HW_LAB 04

Ryan Waterman

2025-02-01

HW/Lab 04

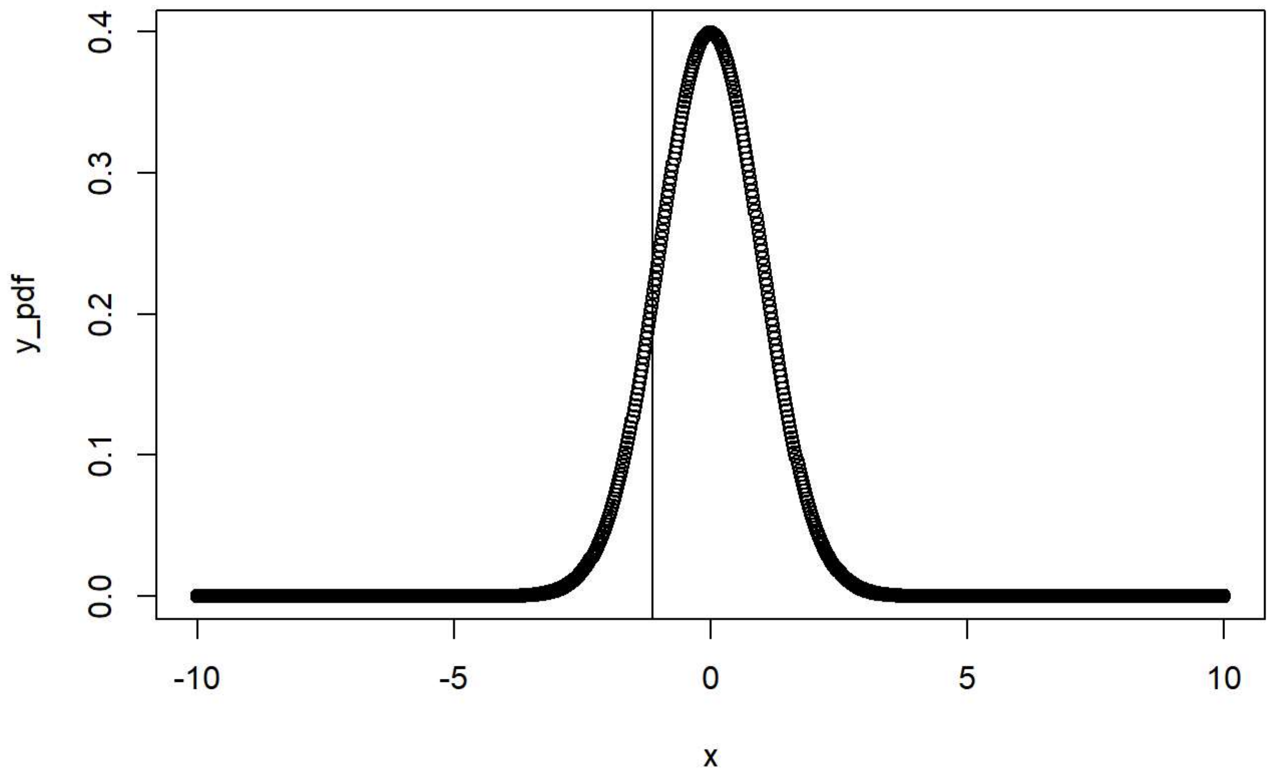
4.2 Area under the curve, Part II.

> What percent of a standard normal distribution N(μ = 0, σ = 1) is found in each region? Be sure to draw a graph.

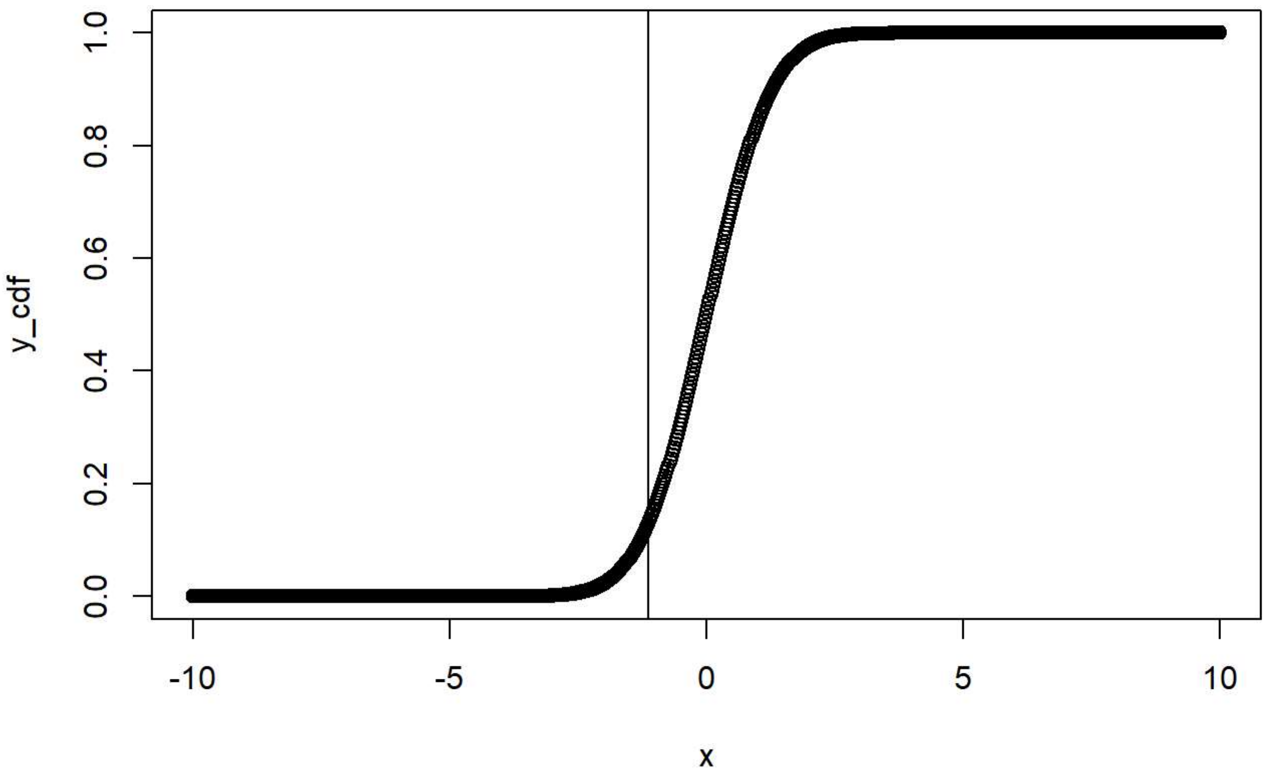
(a) $Z > -1.13$

```
#Set up variables
x <- seq(-10, 10, length=1000)
y_pdf <- dnorm(x, mean=0, sd=1)
y_cdf <- pnorm(x, mean=0, sd=1)
z <- -1.13

#Plot pdf
plot(x, y_pdf)
abline(v=z)
```



```
#Plot cdf
plot(x, y_cdf)
abline(v=z)
```



```
#calculate the percentage
percentage <- 1 - pnorm(z, mean = 0, sd=1)
percentage
```

