# Lab 05

Jennifer Lin

jenniferyjlin@berkeley.edu (mailto:jenniferyjlin@berkeley.edu)

# 1

### From the example in class, what does the y-axis of the histogram respresent (probability of )

### and how does it relate to the binomial distribution?

```
# The y-axis of the histogram:
#    The y-axis of the histogram respresents how frequent do the events (defined by x
-axis) happen.
#    For example, the height of x=10-20 means the amount of the times that x=10-20 ha
ppens during the set of tests.
# Binomial distribution:
#    There are n trials each of which can have two possible outcomes.
#    X is the number of success and n - X would be the number of failures.
#    The number of heads observed on n tosses of a coin
#    The X is the number of heads we obsearve
#    The binomial distribution would be a good model for the number of heads observed
on n tosses of a coin.
#    Thus, for example, if the probability that the coin comes up head on is p=0.2, g
iven n=10 trials,
#    the probablily of getting exactly 2 heads is (0.2^2 * 0.8^8) * choose(10,2) = 0.
3019899.
#    Consequently, in the plot, the height of X=2 is  0.3019899.
```

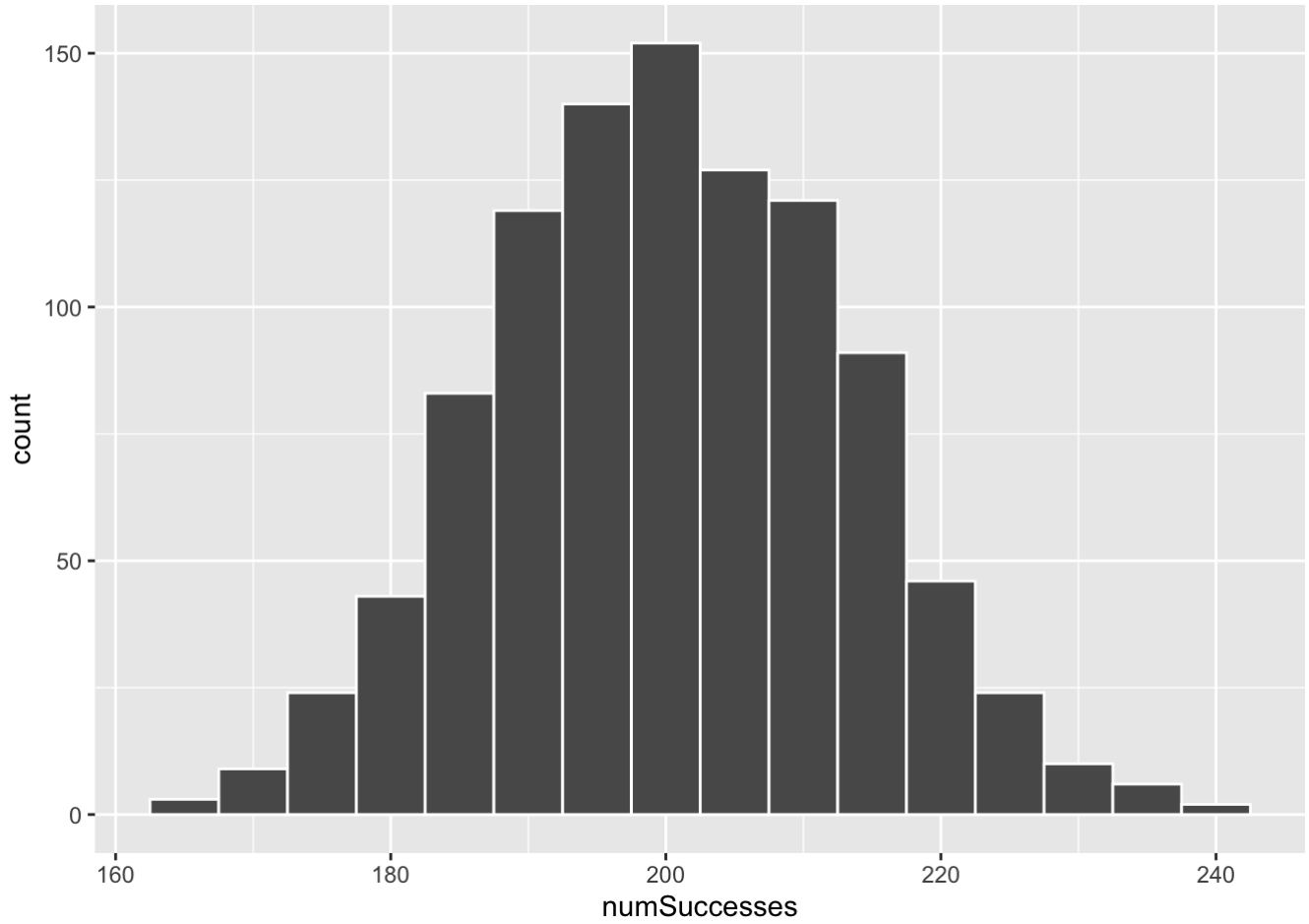
# 2

### From the example in class, please make a plot of the probabilities (exactly how you define that is up to you) as a function of θ.

### Please set n (the number of draws) to be sufficiently large (e.g. 1000) and plot the results.

### What does this plot represent?

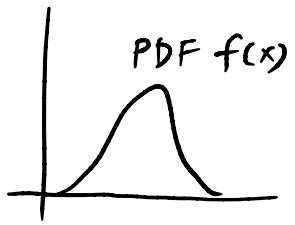
```
library(ggplot2)
theta <- 0.2
numSuccesses <- rep(NaN, 1000)
for (i in 1:1000){
  randomDraw <- runif(1000, 0, 1)
  x <- randomDraw < theta
  successes <- sum(x)
