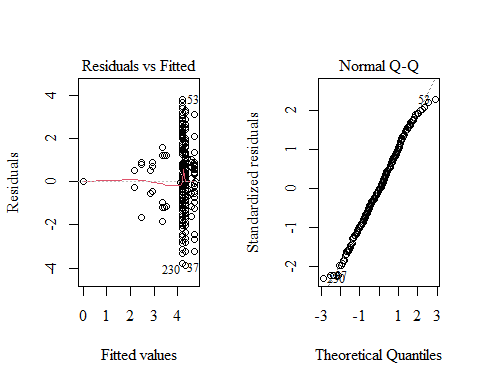
resid1 <- resid(bonegroup.mod1)  
bone.residmod1 <- lm(abs(resid1)~block.f2+geo.f2, data=data1)  
par(family="serif")  
par(mfrow=c(1,2))  
plot(bone.residmod1, which=1:2)

## Warning: not plotting observations with leverage one:  
## 4, 192, 264



bone.residmod2 <- lm(sqrt(abs(resid1))~block.f2+geo.f2, data=data1)  
plot(bone.residmod2, which=1:2)

## Warning: not plotting observations with leverage one:  
## 4, 192, 264



outlierTest(bone.residmod2)

## No Studentized residuals with Bonferroni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferroni p  
## 230 -2.33697 0.020251 NA

The same methods apply for this ANOVA test - we look at residual plots, make transformations to meet our assumptions, and test for outliers. There is no significant evidence for outliers in the residual model. The square root transformation made the normality much better, and the absolute value of the residuals looks like the original distribution: this is acceptable for analysis.

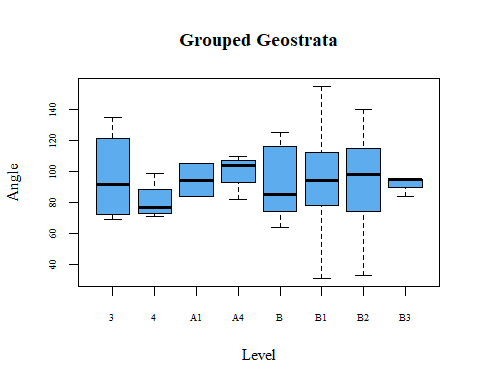
anova(bone.residmod2)

## Analysis of Variance Table  
##   
## Response: sqrt(abs(resid1))  
## Df Sum Sq Mean Sq F value Pr(>F)   
## block.f2 9 54.20 6.0220 2.1376 0.027111 \*   
## geo.f2 9 76.10 8.4552 3.0013 0.002055 \*\*  
## Residuals 245 690.22 2.8172   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

This ANOVA was conducted specifically to test the equality of variance between the factor levels of geo.f2 and block.f2. Because they are significant in this test, we do have unequal variance. This means that an ANOVA test on the model with the factors included will not be accurate: we need to pursue a model that fits variances separately.

## Separate Variance Model

dataG<- read\_excel("C:/Users/hanna/OneDrive/Documents/Hannah2.xlsx")  
geo.g <- as.factor(dataG$Geo.G)  
par(family = "serif")  
boxplot(break1~geo.g, data=dataG, main="Grouped Geostrata", xlab="Level", ylab="Angle", cex.axis=0.6, col="steelblue2")



geo.g1 <- gls(break1~geo.g, data=dataG, weights=varIdent(form=~1|geo.g))  
summary(geo.g1)