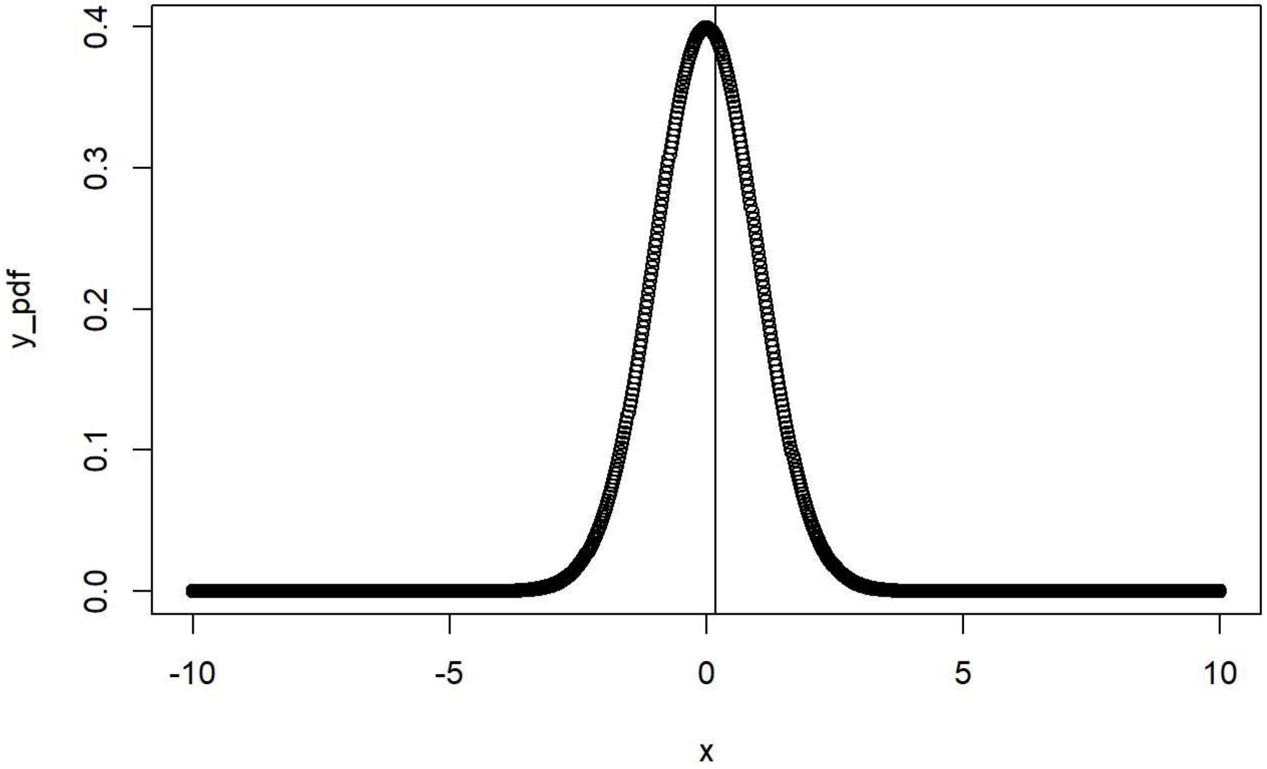
```
## [1] 0.8707619
```

The percentage is ~87%

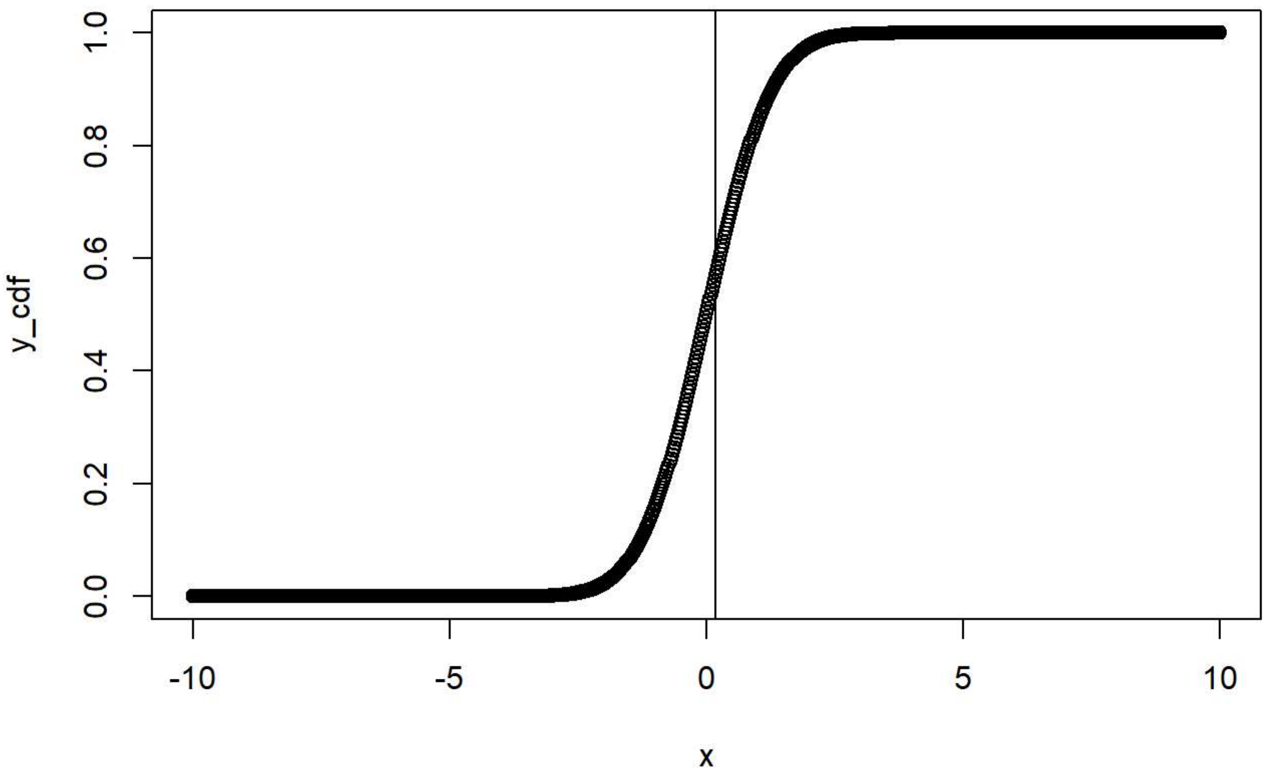
(b) Z < 0.18

```
#Set up variables
x <- seq(-10, 10, length=1000)
y_pdf <- dnorm(x, mean=0, sd=1)
y_cdf <- pnorm(x, mean=0, sd=1)
z <- 0.18

#Plot pdf
plot(x, y_pdf)
abline(v=z)
```



```
#Plot cdf
plot(x, y_cdf)
abline(v=z)
```



```
#calculate the percentage
percentage <- pnorm(z, mean = 0, sd=1)
percentage
```

