Normal.distribution.chp3

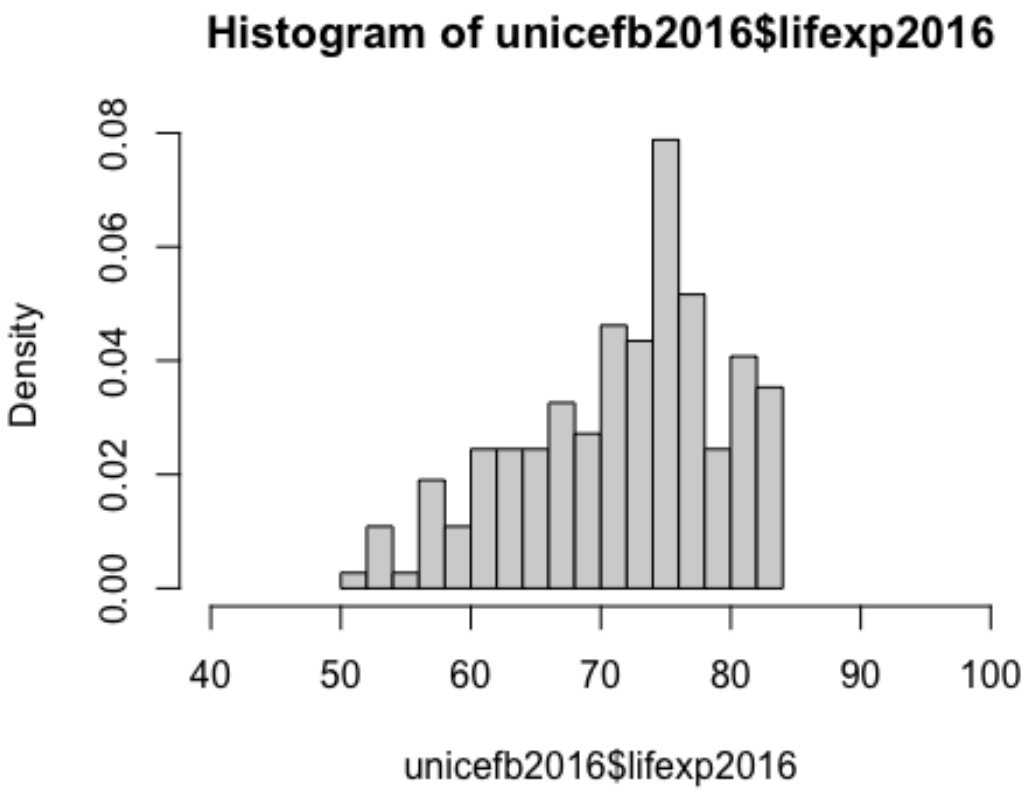
Astudent

9/6/2023

```
load("~/Desktop/Rnotes/unicefbas2016.rda")
```

Going from distribution of data to fitting idealized model with density curve to the data.

```
hist(unicefb2016$lifexp2016,xlim=c(40,100),freq=F,breaks=20)
```




```
#Note the change to freq =F. This is so that the total area inside the boxes is 1.
```