  numSuccesses[i] <- successes
}
numSuccesses_df <- data.frame(cbind(1:length(numSuccesses),numSuccesses))
ggplot(numSuccesses_df, aes(x=numSuccesses)) + geom_histogram(color="white", binwidth
= 5)
```



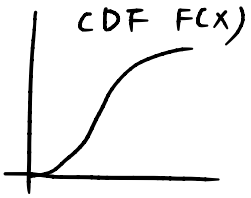

```
# Every random draw, I draw 1000 random values from 0 to 1 (i.e. [randomDraw]).
# Given theta=0.2, in each random draw [randomDraw], I calculate how many values are
 smaller than theta=0.2 and save the number as [success].
# I repeat these steps for 1000 times (according to i in 1:1000) and save the amount
 of successes into numSuccesses.
# Finally, I plot the results.
# For example, the height of X=200±2.5 is around 140.
# This means the in the 1000 repeats of the 1000 trials, around 140 repeats of the ex
periments get 200±2.5 successes during the 1000 trials
# (i.e. around 140 repeats of the experiments get 200±2.5 times of randon draw number
s that is smaller than theta=0.2 during the 1000 trials) .
```

# 3

**If f(x) is a PDF of a certain distribution and F(x) is its corresponding CDF, please define F(x) in terms of f(x).**




$f(x) = F'(x)$




$F(x) = \int_0^{\infty} f(x)\,dx$

# 4

## Please list the parameters of the normal distribution and what each of them determine.

```
# Normal distributions are defined by two parameters, the mean (μ) and the standard d
eviation (σ).
# Mean (μ) determines the middle of the distribution.
# Standard deviation (σ) determines the width of the distribution. 68% of the area of
a normal distribution is within one standard deviation of the mean. Approximately 95%
of the area of a normal distribution is within two standard deviations of the mean.
```

