Week\_9

## *Or, technically week 8, part II*

### First let’s load the libraries we’ll need:

library(dplyr)  
library(ggplot2)

**Note:** to hide the output of these two commands, add message=FALSE to the curvy brackets with {r}. Press the “run current chunk” button, *not* Ctrl-Enter.

#### Loading the data frame:

parentage\_data <- read.csv("../data\_8/parentage\_with\_phenotypes\_CSV.csv")

### Next, we’ll manipulate the data:

#Filtering out NAs and blanks by selecting sex columns only with m or f  
filtered\_parentage\_data <- parentage\_data %>%  
 select(sex, length, depth, sex.1, length.1, depth.1, sex.2, length.2, depth.2) %>%  
 filter(sex == "m" | sex== "f") %>%   
 filter(sex.1 == "m" | sex.1 =="f") %>%   
 filter(sex.2 =="m" | sex.2 == "f") %>%   
   
#Creating length:depth ratio columns for offspring and each parent   
 mutate(LD\_ratio = length/depth) %>%   
 mutate(LD.1\_ratio = length.1/depth.1) %>%   
 mutate(LD.2\_ratio = length.2/depth.2) %>%   
 na.omit() #ensuring all NAs are removed

##### *Notice a pipe was used for this entire chunk*

#### Let’s see what we’ve done so far:

head(filtered\_parentage\_data)

## sex length depth sex.1 length.1 depth.1 sex.2 length.2 depth.2 LD\_ratio  
## 1 f 388 90 m 400 116 f 428 111 4.311111  
## 2 f 398 99 m 400 116 f 428 111 4.020202  
## 3 f 425 94 m 462 157 f 447 114 4.521277  
## 4 m 395 109 m 400 103 f 387 111 3.623853  
## 5 m 415 128 m 431 124 f 410 91 3.242188  
## 6 f 418 104 m 432 137 f 409 99 4.019231  
## LD.1\_ratio LD.2\_ratio  
## 1 3.448276 3.855856  
## 2 3.448276 3.855856  
## 3 2.942675 3.921053  
## 4 3.883495 3.486486  
## 5 3.475806 4.505495  
## 6 3.153285 4.131313

View(filtered\_parentage\_data)

### Next we’ll be calculating the **z-scores** of the parents and offspring.

#### A z-score is the distance of a value from the mean in terms of standard deviations. Statistically, values with z-scores greater than 2 (more than 2 standard deviations from the mean) are considered unusual.

#### Before we can calculate z-scores, though, we need the mean and standard deviation of the length:depth columns.

#standard deviations  
sd\_dad <- sd(filtered\_parentage\_data$LD.1\_ratio)  
sd\_mom <- sd(filtered\_parentage\_data$LD.2\_ratio)  
sd\_offspring <- sd(filtered\_parentage\_data$LD\_ratio)   
  
 #means  
mean\_dad <- mean(filtered\_parentage\_data$LD.1\_ratio)  
mean\_mom <- mean(filtered\_parentage\_data$LD.2\_ratio)  
mean\_offspring <- mean(filtered\_parentage\_data$LD\_ratio)

#### Now for the z-scores, featuring a fancy for loop

#z-scores  
 #dad  
for (i in seq(1, length(filtered\_parentage\_data$LD.1\_ratio))) { #establishing the data to which the loop will apply (the dad length:depth (LD.1) ratio column)  
 z\_score\_dad <- (filtered\_parentage\_data$LD.1\_ratio[i] - mean\_dad)/ sd\_dad #i is used as a generic variable. We're subtracting each entry in LD.1\_ratio from the mean and dividing that difference by the standard deviation  
 filtered\_parentage\_data$z\_score\_dad[i] <- as.numeric(z\_score\_dad) #assigning the results of the above line to the column "z\_score\_dad"  
   
}  
  
 #double check that it worked properly...  
head(filtered\_parentage\_data$z\_score\_dad)

## [1] -0.08279160 -0.08279160 -2.49565180 1.99418987 0.04859165 -1.49056754

(filtered\_parentage\_data$LD.1\_ratio[1] - mean\_dad)/sd\_dad # If the loop worked correctly, this should yield the same value as the first output value from the loop

## [1] -0.0827916

#mom  
for (i in seq(1, length(filtered\_parentage\_data$LD.2\_ratio))) {  
 z\_score\_mom <- (filtered\_parentage\_data$LD.2\_ratio[i] - mean\_mom)/ sd\_mom   
 filtered\_parentage\_data$z\_score\_mom[i] <- as.numeric(z\_score\_mom)  
   
}  
head(filtered\_parentage\_data$z\_score\_mom)

## [1] -0.37144982 -0.37144982 -0.09387941 -1.94401240 2.39433833 0.80128901

#offspring  
for (i in seq(1, length(filtered\_parentage\_data$LD\_ratio))) {  
 z\_score\_offspring <- (filtered\_parentage\_data$LD\_ratio[i] - mean\_offspring)/ sd\_offspring  
 filtered\_parentage\_data$z\_score\_offspring[i] <- as.numeric(z\_score\_offspring)  
   
}  
head(filtered\_parentage\_data$z\_score\_offspring)

