# StatR 101: Fall 2012

Homework 8

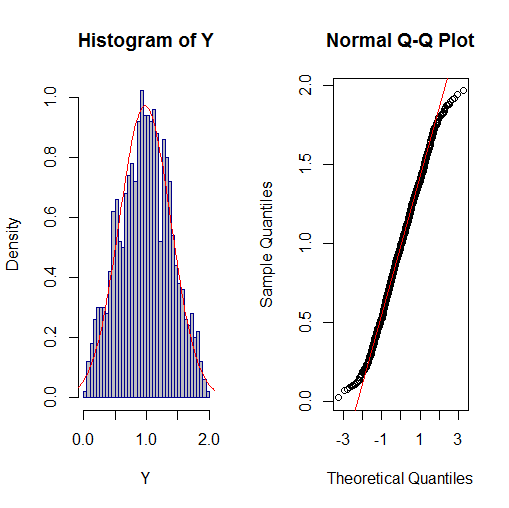
Rod Doe

Wedneday, November 21, 2012

# Central Limit Theorem

## New and Improved CLT Function Superimposes a Normal Curve

### Output



### Input

CLT(runif, n=2, hist.args=list(breaks=30, col="grey", bor="darkblue"))

### Code

CLT <- function(FUN, n, k=1000, ..., hist.args = NULL)

{

Y <- rep(0, k)

for(i in 1:n) { Y <- Y + FUN(k, ...) }

par(mfrow=c(1,2))

# Make hist indicate density, not counts.

do.call(hist, c(list(x=quote(Y)), prob=TRUE, hist.args))

# Generate data with which to plot a normal distribution around the sampled data.

nx = seq(from = mean(Y) - 3\*sd(Y), to = mean(Y) + 3 \* sd(Y), length.out=100)

ny = dnorm(nx, mean(Y), sd(Y))

# Plot a normal curve on top of the data.

lines(nx, ny, col=2)

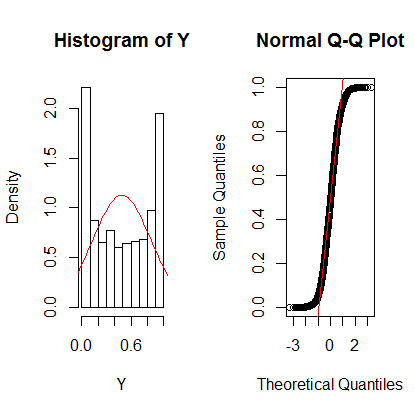
qqnorm(Y); qqline(Y, col=2)

}

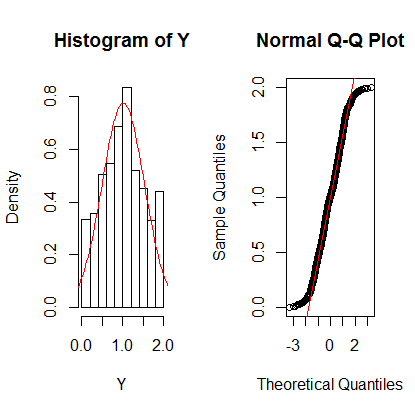
## Exploring the Central Limit Theorem with the Beta Distribution (alpha,beta) = (0.5, 0.5)

With the two shape values at 0.5, the central limit theorem kicks in at around n=10. The following four graphs illustrate this.

### CLT(rbeta, n=1, shape1=0.5, shape2=0.5)



### CLT(rbeta, n=2, shape1=0.5, shape2=0.5)



### CLT(rbeta, n=10, shape1=0.5, shape2=0.5)

### 

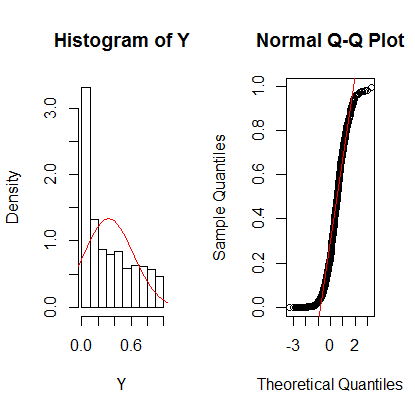
### CLT(rbeta, n=20, shape1=0.5, shape2=0.5)

## CLT(rbeta, n=1, shape1=0.5, shape2=1.0)Beta

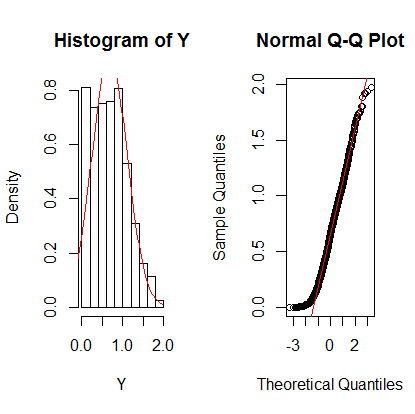
## Beta Distribution with (alpha, beta) = (0.5, 1.0)

This distribution is skewed to the, and as less compliant than its symmetric brother. It was quite normal with n = 20.

