Probability

Aleksandr Fisher

4/20/2020

Sample Space and events

- Probability formalizes chance variation or uncertainty in outcomes.
  - It might rain or be sunny today, we don't know which.
  - To formalize, we need to define the set of possible outcomes.
- **Sample space:** Ω the set of possible outcomes.
- **Event:** any subset of outcomes in the sample space

What is probability?

$$\text{P(event)} = \frac{\text{number of outcomes in the event}}{\text{total number of outcomes in the sample space}}$$

- Consider tossing a fair coin:
  - There are two possible outcomes - tossing a head or tossing a tail.
  - The sample space is the set of all possible outcomes, so the sample space is $\{\text{H, T}\}$.
  - Since the coin is fair, the outcomes are equally likely. The probability of the event toss a head is 0.5 . In symbols, we can write this as $\text{P(H)} = 0.5$

Flipping Coins in R

- With an existing data structure, we want to build an underlying models.
- First argument is the number of trials, second is number of coins, and third is probability of a positive (heads) outcome

```
rbinom (10, 1, .5)
```

```
##  [1] 1 0 1 1 0 1 1 0 1 0
```

```
rbinom (10, 10, .5)
```

```
##  [1] 1 6 7 5 4 7 7 5 3 3
```

```
rbinom (10, 10, .8)
```

```
##  [1]  8  9  8 10  9  8  9  8 10 10
```

Flipping Coins in R

- Generate 100 occurrences of flipping 10 coins, each with 30% probability

```
#Generate 100 occurrences of flipping 10 coins, each with 3
rbinom(100, 10, 0.3)
```

```
##   [1] 0 4 4 3 4 3 7 1 3 3 1 2 2 2 3 5 2 3 1 3 2 1 4 3 3
##  [38] 3 2 3 2 2 2 4 0 2 3 2 3 1 6 1 3 3 2 5 1 4 5 3 2 4
##  [75] 2 4 1 3 5 4 3 4 1 2 1 1 6 4 0 2 2 2 1 3 3 2 4 6 2
```

- The two latter result tell us the number of heads outcomes in the series.

Density and Cumulative Density

- One can use the `dbinom()` function. This function takes almost the same arguments as `rbinom()`. The second and third arguments are `size` and `prob`, but now the first argument is `x` instead of `n`. Use `x` to specify where you want to evaluate the binomial density.
- Confirm your answer using the `rbinom()` function by creating a simulation of `10,000` trials. Put this all on one line by wrapping the `mean()` function around the rbinom() function.
- If you flip 10 coins each with a 30% probability of coming up heads, what is the probability exactly 2 of them are heads?
- Calculate the probability that at least five coins are heads. Note that you can compute the probability that the number of heads is less than or equal to 4, then take 1 - that probability.

Density and Cumulative Density

- The `dbinom()` function takes almost the same arguments as rbinom(). The second and third arguments are size and prob, but now the first argument is x instead of n. Use x to specify where you want to evaluate the binomial density.

```
# Calculate the probability that 2 are heads using dbinom
dbinom(2, 10, .3)
```

```
## [1] 0.2334744
```

```
# Confirm your answer with a simulation using rbinom.
mean(rbinom(10000, 10, .3) == 2)
```

```
## [1] 0.2336
```

Density and Cumulative Density

- If you flip ten coins that each have a 30% probability of heads, what is the probability at least five are heads?

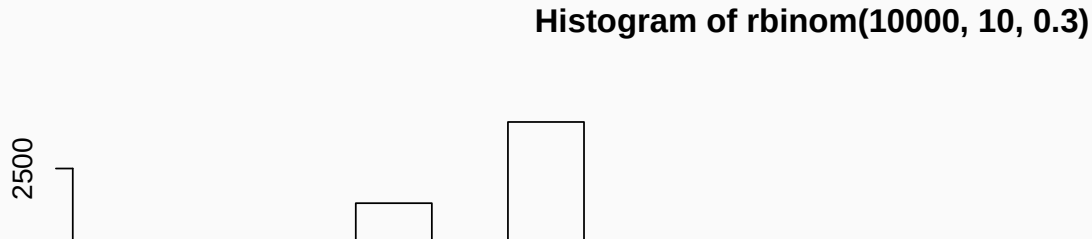
```
# Calculate the probability that 5 are heads using pbinom
pbinom(5, 10, 0.7)
```

```
## [1] 0.1502683
```

```
# Confirm your answer with a simulation of 10,000 trials
mean(rbinom(10000, 10, 0.3) >=5)
```

```
## [1] 0.1537
```

```
hist(rbinom(10000, 10, 0.3))
```



Density and Cumulative Density

-If you flip ten coins that each have a 30% probability of heads, what is the probability at least five are heads?

```
# Try now with 100, 1000, 10,000, and 100,000 trials
mean(rbinom(100, 10, .3) >= 5)
```

```
## [1] 0.14
```

```
mean(rbinom(1000, 10, .3) >= 5)
```

```
## [1] 0.15
```

```
mean(rbinom(10000, 10, .3) >= 5)
```

```
## [1] 0.1512
```

```
mean(rbinom(100000, 10, .3) >= 5)
```

```
## [1] 0.15008
```

Expected Values and Variance

- Most of the time we want to know what the expected value of a distribution is and its variance.
- The expected value of the binomial is the mean of the distribution.

$$E(X) = size * p$$

- The variance is defined as: The average of the squared differences from the Mean.
- To calculate the variance follow these steps: Work out the Mean (the simple average of the numbers) Then for each number: subtract the Mean and square the result (the squared difference).

