PairProgramming_Module_04_Distributions

HD Sheets

2024-08-26

Pair Programming, DSE5001, Module 4

Distributions

HD Sheets, Aug 26, 2024 checked 01/03/2025

Student Name:Ryan Waterman Partner Name: Date Completed:2/5/25

Looking at questions from Open Intro text by Diaz et al.

In looking at problems like this, I seek to create a *solution*, which is really a complete description of the problem that I can use to answer any question about the problem. Develop your understanding first, the *solution*, then answer the question

#Question 4.1

4.1 Area under the curve, Part I. What percent of a standard normal distribution N(mu = 0; sigma = 1) is found in each region? Be sure to draw a graph. (a) Z < -1.35 (b) Z > 1.48 (c) -0.4 < Z < 1.5 (d) |Z| > 2

Discussion of the problem by HDS

This question is asking about a normal distribution, with mu=mean=0, standard deviation= sigma=1

This is the “standardized” form of a normal distribution,

where Z= (x=xmean)/(stdx)

In R, we have a family of functions for working with the normal distribution

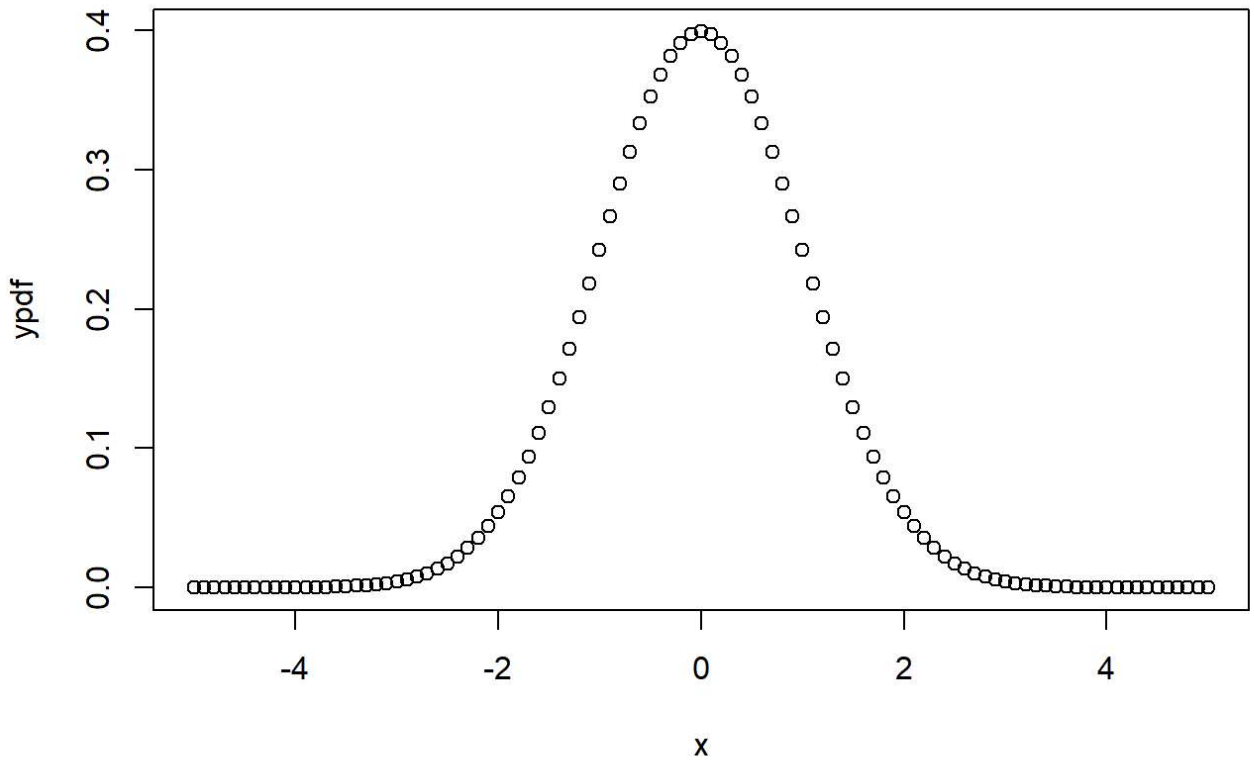
```
rnorm, pnorm, dnorm and qnorm
```

so I’ll be using these functions

Graphs- The problem does ask for graphs of the distribution, probably wise to create them.