# 5

## Come up with an example of when it is apt to use the quantile function.

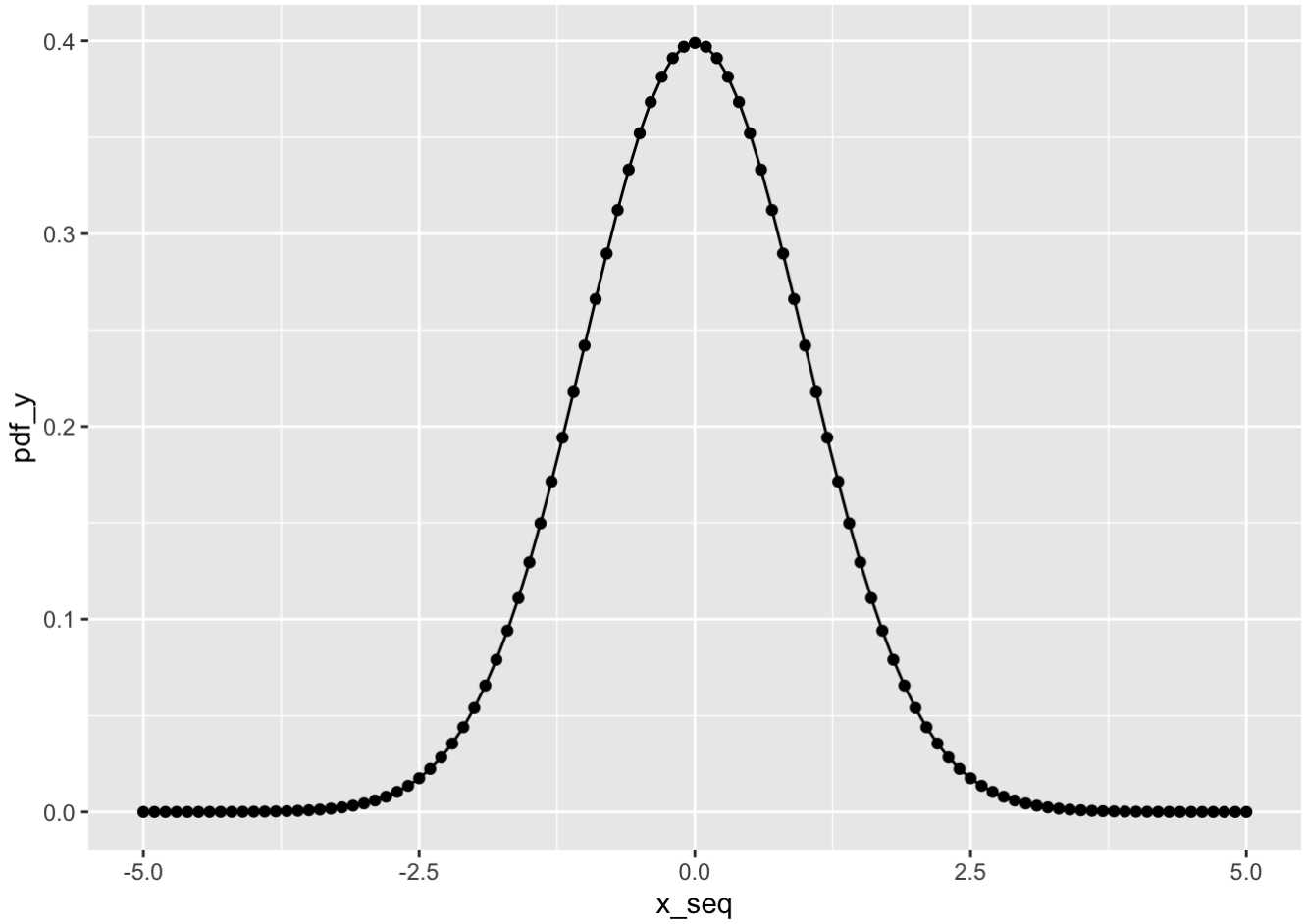
```
# The quantile function is a inverse CDF function.
# The CDF function takes input x and returns values from the [0,1] interval (p=probab
ilities).
# The quantile function (inverse of the cumulative distribution function) tells us wh
at x would make F(x) return some value p.
# For instance, if a boy is taller than or as tall as 75% of his classmates then the
 percentile rank of his height is 75,
# i.e.he is in the 79th percentile of heights in his class.
```