## Generalized least squares fit by REML  
## Model: break1 ~ geo.g   
## Data: dataG   
## AIC BIC logLik  
## 2441.589 2498.374 -1204.794  
##   
## Variance function:  
## Structure: Different standard deviations per stratum  
## Formula: ~1 | geo.g   
## Parameter estimates:  
## B1 4 A4 B3 B2 3 B A1   
## 1.0000000 0.4828989 0.5695212 0.2453407 1.0104897 1.1870743 0.9401307 0.5735874   
##   
## Coefficients:  
## Value Std.Error t-value p-value  
## (Intercept) 93.18978 2.975145 31.322769 0.0000  
## geo.g1 3.56022 13.634357 0.261121 0.7942  
## geo.g2 -12.43978 6.176528 -2.014040 0.0450  
## geo.g3 1.31022 9.566758 0.136956 0.8912  
## geo.g4 5.47689 7.949030 0.689001 0.4914  
## geo.g5 0.47689 7.629233 0.062508 0.9502  
## geo.g6 2.62331 3.346623 0.783867 0.4338  
## geo.g7 0.84868 5.346800 0.158728 0.8740  
##   
## Correlation:   
## (Intr) geo.g1 geo.g2 geo.g3 geo.g4 geo.g5 geo.g6  
## geo.g1 0.509   
## geo.g2 -0.216 -0.303   
## geo.g3 0.173 -0.264 -0.166   
## geo.g4 0.009 -0.274 -0.104 -0.184   
## geo.g5 -0.028 -0.278 -0.090 -0.180 -0.139   
## geo.g6 -0.850 -0.461 0.173 -0.166 -0.022 0.009   
## geo.g7 -0.350 -0.328 0.021 -0.161 -0.082 -0.065 0.289  
##   
## Standardized residuals:  
## Min Q1 Med Q3 Max   
## -2.50380439 -0.70717579 -0.03141038 0.68667201 2.28645900   
##   
## Residual standard error: 25.88584   
## Degrees of freedom: 265 total; 257 residual

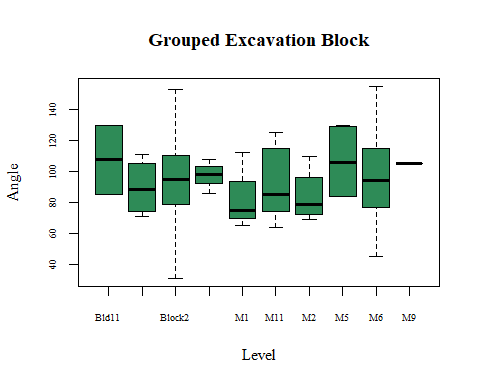
anova(geo.g1)

## Denom. DF: 257   
## numDF F-value p-value  
## (Intercept) 1 4437.245 <.0001  
## geo.g 7 0.912 0.4979

For the separate variance model of geostrata, the levels of “A” and “2” were removed because they each had too few data points for this model to work. The remaining levels of geostrata are shown in the boxplot.

The ANOVA for this model has a high p-value of 0.4979, which means that significant differences between the levels of geostrata were not detected in this sample. The break angles do not significantly differ in the separate geostrata classifications and can be analyzed together, as one group, in the future.

block.g <- as.factor(dataG$Excav.G)  
par(family="serif")  
boxplot(break1~block.g, data=dataG, main="Grouped Excavation Block", xlab="Level", ylab="Angle", cex.axis=0.6, col="seagreen")



block.g1 <- gls(break1~block.g, data=dataG, weights=varIdent(form=~1|block.g))  
anova(block.g1)

## Denom. DF: 255   
## numDF F-value p-value  
## (Intercept) 1 3896.969 <.0001  
## block.g 9 0.442 0.9112

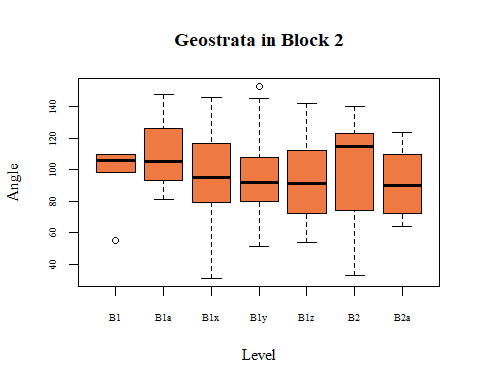
Similarly, the ANOVA for the excavation block levels has a very high p-value. No significant differences were found between the levels, so they can be used as one group in further analysis.

