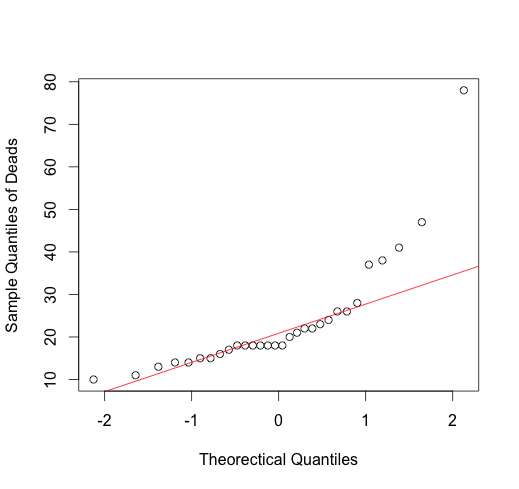
Problem #1:

(a) graph from myqqplot()



R CODE:

Deads <- dat$Dead[dat$CensusNo == 2]

myqqplot <- function(n, data, xlab, ylab){

q <- (1:n - 0.5)/n

qZ <- qnorm(q, mean = 0, sd = 1)

plot(qZ, sort(data), xlab=xlab, ylab=ylab)

q1 <- c(qnorm(0.25, mean=0, sd=1),quantile(data, 0.25))

q3 <- c(qnorm(0.75, mean=0, sd=1),quantile(data, 0.75))

b <- (q3[[2]] - q1[[2]])/(q3[[1]] - q1[[1]]) # the slope

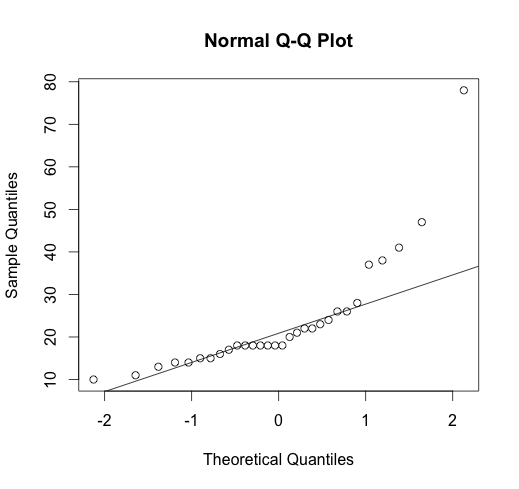
a <- q1[[2]] - b\*q1[[1]] # the intercept

abline(a, b, col="red")

}

myqqplot(n=30, data=Deads, xlab="Theorectical Quantiles", ylab="Sample Quantiles of Deads")

(b) graph from qqnorm() and qqline()

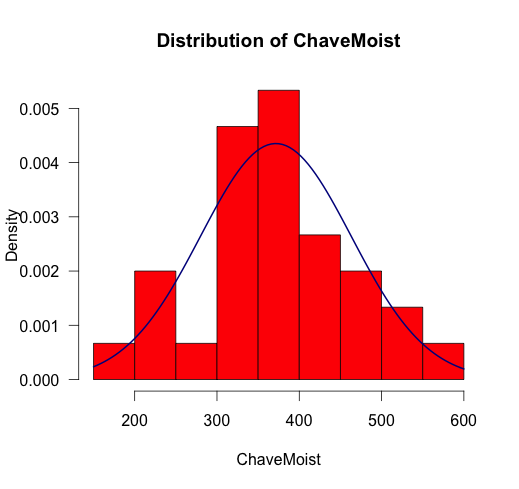


How q-q plot is plotted: First I sort the Deads data from lowest to highest, each point of which naturally stands for the (i/n)th quantile of the data; then I create a vector containing the (i/n)th quantile values, and I intentionally subtract each 0.5 to avoid the “expected” quantile going to infinity when quantile reaches 1; then I compute the “expected” quantile value using qnorm(); finally I scatterplot the “expected” i.e. theoretical quantiles against the sorted Deads data.

