CAAP Statistics - Lec07

R Session3

Jul 15, 2022

Review

- Normal Distribution
  - Standard Normal Distribution and Z-score
  - How to get normal probabilities?
  - How to get normal percentile?
- Bernoulli distribution
- Binomial distribution

Learning Objectives

- Binomial distribution(continued)
- Calculate the normal probabilities in R
- Calculate the normal percentile in R
- Simulate Bernoulli random variable
- Normal approximation of binomial distribution

Normal Probabilities and Percentiles

Load packages

```
library(openintro)
library(tidyverse)
library(ggplot2)
```

Load the data

Body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, are given for 507 physically active individuals - 247 men and 260 women.

```
head(bdims)
```

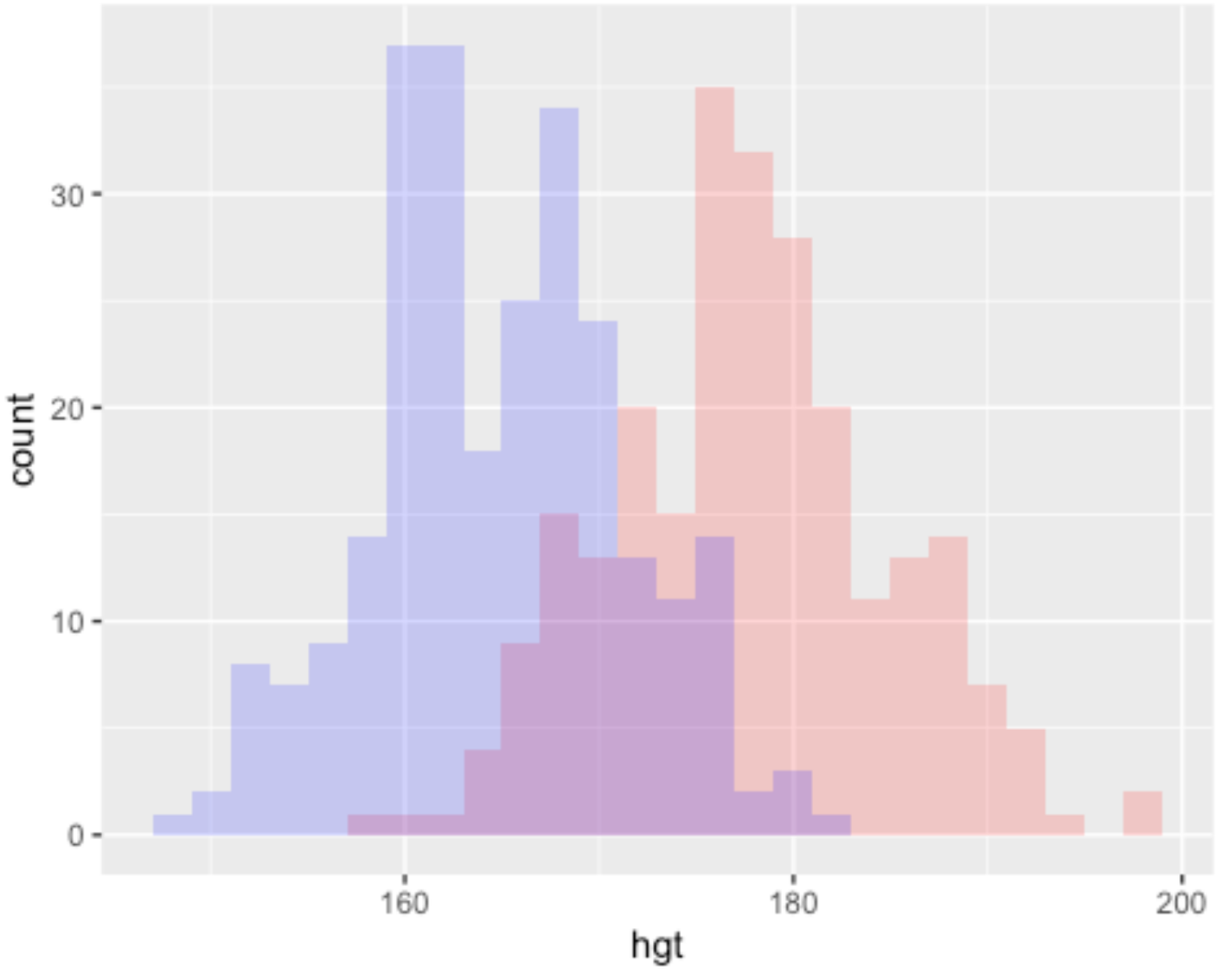
```
## # A tibble: 6 × 25
##   bia_di bii_di bit_di che_de che_di elb_di wri_di kne_di ank_di sho_gi che_gi
##    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1   42.9   26     31.5   17.7   28     13.1   10.4   18.8   14.1   106.   89.5
## 2   43.7   28.5   33.5   16.9   30.8   14     11.8   20.6   15.1   110.   97
## 3   40.1   28.2   33.3   20.9   31.7   13.9   10.9   19.7   14.1   115.   97.5
## 4   44.3   29.9   34     18.4   28.2   13.9   11.2   20.9   15     104.   97
## 5   42.5   29.9   34     21.5   29.4   15.2   11.6   20.7   14.9   108.   97.5
## 6   43.3   27     31.5   19.6   31.3   14     11.5   18.8   13.9   120.   99.9
## # … with 14 more variables: wai_gi <dbl>, nav_gi <dbl>, hip_gi <dbl>,
## #   thi_gi <dbl>, bic_gi <dbl>, for_gi <dbl>, kne_gi <dbl>, cal_gi <dbl>,
## #   ank_gi <dbl>, wri_gi <dbl>, age <int>, wgt <dbl>, hgt <dbl>, sex <int>
```