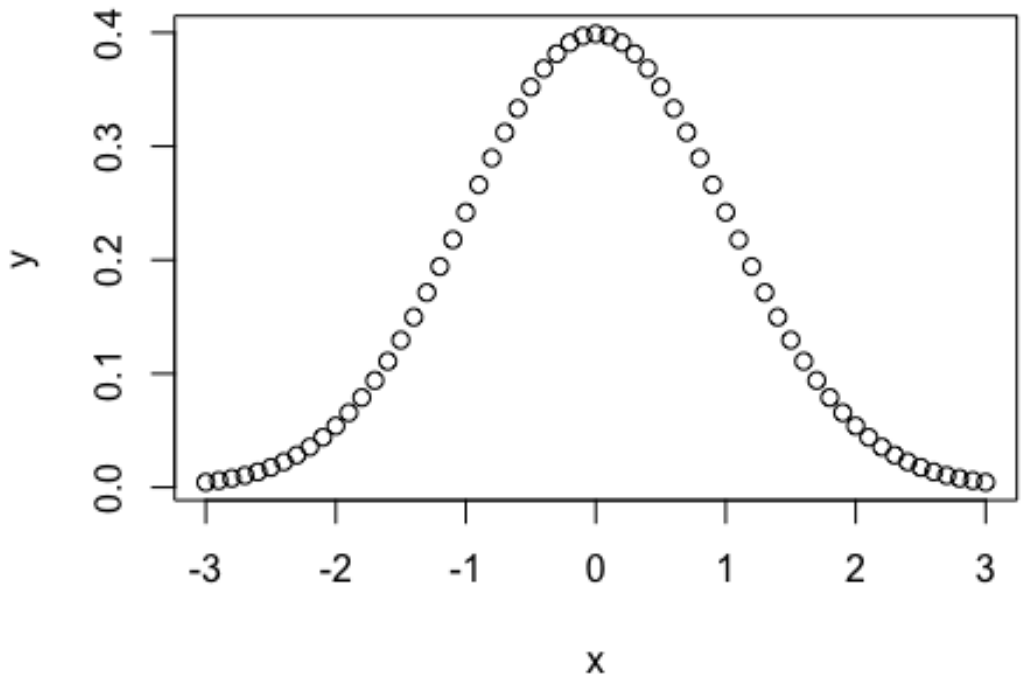
Looks sort of normal except for the chop off at the right.(Why is that chop off there?) Try to fit a density curve.

A density curve is always nonnegative, has area 1 under it. Later see this has to do with probability. Density curves come in all kinds of shapes. (some pictures). Can talk about their

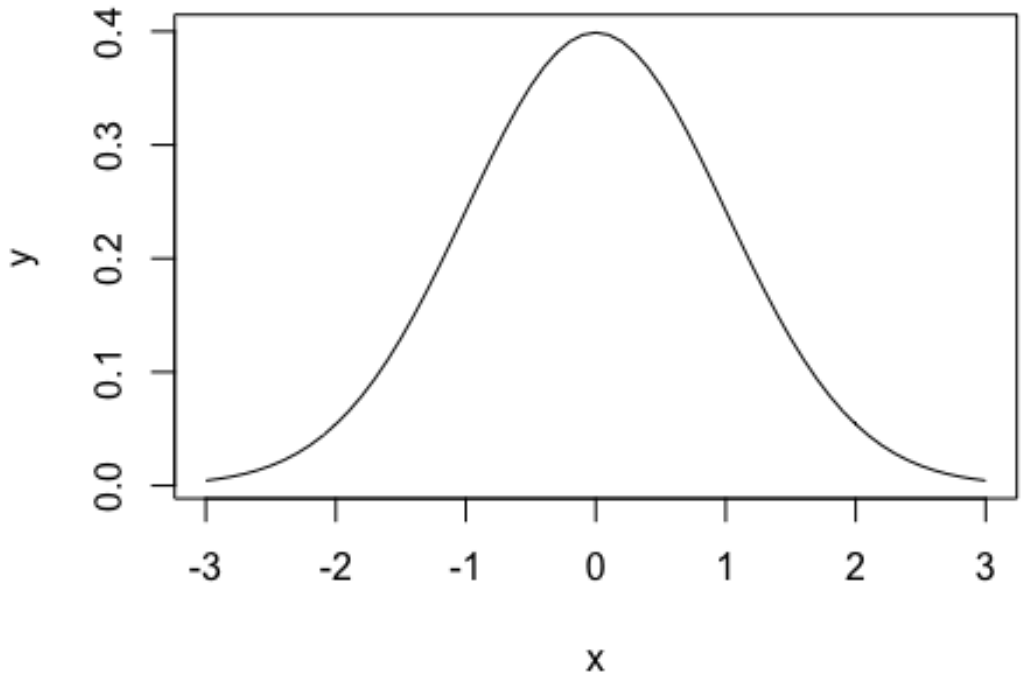
median (equal areas) and mean (balance point). Use $\bar{x}$ for data and μ for density curve. Also σ for the standard deviation instead of s.