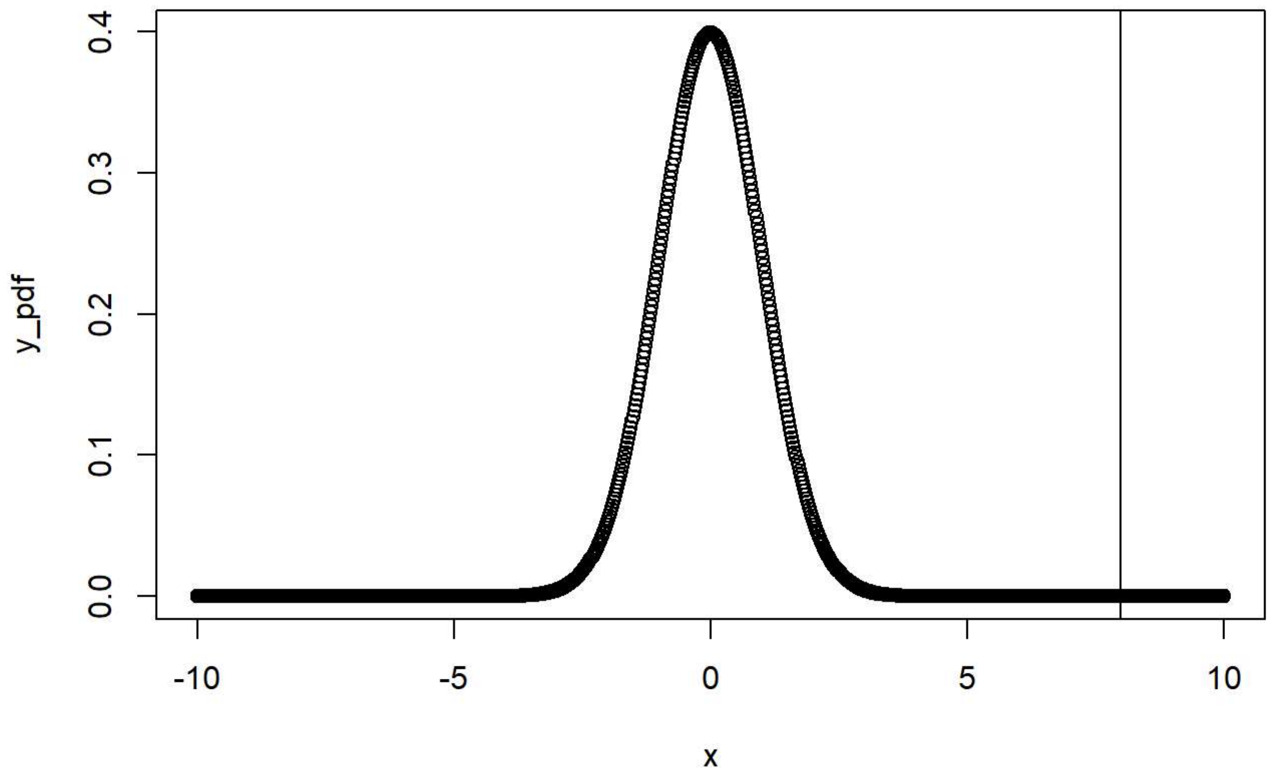
```
## [1] 0.5714237
```

The percentage is ~57%

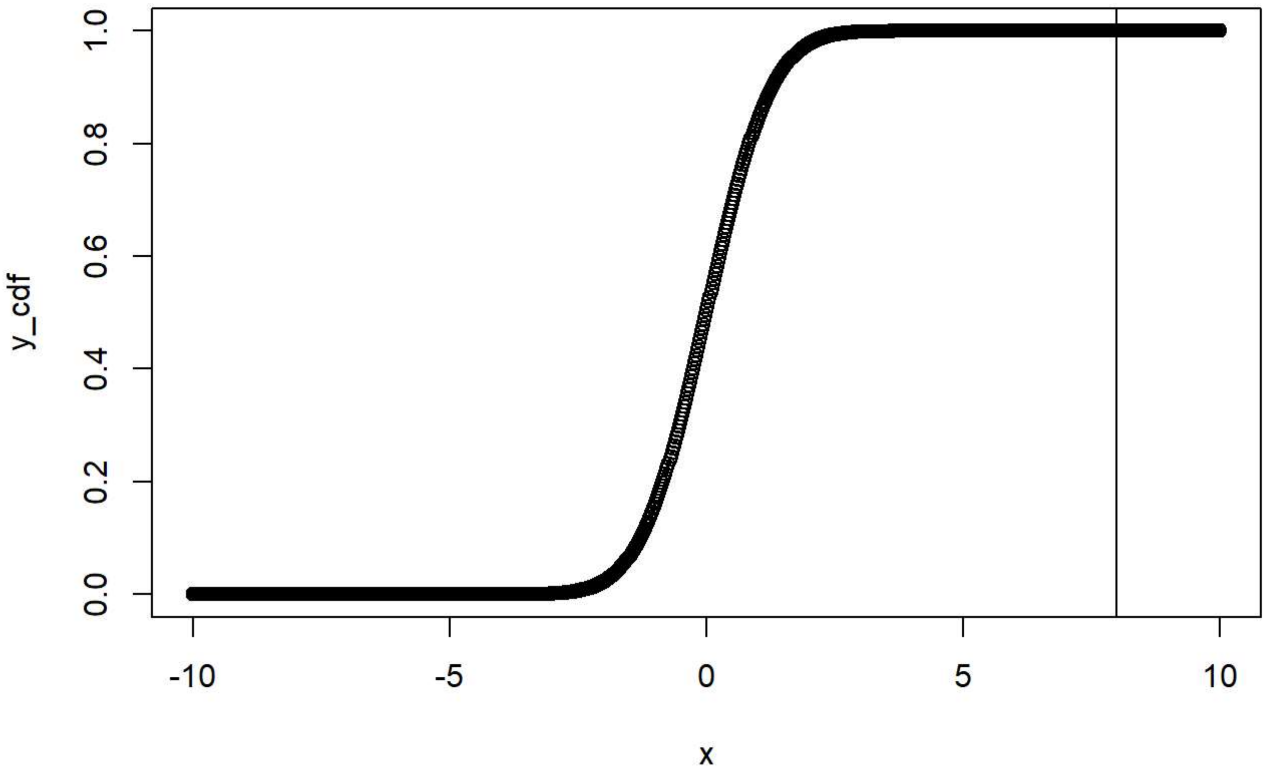
(c) Z > 8

```
#Set up variables
x <- seq(-10, 10, length=1000)
y_pdf <- dnorm(x, mean=0, sd=1)
y_cdf <- pnorm(x, mean=0, sd=1)
z <- 8

#Plot pdf
plot(x, y_pdf)
abline(v=z)
```



```
#Plot cdf
plot(x, y_cdf)
abline(v=z)
```



```
#calculate the percentage
percentage <- 1 - pnorm(z, mean = 0, sd=1)
percentage
```

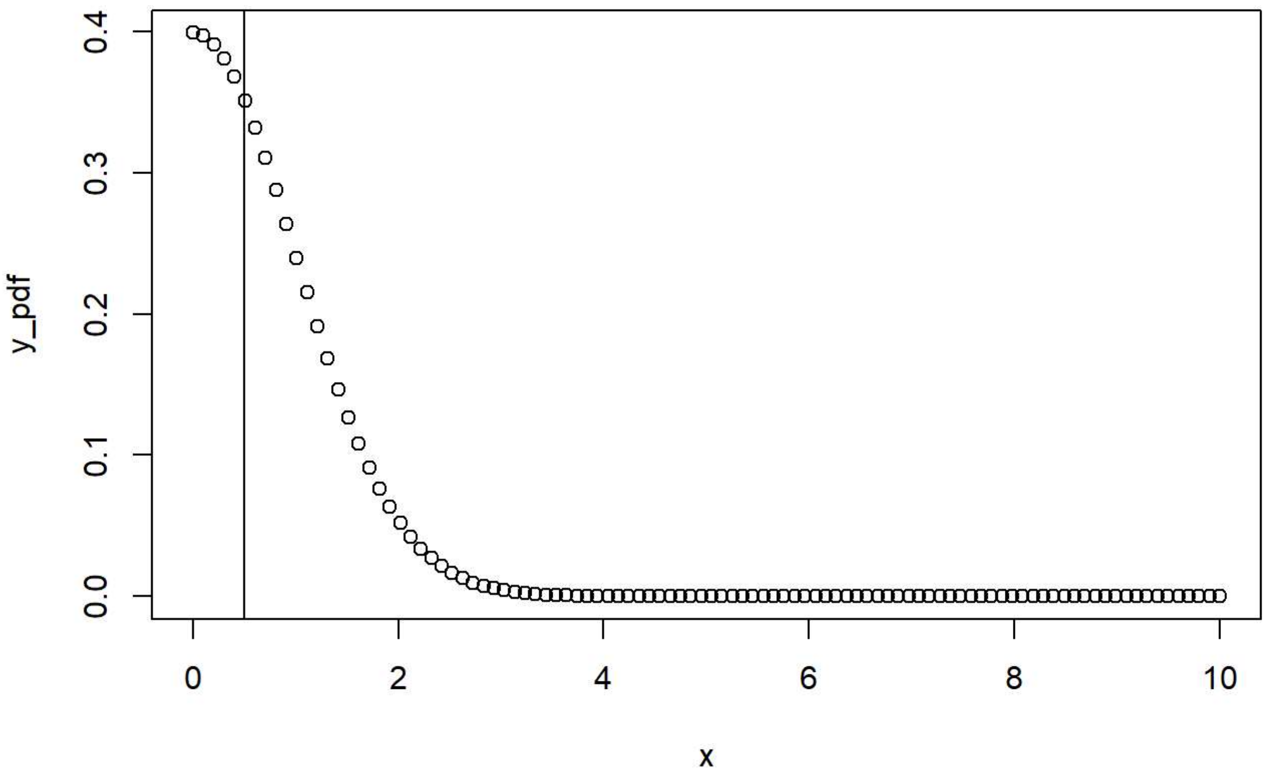
```
## [1] 6.661338e-16
```

The percentage is ~6.6e-14%, which is probably somewhere in the floating point error domain.

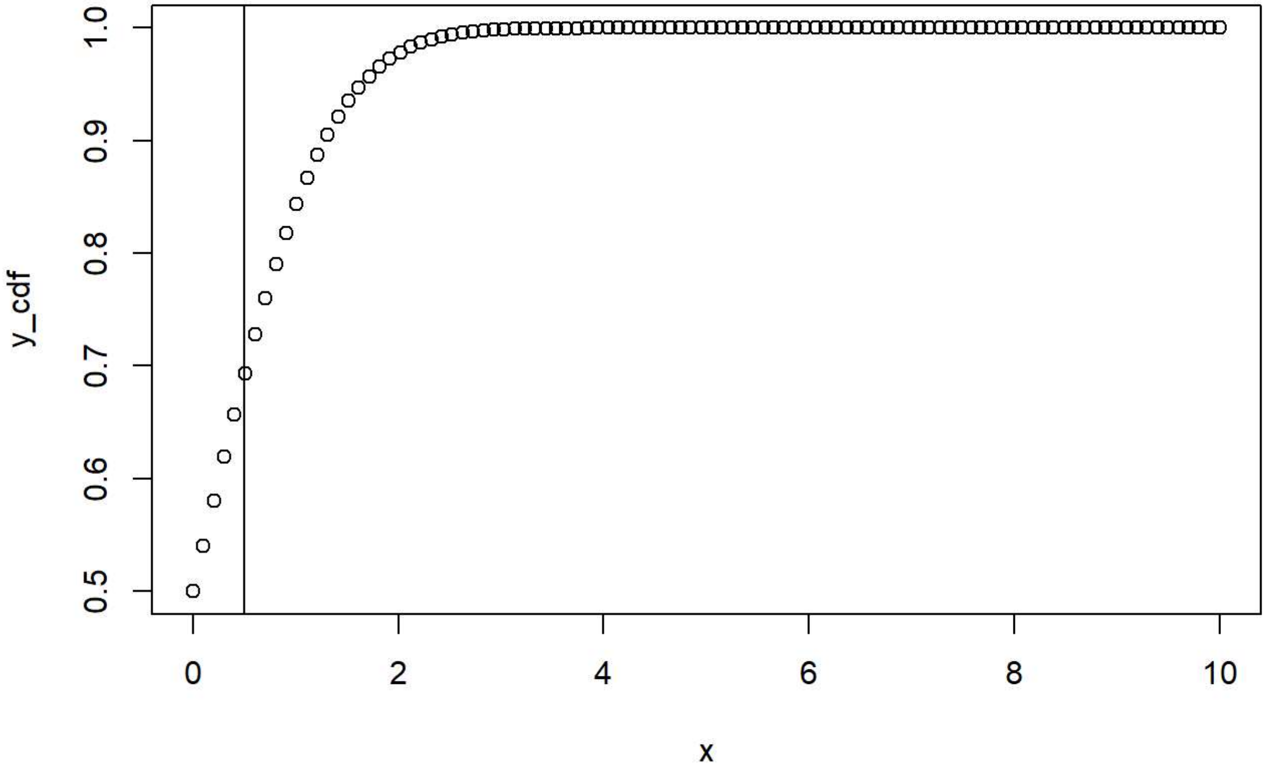
(d) |Z| < 0.5

```
#Set up variables
x <- seq(0, 10, length=100)
y_pdf <- dnorm(x, mean=0, sd=1)
y_cdf <- pnorm(x, mean=0, sd=1)
z <- 0.5

#Plot pdf
plot(x, y_pdf)
abline(v=z)
```



```
#Plot cdf
plot(x, y_cdf)
abline(v=z)
```



```
#calculate the percentage
percentage <- pnorm(z, mean = 0, sd=1)
percentage
```

```
## [1] 0.6914625
```