# 6

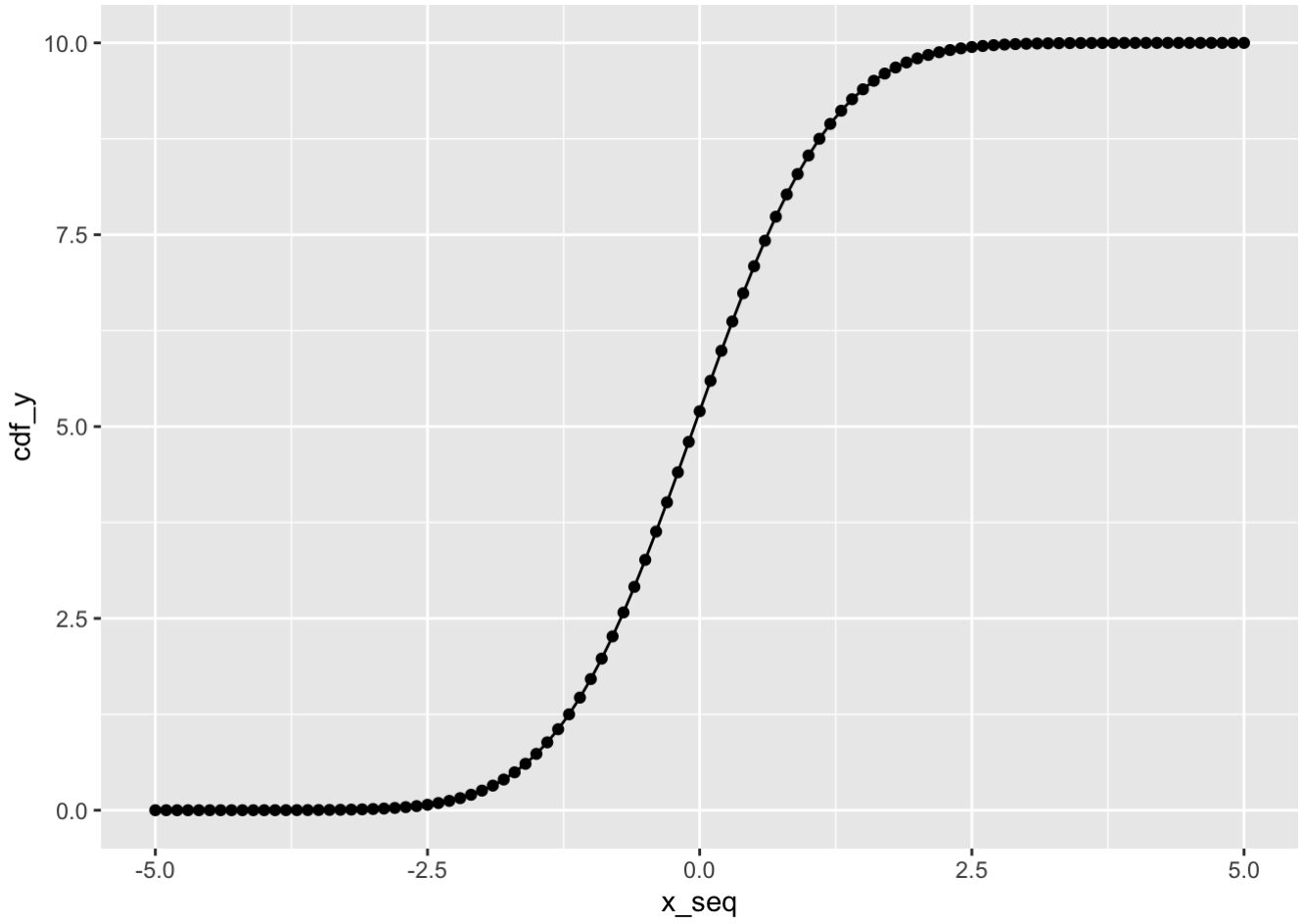
## Instead of using R's built-in function for a cumulative distribution function,

## please code up the CDF as a function of any probability density function of your choosing and its associated parameters and plot both the PDF and CDF.

```
myCDF <- function(density){
  cdf <- c()
  cdf[1] <- density[1]
  for(i in 2:length(density)){
    cdf[i] <- density[i]+cdf[i-1]
  }
  return(cdf)
}
x_seq <- seq(-5, 5, by=0.1)
pdf_y <- dnorm(x_seq)
cdf_y = myCDF(pdf_y)
# plot PDF
pdf_y_df <- data.frame(x=x_seq,pdf_y=pdf_y)
ggplot(pdf_y_df,aes(x=x_seq,y=pdf_y)) + geom_line() + geom_point()
```