The normal curves:

```
#example:
x=seq(-3,3,by = 0.1) # numbers from -3 to 3 in increments of 0.1
 y=dnorm(x)
plot(x,y)  # get circles
```

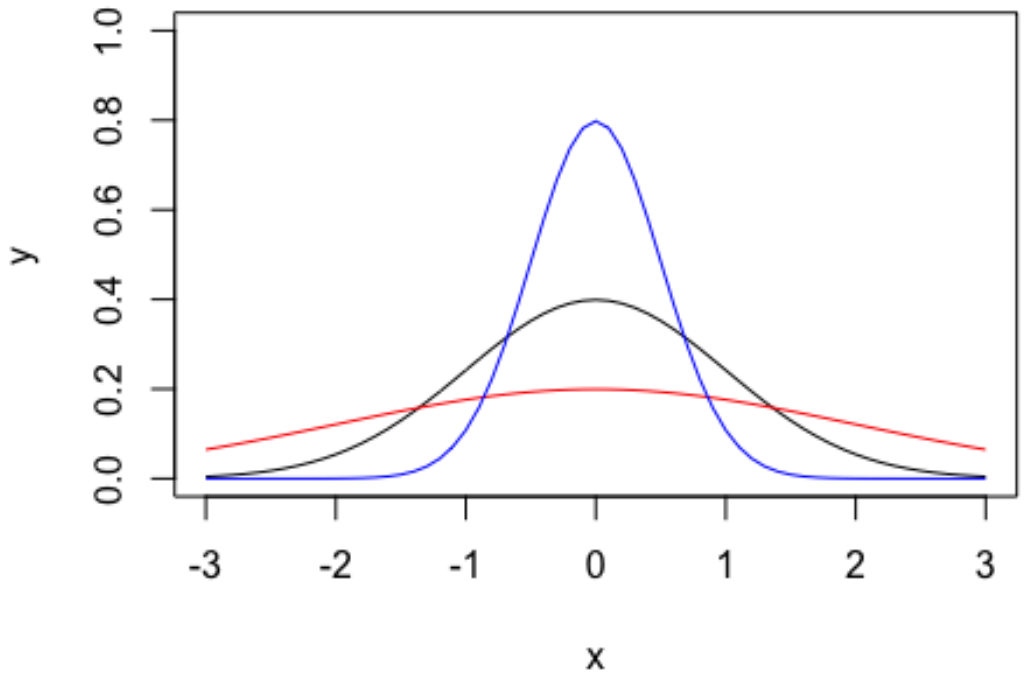



```
plot(x,y,'l') #get line segments
```




```r
#This is the standard normal distribution

plot(x,y,'l',ylim=c(0,1))
y2=dnorm(x,0,2)
y3=dnorm(x,0,0.5)
lines(x,y2,col='red') # to add anothe plot to the first graph.
lines(x,y3,col='blue') #try this without the ylim
```



