Assignment 1

Fenglin Chen

21/09/2021

### Question 1 a)

is location invariant if .

First, substitute into the equation.

Then, add a constant to all elements. Since the population mean is location equivariant, it also gains .

The additions cancel out.

Therefore, excess kurtosis is location invariant.

### Question 1 b)

is scale invariant if .

Similar to part a), begin by substituting into the equation.

Then, multiply all elements by . Since the population mean is scale equivariant, it is also multiplied by .

Factor out the multiplier in both the numerator and denominator.

is a constant, so it can be factored out of the sum and cancelled.

Therefore, excess kurtosis is scale invariant.

### Question 1 c)

Since excess kurtosis is both location invariant (part a) and scale invariant (part b), then by definition, it is location-scale invariant.

### Question 1 d)

is scale invariant if . Population is duplicated times, and the new population size is .

The summing operation is applied times, but the population is times larger, so the factors cancel out.

Therefore, excess kurtosis is replication invariant.

### Question 1 e)

In this case, the equation for is changed.

For part b), the equation becomes:

Since the multiplier does not relate to the size of the population , the attribute is still scale invariant (no change). However, for part d), the equation becomes:

Therefore, is no longer replication invariant, but rather neither invariant nor equivariant.

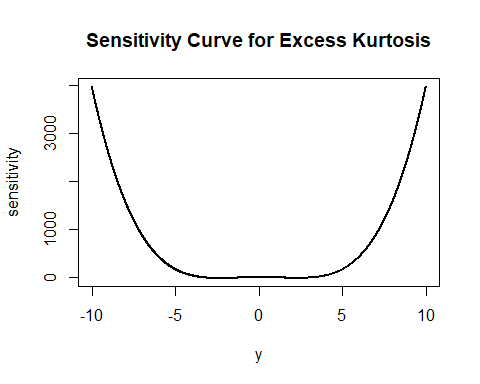
### Question 1 f)

The sensitivity curve is defined as:

Combined, the equation becomes:

### Question 1 g)

library(e1071)  
set.seed(341)  
pop <- rt(1000, 10)  
y <- seq(-10, 10, length.out=1001)  
  
sc <- function(y.pop, y, attr, ...) {  
 N <- length(y.pop) + 1  
 sapply(y, function(y.new) {  
 N \* (attr(c(y.new, y.pop), ...) - attr(y.pop, ...))  
 })  
}  
  
plot(y, sc(pop, y, kurtosis), type="l", lwd=2,  
 main="Sensitivity Curve for Excess Kurtosis",  
 ylab="sensitivity", xlab="y")



Based on this plot, kurtosis is robust to additions because its sensitivity curve is relative flat within [-5, 5]. However, it is unbounded because the two ends go to infinity, so it can be easily influenced by extreme outliers in the population.

### Question 1 h)

1. measures how close values lie in relation to each other, not in relation to some constant like 0, so it should be location invariant. A cluster of points are just as spread out at as .
2. should be scale equivariant. When a population of points are each multiplied by , the distance/spread between any two points are also multiplied by , so should reflect this change linearly.
3. should not be location-scale invariant nor equivariant, because location and scale have different ideal properties as seen in points 1 and 2. Choosing either would compromise location invariance or scale equivariance.

### Question 2 a)

filename <- paste("data/iris.csv", sep="/")  
iris <- read.csv(filename, header=TRUE)  
nrow(iris)

## [1] 150

ncol(iris)

## [1] 5

### Question 2 b)

library(pander)  
iris.freq <- table(iris$Species)  
pander(iris.freq, type = 'grid')

|  |  |  |
| --- | --- | --- |
| Iris-setosa | Iris-versicolor | Iris-virginica |
| 50 | 50 | 50 |

### Question 2 c)

print(paste0("Species with the largest sepal widths: ", iris$Species[iris$SepalWidth == max(iris$SepalWidth)][1]))

## [1] "Species with the largest sepal widths: Iris-setosa"

print(paste0("Species with the smallest sepal widths: ", iris$Species[iris$SepalWidth == min(iris$SepalWidth)][1]))

## [1] "Species with the smallest sepal widths: Iris-versicolor"

### Question 2 d)

for (s in unique(iris$Species)){  
 print(paste0("Average sepal length for ", s, ": ", mean(iris$SepalLength[iris$Species == s])))  
}

## [1] "Average sepal length for Iris-setosa: 5.006"  
## [1] "Average sepal length for Iris-versicolor: 5.936"  
## [1] "Average sepal length for Iris-virginica: 6.588"

### Question 2 e)

iris$PetalRatio <- iris$PetalWidth/iris$PetalLength  
  
# Iris species with largest and smallest PetalRatios  
print(paste0("Species with the largest petal ratio: ", iris$Species[iris$PetalRatio == max(iris$PetalRatio)][1]))

## [1] "Species with the largest petal ratio: Iris-virginica"

print(paste0("Species with the smallest petal ratio: ", iris$Species[iris$PetalRatio == min(iris$PetalRatio)][1]))