Problem #2:

For ChaveMoist in Census 1:

(a) histogram with normal curve



R CODE:

CM <- dat$ChaveMoist[dat$CensusNo == 1]

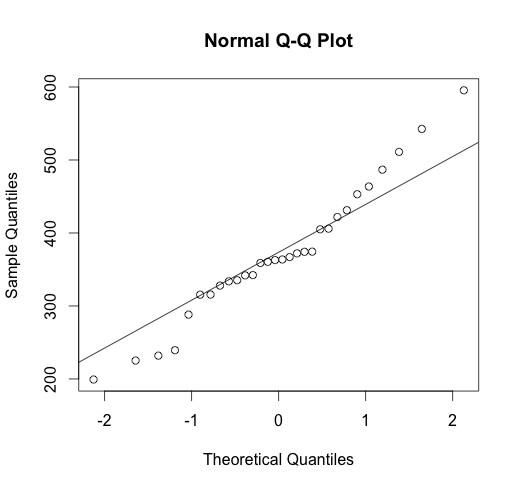
hist(CM, las = 1, main= "Distribution of ChaveMoist", xlab = "ChaveMoist",

col = 2, prob = T)

curve(dnorm(x, mean = mean(CM), sd = sd(CM)),

add=T, col = "darkblue", lwd = 2)

(b) q-q plot



R CODE:

qqnorm(CM)

qqline(CM)

(c) Shapiro test

> shapiro.test(CM)

Shapiro-Wilk normality test

data: CM

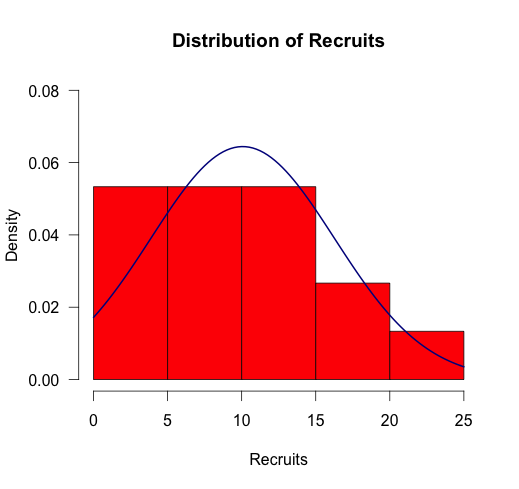
W = 0.9688, p-value = 0.5059

Interpretation:

The p-value signifies there is no evidence to assert that the distribution is NOT normal. And from the qq plot and histogram I think distribution of ChaveMoist is normally distributed.

For Recruits in CensusNo 2:

(a)



R CODE:

# Recruits from CensusNo == 2

rec <- dat$Recruits[dat$CensusNo == 2]

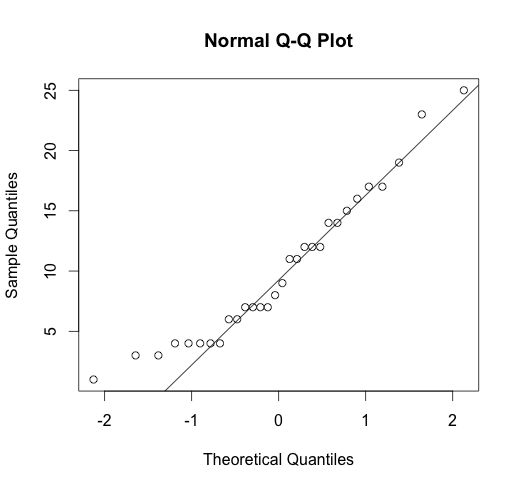
hist(rec, las = 1, main= "Distribution of Recruits", xlab = "Recruits",

col = 2, prob = T, ylim=c(0, 0.08))

curve(dnorm(x, mean = mean(rec), sd = sd(rec)),

add=T, col = "darkblue", lwd = 2)

(b)



(c)

> shapiro.test(rec)

Shapiro-Wilk normality test

data: rec

W = 0.9377, p-value = 0.0787

Interpretation:

From Shapiro test the p value is not significant enough to reject the normal distribution assumption. Q-Q plot somehow leads me to believe that the distribution of Recruits is normal.