## Within Block 2

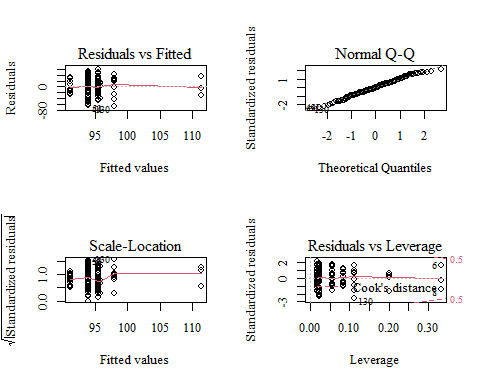
Because Block 2 is the largest contributor to our sample, geostrata differences within this block were of interest. The process is the same as for the larger models.

dataC <- read\_excel("C:/Users/hanna/OneDrive/Documents/Hannah1.xlsx", sheet="C")  
dataC1 <- na.exclude(dataC)

block2.mod <- lm(Break~Geostrat, data=dataC1)  
par(family="serif")  
boxplot(Break~Geostrat, data=dataC1, main="Geostrata in Block 2", xlab="Level", ylab="Angle", cex.axis=0.6, col="sienna2")



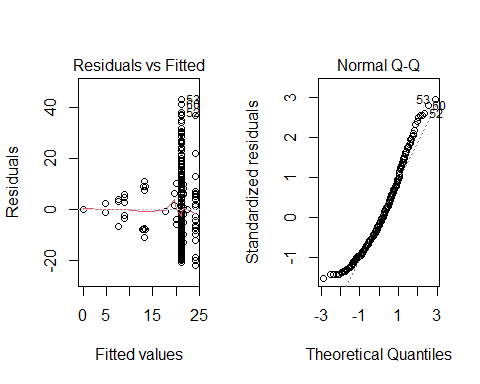
par(mfrow=c(2,2))  
plot(block2.mod)



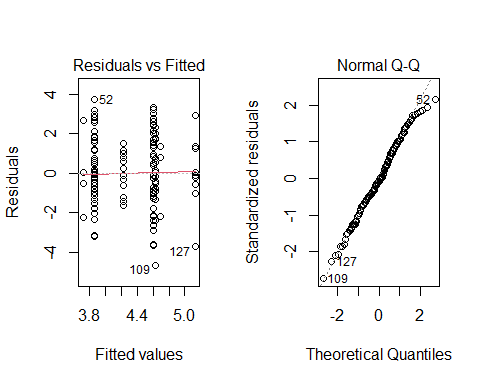
Looking at the boxplot and Scale-Location plot, the unequal variance problem from the earlier models seems to also extend here: we will use the same technique of testing for unequal variance and then fitting variances separately if we do indeed have problems meeting the equal variance assumption.

resid2 <- resid(block2.mod)  
block2.residmod <- lm(abs(resid2)~Geostrat, data=dataC1)  
par(mfrow=c(1,2))  
plot(bone.residmod1, which=1:2)

## Warning: not plotting observations with leverage one:  
## 4, 192, 264



block2.residmod2 <- lm(sqrt(abs(resid2))~Geostrat, data=dataC1)  
plot(block2.residmod2, which=1:2)



outlierTest(block2.residmod2)

## No Studentized residuals with Bonferroni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferroni p  
## 109 -2.787237 0.0060829 0.86985

anova(block2.residmod2)

## Analysis of Variance Table  
##   
## Response: sqrt(abs(resid2))  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Geostrat 6 24.26 4.0437 1.3172 0.2536  
## Residuals 136 417.50 3.0698

With Geostrat plotted against break angles, we see that the square root transformation really helps the data to meet the assumptions necessary for ANOVA. The second group of plots shows much better equal variance and normality than the first group.

This time, the p-value is not significant, and shows that equal variance is not an issue for this model. Because of this, we can use a regular ANOVA without variances fitted separately.

anova(block2.mod)

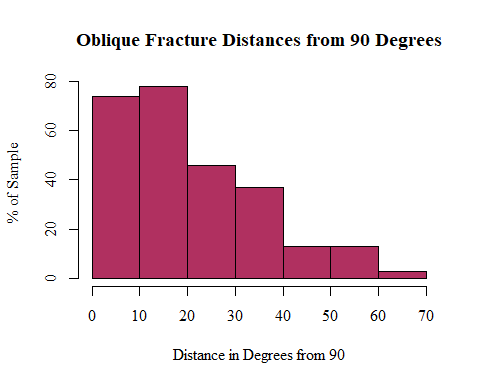
## Analysis of Variance Table  
##   
## Response: Break  
## Df Sum Sq Mean Sq F value Pr(>F)  
## Geostrat 6 1134 188.95 0.2525 0.9576  
## Residuals 136 101771 748.32

We get a very high p-value: within Excavation Block 2, the break angles are not significantly different in each level of geostrata.