Filter out the data

```
male = bdims %>%
   filter(sex == 1) # Male
female = bdims %>%
  filter(sex ==0) # Female
```

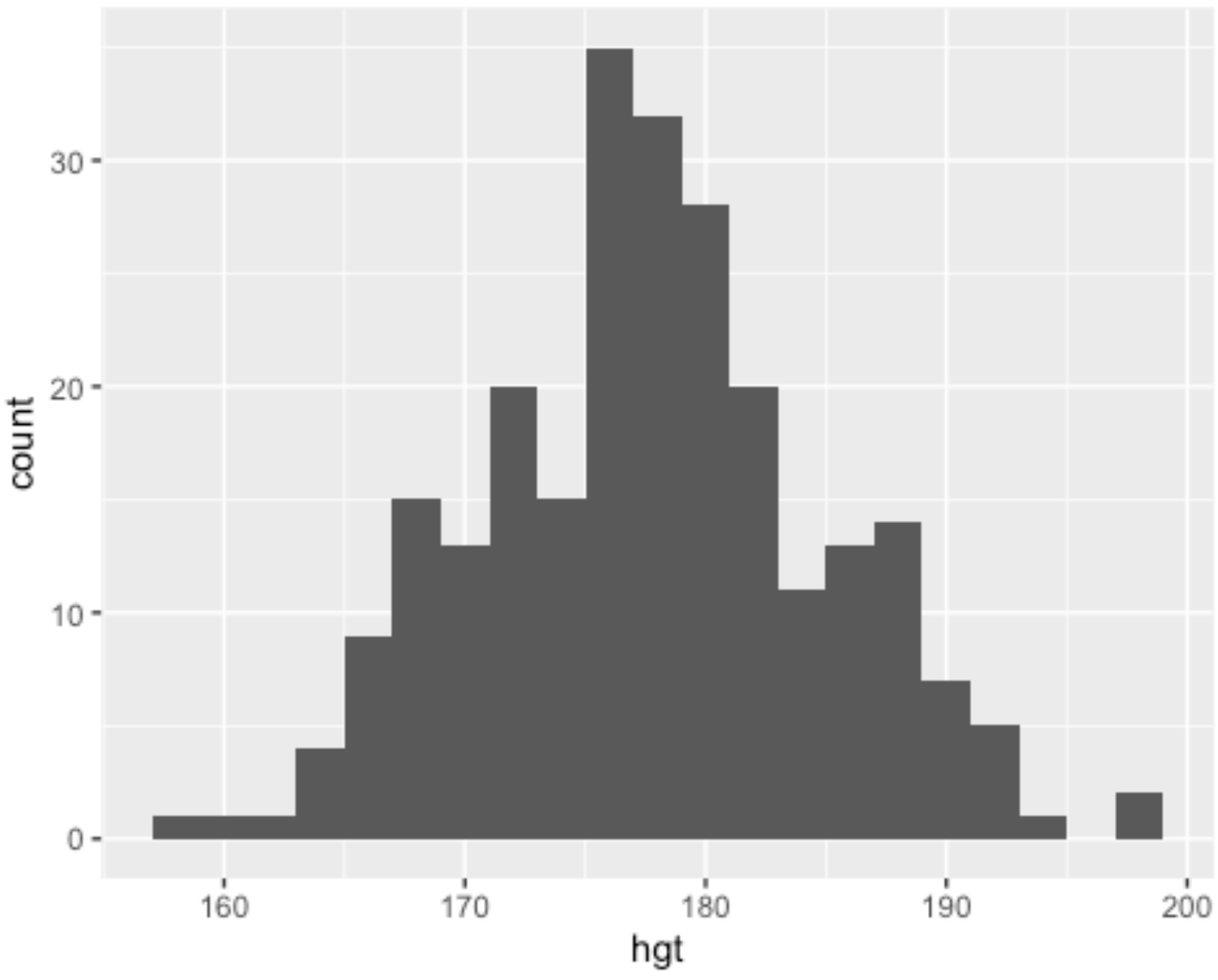
Visualize the distribution of hgt

```
ggplot(bdims,aes(x=hgt)) +
    geom_histogram(data=subset(bdims,sex == 1),fill = "red", alpha = 0.2, binwidth = 2) +
    geom_histogram(data=subset(bdims,sex == 0),fill = "blue", alpha = 0.2, binwidth = 2)
```