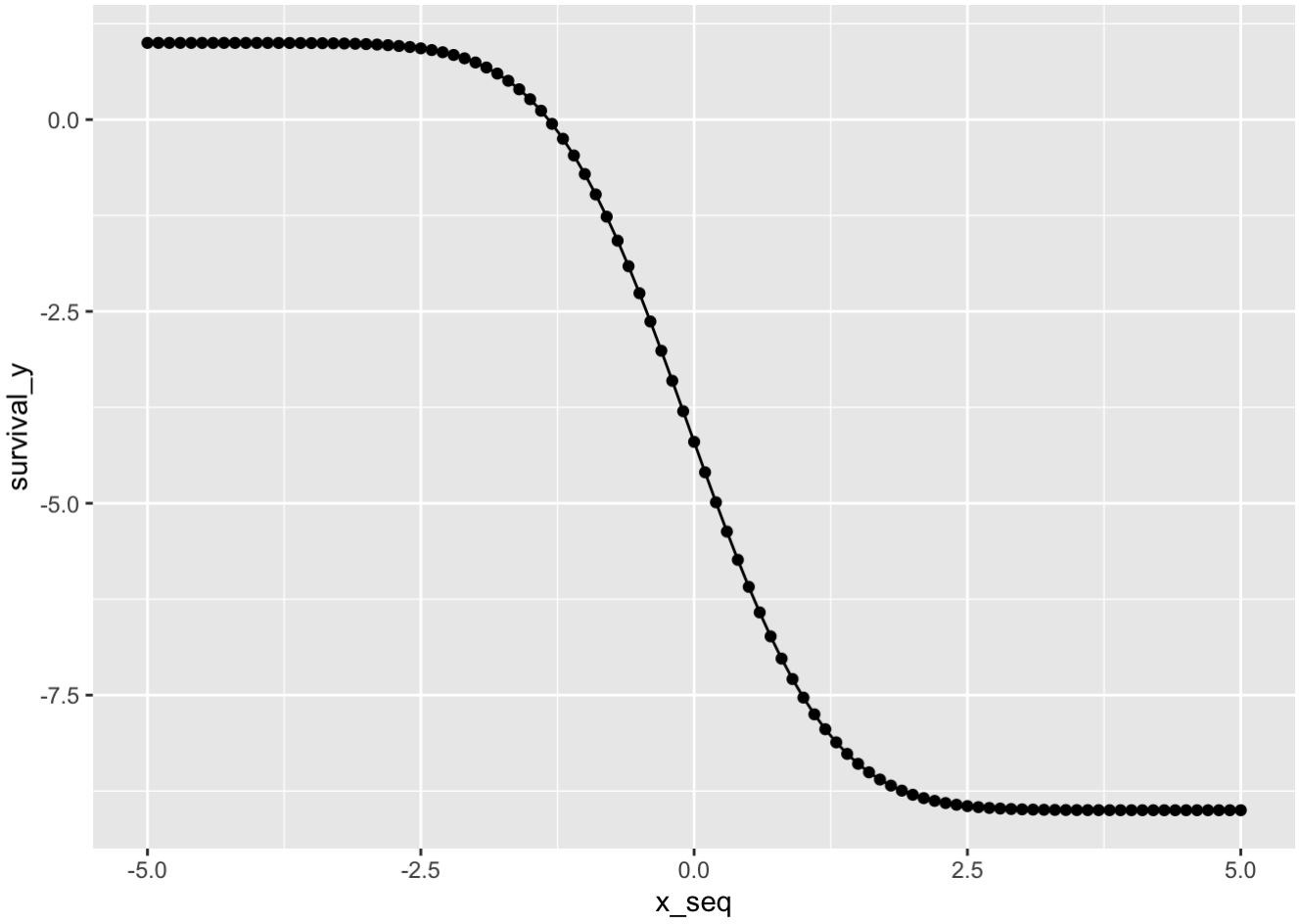

```
# plot CDF
cdf_y_df <- data.frame(x=x_seq,cdf_y=cdf_y)
ggplot(cdf_y_df,aes(x=x_seq,y=cdf_y)) + geom_line() + geom_point()
```




# 7

## Please explain in your own words what a survival function is.

```
# The x-axis is time. The y-axis is the surviving amount of the subjects.
# Because the death porpotion of the population is growing and accumulating along the time, which is a CDF,
# the surviving population is 1 - death porpotion of the population.
# Thus, a survival function can be represented as 1 - CDF.
survival_y <- rep(1,length(cdf_y))-cdf_y
survival_y_df <- data.frame(x=x_seq,survival_y=survival_y)
ggplot(survival_y_df,aes(x=x_seq,y=survival_y)) + geom_line() + geom_point()
```