## Similarity to Hyenas

Following a procedure set out by Coil et al. (2017) and using data from that project, this histogram shows the absolute value of the angle distances from 90.

A <- abs(data1$break1-90)  
par(family="serif")  
hist(A, breaks=7, col="maroon", main="Oblique Fracture Distances from 90 Degrees", xlab="Distance in Degrees from 90", ylab="% of Sample")



data2 <- read\_excel("C:/Users/hanna/OneDrive/Documents/Hyena Breakage Angles.xlsx", sheet= "Obl Angles")

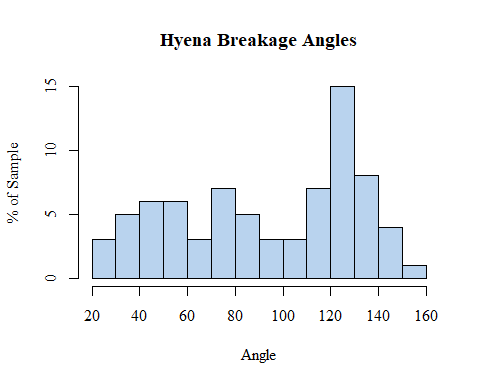
To compare the angles of hyena breakage to the site, we would want to use a t test. However, assumptions still apply, so we have to look into those.

We know that the samples are independent of each other, and we will use a simplistic measure of equal variance for the lists of data. The normality can be assessed using histograms.

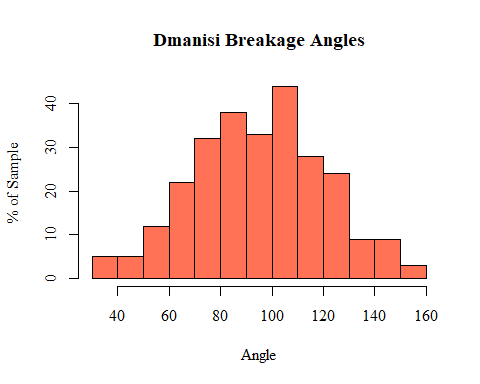
(sd(data2$Break1A)^2)/(sd(data1$break1)^2)

## [1] 2.158764

par(family="serif")  
hist(data2$Break1A, breaks=13, main="Hyena Breakage Angles", xlab="Angle", ylab="% of Sample", col="slategray2")



hist(data1$break1, main="Dmanisi Breakage Angles", xlab="Angle", ylab="% of Sample", col="coral1")

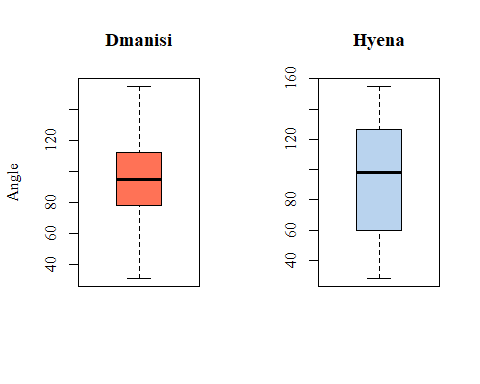


The ratio of the hyena variance over the site variance is less than 4, so we can assume equal variance. The histograms, however, show that the hyena data is not normally distributed, so we need to use a test similar to a t-test, but one that does not need normality as an assumption.

stats::wilcox.test(data2$Break1A,data1$break1)

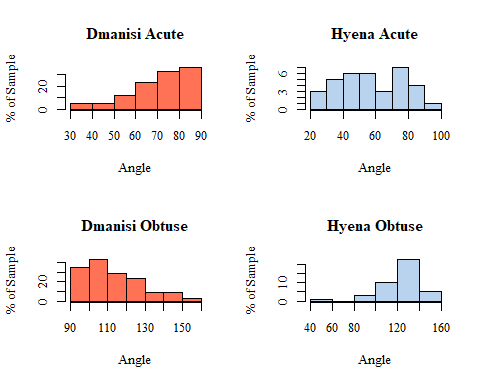
##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: data2$Break1A and data1$break1  
## W = 10164, p-value = 0.8623  
## alternative hypothesis: true location shift is not equal to 0

par(mfrow=c(1,2))  
par(family="serif")  
boxplot(data1$break1, main="Dmanisi", ylab="Angle", col="coral1")  
boxplot(data2$Break1A, main="Hyena", col="slategray2")