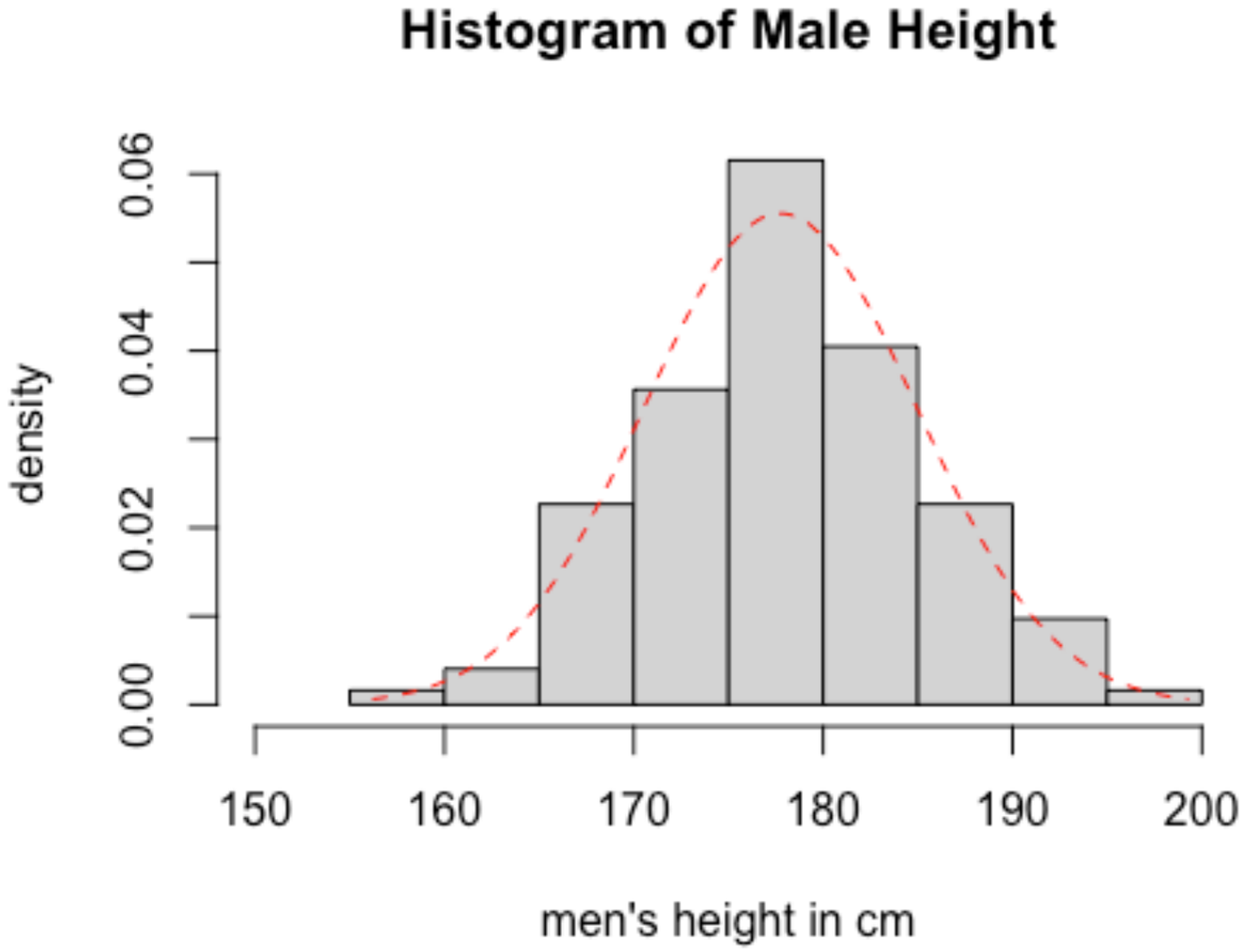
Let's focus on `male` data

```
ggplot(male, aes(x = hgt))+
  geom_histogram(binwidth = 2)
```



Normal Distribution

```
mu = mean(male$hgt)
sigma = sd(male$hgt)
x = seq(-3,3, length = 100)
hist(male$hgt, freq = FALSE, xlab="men's height in cm", ylab="density",
     main = "Histogram of Male Height", xlim = c(150, 200))
lines(x*sigma+mu, dnorm(x*sigma+mu, mean = mu, sd = sigma), col="red", lty=2)
```



Normal Probabilities

What is the probabilities of the male height being less than 175 cm?

```
(zscore = (175-mu)/sigma)
## [1] -0.3821668
pnorm(175, mean = mu, sd = sigma)
## [1] 0.3511688
pnorm(zscore)
## [1] 0.3511688
```

Normal Probabilities

What is the probabilities of the male height being taller than 185 cm?

```
(zscore = (185-mu)/sigma)
```

```
## [1] 1.009887
```

```
1-pnorm(185, mean = mu, sd = sigma)
```

```
## [1] 0.1562746
```

```
1-pnorm(zscore)
```

```
## [1] 0.1562746
```

Normal Percentile

The n-th percentile is defined as the value where n percent of the data are below its value. What is the 97% percentile of this distribution?

```
qnorm(0.97, mean = mu, sd = sigma)
## [1] 191.2563
qnorm(0.97)*sigma + mu
## [1] 191.2563
quantile(male$hgt, 0.97)
##     97%
 ## 191.43
```

Bernoulli Trial

Recall the Milgram experiment.

```
p = 0.35 #probability of success
outcome = c(1,0) # success or failure
sample(outcome, size = 1, prob = c(p, 1-p))
## [1] 0
```

```
nexp = 1000
set.seed(1004)
result_ber = matrix(0, nrow = nexp, ncol = 1)
for (i in 1:nexp){
  result_ber[i,] = sample(outcome, size = 1, prob = c(p,1-p))
}
table(result_ber)
## result_ber
##   0   1
## 675 325
```