Expected Values and Variance

- What is the expected value of a binomial distribution where 25 coins are flipped, each having a 30% chance of heads?

```
#Calculate the expected value using the exact formula
print(25*0.3)
```

```
## [1] 7.5
```

```
# Confirm with a simulation using rbinom
X<-rbinom(10000,25, 0.3)
mean(X)
```

```
## [1] 7.496
```

```
# Calculate the variance using the exact formula
print(25*0.3*0.7)
```

```
## [1] 5.25
```

Probability of event A and event B

- Coin represents a yes or no outcome - very common in political science research. What we want to know are the mathmatical laws surrounding random events in order to make better predictions about outcomes we care about.
- Probability of A * Probability of B. This only holds true for independent events (which may or may not be realistic in poltical science)
- If events A and B are independent, and A has a 40% chance of happening, and event B has a 20% chance of happening, what is the probability they will both happen?

```
# Simulate 100,000 flips of a coin with a 40% chance of he
A <- rbinom(100000, 1, 0.4)

# Simulate 100,000 flips of a coin with a 20% chance of he
```

Probability of event A and event B

- Randomly simulate 100,000 flips of A (40% chance), B (20% chance), and C (70% chance). What fraction of the time do all three coins come up heads?

```
# You've already simulated 100,000 flips of coins A and B
A <- rbinom(100000, 1, .4)
B <- rbinom(100000, 1, .2)

# Simulate 100,000 flips of coin C (70% chance of heads)
C <- rbinom(100000, 1, .7)

# Estimate the probability A, B, and C are all heads
mean(A&B&C)
```

```
## [1] 0.05648
```

Probability of event A or event B

- Probability of A + Probability of B - (Probability of A & B)
- Think about this as overlapping circles in a ven diagram.
- Pr(A or B) = Pr(A) + Pr(B) - Pr(A and B)
- Pr(A or B) = Pr(A) + Pr(B) - Pr(A) X Pr(B)
- If coins A and B are independent, and A has a 60% chance of coming up heads, and event B has a 10% chance of coming up heads, what is the probability either A or B will come up heads?

Probability of event A or event B

- In the last exercise you found that there was a ____ chance that either coin A (60% chance) or coin B (10% chance) would come up heads. Now you'll confirm that answer using simulation.

```
# Simulate 100,000 flips of a coin with a 60% chance of he
A <- rbinom(100000, 1, 0.6)

# Simulate 100,000 flips of a coin with a 10% chance of he
B <- rbinom(100000, 1, 0.1)

# Estimate the probability either A or B is heads
mean (A | B)
```

```
## [1] 0.63943
```

Probability of event A or event B

- Suppose X is a random Binom(10, .6) variable (10 flips of a coin with 60% chance of heads) and Y is a random Binom(10, .7) variable (10 flips of a coin with a 70% chance of heads), and they are independent.
- What is the probability that either of the variables is less than or equal to 4?

```
# Use rbinom to simulate 100,000 draws from each of X and
X <- rbinom(100000, 10, .6)
Y <- rbinom(100000, 10, .7)

# Estimate the probability either X or Y is <= to 4
mean(X <= 4 | Y <= 4)
```

```
## [1] 0.20675
```

```
# Use pbinom to calculate the probabilities separately
```