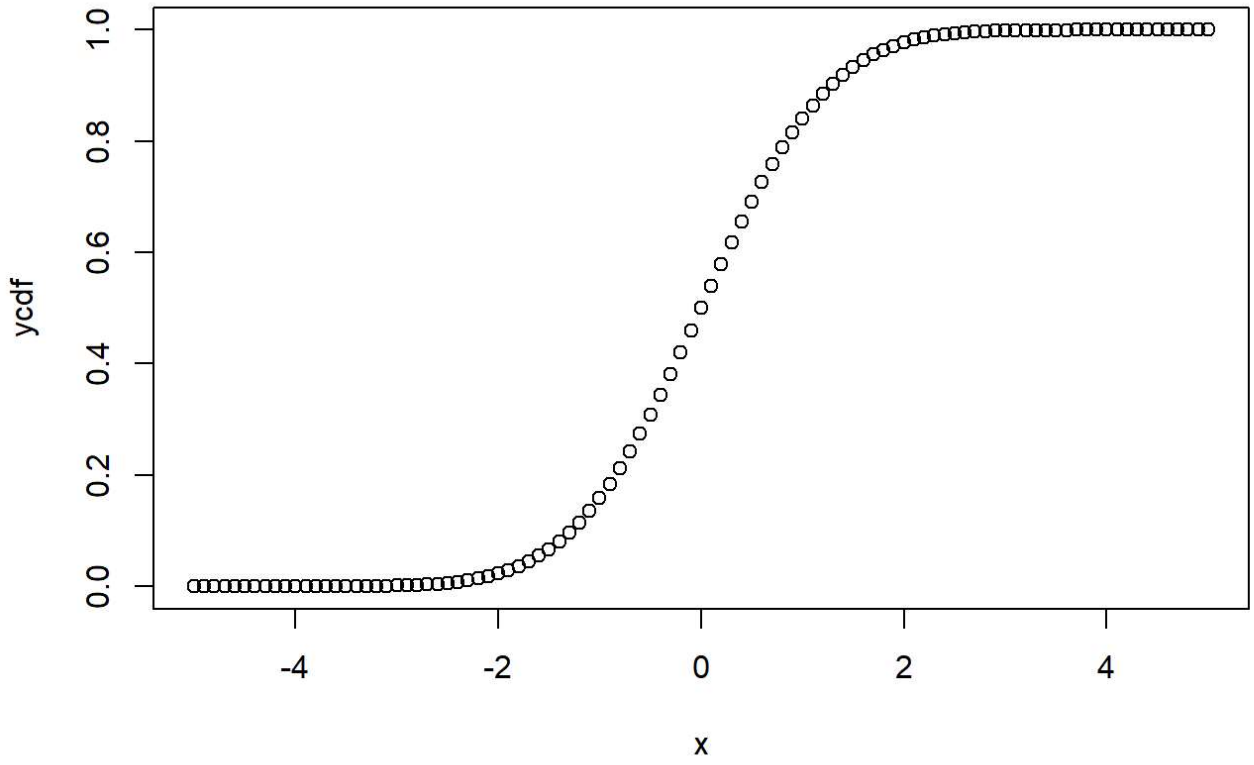
Let’s create graphs of the pdf and cdf, from x=-5 to 5 say

```
#set up a list of values x to use in the graph, from -5 to 5 in steps of 5

x=seq(-5,5,0.1)

# get the pdf at each x, using the function dnorm
# we ae calculating the p(x) value at each value of x, for a gaussian of mean 0 and std dev =1

ypdf=dnorm(x,mean=0,sd=1)

#create the plot

plot(x,ypdf)
```



Here is the cdf

```
# use pnorm to calculate the cdf and then plot it

# pnorm gives us p(z<x) for each value in x,   ie the cummulative distribution at mean=0, std dev
=1

ycdf=pnorm(x, mean=0, sd=1)

plot(x, ycdf)
```



Okay, so that is the distribution we are working with

Let's look at the first question, a

Area under the curve, Part I. What percent of a standard normal distribution N(mu = 0; sigma = 1) is found in each region? Be sure to draw a graph. (a) Z < -1.35