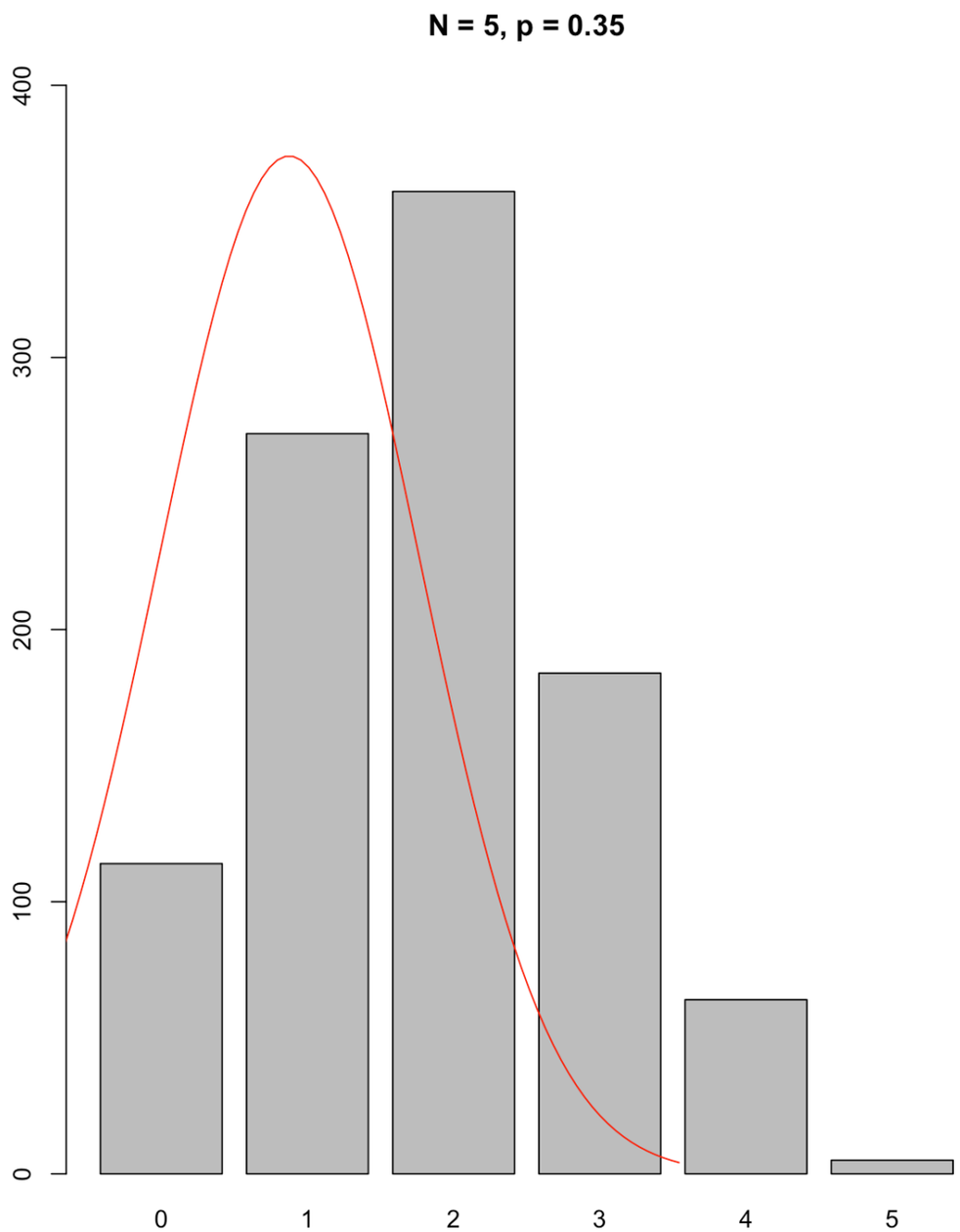
Binomial Trial - Normal Approximation

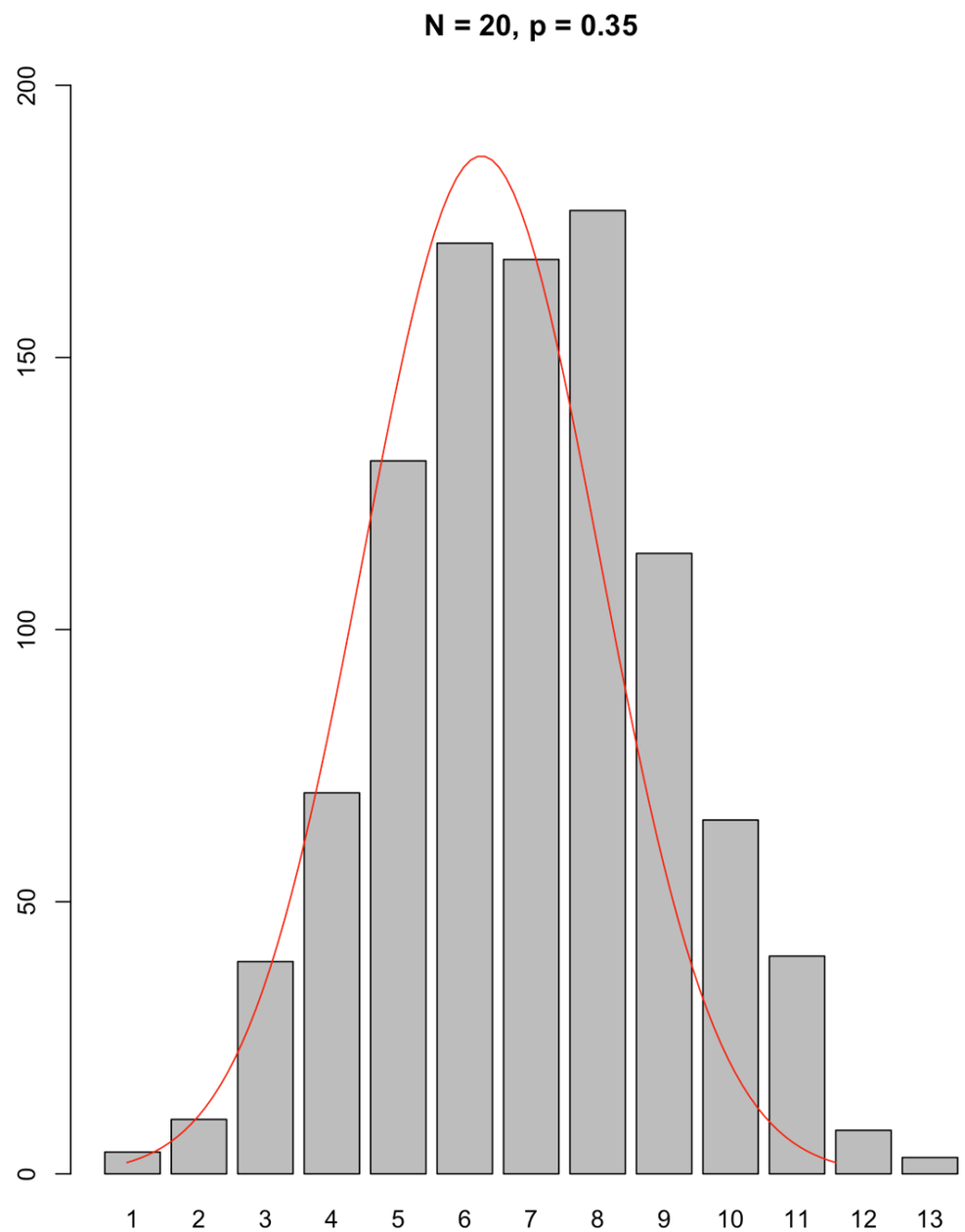
```
n_small = 5
n_large = 20
set.seed(1004)
result_bin_sm = matrix(0, nrow = nexp, ncol = n_small)
result_bin_lg = matrix(0, nrow = nexp, ncol = n_large)
for (i in 1:nexp){
  result_bin_sm[i,] = sample(outcome, size = n_small, prob =
c(p,1-p), replace = TRUE)
  result_bin_lg[i,] = sample(outcome, size = n_large, prob =
c(p,1-p), replace = TRUE)
}
```