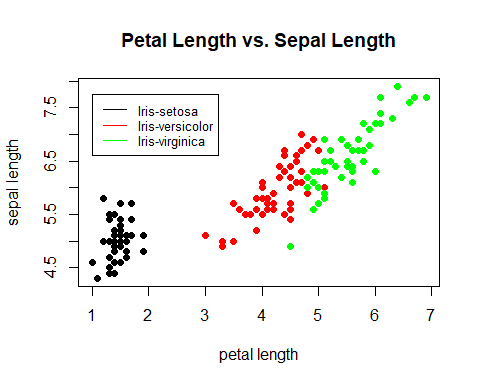
## [1] "Species with the smallest petal ratio: Iris-setosa"

# Proportion of samples with PetalRatio > 0.3  
for (s in unique(iris$Species)){  
 print(paste0("Proportion of samples in ", s, " with PetalRatio > 0.3: ", mean(iris$PetalRatio[iris$Species == s] > 0.3)))  
}

## [1] "Proportion of samples in Iris-setosa with PetalRatio > 0.3: 0.04"  
## [1] "Proportion of samples in Iris-versicolor with PetalRatio > 0.3: 0.66"  
## [1] "Proportion of samples in Iris-virginica with PetalRatio > 0.3: 0.86"

### Question 2 f)

# Match each species name to a colour  
species <- unique(iris$Species)  
iris$SpeciesNum <- match(iris$Species, species)  
colours <- c("black", "red", "green")  
  
# Plot the points and add a legend  
plot(iris$PetalLength, iris$SepalLength, pch=19,  
 col=colours[as.numeric(iris$SpeciesNum)],   
 xlab="petal length", ylab="sepal length",   
 main="Petal Length vs. Sepal Length")  
legend(1, 7.75, legend=species, col=colours, lty=1, cex=0.8)

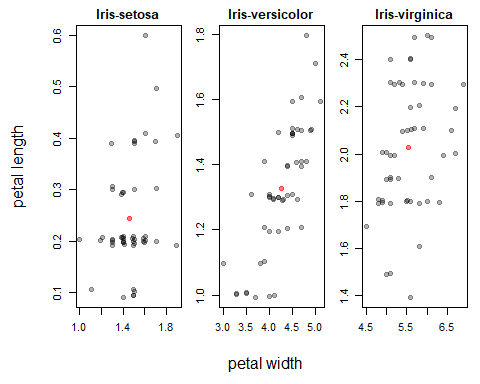


From the plot, it is clear that Iris-setosas can be distinguished from Iris-versicolor and Iris-virginica by petal length alone, as there is a clear separation between the groups.

In addition, there is a positive linear trend between petal length and sepal length in Iris-versicolor and Iris-virginica.

### Question 2 g)

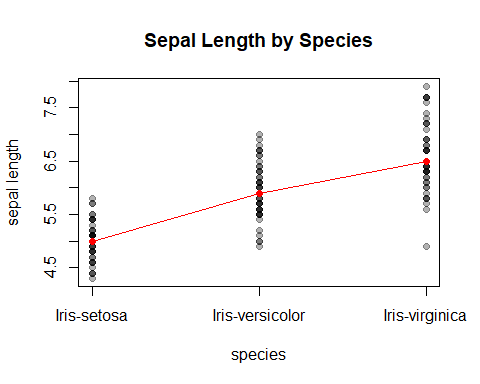
par(mfrow=c(1,3), mar=c(2, 2, 2, 1), oma=c(4, 4, 0, 0))  
for (s in species) {  
 plot(jitter(PetalWidth, factor=0.5) ~ jitter(PetalLength, factor=0.5),   
 data=iris[iris$Species == s, ],  
 col=adjustcolor("black", 0.3), pch=19, main=s)  
 points(mean(PetalWidth) ~ mean(PetalLength), pch=19,  
 data=iris[iris$Species == s, ],   
 col=adjustcolor("red", 0.5))  
}  
mtext('petal width', side=1, outer=TRUE, line=2)  
mtext('petal length', side=2, outer=TRUE, line=2)



The plots show that there is a slight positive linear relationship between petal width and petal length, especially in the plot of Iris-versicolor. The relationship might be clearer if the values were not discrete/overlapping.