$N(\mu, \sigma)$ means

It is (student answers) about μ It has standard deviation σ

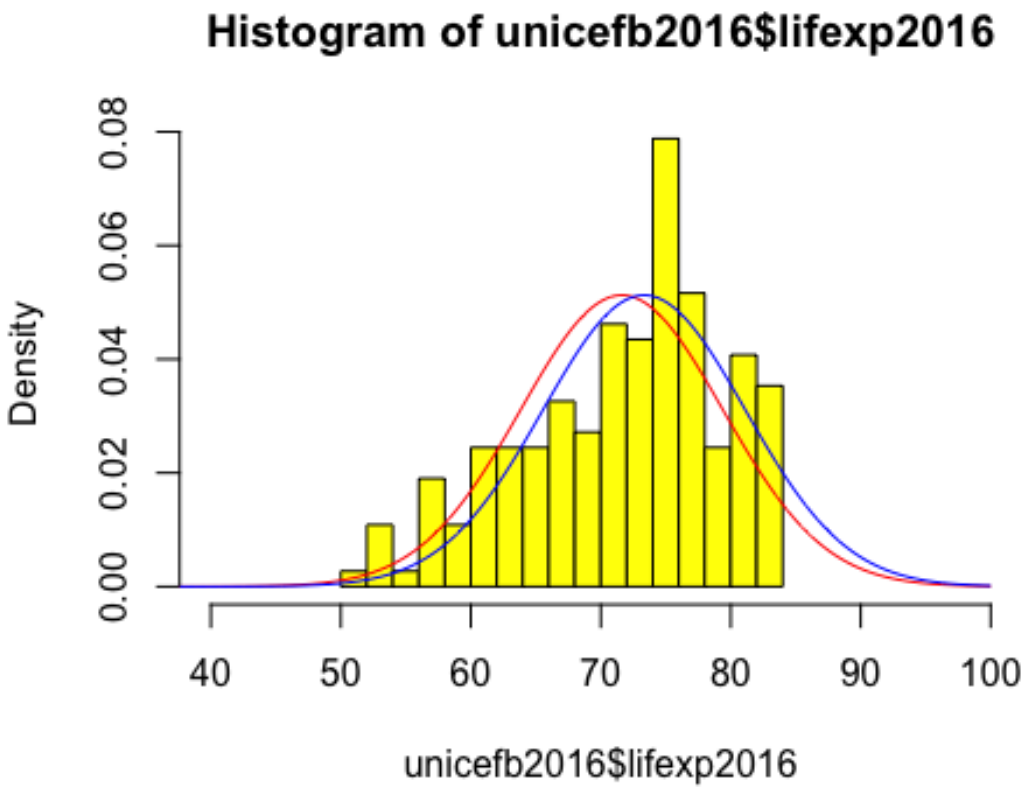
It has (student answers) peaks

One standard deviation away from the center are the (i-) points of the curve.

The function for the curve of $N(\mu, \sigma)$ is (write on board). We won't use that is this class.

Try fitting a normal curve to our above distribution

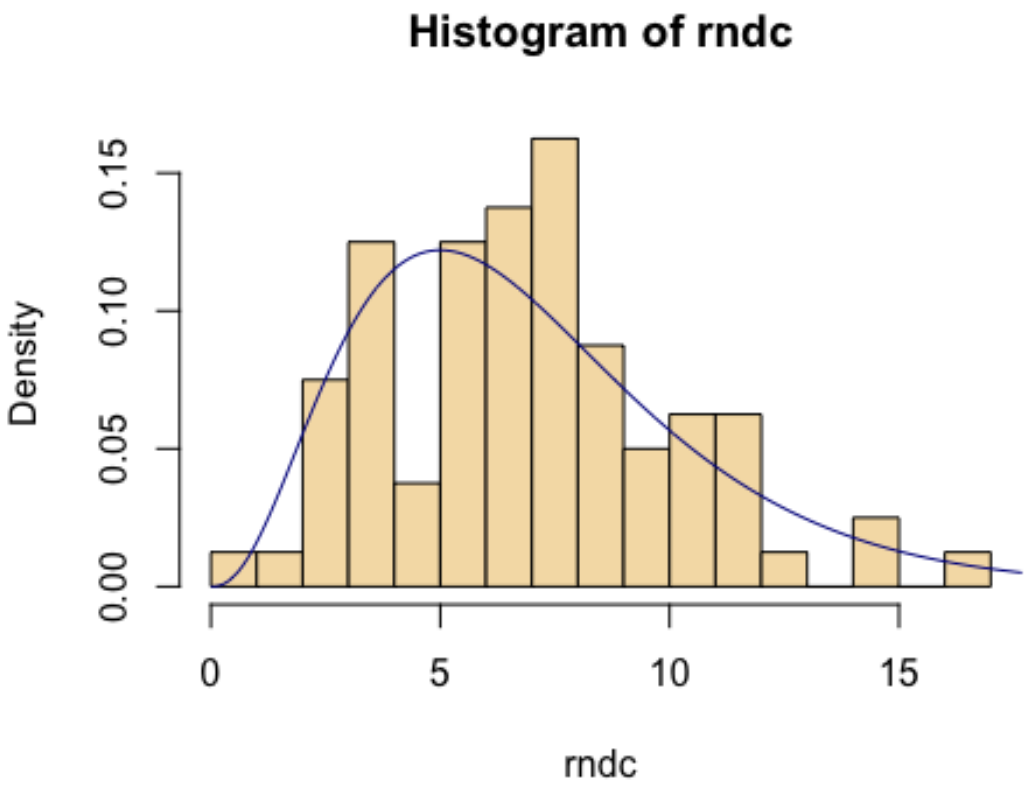
```
hist(unicefb2016$lifexp2016,xlim=c(40,100),freq=F,col='yellow',breaks=20)
m=mean(unicefb2016$lifexp2016,na.rm=TRUE)
s=sd(unicefb2016$lifexp2016,na.rm=TRUE)
md=median(unicefb2016$lifexp2016,na.rm=TRUE)
s=sd(unicefb2016$lifexp2016,na.rm=TRUE)
x=seq(20,100,0.1)
y=dnorm(x,m,s)
yd=dnorm(x,md,s)
lines(x,y,col='red')
lines(x,yd,col='blue')
```