Multiplying Random Variables

$X \sim Binomial(10, .5)$

```
X <- rbinom(100000, 10, .6)
```

$Y \sim 3 * x$

```
Y<-3*X
```

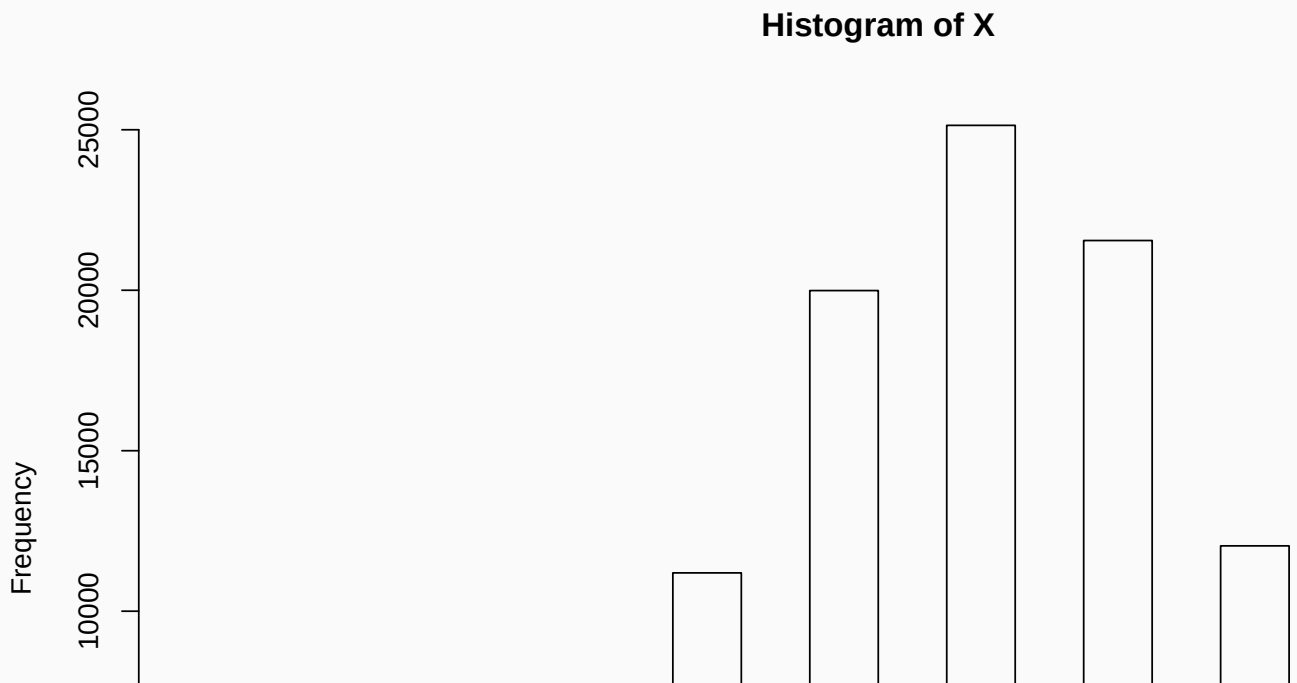
Multiplying Random Variables

$E[k * x] = k * E[X]$

$Var(k * x) = k^2 * Var(X)$

- Compare the historgrams of the two figures. Both the expected value and the variance should increase.

```
hist(X)
```



Multiplying Random Variables

```r
# Simulate 100,000 draws of a binomial with size 20 and p =
X <- rbinom(100000, 20, 0.1)

# Estimate the expected value of X
mean(X)
```

```
## [1] 2.00855
```

```r
# Estimate the expected value of 5 * X
Y= 5*X
mean(Y)
```

```
## [1] 10.04275
```

```r
# X is simulated from 100,000 draws of a binomial with size
X <- rbinom(100000, 20, .1)
```

Adding two random variables

$$X \ Binomial(10, .5)$$

$$Y \ Binomial(100, .2)$$

$$Z \sim X + Y$$

- Z is both larger and more spread out

$$E[X + Y] = E[X] + E[Y]$$

$$Var[X + Y] = Var[X] + Var[Y]$$

Adding two random variables

- If X is drawn from a binomial with size 20 and p = .3, and Y from size 40 and p = .1, what is the expected value (mean) of X + Y?

```
# Simulate 100,000 draws of X (size 20, p = .3) and Y (size
X <- rbinom(100000, 20, 0.3)
Y <-rbinom(100000, 40, 0.1)

# Estimate the expected value of X + Y
mean(X+Y)
```

```
## [1] 10.00815
```

```
# Simulation from last exercise of 100,000 draws from X and
X <- rbinom(100000, 20, .3)
Y <- rbinom(100000, 40, .1)
```

Updating with evidence

Updating is the heart of Bayesian Statistics. If we see 14 heads out of 20, how likely is that outcome? How likely is it that the coin is biased?

- Suppose you have a coin that is equally likely to be fair (50% heads) or biased (75% heads). You then flip the coin 20 times and see 11 heads.
- Without doing any math, which do you now think is more likely- that the coin is fair, or that the coin is biased?

Updating with evidence

- We see 11 out of 20 flips from a coin that is either fair (50% chance of heads) or biased (75% chance of heads). How likely is it that the coin is fair?

```
# Simulate 50000 cases of flipping 20 coins from fair and
fair <- rbinom(50000, 20, 0.5)
biased <- rbinom(50000, 20, 0.75)

# How many fair cases, and how many biased, led to exactly
fair_11 <- sum(fair==11)
biased_11 <- sum(biased==11)

# Find the fraction of fair coins that are 11 out of all c

#This is the posterior probability that a coin with 11/20
fair_11/(fair_11+biased_11)
```

Updating with evidence

- Suppose that when you flip a different coin (that could either be fair or biased) 20 times, you see 16 heads.
- Without doing any math, which do you now think is more likely- that this coin is fair, or that it's biased?
- We see 16 out of 20 flips from a coin that is either fair (50% chance of heads) or biased (75% chance of heads). How likely is it that the coin is fair?

```
# Simulate 50000 cases of flipping 20 coins from fair and
fair <-rbinom(50000, 20, 0.5)
biased <-rbinom(50000, 20, 0.75)

# How many fair cases, and how many biased, led to exactly
fair_16 <- sum(fair==16)
biased_16 <- sum(biased==16)
```

Prior Probability

- We see 14 out of 20 flips are heads, and start with a 80% chance the coin is fair and a 20% chance it is biased to 75%.
- You'll solve this case with simulation, by starting with a "bucket" of 10,000 coins, where 8,000 are fair and 2,000 are biased, and flipping each of them 20 times.