This one is a bit tricky, because the absolute value of Z constrains the graph between 0 and 10 (in this case), as seen above. This means that the maximum area under the PDF curve is 0.5 instead of 1, which changes the percentage calculation to what is seen in the code block, above, resulting in a percentage of ~69%

4.4 Triathlon times, Part I.

> In triathlons, it is common for racers to be placed into age and gender groups. Friends Leo and Mary both completed the Hermosa Beach Triathlon, where Leo competed in the Men, Ages 30 - 34 group while Mary competed in the Women, Ages 25 - 29 group. Leo completed the race in 1:22:28 (4948 seconds), while Mary completed the race in 1:31:53 (5513 seconds). Obviously Leo finished faster, but they are curious about how they did within their respective groups. Can you help them? Here is some information on the performance of their groups: • The finishing times of the Men, Ages 30 - 34 group has a mean of 4313 seconds with a standard deviation of 583 seconds. • The finishing times of the Women, Ages 25 - 29 group has a mean of 5261 seconds with a standard deviation of 807 seconds. • The distributions of finishing times for both groups are approximately Normal. Remember: a better performance corresponds to a faster finish.

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

Leo: N(μ=4313,σ=583), Mary: N(μ=5261,σ=807)

(b) What are the Z-scores for Leo’s and Mary’s finishing times? What do these Z-scores tell you?

```
#Compute Leo's z-score:
x_leo <- 4948
μ_leo <- 4313
σ_leo <- 583

Z_leo <- (x_leo-μ_leo)/σ_leo

#Compute Mary's z-score:
x_mary <- 5513
μ_mary <- 5261
σ_mary <- 807

Z_mary <- (x_mary-μ_mary)/σ_mary

Z_leo
```

```
## [1] 1.089194
```

```
Z_mary
```

```
## [1] 0.3122677
```

The both Z-scores are positive, indicating that both Leo and Mary had higher times than the mean. In this normalized format, it is clear that Leo performed worse than mary, as his Z-score is greater.

(c) Did Leo or Mary rank better in their respective groups? Explain your reasoning.

Mary ranked better because she had a lower Z-score. The Z-score is a measure of standard deviations from the mean, therefore a smaller number, including negatives, would indicate a better performance relative to the mean.

(d) What percent of the triathletes did Leo finish faster than in his group?

```
percentage <- 1-pnorm(x_leo, μ_leo, σ_leo)
percentage
```

```
## [1] 0.1380342
```

Leo finished faster than 13.8% of athletes in his group.

(e) What percent of the triathletes did Mary finish faster than in her group?

```
percentage <- 1-pnorm(x_mary, μ_mary, σ_mary)
percentage
```

```
## [1] 0.3774186
```

Mary finished faster than 37.8% of athletes in her group.

(f) If the distributions of finishing times are not nearly normal, would your answers to parts (b) - (e) change? Explain your reasoning.

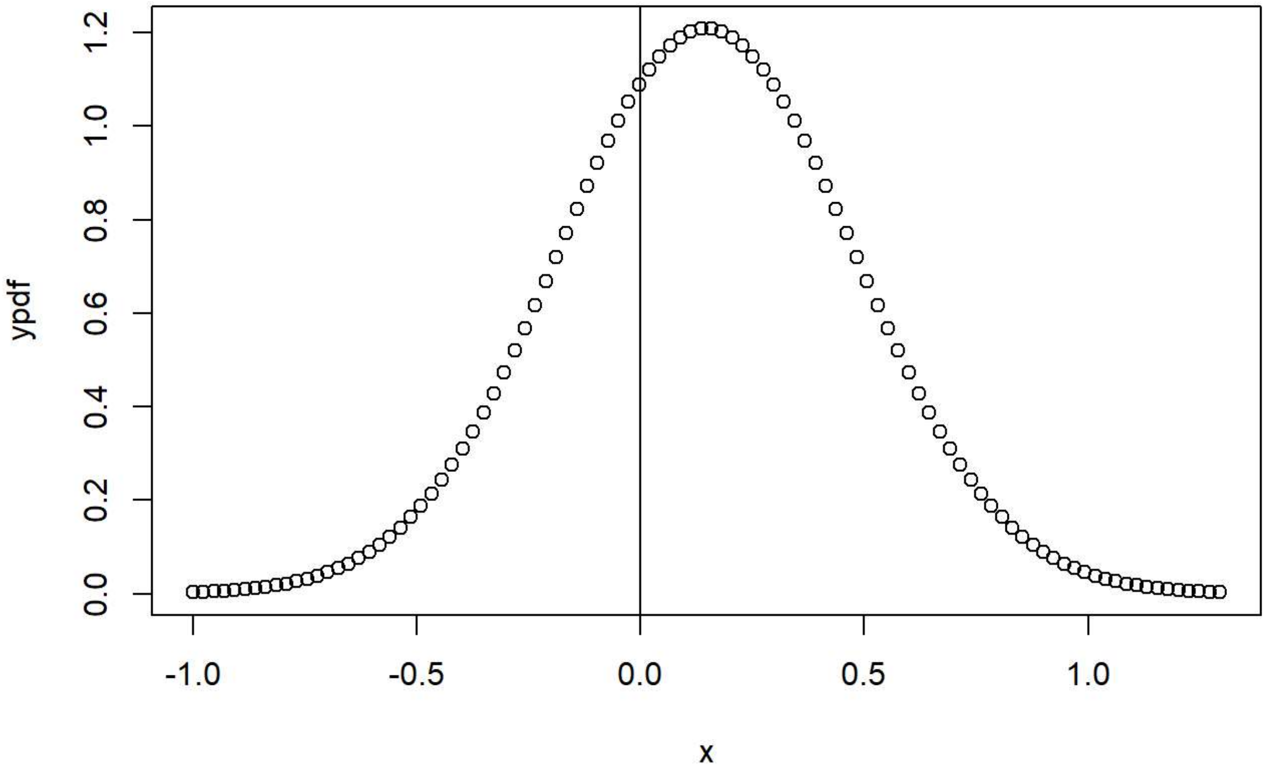
Yes, because the functions used assume a normal distribution centered about the mean. Distributions with skew would disproportionately affect percentages on one side of the mean.

4.8 CAPM.

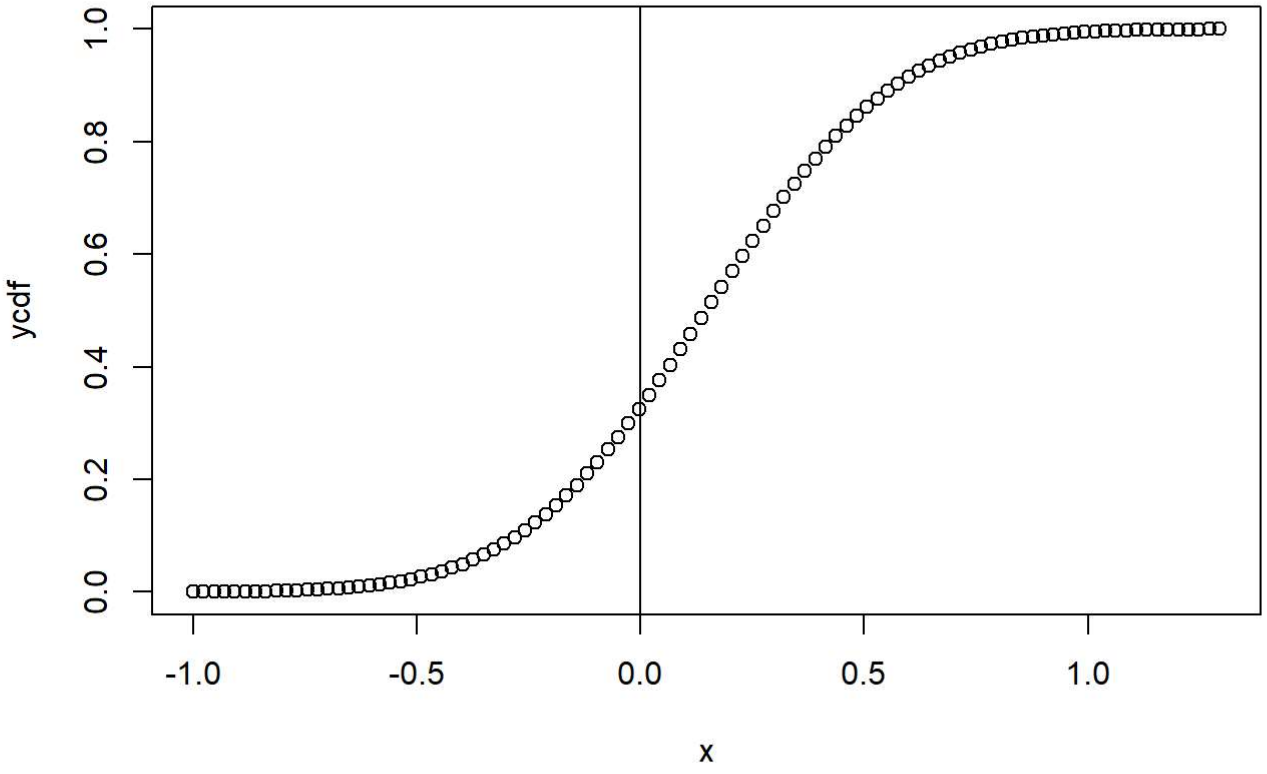
> The Capital Asset Pricing Model (CAPM) is a financial model that assumes returns on a portfolio are normally distributed. Suppose a portfolio has an average annual return of 14.7% (i.e. an average gain of 14.7%) with a standard deviation of 33%. A return of 0% means the value of the portfolio doesn't change, a negative return means that the portfolio loses money, and a positive return means that the portfolio gains money. Plot the PDF and CDF.