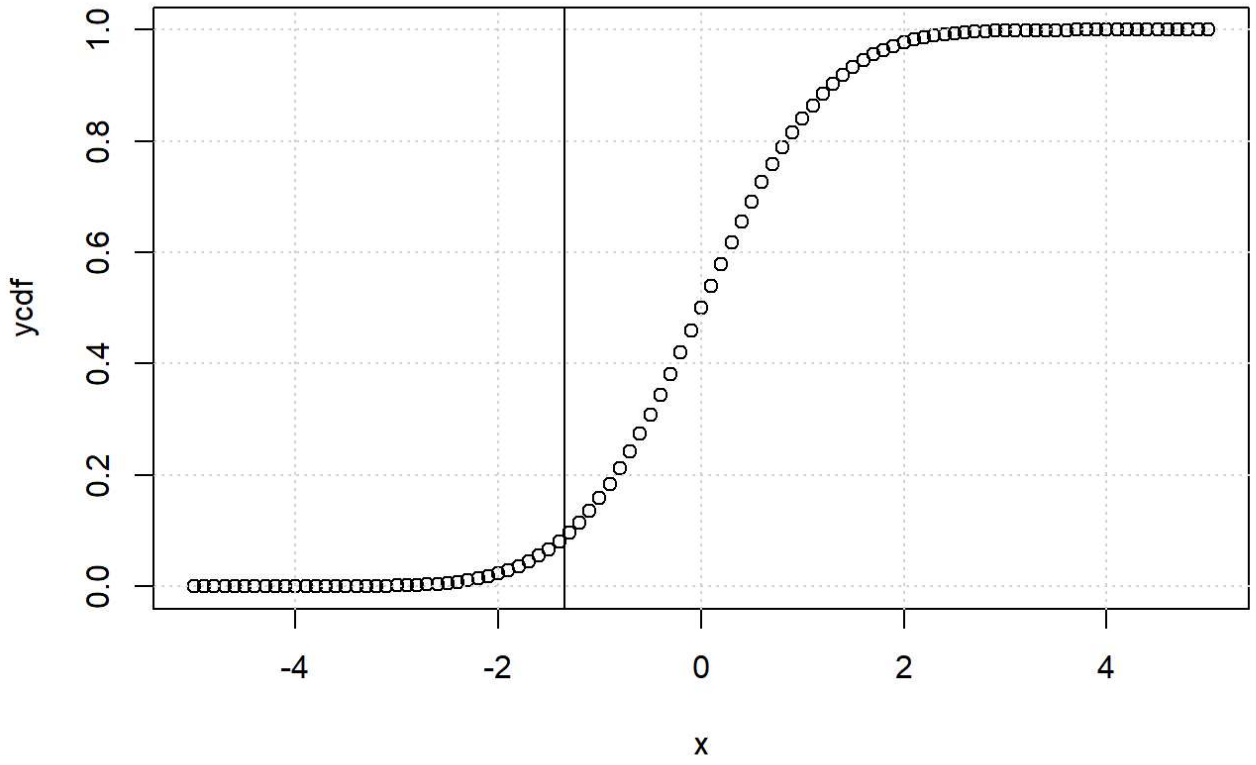
*The area under the curve means the area under the pdf from - infinity up to -1.35

The cdf is just the area under the pdf from -infinity to x

so the cdf value at x=-1.35 is P(x<-1.35)

We can look at the graph to see this, we will add a line at x= -1.35

```
#plot the cdf, using the data we created earlier

plot(x,ycdf)

# add a vertical line at x=-1.35,

abline(v=-1.35)

grid()
```



We could try to read the value off the graph, it looks like the ycdf value is at about 0.1, meaning that 0.1 or 10% of the area under a normal distribution with mean 0 and std dev 1 is below -1.35

We can get the exact value from the cdf function pnorm

```
pnorm(-1.35, mean=0, sd=1)
```

```
## [1] 0.08850799
```

The exact value for p(x<-.135) is 0.08850799 (about 8.9%)

Use the graph to get a general sense of what is happening, use the function to get an exact value

Next question, 4.1b

4.1 Area under the curve, Part I. What percent of a standard normal distribution N(mu = 0; sigma = 1) is found in each region? Be sure to draw a graph.

b. Z > 1.48

Okay, we want p(x>1.48)