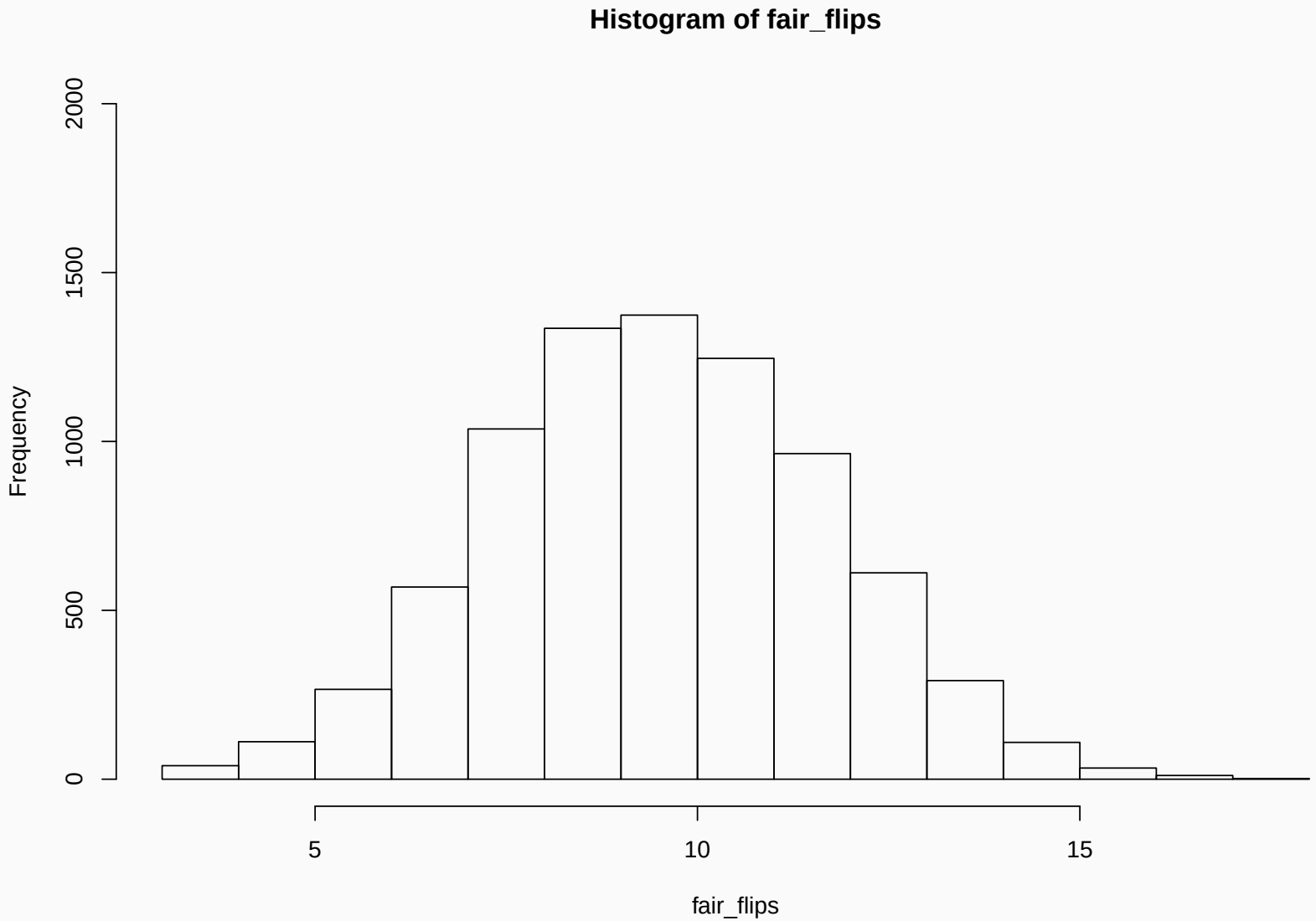
```
# Simulate 8000 cases of flipping a fair coin, and 2000 of
fair_flips <- rbinom (8000, 20, 0.5)
biased_flips <-rbinom (2000, 20, 0.75)

# Find the number of cases from each coin that resulted in
fair_14 <- sum(fair_flips==14)
biased_14 <-sum(biased_flips==14)

# Use these to estimate the posterior probability
```

Updating with evidence

```
hist(fair_flips, ylim=c(0,2000))
```



Updating with evidence

- Suppose instead of a coin being either fair or biased, there are three possibilities: that the coin is fair (50% heads), low (25% heads), and high (75% heads). There is a 80% chance it is fair, a 10% chance it is biased low, and a 10% chance it is biased high.
- You see 14/20 flips are heads. What is the probability that the coin is fair?
- Use the `rbinom()` function to simulate 80,000 draws from the fair coin, 10,000 draws from the high coin, and 10,000 draws from the low coin, with each draw containing 20 flips. Save them as `flips_fair`, `flips_high`, and `flips_low`, respectively.

```
# Simulate 80,000 draws from fair coin, 10,000 from each o
flips_fair <- rbinom(80000, 20, 0.5)
```

Bayes Theorem

$$Pr(14Heads|Fair) * Pr(Fair)$$

$$Pr(14Heads|Biased) * Pr(Biased)$$

$$Pr(Biased|14Heads) = \frac{Pr(14HeadsandBiased)}{Pr(14HeadsandBiased) + Pr(14HeadsandFair)}$$

Bayes Theorem

- More Abstractly we want to find the Probability of event A given event B or . . .