### Question 2 h)

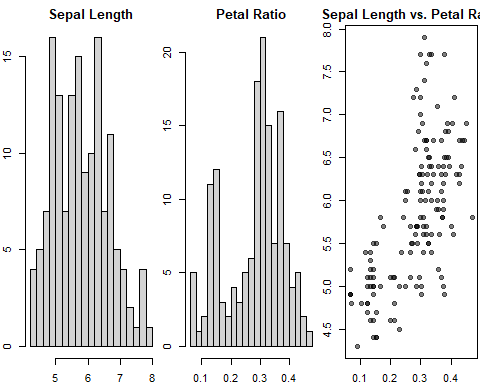
# Plot the base points  
par(mfrow=c(1,1))  
plot(iris$SpeciesNum, iris$SepalLength, pch=19,   
 col=adjustcolor("black", 0.3), xaxt="n",   
 xlab="species", ylab="sepal length", main="Sepal Length by Species")  
axis(side=1, at=c(1.0, 2.0, 3.0), labels=species)  
  
# Calculate, plot, and connect the medians  
medians <- c(median(iris$SepalLength[iris$SpeciesNum == 1]),  
 median(iris$SepalLength[iris$SpeciesNum == 2]),  
 median(iris$SepalLength[iris$SpeciesNum == 3]))  
points(1:3, medians, col="red", pch=19)  
lines(1:3, medians, col="red", pch=19)



### Question 2 i)

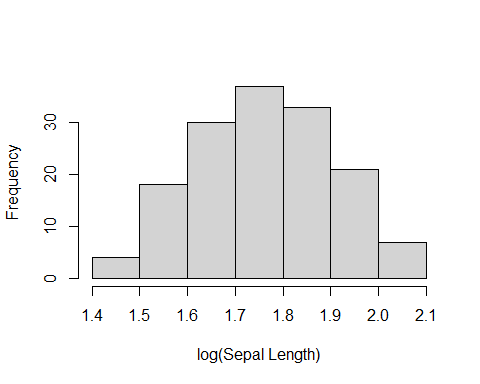
powerfun <- function(x, alpha) {  
 if(sum(x <= 0) > 1) stop("x must be positive")  
 if (alpha == 0)  
 log(x)  
 else if (alpha > 0) {  
 x^alpha  
 } else -x^alpha  
}

# Part i: Histogram and scatterplots of SepalLength and PetalRatio  
par(mfrow=c(1,3), mar=c(2, 2, 2, 0.2))  
hist(iris$SepalLength, breaks=15, main="Sepal Length")  
hist(iris$PetalRatio, breaks=15, main="Petal Ratio")  
plot(SepalLength ~ PetalRatio, data=iris, pch=19,  
 col=adjustcolor("black", 0.5),  
 main="Sepal Length vs. Petal Ratio")



# Part ii: Find alpha to make SepalLength symmetric  
for (i in c(-0.2, -0.1, 0, 0.1, 0.2)) {  
 hist(powerfun(iris$SepalLength, i), breaks=8)  
}

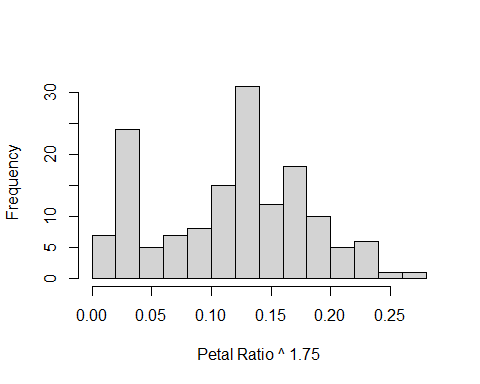
hist(powerfun(iris$SepalLength, 0), breaks=8,  
 xlab="log(Sepal Length)", ylab="Frequency", main="")



After testing, the power 0 (log) makes sepal length most symmetric.

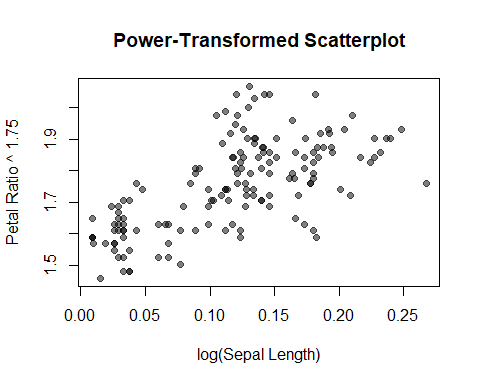
# Part iii: Find alpha to make PetalRatio symmetric  
for (i in c(1.25, 1.5, 1.75, 2, 2.25)) {  
 hist(powerfun(iris$PetalRatio, i), breaks=30)  
}

hist(powerfun(iris$PetalRatio, 1.75), breaks=15,  
 xlab="Petal Ratio ^ 1.75", ylab="Frequency", main="")



A power of 1.75 makes the histogram of petal ratio most symmetric, though it is harder to judge than sepal length due to the second peak on the left.

# Part iv: Find the pair of alphas to make the scatterplot approximately linear  
plot(powerfun(SepalLength, 0) ~ powerfun(PetalRatio, 1.75), data=iris,   
 pch=19, col=adjustcolor("black", 0.5),  
 main="Power-Transformed Scatterplot",   
 xlab="log(Sepal Length)", ylab="Petal Ratio ^ 1.75")



The new scatterplot uses the best alphas for sepal length (0) and petal ratio (1.75), respectively, resulting in a more linear graph than the one in part i).