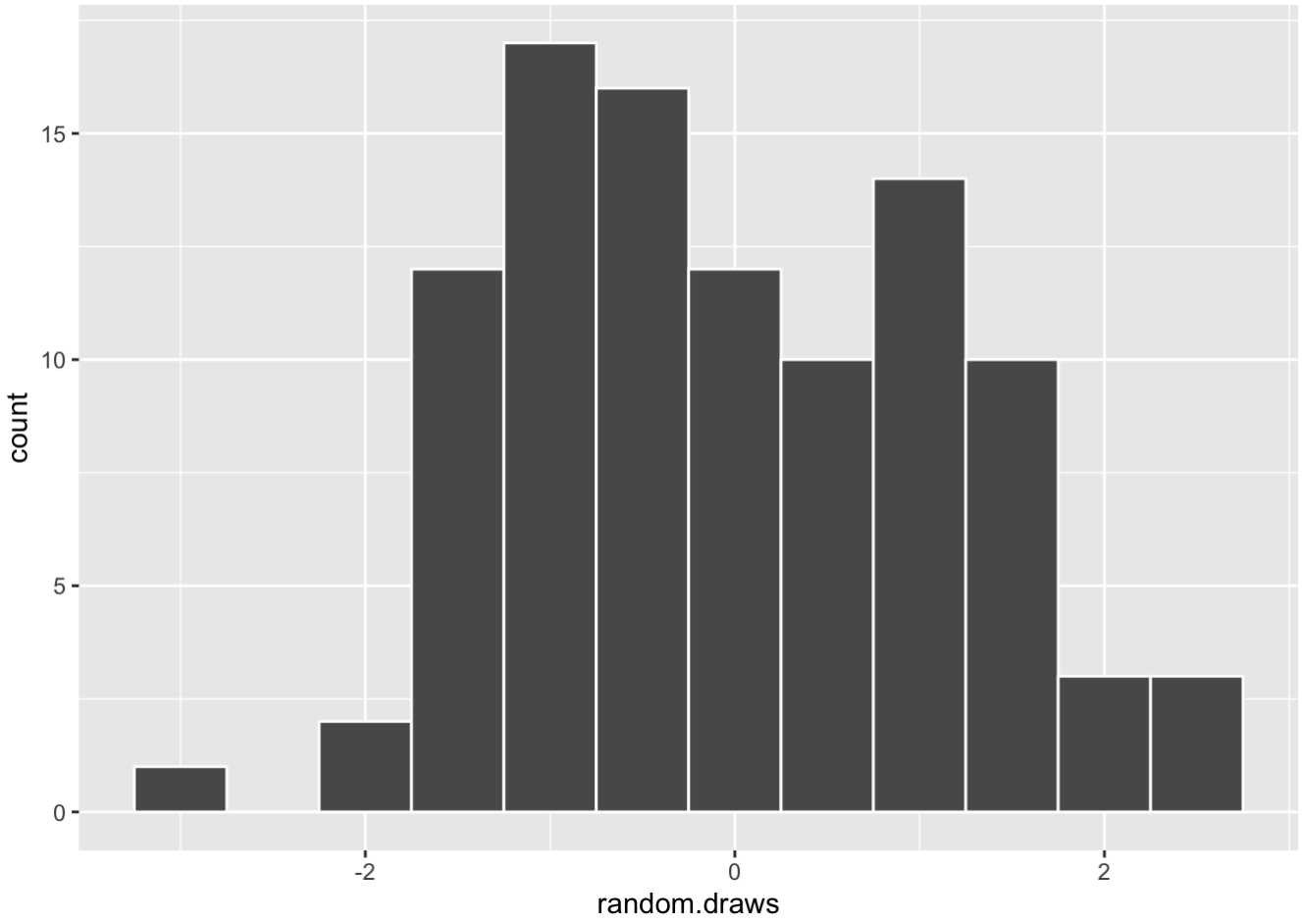



# 8

**Using the rnorm function, please make 100 random draws from the normal distribution and plot the results.**

**Compare it to your plots from questions 1 and 2.**

```
random.draws <- rnorm(100,0,1) # rnorm(n, mean = 0, sd = 1)
random.draws_df <- data.frame(cbind(1:length(random.draws),random.draws))
ggplot(random.draws_df, aes(x=random.draws)) + geom_histogram(color="white", binwidth
= 0.5)
```




```
# The result plot of Q8 is just 1 of the 1000 trials in Q2.
# The Q8 results shows what the result of one trial would be like.
# We can then calculate the amount of the random draws that pass our thershold (for e
xample, smaller than 0.2) and keep that number for this trial.
# Next, we keep doing this kind of trial (Q8) for 1000 times, the result would become
something like Q2.
```