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A) + Pr(B|notA)Pr(notA)}$$

- Applying this to our old example where:
- A = Biased
- B = 14 Heads

Updating with evidence

```
# Use dbinom to calculate the probability of 11/20 heads w
probability_fair <- dbinom(11, 20, 0.5)
probability_biased <-dbinom(11, 20, 0.75)

# Calculate the posterior probability that the coin is fai
probability_fair/(probability_fair+probability_biased)
```

```
## [1] 0.8554755
```

```
# Find the probability that a coin resulting in 14/20 is f

probability_fair <- dbinom(14, 20, 0.5)
probability_biased <-dbinom(14, 20, 0.75)

probability_fair/(probability_fair+probability_biased)
```



```
## [1] 0.170211
```

Updating with evidence

- Now you'll find, using the `dbinom()` approach, the posterior probability if there were two other outcomes.

```
# Find the probability that a coin resulting in 18/20 is fa
probability_fair <- dbinom(18, 20, 0.5)
probability_biased <-dbinom(18, 20, 0.75)

probability_fair/(probability_fair+probability_biased)
```

```
## [1] 0.002699252
```

Updating with evidence

- Suppose we see 16 heads out of 20 flips, which would normally be strong evidence that the coin is biased. However, suppose we had set a prior probability of a 99% chance that the coin is fair (50% chance of heads), and only a 1% chance that the coin is biased (75% chance of heads).
- You'll solve this exercise by finding the exact answer with dbinom() and Bayes' theorem. Recall that Bayes' theorem looks like:

$$Pr(fair|A) = \frac{Pr(A|fair)Pr(fair)}{Pr(A|fair)Pr(fair) + Pr(A|biased)Pr(biased)}$$

```
# Use dbinom to find the probability of 16/20 from a fair c
probability_16_fair <-dbinom(16, 20, 0.5)
probability_16_biased <-dbinom(16, 20, 0.75)
```

Normal Distribution

- When you draw from the binomial with a large size, you approximate a normal distribution.
- Also known as Guassian distribution or bell curve. Measurement errors in scientific experiements take this shape.

$$X \sim Normal(\mu, \sigma)$$

$$\sigma = \sqrt{Var(X)}$$

$$\mu = size * p$$

$$\sigma = \sqrt{size * p * (1 - p)}$$

Normal Distribution

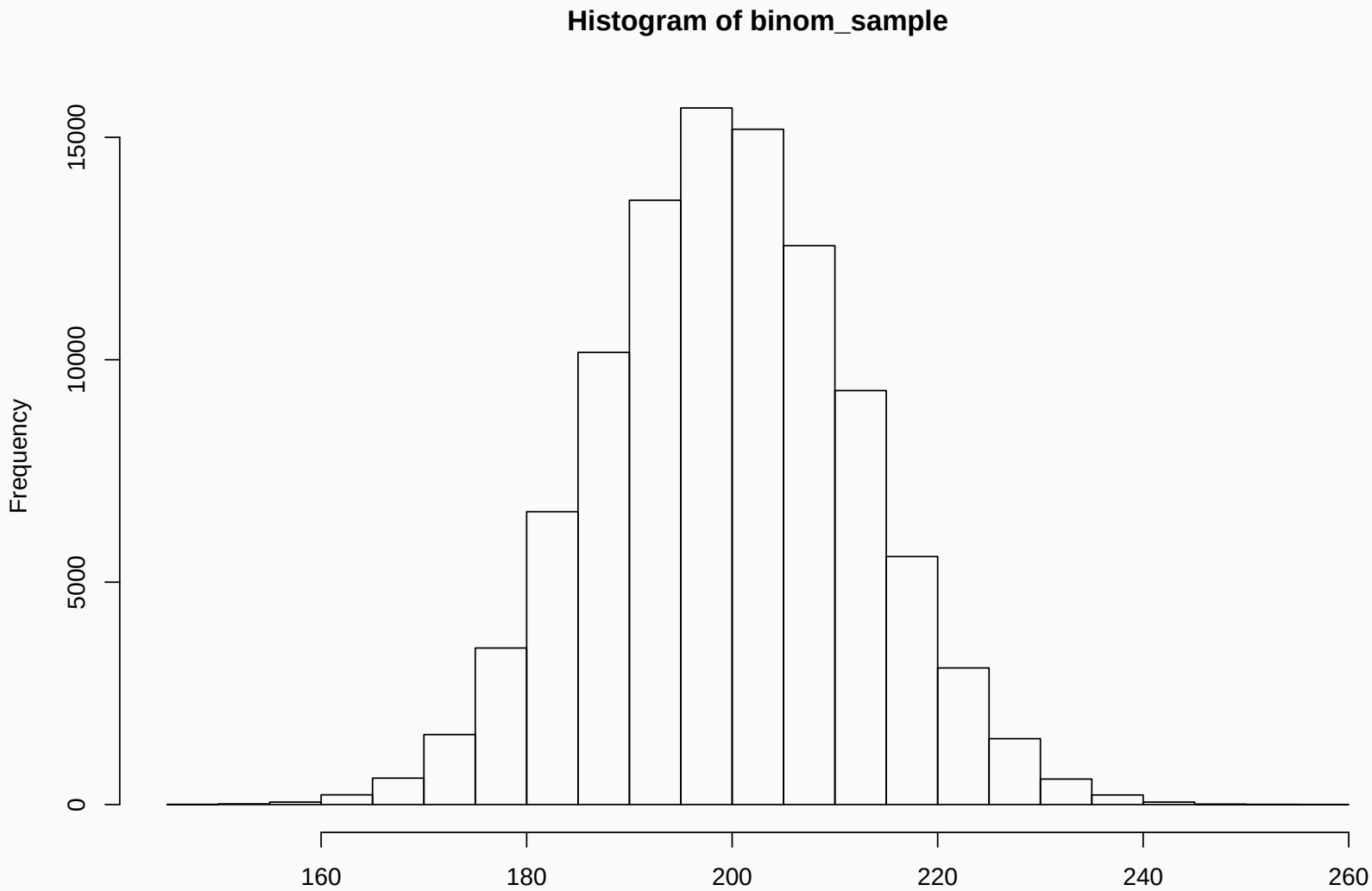
- Suppose you flipped 1000 coins, each with a 20% chance of being heads. What would be the mean and variance of the binomial distribution?
- In this exercise you'll see for yourself whether the normal is a reasonable approximation to the binomial by simulating large samples from the binomial distribution and its normal approximation and comparing their histograms.

```
# Draw a random sample of 100,000 from the Binomial(1000,
binom_sample <- rbinom(100000, 1000, 0.2)

# Draw a random sample of 100,000 from the normal approxim
expected_value <- 1000*0.2
variance <- 1000*0.2*0.8
stdev <- sqrt(variance)
```