(a) What percent of years does this portfolio lose money, i.e. have a return less than 0%?

```
#Plot pdf and cdf
x <- seq(-1, 0.147+1.147, length=100)
ypdf <- dnorm(x, mean = 0.147, sd=0.33)
ycdf <- pnorm(x, mean = 0.147, sd=0.33)
plot(x, ypdf)
abline(v=0)
```



```
plot(x, ycdf)
abline(v=0)
```



```
#Determine the percentage
pnorm(0, mean=0.147,sd=0.33)
```

```
## [1] 0.3279957
```

This portfolio loses money in 32.8 percent of the time.

(b) What is the cutoff for the highest 15% of annual returns with this portfolio?

I am calculating this based on the assumption that the data is normally distributed and centered about the mean. In this case, the max loss would be 100% (i.e. no more money), and therefore the max gain would be 14.7% + 114.7%, or 129.4.

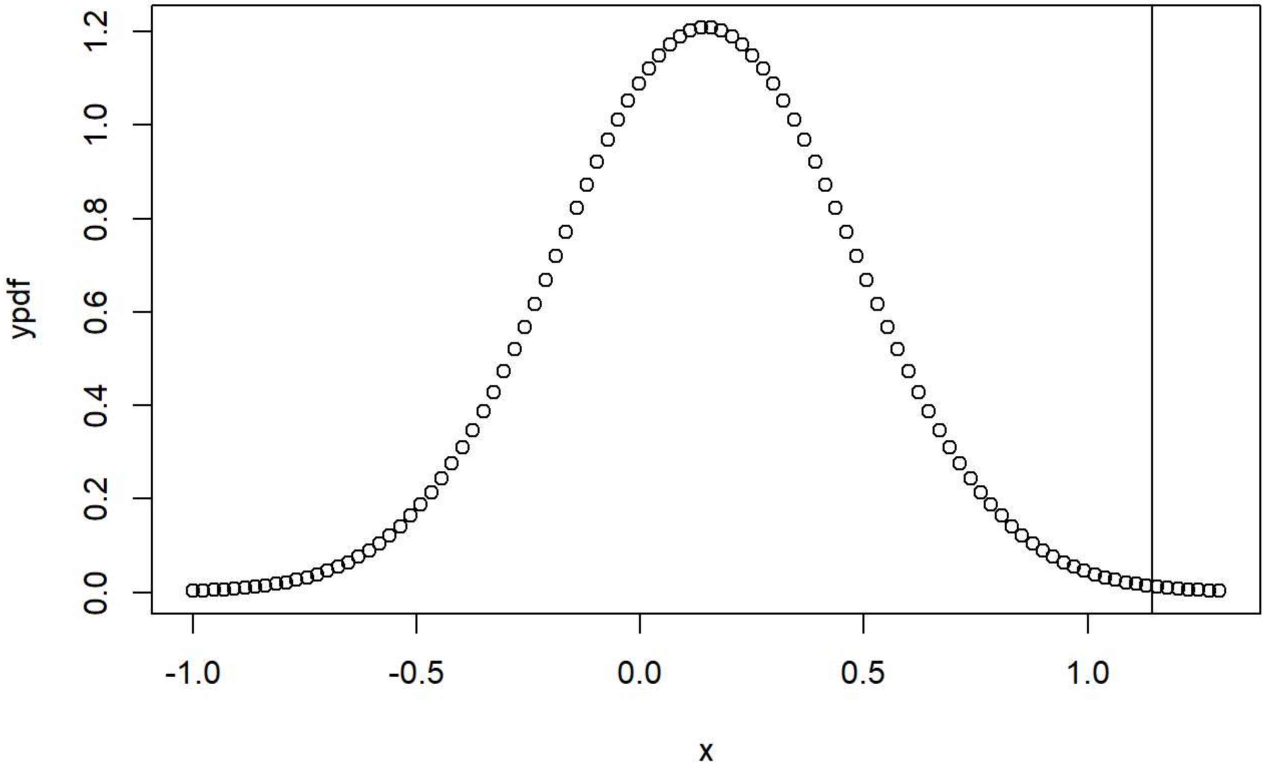
NOTE: All revised work is in bold or denoted with comments in the code block.

Revision on 2/6: I took another look at the pair programming and realized I was both thinking about this the wrong way, and over complicating it. I can just use the qnorm function at 0.85 to get the highest 15% of annual returns.

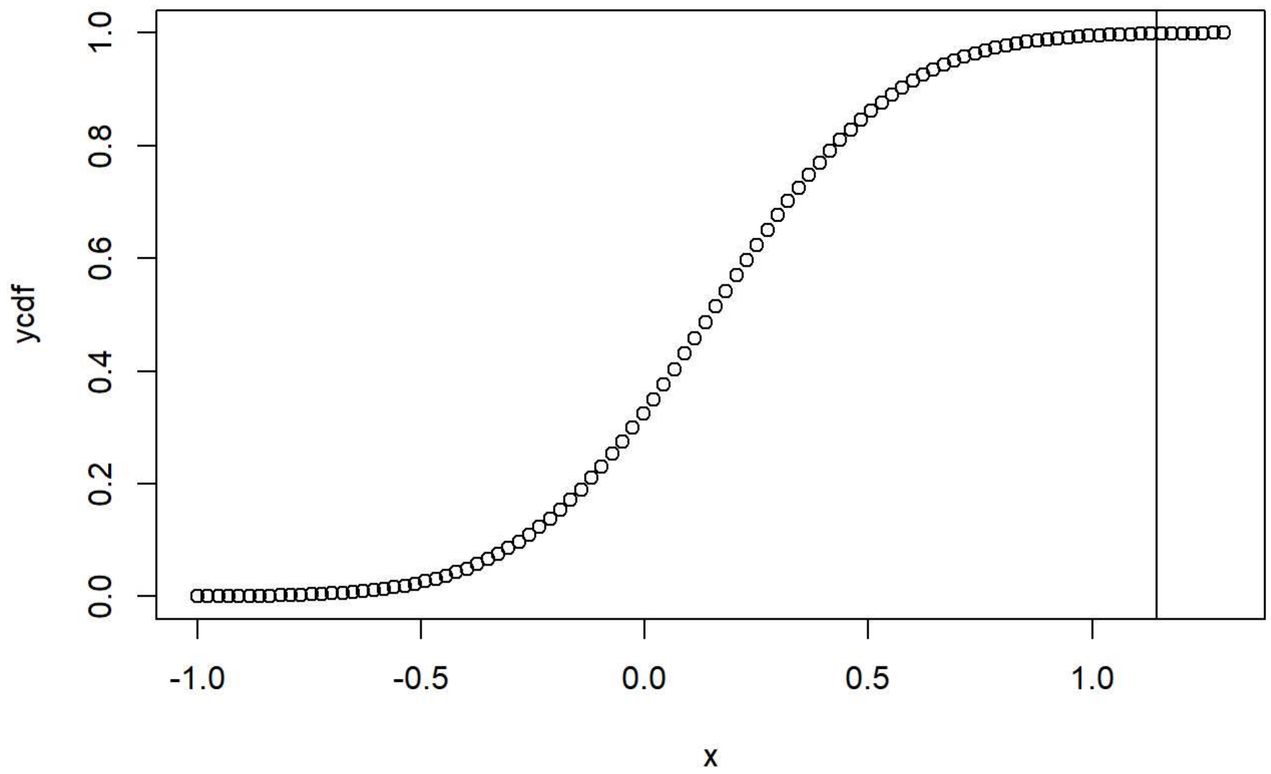
```
highest_15_percent <- 1.294-0.15

#Set pdf and cdf
x <- seq(-1, 1.294, length=100)
ypdf <- dnorm(x, mean = 0.147, sd=0.33)
ycdf <- pnorm(x, mean = 0.147, sd=0.33)

#plot pdf
plot(x, ypdf)
abline(v=highest_15_percent)
```



```
#plot cdf
plot(x, ycdf)
abline(v=highest_15_percent)
```



```
#Determine the percentage
dnorm(highest_15_percent, mean=0.147,sd=0.33)
```

```
## [1] 0.01259901
```

```
pnorm(highest_15_percent, mean=0.147,sd=0.33)
```

```
## [1] 0.9987412
```

```
#REVISION 2/6:
qnorm(0.85, mean=0.147,sd=0.33)
```

```
## [1] 0.489023
```

Looks like about 99.9% of the time the portfolio returns less than the highest 15%. I feel like I did this very wrong.