Binomial - Normal Approximation(Code)

```
bin_dist_sm = table(rowSums(result_bin_sm))
bin_dist_large = table(rowSums(result_bin_lg))
par(mfrow=c(1,2))
barplot(bin_dist_sm, main = "N = 5, p = 0.35", ylim = c(0,400))
lines(x*sqrt(5*0.35*0.65)+5*0.35,
dnorm(x*sqrt(5*0.35*0.65)+5*0.35, 5*0.35,
sqrt(5*0.35*0.65))*nexp,col="red")
barplot(bin_dist_large, main = "N = 20, p = 0.35", ylim
=c(0,200))
lines(x*sqrt(20*0.35*0.65)+20*0.35,
dnorm(x*sqrt(20*0.35*0.65)+20*0.35, 20*0.35,
sqrt(20*0.35*0.65))*nexp,col="red")
```

Binomial - Normal Approximation(Plot)





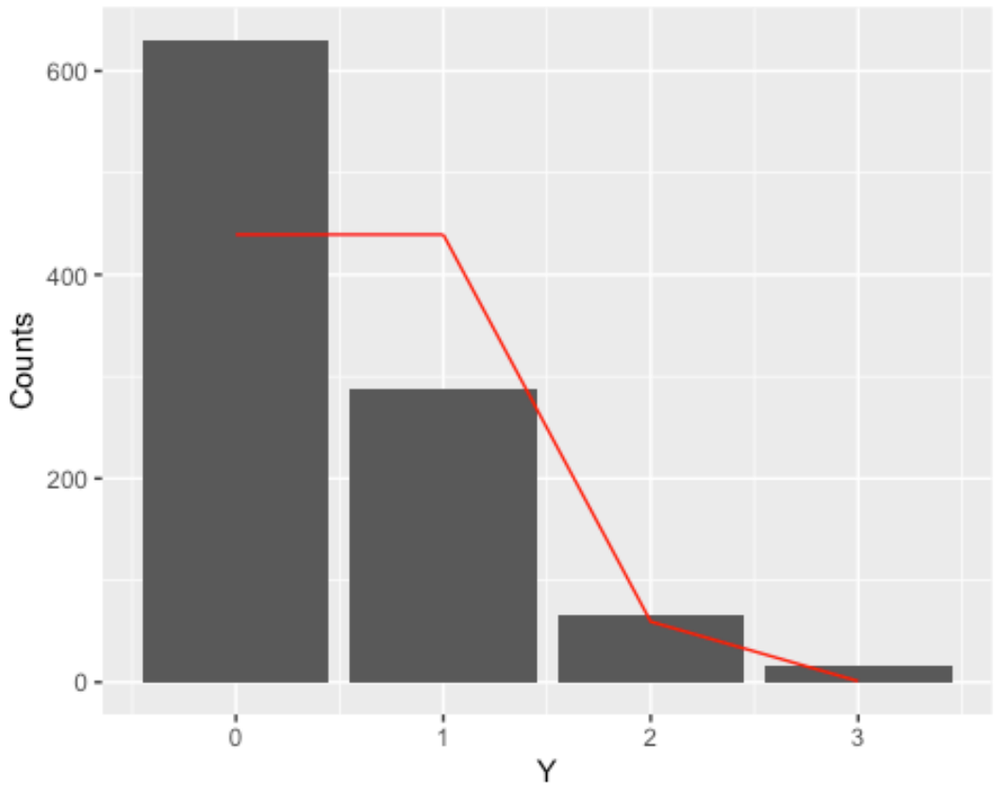
Poisson - Normal Approximation(Code)