Normal Distribution

```r
# Compare the two distributions with the compare_histograms
hist(binom_sample)
```



Normal Distribution

- If you flip 1000 coins that each have a 20% chance of being heads, what is the probability you would get 190 heads or fewer?
- You'll get similar answers if you solve this with the binomial or its normal approximation. In this exercise, you'll solve it both ways, using both simulation and exact calculation
- A binomial distribution is different from a normal distribution, and yet if the sample size is large enough, the shapes will be quite similar.

```
# Simulations from the normal and binomial distributions
binom_sample <- rbinom(100000, 1000, .2)
normal_sample <- rnorm(100000, 200, sqrt(160))

# Use binom_sample to estimate the probability of <= 190 h
```

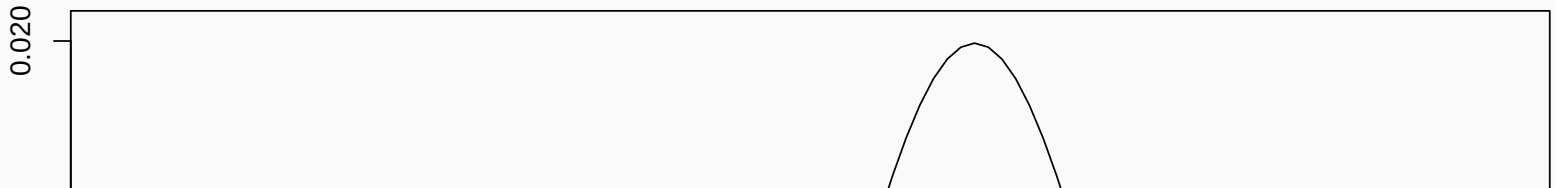
Normal Distribution

```r
#Probability of having a value *lower* than 90 in a normal

pnorm(q = 90, mean = 124, sd = 20, lower.tail = TRUE)

## [1] 0.04456546

curve(dnorm(x, mean = 124, sd = 20), xlim = c(0, 200))
abline(h = 0)
sequence <- seq(0, 90, 0.1)
polygon(x = c(sequence,90,0),
        y = c(dnorm(c(sequence),124,20),0,0),
        col = "grey")
```