^^^ Well past Ryan, you did. The highest 15% of annual returns is cut off at an annual return of about 48.9%.

4.14 Defective rate.

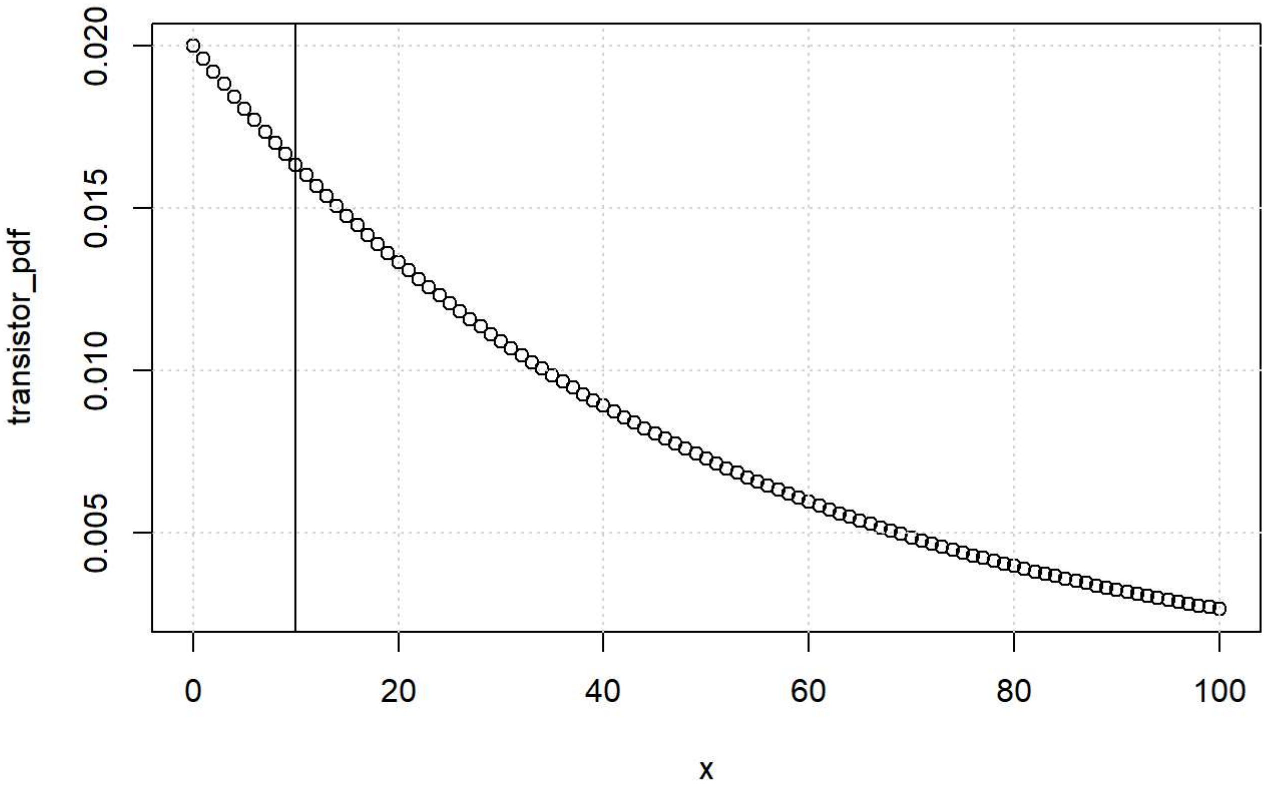
> A machine that produces a special type of transistor (a component of computers) has a 2% defective rate. The production is considered a random process where each transistor is independent of the others.

(a) What is the probability that the 10th transistor produced is the first with a defect?

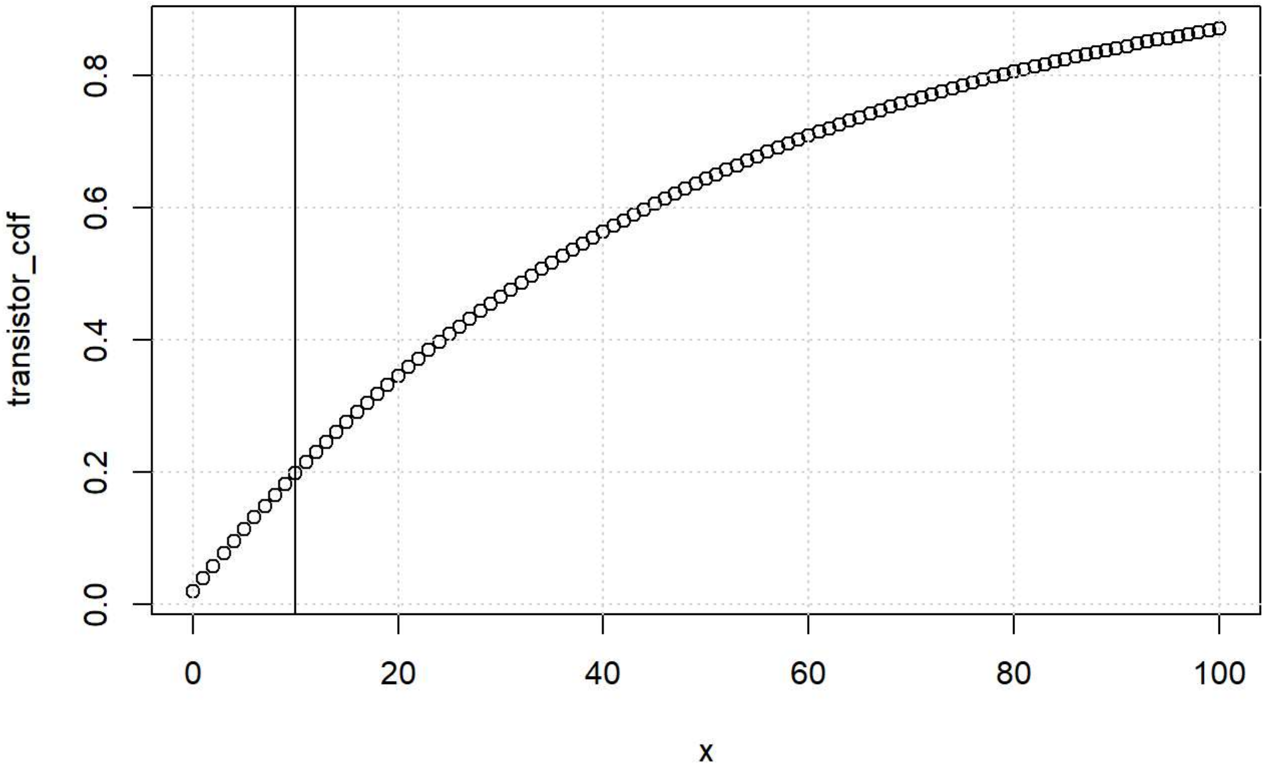
```
x=0:100

#Get pdf and cdf
transistor_pdf=dgeom(x, prob=0.02)
transistor_cdf=pgeom(x, prob=0.02)

#plot pdf
plot(x,transistor_pdf)
grid()
abline(v=10)
```



```
#plot cdf
plot(x,transistor_cdf)
grid()
abline(v=10)
```



```
#find the probability of a failure at the tenth transistor
dgeom(10,prob=0.02)
```

```
## [1] 0.01634146
```

The probability is 1.63%

(b) What is the probability that the machine produces no defective

transistors in a batch of 100?

```
#The probability of a non-defective transistor is 1-p
p_non_def <- 1-0.02

#We don't expect a single success in the whole set, so we can calculate the probability of 100 c
onsecutive non-defective
#this probability is p = p_non_def1*pp_non_def2*p_non_def3... or p = p_non_def^100
p_100_non_def <- p_non_def^100
p_100_non_def
```

```
## [1] 0.1326196
```

The probability is about 13.26%

(c) On average, how many transistors would you expect to be produced before the first with a defect? What is the standard deviation?

```
p=0.02

#This can be derived with 1/p
expected_value=1/0.02
expected_value
```

```
## [1] 50
```

```
#equation found in the textbook for geometric distributions
sd=((1-p)/p^2)^0.5
sd
```

```
## [1] 49.49747
```

I would expect 50 transistors on average, and the standard deviation is ~49.5

(d) Another machine that also produces transistors has a 5% defective rate where each transistor is produced independent of the others. On average how many transistors would you expect to be produced with this machine before the first with a defect? What

is the standard deviation?

```
p=0.05

#This can be derived with 1/p
expected_value=1/p
expected_value
```

```
## [1] 20
```

```
#equation found in the textbook for geometric distributions
sd=((1-p)/p^2)^0.5
sd
```

```
## [1] 19.49359
```

The first defect would be on transistor 20, and the standard deviation is ~19.49

(e) Based on your answers to parts (c) and (d), how does increasing the probability of an event affect the mean and standard deviation of the wait time until success?