Which curve fits better?

Another example of fitting a density curve to a histogram

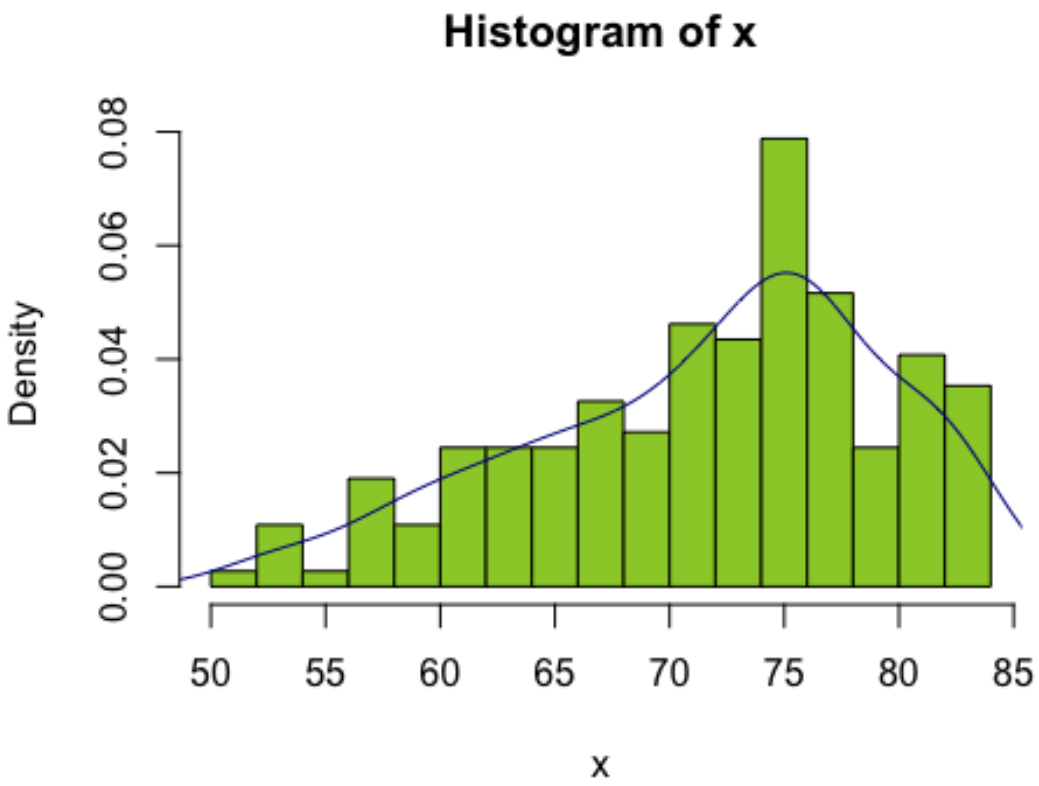
```
rndc=rchisq(80, df=7)
x=seq(0,20,by = 0.1)
 y=dchisq(x,df=7)
hist(rndc,freq=FALSE,col='wheat', breaks=20)
lines(x,y,col='darkblue') #plots the chi square distribution  over the
histogram. We won't study this distribution until later.
```