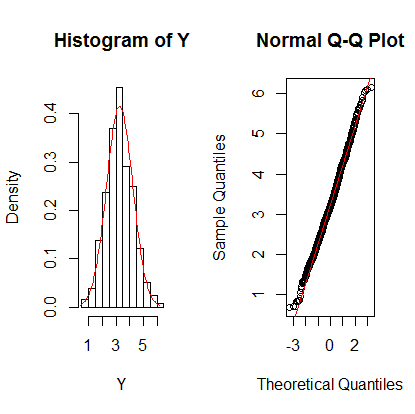
### CLT(rbeta, n=1, shape1=0.5, shape2=1.0)



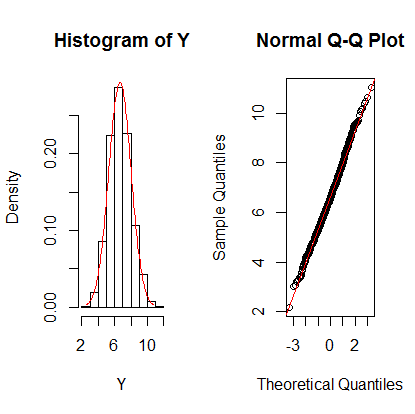
### CLT(rbeta, n=2, shape1=0.5, shape2=1.0)



### CLT(rbeta, n=10, shape1=0.5, shape2=1.0)



### CLT(rbeta, n=20, shape1=0.5, shape2=1.0)



# Sampling Distribution

## Tweak the SampleMean() function

This was a simple tweak that borrowed heavily from the previous simple tweak.

SampleMean <- function(X, n=30, k=10000, ...)

{

X.bar <- c()

for(i in 1:k) { X.bar <- c(X.bar, mean(sample(X, n))) }

hist(X.bar, prob=TRUE, ...)

nx = seq(from = mean(X) - 3\*sd(X), to = mean(X) + 3 \* sd(X), length.out=100)

ny = dnorm(nx, mean(X), sd(X))

# Plot a normal curve on top of the data.

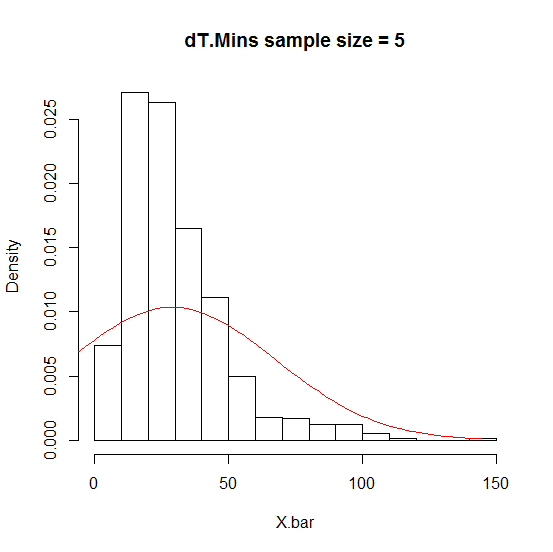
lines(nx, ny, col=2)

}

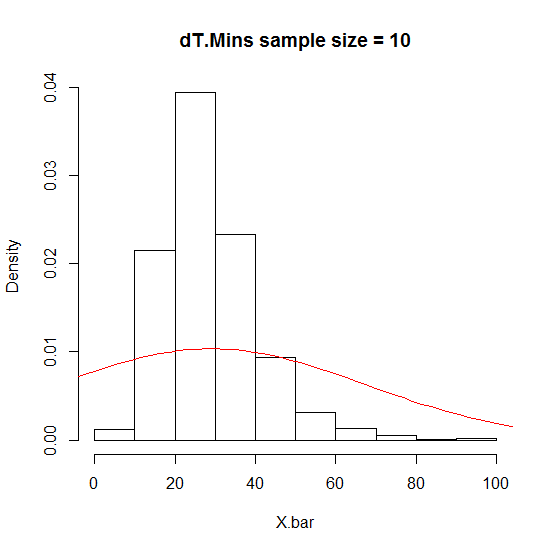
We have one small problem here – since we are plotting time consecutive differences, barring a warp in the time-space continuum, there will be no negative consecutive time differences. However, the normal approximation will lose much of its left side mass as it gets truncated at zero.