Normal Distribution

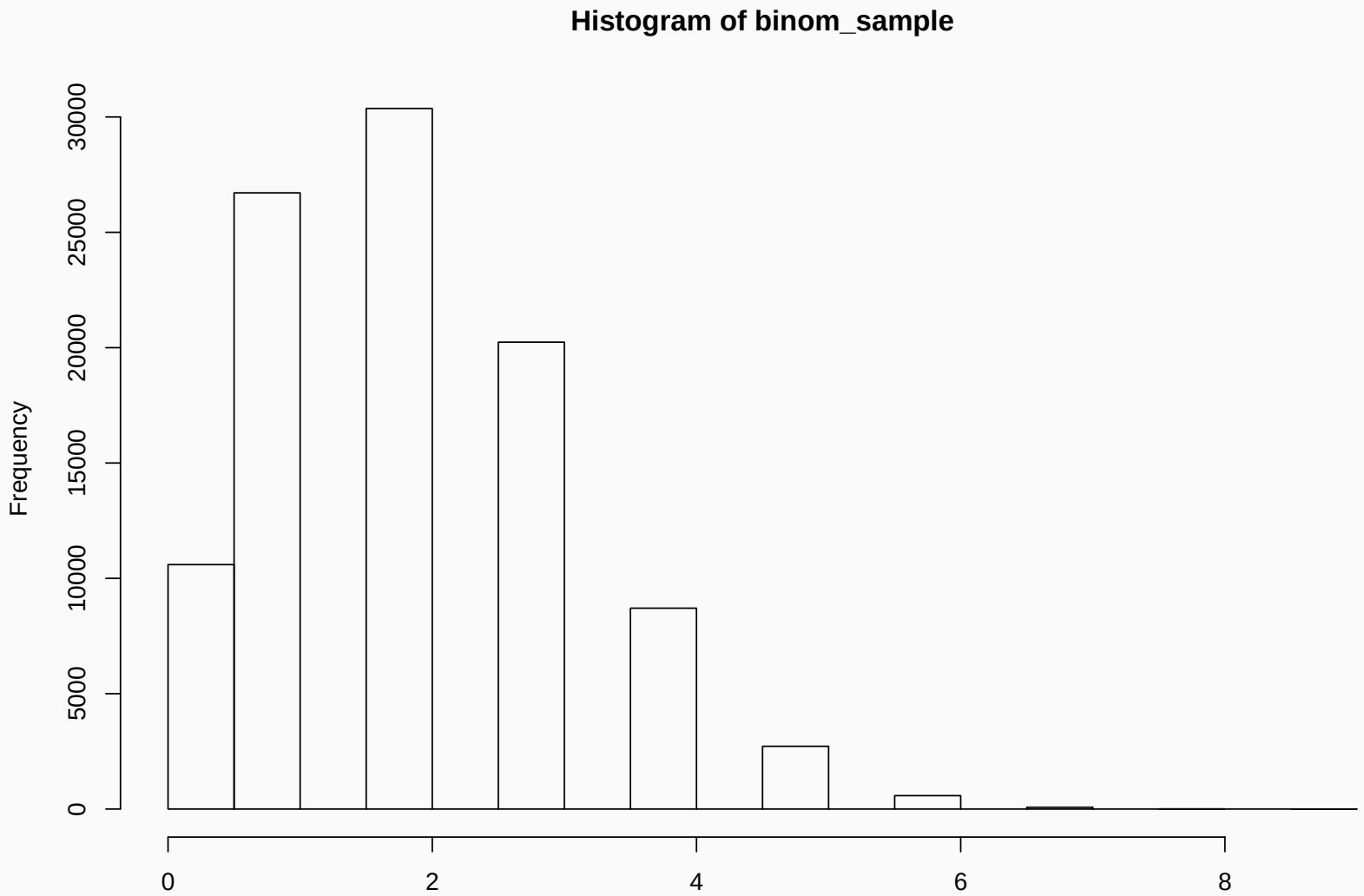
- When we flip a lot of coins, it looks like the normal distribution is a pretty close approximation. What about when we flip only 10 coins, each still having a 20% chance of coming up heads? Is the normal still a good approximation?

```
# Draw a random sample of 100,000 from the Binomial(10, .2)
binom_sample <- rbinom(100000, 10, .2)

# Draw a random sample of 100,000 from the normal approxim
normal_sample <- rnorm(100000, 2, sqrt(1.6))
```

Normal Distribution

```r
# Compare the two distributions with the compare_histograms
hist(binom_sample)
```



Normal distribution

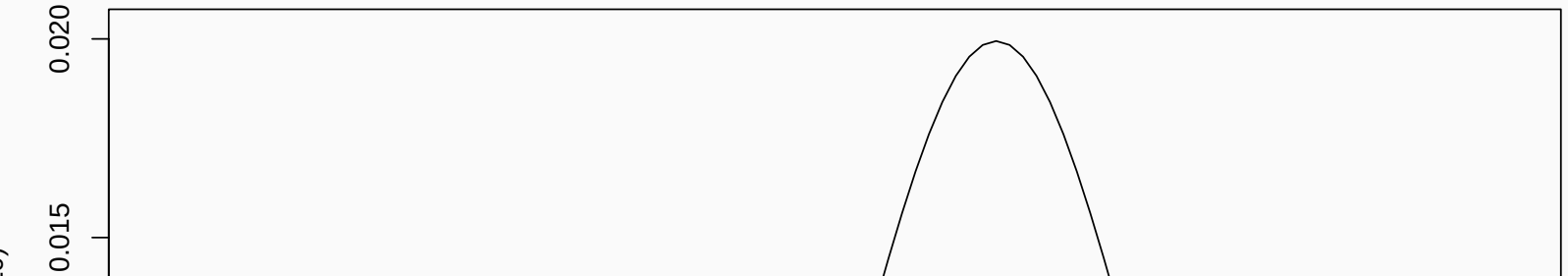
```r
# Density at 90 in a normal distribution with a mean of 12
dnorm(x = 90, mean = 124, sd = 20)
```

```
## [1] 0.004702454
```

```r
# Density at c(10,20,30) in a normal distribution with a m
dnorm(x = c(10,20,30), mean = 124, sd = 20)
```

```
## [1] 1.756978e-09 2.680518e-08 3.184913e-07
```

```r
# Graphing the distribution: Do not give a value to x, the
curve(dnorm(x, mean = 124, sd = 20), xlim = c(0, 200))
```



Poisson distribution

- Used for when N is large and probability is small. This is useful for both count data and rare events.

$$X \sim Poisson(\lambda)$$

$$E[X] = \lambda$$

- Modeling how many protesters, number of events, etc.
- If you were drawing from a binomial with size = 1000 and p = .002, what would be the mean of the Poisson approximation?