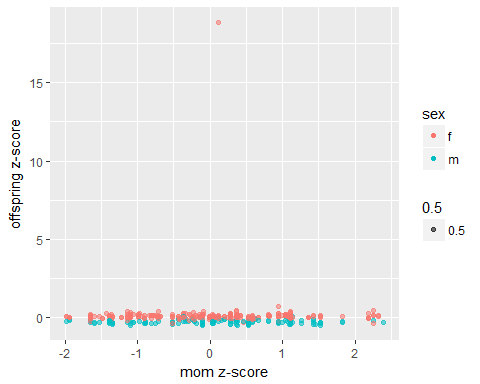
## [1] 0.28482151 0.12709609 0.39876926 -0.08779674 -0.29472868 0.12656950

##### *These all look like reasonable z-scores!*

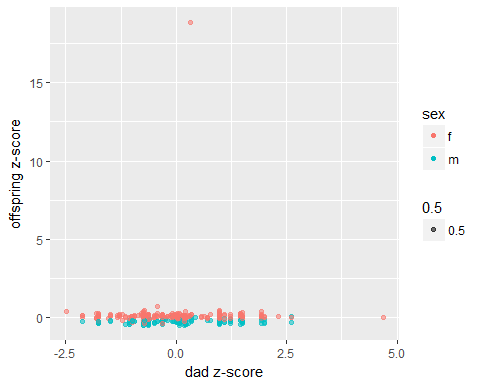
### Now that we have the z-scores filled out, let’s graph them to see if there are any correlations.

#### More specifically, We’ll make scatter plots for mom z-score vs. ofspring z-score and dad z-score vs. offspring z-score.

##graphing z-scores to compare between offspring and each parent  
 #mom and offspring  
ggplot(data = filtered\_parentage\_data) +  
 geom\_point(mapping = aes(x = filtered\_parentage\_data$z\_score\_mom, y = filtered\_parentage\_data$z\_score\_offspring, color = sex, alpha = 0.5))+  
 labs(x = "mom z-score", y = "offspring z-score")

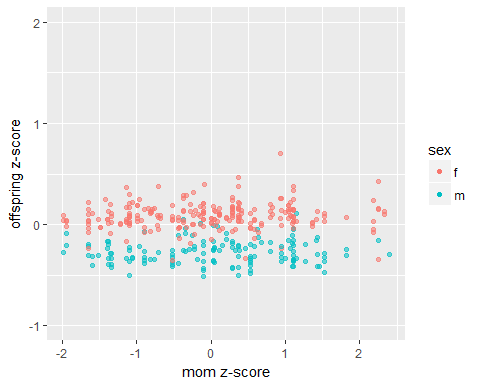


#dad and offspring  
ggplot(data = filtered\_parentage\_data) +  
 geom\_point(mapping = aes(x = filtered\_parentage\_data$z\_score\_dad, y = filtered\_parentage\_data$z\_score\_offspring, color = sex, alpha = 0.5))+  
 labs(x = "dad z-score", y = "offspring z-score")

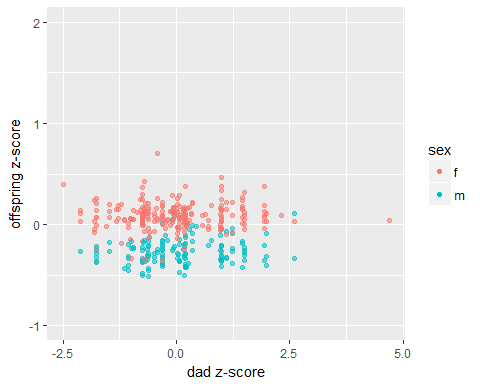


#### We can see that there’s one obvious outlier in both plots throwing off the scale and obscuring any trends that may be there. Let’s fix that (and also take out that “0.5” legend that makes no sense):

#mom and offspring - outlier removed  
 ggplot(data = filtered\_parentage\_data) +  
 geom\_point(mapping = aes(x = filtered\_parentage\_data$z\_score\_mom, y = filtered\_parentage\_data$z\_score\_offspring, color = sex, alpha = 0.5))+  
 labs(x = "mom z-score", y = "offspring z-score") +  
 ylim(-1, 2)+  
 guides(alpha = FALSE)



#dad and offspring - outler removed  
 ggplot(data = filtered\_parentage\_data) +  
 geom\_point(mapping = aes(x = filtered\_parentage\_data$z\_score\_dad, y = filtered\_parentage\_data$z\_score\_offspring, color = sex, alpha = 0.5))+  
 labs(x = "dad z-score", y = "offspring z-score")+  
 ylim(-1,2)+  
 guides(alpha = FALSE)



#### That’s better! From these graphs we can see:

##### \* There appears to be no correlation between parent z-score (either mom or dad) and offspring z-score

##### \* Most offspring z-scores are between -0.5 and 0.5

##### \* Mst mom z-scores range from -2 to 2, and most dad z-scores range from -2.5 to 2.5

##### \* Female offspring tend to have more positive z-score values, and male offspring tend to have more negative z-score values

[[1]](#footnote-44) This will correspond to a footnote

1. This is a footnote [↑](#footnote-ref-44)