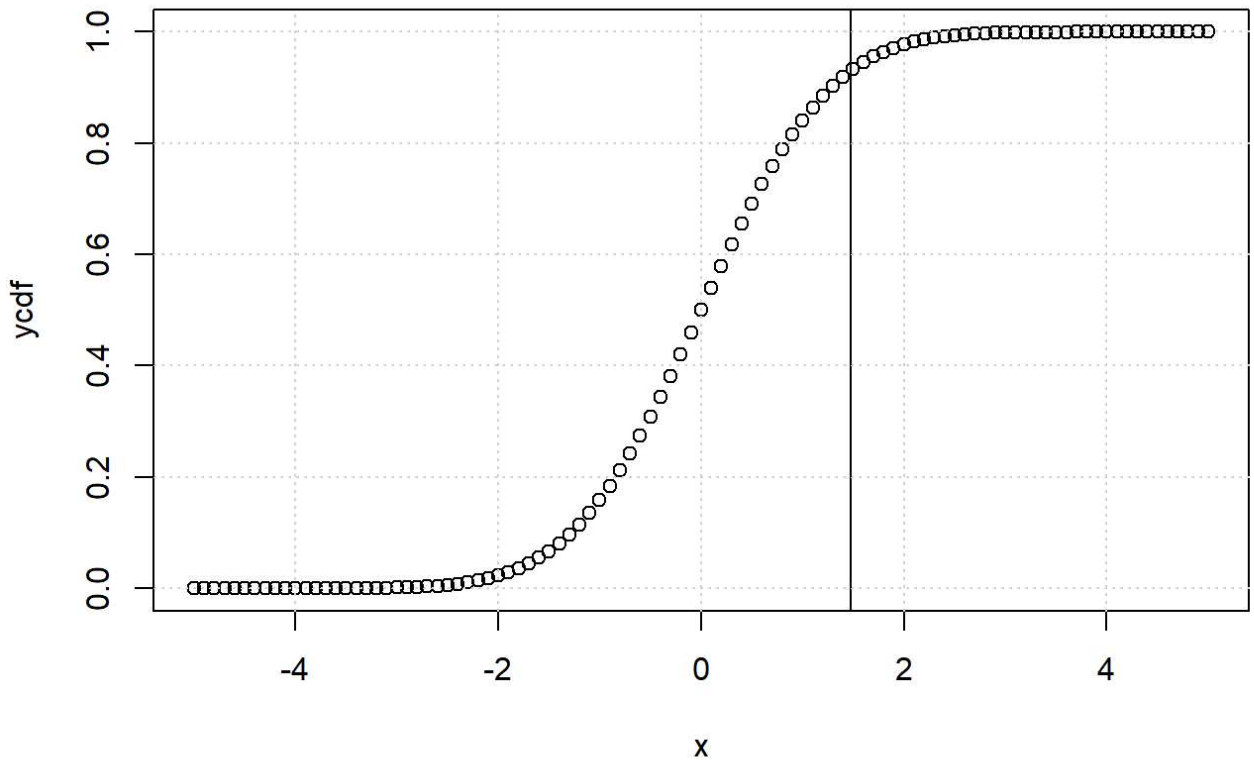
the cdf can give us p(x<1.48), as we saw above

But we know p(x<1.48) plus p(x>1.48) must add to 1, so

```
p(x<1.48) + p(x>1.48) = 1

p(x>1.48) = 1 - p(x<1.48)
```

Here's a quick visual depiction

```
#plot the cdf, using the data we created earlier

plot(x,ycdf)

# add a vertical line at x=-1.35,

abline(v=1.48)

grid()
```



Looking at the plot, it looks like p(x<1.48) is around 0.94 (94%) or so, though it's hard to be exact

so p(x>1.48) = 1-0.94 = 0.06, or about 6%

To get the exact value, we can use pnorm again

```
1-pnorm(1.48,mean=0,sd=1)
```

```
## [1] 0.06943662
```

and the correct value is 0.06943663, or about 6.94%

#Question 4.1 c

4.1 Area under the curve, Part I. What percent of a standard normal distribution N(mu = 0; sigma = 1) is found in each region? Be sure to draw a graph.

c. -0.4 < Z < 1.5

In this question we would the region bounded by -0.4 and 1.5

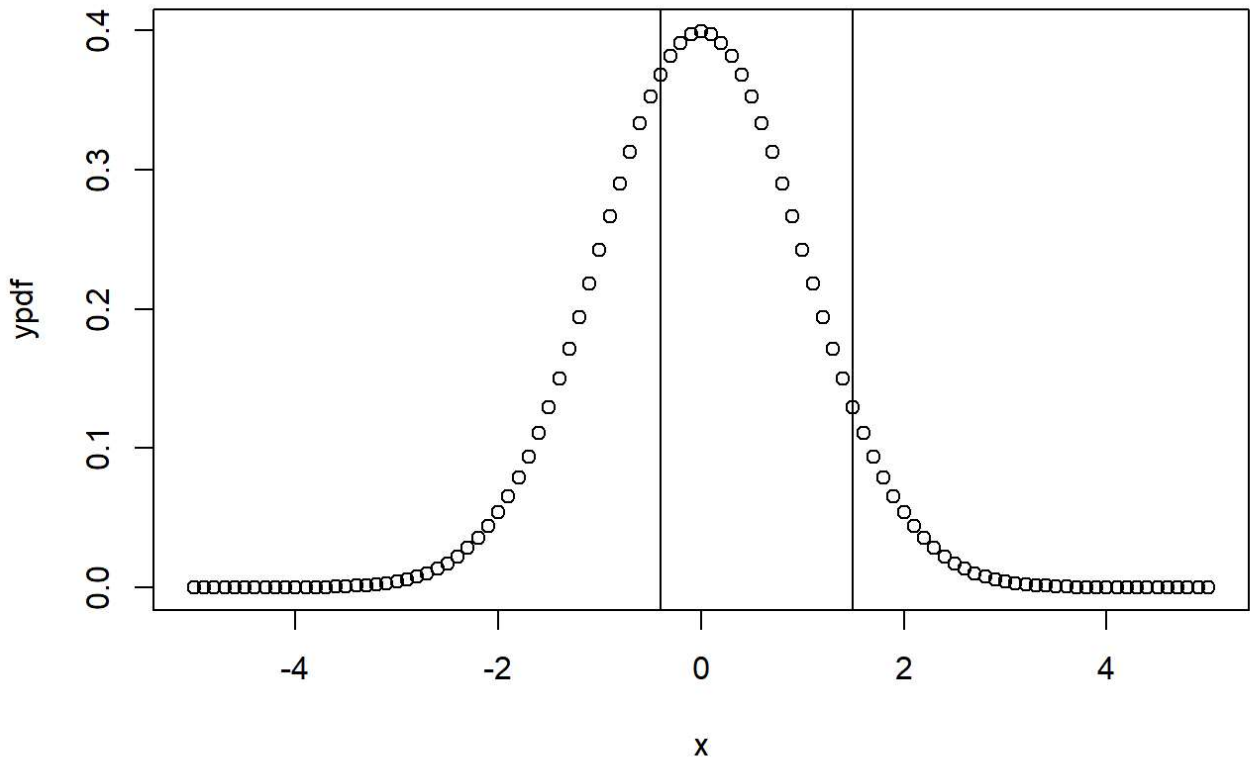
The region below 1.5 would be pnorm(1.5,mean=0, sd=1)