The Wilcoxon Test shows no significant difference between the full means of our site versus hyena breakage. However, because the hyena breakage data has such a large range, it could be useful to separate the acute and obtuse angles from each sample and look at their comparisons separately.

par(mfrow=c(2,2))  
par(family="serif")  
hist(dataB$Acute1, main="Dmanisi Acute", xlab="Angle", col="coral1", ylab="% of Sample")  
hist(data2$Acute, main="Hyena Acute", xlab="Angle", col="slategray2", ylab="% of Sample")  
hist(dataB$Obtuse1, main="Dmanisi Obtuse", xlab="Angle", col="coral1", ylab="% of Sample")  
hist(data2$Obtuse, main="Hyena Obtuse", xlab="Angle", col="slategray2", ylab="% of Sample")

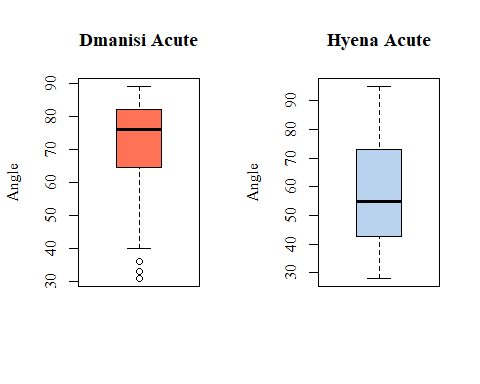


The histograms show that we still don’t meet the normality assumption, so we will still use the Wilcoxon Test.

wilcox.test(dataB$Acute1, data2$Acute)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: dataB$Acute1 and data2$Acute  
## W = 2873.5, p-value = 0.0001307  
## alternative hypothesis: true location shift is not equal to 0

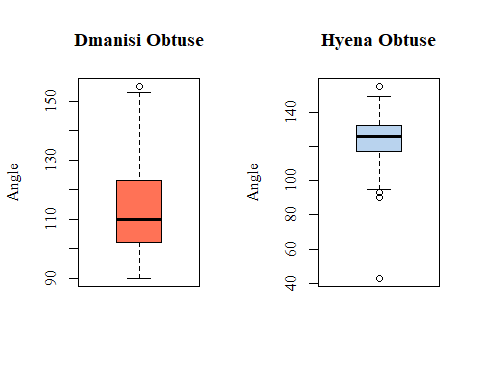
par(mfrow=c(1,2))  
par(family="serif")  
boxplot(dataB$Acute1, main="Dmanisi Acute", ylab="Angle", col="coral1")  
boxplot(data2$Acute, main="Hyena Acute", ylab="Angle", col="slategray2")



wilcox.test(dataB$Obtuse1, data2$Obtuse)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: dataB$Obtuse1 and data2$Obtuse  
## W = 1967.5, p-value = 0.000121  
## alternative hypothesis: true location shift is not equal to 0

par(mfrow=c(1,2))  
boxplot(dataB$Obtuse1, main="Dmanisi Obtuse", ylab="Angle", col="coral1")  
boxplot(data2$Obtuse, main="Hyena Obtuse", ylab="Angle", col="slategray2")



When comparing the acute angles from our site to the acute angles of the hyena site, the test is significant. The acute breakage angles from each sample are significantly different.

Similarly, the obtuse angles from each sample were significantly different.

These differences are highly visible on the box plots – in both cases, the means are very far removed from each other, and in the Obtuse comparison, the ranges do no overlap.

These results suggest that the breakage angles found at the site are not statistically similar to hyena breakge from the other study.

References

Coil, R., 2016. Spatial approaches to site formation and carnivore-hominin interaction at

Dmanisi, Georgia. Ph.D. Dissertation, University of Minnesota.

Coil, R., Tappen, M., Yezzi-Woodley, K., 2017. New analytical methods for comparing bone

fracture angles: a controlled study of hammerstone and hyena (*Crocuta Crocuta*) long

bone breakage. Archaeometry 59, 900-917.

Tappen, M. et al., Press. Journal of Human Evolution.