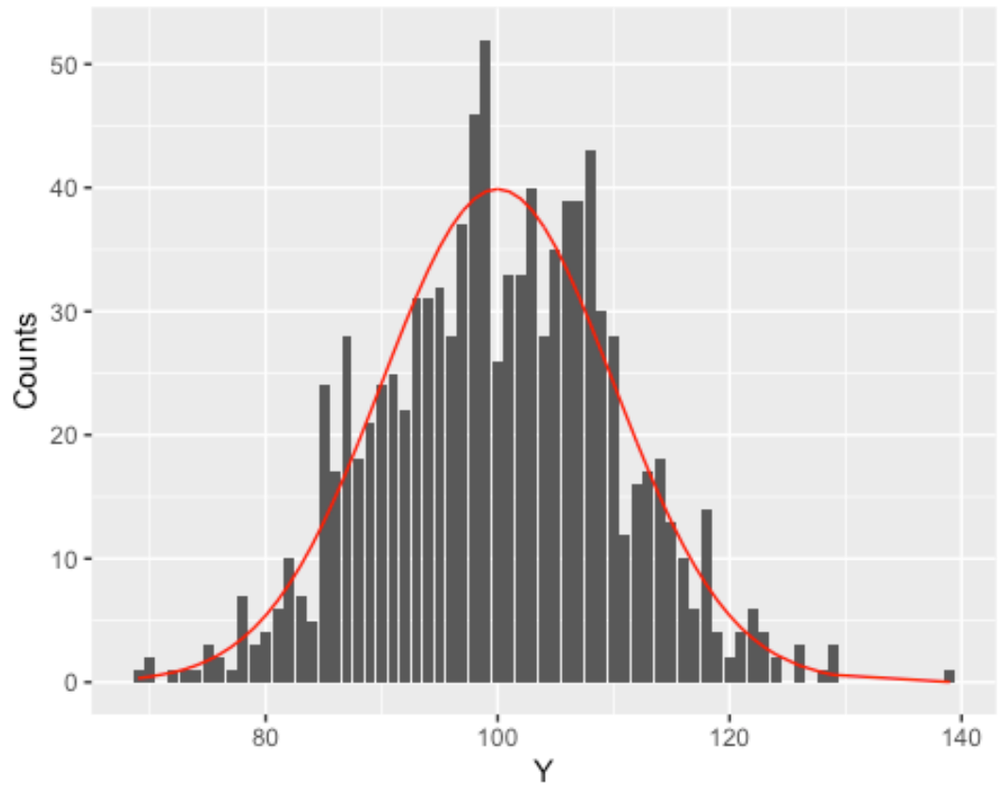
```
lambda_small = 0.5
lambda_large = 100
set.seed(1004)
result_poi_sm = data.frame(Y = rpois(nexp, lambda_small))
result_poi_lg = data.frame(Y = rpois(nexp, lambda_large))

result_poi_sm %>%
  ggplot()+
  geom_bar(aes(x=Y))+
  geom_line(aes(x = Y, dnorm(Y, mean = lambda_small, sd = sqrt(lambda_small))*nexp), color="red")+
  ylab(“Counts")

result_poi_lg %>%
  ggplot()+
  geom_bar(aes(x=Y))+
  geom_line(aes(x = Y, dnorm(Y, mean = lambda_large, sd = sqrt(lambda_large))*nexp), color="red")+
  ylab("Counts")
```

Poisson - Normal Approximation(Plot)





First Quiz on Thursday

- **Quiz**
  - OpenIntro Chapter 1-4
  - Lecture 1-9
  - Code from R sessions will be on the exam.
    You need to know the meaning of each code — eg: what does `sample(1:6, size =2)` do?
- Office hour from **7pm** via Zoom.