if we took this value and subtracted the region below -0.4, that would give us the region between the two boundaries, calculated as

pnorm(1.5, mean =0, sd=1)-pnorm(-0.4, mean=0, sd=1)

Visually, we want the area under the curve between the lines on this plot

```
plot(x,ypdf)
abline(v=-0.4)
abline(v=1.5)
```




```
pnorm(1.5,0,1)-pnorm(-0.4,0,1)
```

```
## [1] 0.5886145
```

#Question 4.1 d

4.1 Area under the curve, Part I. What percent of a standard normal distribution N(mu = 0; sigma = 1) is found in each region? Be sure to draw a graph.

d. p(|Z|) > 2

so |z|>2 means regions where z>2 and where x<-2

p(x<-2) is pnorm(-2,mean=0,sd=1),

p(x>2) is 1-p(x<2) = 1-pnorm(2,mean=0,sd=1)

P(|x|>2)= P(x<-2)+ P(x>2)= pnorm(-2,mean=0,sd=1)+1-pnorm(2,mean=0,sd=1)

```
pnorm(-2,mean=0,sd=1)+1-pnorm(2,mean=0,sd=1)
```

```
## [1] 0.04550026
```

#Graphics or Visualization- Do we really need it

We did not have to draw the graphs of the distribution to do this problem. All we really needed was pnorm(), the function to calculate the cdf

But

a.) Mentally, we have to connect mathematics to visualizations to works to talk with other people

b.) It is a sanity check on our answers, if the graphs look crazy, something is wrong

c.) We need to learn these tools

When doing the homework this week, show me the pdf and cdf for the distribution, but you don't need to do more graphing than that.

On to the next problem, 4.3 from Open Intro

4.3 GRE scores, Part I.

Sophia who took the Graduate Record Examination (GRE) scored 160 on the Ver- bal Reasoning section and 157 on the Quantitative Reasoning section. The mean score for Verbal Reasoning section for all test takers was 151 with a standard deviation of 7, and the mean score for the Quantitative Reasoning was 153 with a standard deviation of 7.67. Suppose that both distributions are nearly normal.

a. Write down the short-hand for these two normal distributions.

   n(mu=151, sigma=7) or dnorm(x,mean=151,sd=7) -for Verbal n(mu=153, sigma=7.67) or dnorm(x,mean=153,sd=153) -for quantitative

b. What is Sophia's Z-score on the Verbal Reasoning section? On the Quantitative Reasoning section? Draw a standard normal distribution curve and mark these two Z-scores.

the formula is z= (x-mu)/sigma

```
#find z for verbal

xVerbal=160
muVerbal=151
sdVerbel=7

zVerbal=(xVerbal-muVerbal)/sdVerbel

print(zVerbal)
```

```
## [1] 1.285714
```