The mean and standard deviation are linearly proportional with the wait time until success.

4.18 Chickenpox, Part I.

> Boston Children’s Hospital estimates that 90% of Americans have had chickenpoxby the time they reach adulthood.

(a) Suppose we take a random sample of 100 American adults. Is the use of the binomial distribution appropriate for calculating the probability that exactly 97 out of 100 randomly sampled American adults had chickenpox during childhood? Explain.

The Binomial distribution would be appropriate for calculating the probability. The binomial distribution describes the probability of have some number of successes (k) in a set of independent trials (n). In this case, k=97 and n=100.

(b) Calculate the probability that exactly 97 out of 100 randomly

sampled American adults had chickenpox during childhood.

```
#Assign the variables
successes=97
trials=100
prob=0.9

#Use the dbinom function to compute exactly 97 successes out of 100 with 90% probability
prob_97_100<-dbinom(successes, size=trials, prob=prob)
prob_97_100
```

```
## [1] 0.005891602
```

~0.59%

(c) What is the probability that exactly 3 out of a new sample of 100 American adults have not had chickenpox in their childhood?

Three adults not having it is equivalent to 97/100 having it…

```
#Assign the variables
successes=97
trials=100
prob=0.9

#Use the dbinom function to compute exactly 97 successes out of 100 with 90% probability
prob_97_100<-dbinom(successes, size=trials, prob=prob)
prob_97_100
```

```
## [1] 0.005891602
```

Hmmm… Dejavu. Just to make sure I am not misinterpreting this question because it defines 3 out of a new set of 100, calculate again with 194 successes out of 200 trials.

```
#Assign the variables
successes=194
trials=200
prob=0.9

#Use the dbinom function to compute exactly 97 successes out of 100 with 90% probability
prob_194_200<-dbinom(successes, size=trials, prob=prob)
prob_194_200
```

```
## [1] 0.0001094006
```

This is significantly lower, 0.01% versus 0.5%. This is likely a result of the factorials in the denominator if the binomial function.

(d) What is the probability that at least 1 out of 10 randomly sampled American adults have had chickenpox?

The probability that at least one will have had chicken pox is equal to 1-P(x<1), or 1-P(x==0). Source: Pair programming.

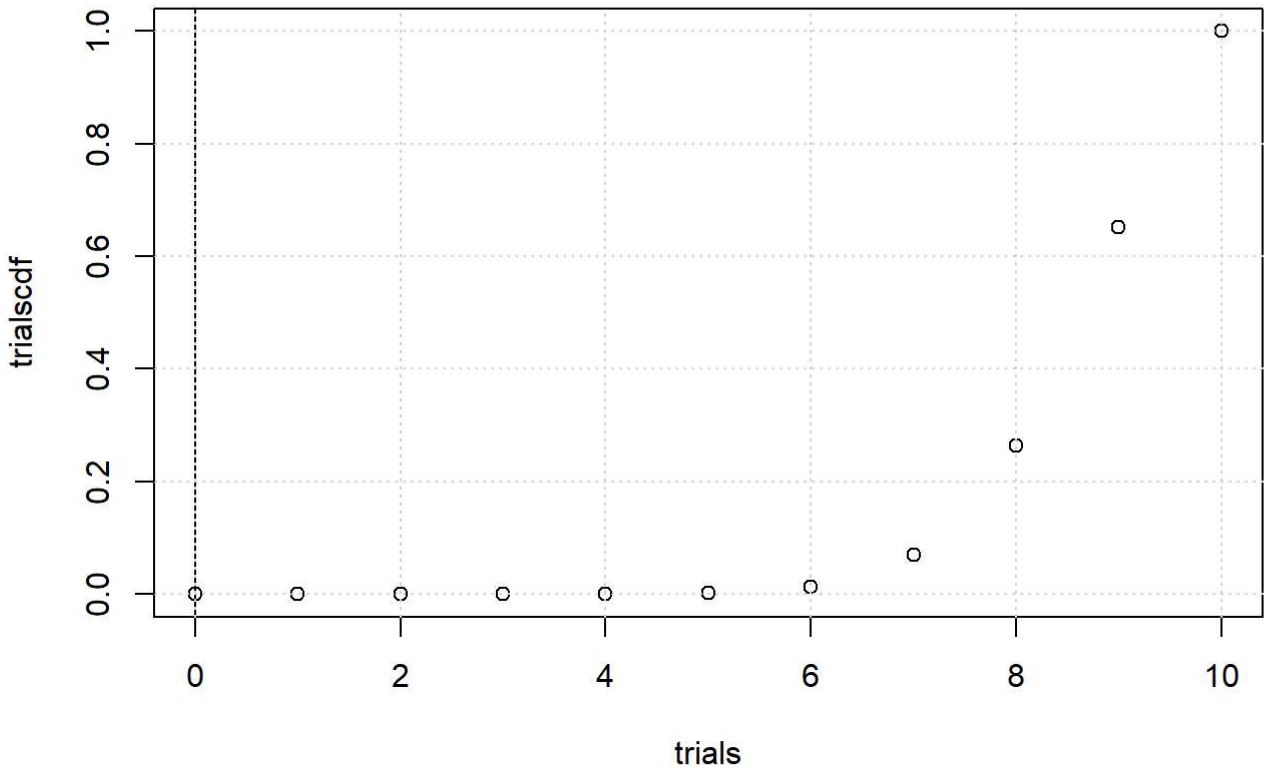
```
successes=0
trials=10
prob=0.9
prob_1_10 <- 1-pbinom(successes, size=trials, prob=prob)
prob_1_10
```

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

100%… That seems fishy. I am going to verify with the CDF plot.

```
#variables
successes=0
trials=0:10
prob=0.9

#plot it
trialscdf=pbinom(trials,size=max(trials), p=prob)
plot(trials,trialscdf)
abline(v=successes)
grid()
```



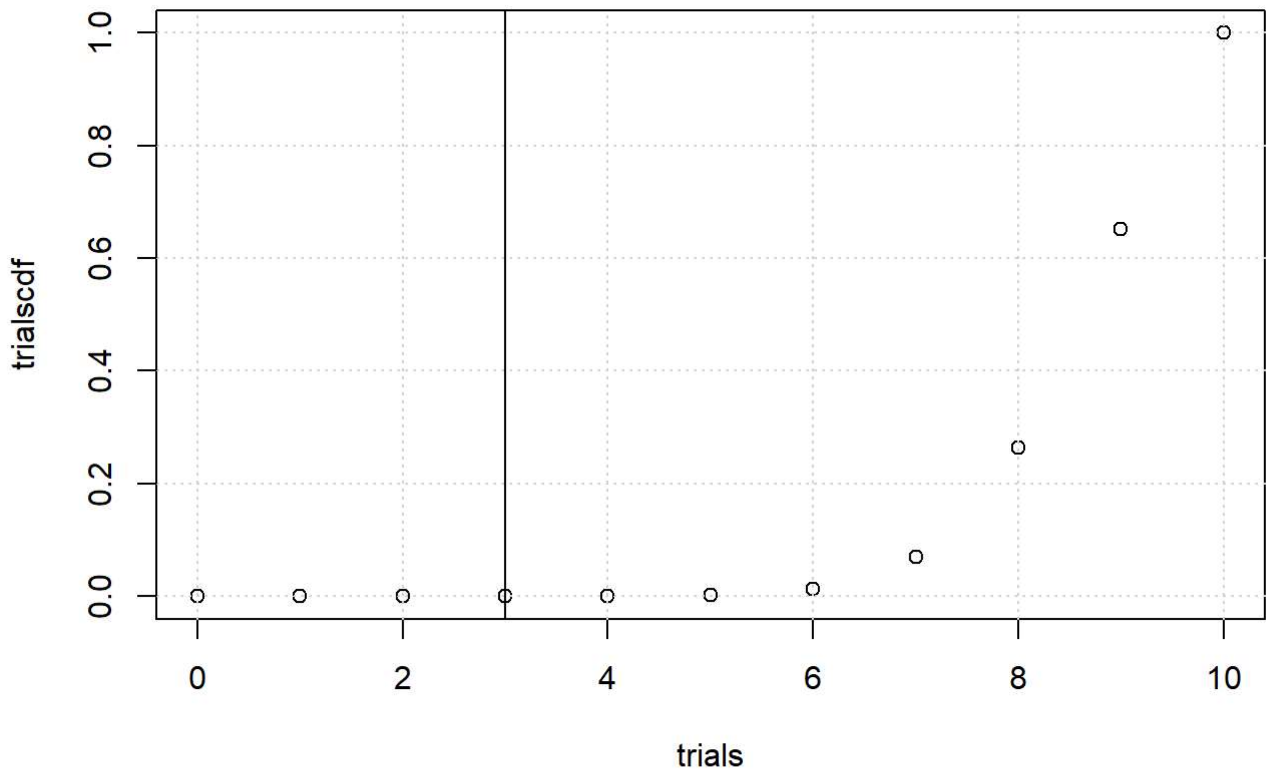
Whelp, I need to work on my intuition.