There is a built in function in R to fit a density curve to data.This graph below is just to show you we can get a better fitting density curve than a normal curve. We stick with the normal distribution, when appropriate, because, as we will see, a normal density has features which make it is easier to deal with theoretically.

```
x=unicefb2016$lifexp2016[!is.na(unicefb2016$lifexp2016)]
hist(x,breaks=20, col='yellowgreen',freq=F)

lines(density(x),col='darkblue')
```




```
str(density(x)) #height at point x
```

```
## List of 7
##  $ x        : num [1:512] 44.4 44.5 44.6 44.7 44.8 ...
##  $ y        : num [1:512] 2.13e-05 2.40e-05 2.69e-05 3.01e-05 3.38e-05 ...
##  $ bw       : num 2.47
##  $ n        : int 184
##  $ call     : language density.default(x = x)
##  $ data.name: chr "x"
##  $ has.na   : logi FALSE
##  - attr(*, "class")= chr "density"
```

```
sum(density(x)$y)*0.1 #The area under the density curve is supposed to be 1.
```

```
## [1] 1.094244
```

Rule of thumb (one good feature of the normal distribution) For the normal curve roughly 6_% (student answers) of the area under the curve is between and
The middle 9_% (student answers) is roughly between

For life expectancy of countries: Using the normal distribution with mean

```
round(mean(unicefb2016$lifexp2016,na.rm=TRUE))
```

```
## [1] 72
```

and standard deviation

```
round(sd(unicefb2016$lifexp2016,na.rm=TRUE))
```

```
## [1] 8
```

to approximate countries' life expectancy distribution.

a.Use this and the 68-95-99.7 rule of thumb to determine where the middle 68% is.

b.What percent is above 64 years?

c.What percent is below 48?

d. Where is the upper 97.5%?

e. Another example for this.

f. How many standard deviations is 85 above the mean? (z - score)?

Let's compare b. with the actual proportion of countries with life expectancy at or above 64.

```
tot=sum(!is.na(unicefb2016$lifexp2016)) #total number of non missing data
points for lifexp2016
numbelow64=sum(unicefb2016$lifexp2016[!is.na(unicefb2016$lifexp2016)]>=64)
numbelow64/tot
```

```
## [1] 0.8097826
```

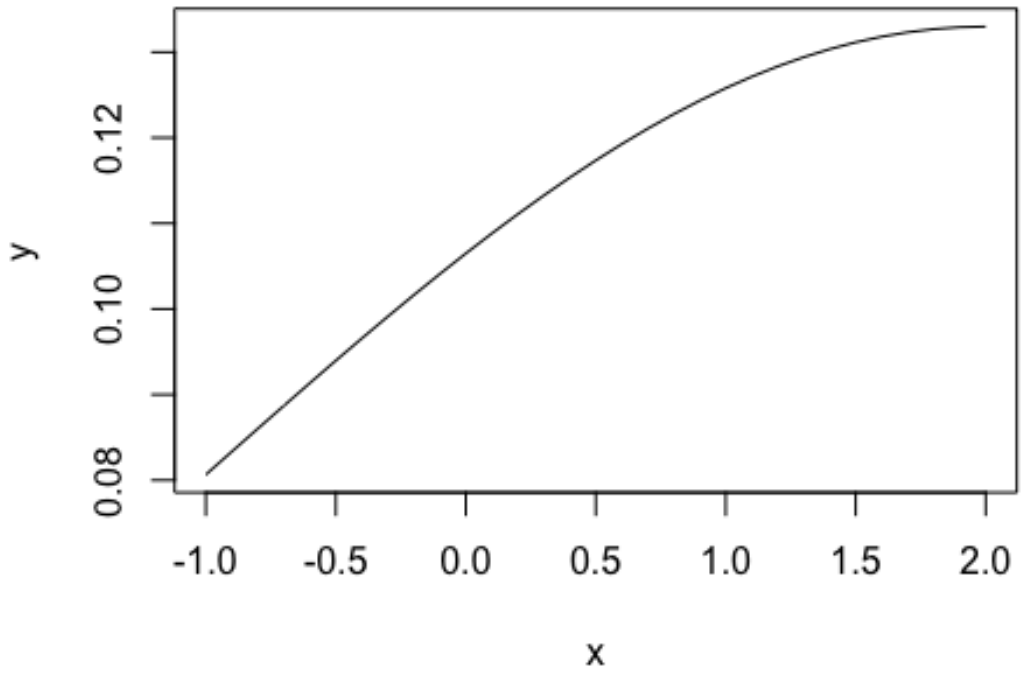
Off by how much?

```
#1) dnorm(a,mean=m, sd= s) gives the value of the probability density
function at a for the #normal distribution with mean m and standard deviation
s. Default for mean is   and  sd is  .
#We often use this for graphing the normal distribution, as in some of the
above code.

dnorm(0)
```

```
## [1] 0.3989423
```

```
x=seq(-1,2,by=0.1)
y=dnorm(x,2,3)
plot(x,y,'l')
```




```
#What is the approximate area under this curve?
```

