#two group testing

#taken from URL https://data.gov.uk/dataset/573ff7a7-eb35-4d4e-9638-bad7e76d3aa5/bis-sickness-absence

sickness <- read.csv("BIS\_SICKNESS.csv", sep=",", header=T)

typeof(sickness$Days\_lost\_due\_to\_sickness\_2013)

#H0 is that there is no signficant differences between the means of the two years

#ho = median differneces are the same

#lets get the standard deviation of

sd(sickness$Days\_lost\_due\_to\_sickness\_2013) # sd is 5265.69

sd(sickness$Days\_lost\_due\_to\_sickness\_2014) #sd is 5784.987

#lets now try normalise 2014 data

plot(sickness$Days\_lost\_due\_to\_sickness\_2014)

qqnorm(sickness$Days\_lost\_due\_to\_sickness\_2014)

qqline(sickness$Days\_lost\_due\_to\_sickness\_2014)

#lets now try normalise 2013 data

plot(sickness$Days\_lost\_due\_to\_sickness\_2013)

qqnorm(sickness$Days\_lost\_due\_to\_sickness\_2013)

qqline(sickness$Days\_lost\_due\_to\_sickness\_2013)

#lets illustrate this on the box plot

boxplot(sickness$Days\_lost\_due\_to\_sickness\_2013)

boxplot(sickness$Days\_lost\_due\_to\_sickness\_2014)

boxplot(sickness$Days\_lost\_due\_to\_sickness\_2013, sickness$Days\_lost\_due\_to\_sickness\_2014, names = c("2013", "2014"))

#lets find out the means

mean(sickness$Days\_lost\_due\_to\_sickness\_2013) #5092.73

mean(sickness$Days\_lost\_due\_to\_sickness\_2014) #5900.227

#lets illustrate non-normality via histograms

hist(sickness$Days\_lost\_due\_to\_sickness\_2014)

hist(sickness$Days\_lost\_due\_to\_sickness\_2013)

#shapiro testing

shapiro.test(sickness$Days\_lost\_due\_to\_sickness\_2013)

shapiro.test(sickness$Days\_lost\_due\_to\_sickness\_2014)

#we can reject the null hypothesis

#lets now do a wilcoxon test

wilcox.test(sickness$Days\_lost\_due\_to\_sickness\_2013, sickness$Days\_lost\_due\_to\_sickness\_2014, paired = TRUE)

#there is a significant difference between the two data sets.

#lets do some power work

install.packages("effsize")

library("effsize")

install.packages("pwr")

library("pwr")

#lets find out the effect size between the groups

cohen.d(sickness$Days\_lost\_due\_to\_sickness\_2013, sickness$Days\_lost\_due\_to\_sickness\_2014)

# d is -0.1459833

#lets now do a power test to find out power values

pwr.t.test(n = 14, d = -0.1459833, power =NULL, alternative = "two.sided", type = "two.sample")

#our power is 0.06 or 6% which is terrible

pwr.t.test(n = NULL, d = -0.1459833, power =.80, alternative = "two.sided", type = "two.sample")

#our real n should be 737.5591

#there is a signficant difference between the two but the chance of making a type 2 error

#is huge!!!!

#suggestion would be to introduce many more Ns'. So divide the places up by areas or even by department.

#this is not good statistics

sum(sickness$Days\_lost\_due\_to\_sickness\_2013) #71298.22

sum(sickness$Days\_lost\_due\_to\_sickness\_2014) #82603.18

mean(sickness$Days\_lost\_due\_to\_sickness\_2013) #5092.73

mean(sickness$Days\_lost\_due\_to\_sickness\_2014) #82603.18