Problem #3:

After reducing sample size to 15, I change the conf value to qt(0.975, df=14) based on t distribution, which equals 2.14.

I run it for 10 times, and the approximate proportions of times that the mean is outside of the CI’s are 3%, 5%, 2%, 8%, 2%, 7%, 3%, 8%,7%, 5%.

Then I change the conf back to 1.96, and run again for 10 times. The proportions of times that mean is outside of CI’s are 8%, 7%, 7%, 10%, 4%, 8%, 9%, 4%, 8%.

Conclusion: the proportion of times that the mean is outside of the CI’s increased if I use 1.96. Basically it is because the t statistics 2.14 is larger than z statistics 1.96 thus the confidence interval is larger. I choose to use t statistics since sample size is reduced, t distribution is a more fitting model to describe the data compared to normal distribution.

R CODE:

# Problem #3

mu <- 0

sd <- 1

n <- 50

runs <- 100

conf <- qnorm(0.975)

n2 <- 15

conf2 <- qt(0.975, df=14)

ci <- matrix(nrow=runs, ncol=2)

for (i in 1:runs){

samp <- rnorm(n2, mu, sd)

sx <- mean(samp)

ci[i,] <- c(sx-conf2\*sd(samp)/sqrt(n2), sx+conf2\*sd(samp)/sqrt(n2))

}

par(mfrow=c(1,1))

plot(0,0, xlim=c(-2,2), ylim=c(0,100), type="n",

xlab="CI's of Samples", ylab="Number of Runs", las=1)

abline(v=0, lwd=2)

cnt <- 0

for(i in 1:runs){

clr <- 1

if(ci[i,1]>0)clr=2

if(ci[i,2]<0)clr=2

if(clr == 2)cnt=cnt+1

segments(ci[i,1], i, ci[i,2], i, col=clr)

}

text(-1.75, 95, paste("Count = ", cnt), cex=0.9)

# use old conf of 1.96

mu <- 0

sd <- 1

n <- 50

runs <- 100

conf <- qnorm(0.975)

n2 <- 15

conf2 <- qt(0.975, df=14)

ci <- matrix(nrow=runs, ncol=2)

for (i in 1:runs){

samp <- rnorm(n2, mu, sd)

sx <- mean(samp)

ci[i,] <- c(sx-conf\*sd(samp)/sqrt(n2), sx+conf\*sd(samp)/sqrt(n2))

}

par(mfrow=c(1,1))

plot(0,0, xlim=c(-2,2), ylim=c(0,100), type="n",

xlab="CI's of Samples", ylab="Number of Runs", las=1)

abline(v=0, lwd=2)

cnt <- 0

for(i in 1:runs){

clr <- 1

if(ci[i,1]>0)clr=2

if(ci[i,2]<0)clr=2

if(clr == 2)cnt=cnt+1

segments(ci[i,1], i, ci[i,2], i, col=clr)

}

text(-1.75, 95, paste("Count = ", cnt), cex=0.9)

Problem #4:

The C.I. is (347.2126, 421.2981), which means that 95% of the time, the true mean value will be contained in this range, or I could say for this experiment, the chance that range (347.2126, 421.2981) covers the true mean is 95%.

R CODE:

# Problem #4

cm <- dat$ChaveMoist[dat$CensusNo == 2]

cm.n <- length(cm)

cm.mean <- mean(cm)

cm.se <- sd(cm)/sqrt(cm.n)

Ts <- qt(0.975, df=cm.n-1)

cm.ci <- c(cm.mean-Ts\*cm.se, cm.mean+Ts\*cm.se)