Convert the distribution to standard normal. If x is a value from the $N(\mu, \sigma)$ distribution $z = (x - \mu)/\sigma$ is the z-score for x or the standardization of x.

Using the normal distribution: Example-a country that has life expectancy 75 is how many standard deviations above the mean? Draw a picture.

Which of the following are true.Why?

More than 50% of countries have life expectancy greater than that country?

More than 16% of countries have life expectancy greater than that country?

Less than 2.5% of countries have life expectancy greater than that country? Draw a picture.

Use the standard z-score table to compute what percent of countries have life expectancy less than 75. Note that we cannot use the 68-95-99.7 rule here. Why?

```
# 2)
#pnorm(a,mean=m, sd= s) #gives the  proportion of X<=a, or area under the
curve to the left of a for the normal distribution with mean m #and standard
deviation s.
pnorm(-1)
```

```
## [1] 0.1586553
```

```
    # (proportion below one standard deviation to left of mean is roughly .16
by the 68% rule )
 pnorm(-1, lower.tail = FALSE)    # gets the upper tail
```

```
## [1] 0.8413447
```

```
pnorm(40, 30,10)-pnorm(25, 30,10)  # Sketch what this represents.
```

```
## [1] 0.5328072
```

Do some quick examples from book using pnorm.

```
# 3)
qnorm(.16) # answers the question for what x is the area under the  standard
normal curve to the left of x equal to .16. What should this be -roughly?
```

```
## [1] -0.9944579
```

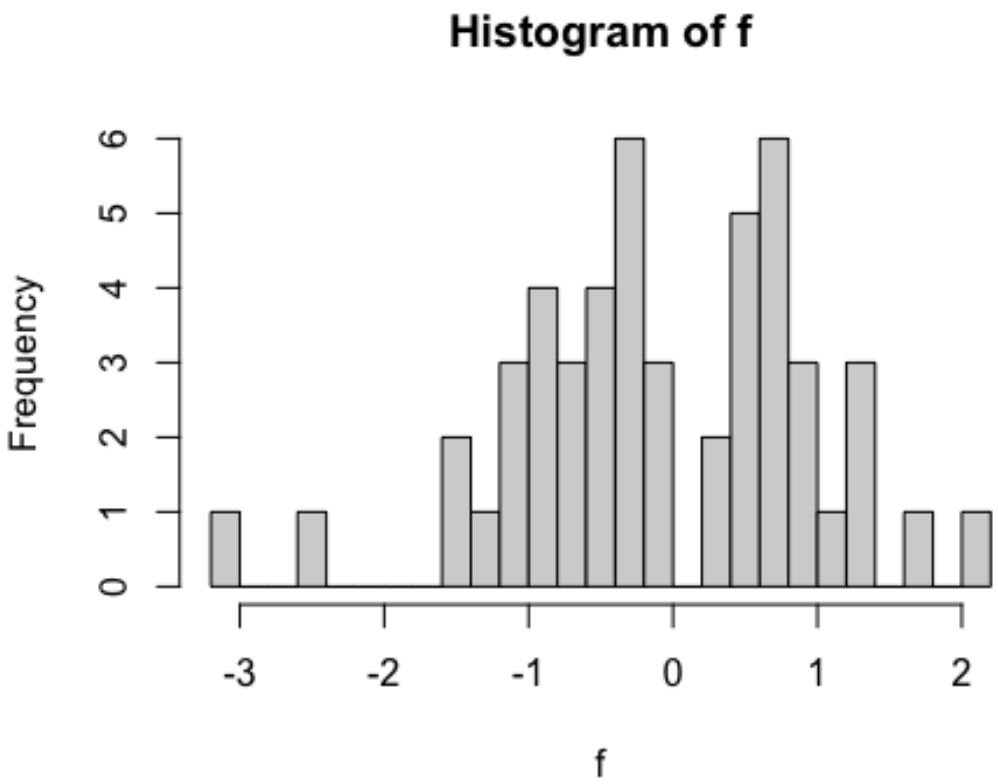
```
qnorm(.16, lower.tail = F)  #What should this be -roughly?
```

```
## [1] 0.9944579
```

Do some problems from the book using qnorm.