(e) What is the probability that at most 3 out of 10 randomly sampled American adults have not had chickenpox?

At most… CDF (thanks again pair programming).

```
#variables
successes=3
trials=0:10
prob=0.9

#plot it
trialscdf=pbinom(trials,size=max(trials), p=prob)
plot(trials,trialscdf)
abline(v=successes)
grid()
```



```
prob_3_10<-pbinom(successes,size=max(trials), p=prob)
prob_3_10
```

```
## [1] 9.1216e-06
```

Not quite zero, but close, about 0.00091%

4.32 Stenographer’s typos.

> A very skilled court stenographer makes one typographical error (typo) per hour on average.

(a) What probability distribution is most appropriate for calculating the probability of a given number of typos this stenographer makes in an hour?

In this case, a Poisson distribution would be most fitting. The population would be every keystroke made by the stenographer in an hour, which is a large population over a time period. Assuming a stenographer has a relatively consistent word-per-minute output, the population size should remain somewhat fixed, and we are concerned with the number of typos, which could be considered an event in the population.

(b) What are the mean and the standard deviation of the number of typos this stenographer makes?

The mean of a Poisson distribution is the average number of historical occurrences in a time frame, in this case 1. The standard deviation is the square root of the mean, or sqrt(1).

(c) Would it be considered unusual if this stenographer made 4 typos in a given hour?

Let’s find out… From the book, P(observe 4 events)=((1$^{(4))(e}$(-1)))/4!

```
prob_4_events <- (1^(4)*exp(-1))/factorial(4)
prob_4_events
```

```
## [1] 0.01532831
```

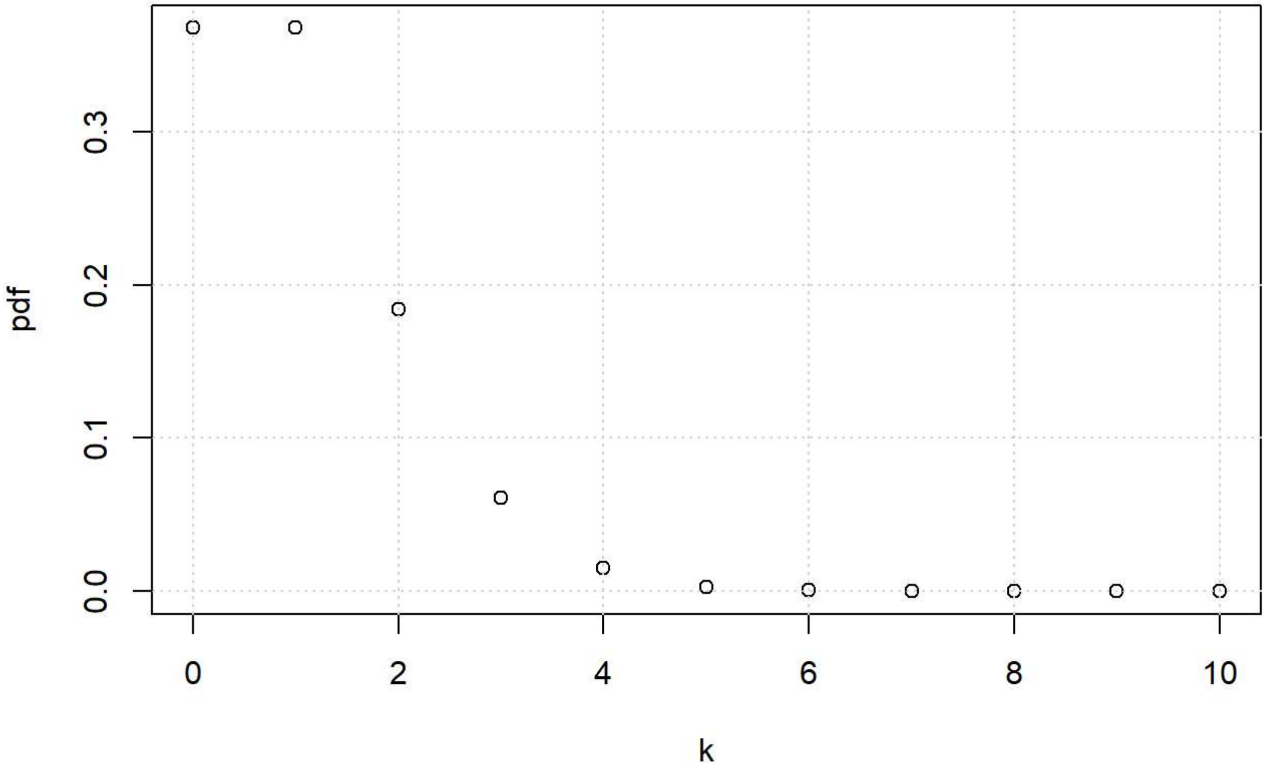
This would be unusual, occurring in only 1.5% of the hour long observational windows.

(d) Calculate the probability that this stenographer makes at most 2 typos in a given hour.

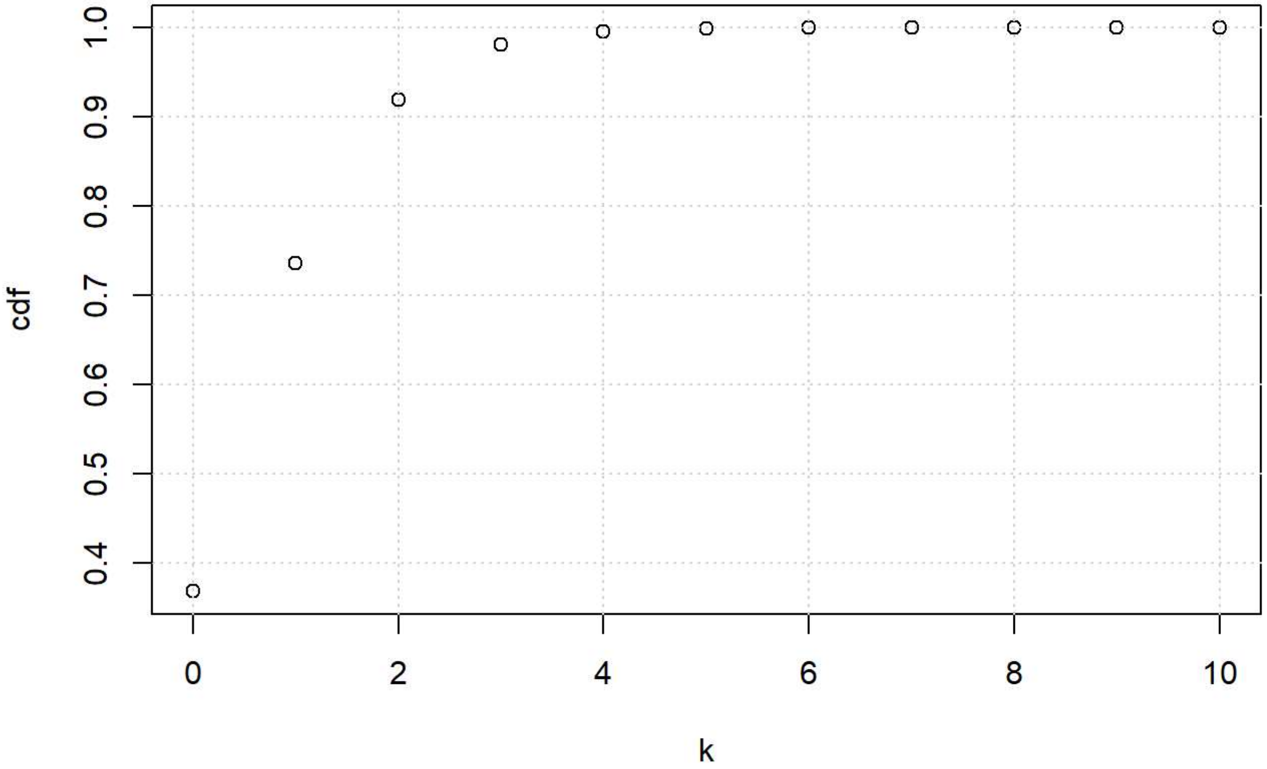
Plot the PDF and CDF

```
#Set up range of occurrences
k=0:10

#Plot the pdf
pdf=dpois(k,lambda=1)
plot(k,pdf)
grid()
```



```
#plot the cdf
cdf=ppois(k,lambda=1)
plot(k,cdf)
grid()
```



At most indicates that we need the CDF… Find the actual value.

```
occurrences=2
cdf=ppois(occurrences,lambda=1)
cdf
```

```
## [1] 0.9196986
```

At most 2 occurrences will occur about 92 percent of the time.