The following graphs show that the normal approximation is more valid for smaller sample sizes. For sample size of 100, the normal approximation is essentially a straight line.

### SampleMean(dT.mins, n=5, k=1000, main="dT.Mins sample size = 5")



### SampleMean(dT.mins, n=10, k=1000, main="dT.Mins sample size = 10")



### SampleMean(dT.mins, n=30, k=1000, main="dT.Mins sample size = 30")

### SampleMean(dT.mins, n=100, k=1000, main="dT.Mins sample size = 100")

# 

## Recommended bonus problems:

I will try these this week. I am being urged to clean the house before guests come over. ☹

# Apocalyptic Inference

Seen on bumper sticker: “Palin 2012 – how did the Mayans know?”

## Probability of 30 earthquakes with a mean interval of 15 minutes or less

We want the area under the curve to the left of the vertical line in the following graph.

eq <- read.csv("http://neic.usgs.gov/neis/gis/qed.asc")

DateTime <- strptime(paste(eq$Date, eq$Time), format="%Y/%m/%d %H:%M:%S")

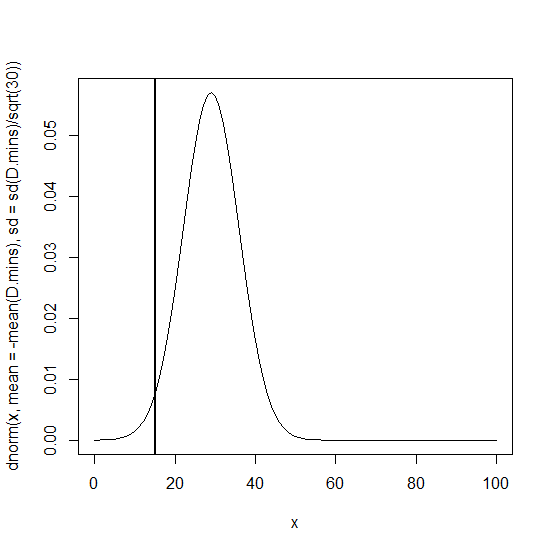
D.mins <- difftime(DateTime[-1], DateTime[-length(DateTime)], units="mins")

curve(dnorm(x, mean =-mean(D.mins), sd=sd(D.mins)/sqrt(30)), xlim=c(0,100))

abline(v=15, lwd=2)

> pnorm(15, -mean(D.mins), sd(D.mins)/sqrt(30))

[1] 0.02326953 # Slightly more than 2 percent probability of 30 earthquakes

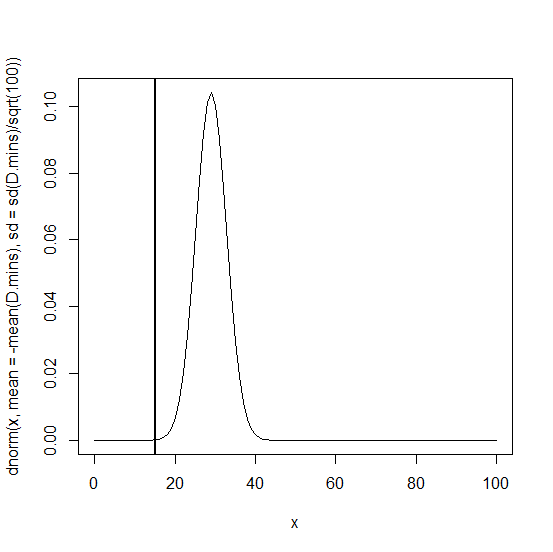


## Probability of 100 earthquakes with a mean interval of 15 minutes or less

This is the area under the curve to the left of the vertical line in the graph below.

> pnorm(15, -mean(D.mins), sd(D.mins)/sqrt(100))

[1] 0.0001394836 # Probability = 0.014 percent.



## When to panic?

The code below indicates that it is statistically appropriate to panic when the 21st earthquake occurs with an average rate of every 15 minutes of less.

> peq = numeric(30)

> for (i in 1:30) { peq[i] = pnorm(15, -mean(D.mins), sd(D.mins)/sqrt(i)) }

> peq

[1] 0.35814982 0.30364786 0.26452978 0.23366894 0.20822187 0.18668844

[7] 0.16815320 0.15200434 0.13780681 0.12523695 0.11404566 0.10403603

[13] 0.09504912 0.08695442 0.07964319 0.07302382 0.06701828 0.06155960

[19] 0.05658985 0.05205860 0.04792171 0.04414034 0.04068020 0.03751086

[25] 0.03460524 0.03193919 0.02949105 0.02724142 0.02517282 0.02326953