For the normal N(72,8) distribution used for life expectancy, where is the upper 75%? What other number does this represent? Where is the lower 50%?

```
#rnorm r for random
f= rnorm(50)  # generates 50 random numbers from the standard normal
distribution
# What that means is since there are more values in the middle we will
usually get more values in the middle. For example, roughly what percent of
our values will be between -1 and 1?
hist(f, breaks=20) #change the number to 200.To 1000. What do you observe?
```

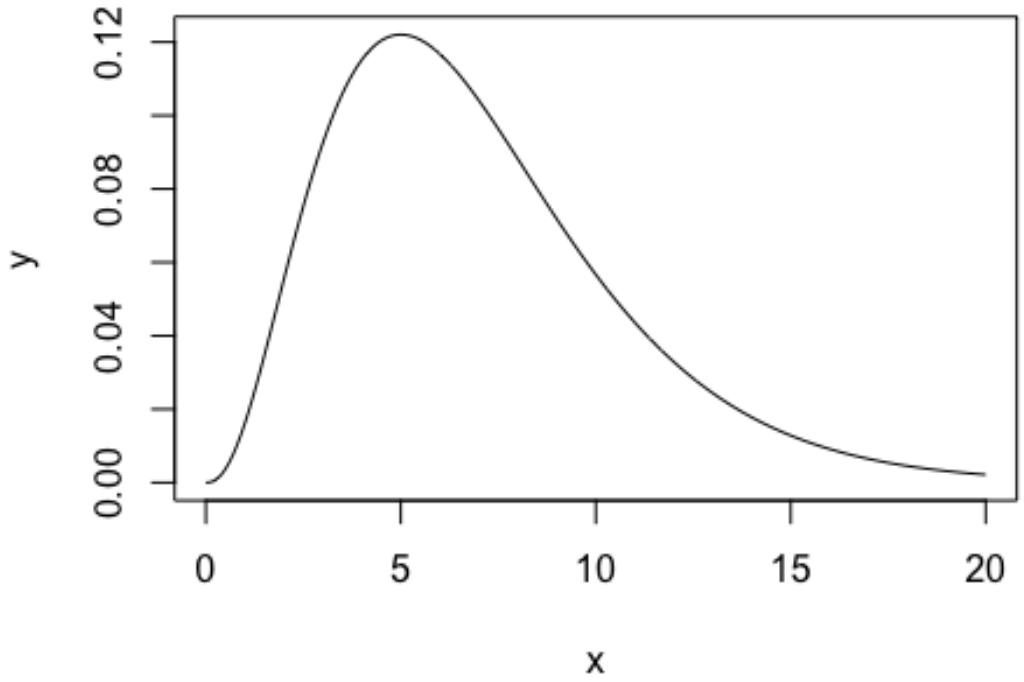


Suppose the distribution of the weight of cheetahs is N(105,25) lbs. Where is the highest 25% ? What percent is below 100? What percent is between 90 and 115? How many standard deviations is 90 below the mean?

The weight of male cheetahs is N(110,25) of females is N(100,20) Which is relatively larger- a female cheetah that weighs 98 lbs or a male that weighs 100 lbs. ?

One can recreate the above p,d,q, r with other types of distributions. Of course there might be restrictions on the values for x. Examples would be chi squared, exponential, F distribution, binomial (which we will study later). A few examples with chi squared are below.

```
x=seq(0,20,by = 0.1)
 y=dchisq(x,df=7)
plot(x,y,type='l')
```




```r
pchisq(5,df=7)
```

```
## [1] 0.3400368
```

```r
qchisq(0.5,df=7)
```

```
## [1] 6.345811
```

```r
#median is actually
7*(1-2/(9*7))^3
```

```
## [1] 6.354273
```