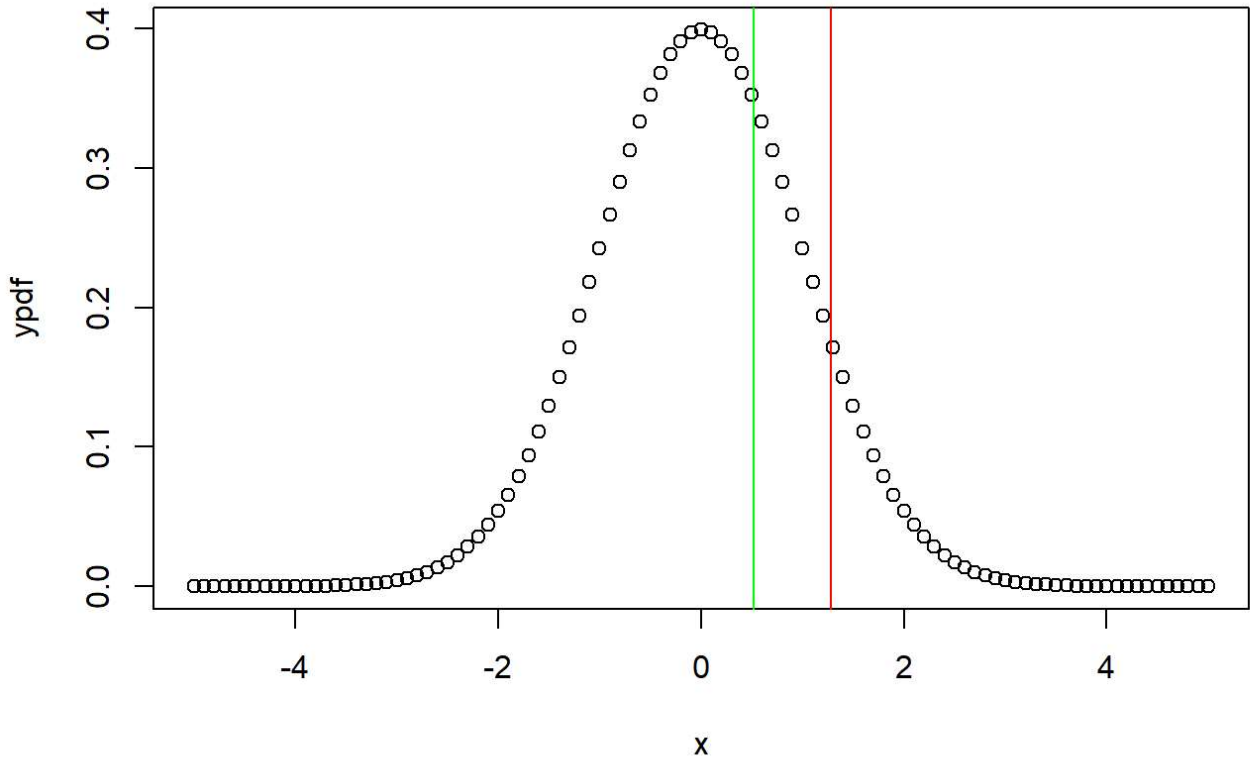
```
#find z for verbal

xQuant=157
muQuant=153
sdQuant=7.67

zQuant=(xQuant-muQuant)/sdQuant

print(zQuant)
```

```
## [1] 0.5215124
```

Plot, from x=-5 to 5 of the n(0,1) distibution

Add the lines at 1.285 and 0.521

```
#set up a list of values x to use in the graph, from -5 to 5 in steps of 5

x=seq(-5,5,0.1)

# get the pdf at each x, using the function dnorm
# we ae calculating the p(x) value at each value of x, for a gaussian of mean 0 and std dev =1

ypdf=dnorm(x,mean=0,sd=1)

#create the plot

plot(x,ypdf)


#verbal in red
abline(v=1.285, col='red')

#quant in green

abline(v=0.521, col="green")
```



c. What do these Z-scores tell you?

Positive z scores, indicate she is above the average, verbal is 1.285 sd above normal while quant is 0.521, which is just above average. Since these are Z scores, they are
directly comparable, as they are distances above or below average, scaled by standard deviation.

d. Relative to others, which section did she do better on?

Higher z is steps of std deviations, so she did better on verbal

e. Find her percentile scores for the two exams.

These are the cdf values, use pnorm

```
pnorm(1.285, mean=0, sd=1)
```

```
## [1] 0.9006039
```

```
pnorm(0.521, mean=0, sd=1)
```

```
## [1] 0.6988166
```

f. What percent of the test takers did better than her on the Verbal Reasoning section? On the Quantitative Reasoning section?

These are related to the cdf values

For verbal, she was at 0.90, or the 90% level, 10% of test takers did better

For quantitative, she was at 0.698, or 69.8%, so 30.2% of test takers did better

g. Explain why simply comparing raw scores from the two sections could lead to an incorrect conclusion as to which section a student did better on.

Comparing the raw scores doesn't take into account the difference in both the mean and std dev values on the two test, verbal and quantitative

h. If the distributions of the scores on these exams are not nearly normal, would your answers to parts (b)

- f. change? Explain your reasoning.
  Yes, my answers assumed the normal distribution described the scores, I used pnorm() to calculate the percentile levels, if another distribution held, we would need a different cdf function.

