Numerical data analysis using normal distributions and Z score transformations in R

Till now, we have explored builtin dataset called iris. We also explored importing datasets from websites. We also explored making our own CSV datasets and importing them into R.

Today, we will explore more builtin datasets and find their distributions and normalize then using Z score transformation method. In the discussion section for relevant task , don’t forget to discuss about z transformations and about normal distributions and how they relate to the particular task.

**TASK 0 Looking through built in datasets**

data()

Look through some of the builtin datasets by typing in

print(name of dataset)

head(name of dataset)

**TASK 1 Original “Trees” dataset**

We will look through the dataset called trees

print(trees)

head(trees)

plot(trees$Height)

plot(trees$Girth)

plot(trees$Volume)

hist(trees$Height, probability=TRUE)

hist(trees$Girth)

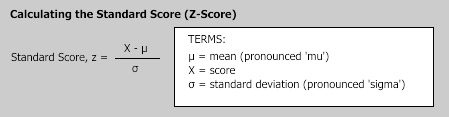
hist(trees$Volume)

Just by eyeballing the histograms which one of these three variables, Height, Volume, and Girth looks like they are normally distributed ?

mean(trees$Height)

sqrt(var(trees$Height)

**TASK 2 Z-transforming the trees dataset**



Z\_transformed\_heights <- (trees$Height-mean(trees$Height))/sqrt(var(trees$Height))

hist(Z\_transformed\_heights)

Z\_transformed\_Girth <- (trees$Girth-mean(trees$Girth))/sqrt(var(trees$Girth))

hist(Z\_transformed\_Girth)

mean(Z\_transformed\_heights)

sqrt(var(Z\_transformed\_heights)

Discuss about how the mean/SD of the original dataset in task 1 if different from the mean/SD of the z-transformed dataset in task 2.

**TASK 3 Comparing the original to a normal dataset.**

hist(trees$Height, probability=TRUE)

lines(density(trees$Height), col="red")

x <- seq(60,90,0.1)

y <- dnorm(x, mean=mean(trees$Height), sd=sqrt(var(trees$Height)))  
lines(x=x,y=y , col = "blue")

**TASK 4 Comparing the z transformed dataset to a normal dataset.**

hist(Z\_transformed\_heights, probability=TRUE)

lines(density(Z\_transformed\_heights), col="red")

x <- seq(-3,3,0.1)

y <- dnorm(x, mean=0, sd=1)  
lines(x=x,y=y , col="blue")

**TASK 5 Reviewing everything again**

bdims <- read.csv("http://www.jkarreth.net/files/bdims.csv")

head(bdims)

summary(bdims)

hist(bdims$hgt, probability = TRUE, breaks = 20)  
lines(density(bdims$hgt), col="red")

x <- 140:190  
y <- dnorm(x = x, mean = mean(bdims$hgt), sd = sqrt(var(bdims$hgt)))  
lines(x = x, y = y, col = "blue")

Z\_transformed\_heights <- (bdims$hgt-mean(bdims$hgt))/sqrt(var(bdims$hgt))

hist(Z\_transformed\_heights, probability=TRUE)

lines(density(Z\_transformed\_heights), col="red")

x <- seq(-3,3,0.1)  
y <- dnorm(x = x, mean = 0, sd = 1)  
lines(x = x, y = y, col = "blue")