#Question 4.7

4.7 LA weather, Part I.

The average daily high temperature in June in LA is 77□F with a standard deviation of 5□F. Suppose that the temperatures in June closely follow a normal distribution.

a. What is the probability of observing an 83□F temperature or higher in LA during a randomly chosen day in June?

b. How cool are the coldest 10% of the days (days with lowest high temperature) during June in LA?

4.7a,

The problem says we can use a normal distribution with mean 77, sd=5

part a asks

"What is the probability of observing an 83□F temperature or higher in LA during a randomly chosen day in June?"

This is a cumulative normal question, so we will use a cdf

p(x>83), which is 1-p(x<83)

we always have to phrase the question in terms of p(x<a) to use the cummulative distribution

We can write the r code now,

we want 1-p(x<83) for a normal distribution of mean 77 and sd=5

```
1-pnorm(83,mean=77,sd=5)
```

```
## [1] 0.1150697
```

So an 11.5% chance of a June day in LA with a max Temp over 83 F

Part b, same distribution

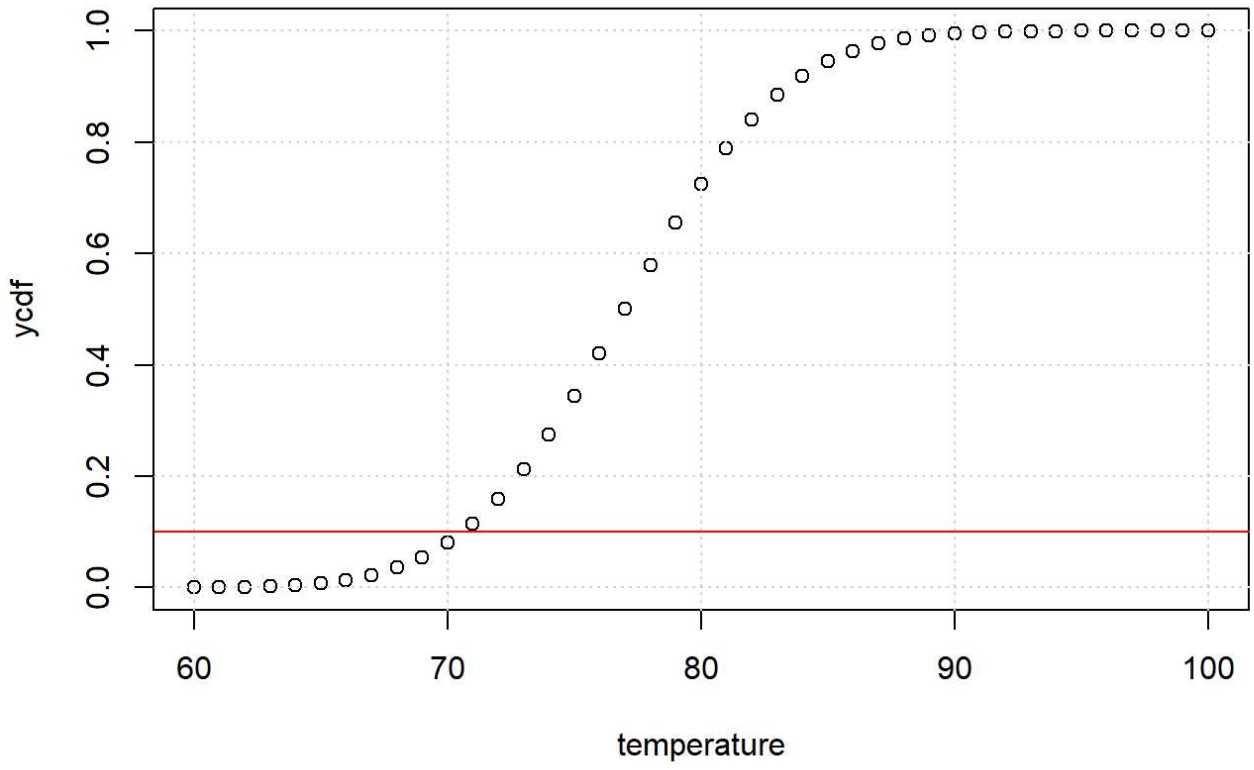
"(b) How cool are the coldest 10% of the days (days with lowest high temperature) during June in LA?"

This is one we haven't seen before, it uses the quantile function, meaning we have the percentile level (10%) here and we want to find a, such that

P(x<a)=10%, or p(x<a)<0.1

Visually, we are again using the cdf, but now we know the percentile

```
#step up temperatures from 60 to 100

temperature=seq(60,100,by=1)

#set up the cd for this temperature range

ycdf=pnorm(temperature,mean=77,sd=5)

plot(temperature, ycdf)
grid()
abline(h=0.1,col='red')
```



Okay,

looking at the graph, the red line at 10% crosses the curve at roughly 70 degrees, so 10% (0.1) of the days have a maximum temp of 70 or less

The quantile function in r for the normal distribution will calculate this for us, we can get it using qnorm

```
qnorm(0.1, mean=77,sd=5)
```

```
## [1] 70.59224
```

So 10% of the days have a max temperature of 70.6 degrees are less,

Question/Action

Use the methods shown above to do problem 4.5

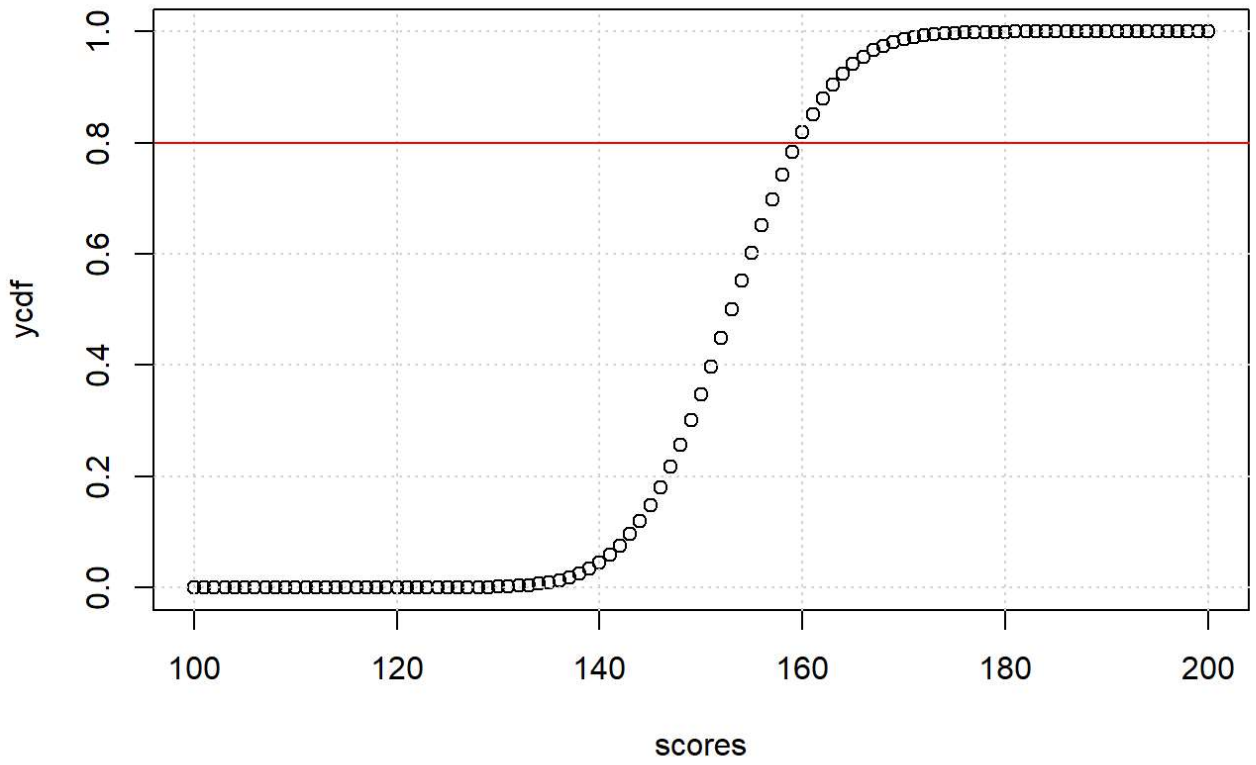
4.5 GRE scores, Part II. In Exercise 4.3 we saw two distributions for GRE scores: N(mean = 151; sd = 7) for the verbal part of the exam and N(mean = 153;sd = 7:67) for the quantitative part. Use this information to compute each of the following:

a. The score of a student who scored in the 80th percentile on the Quantitative Reasoning section. *show a graph of this, with the cdf and 80th percentile line shown*

```
qnorm(0.8, 153, 7.67)
```

```
## [1] 159.4552
```

```
#Create range of scores
scores=seq(100,200,by=1)

#set up the cd for this temperature range
ycdf=pnorm(scores,mean=153,sd=7.67)

plot(scores, ycdf)
grid()
abline(h=0.8,col='red')
```



b. The score of a student who scored worse than 70% of the test takers in the Verbal Reasoning section.

Worse than 70% is 30th percentile…

```
qnorm(0.3, 151, 7)
```

```
## [1] 147.3292
```

```
#Create range of scores
scores=seq(100,200,by=1)

#set up the cd for this temperature range
ycdf=pnorm(scores,mean=151,sd=7)

plot(scores, ycdf)
grid()
abline(h=0.3,col='red')
```