Poisson distribution

- A random variable may follow a Poisson distribution if the event being considered is rare, the population is large, and the events occur independently of each other.
- If we were flipping 100,000 coins that each have a .2% chance of coming up heads, you could use a Poisson(2) distribution to approximate it. Let's check that through simulation.

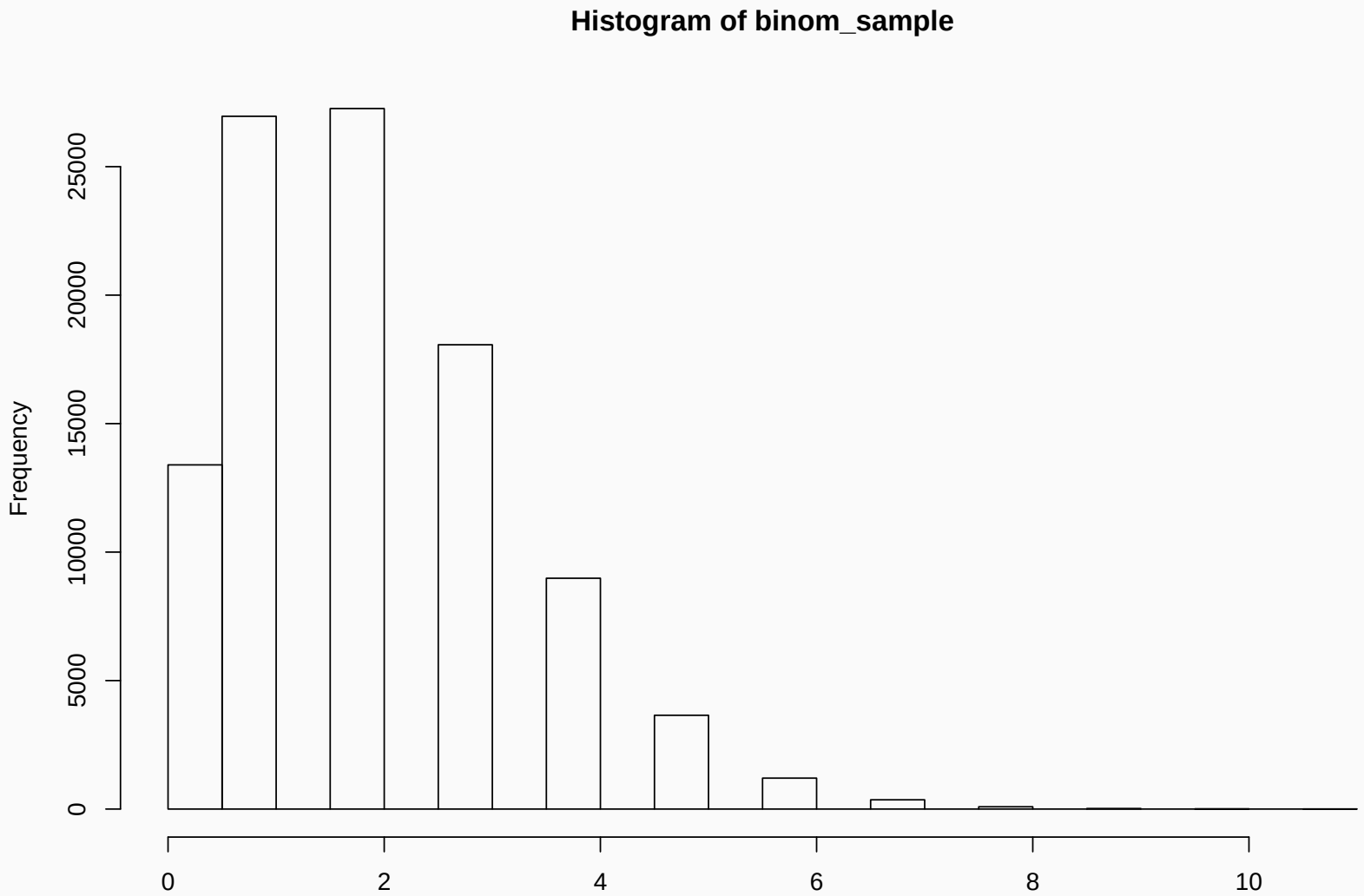
```
#flipping many coins, each with low distribution (N is lar

# Draw a random sample of 100,000 from the Binomial(1000,
binom_sample <- rbinom(100000, 1000, 2/1000)

# Draw a random sample of 100,000 from the Poisson approxi
poisson_sample <- rpois(100000, 2)
```

Poisson distribution

```r
# Compare the two distributions with the compare_histograms
hist(binom_sample)
```



Poisson distribution

- In this exercise you'll find the probability that a Poisson random variable will be equal to zero by simulating and using the `dpois()` function, which gives an exact answer.

```
# Simulate 100,000 draws from Poisson(2)
poisson_sample <- rpois(100000, 2)

# Find the percentage of simulated values that are 0
mean(poisson_sample==0)
```

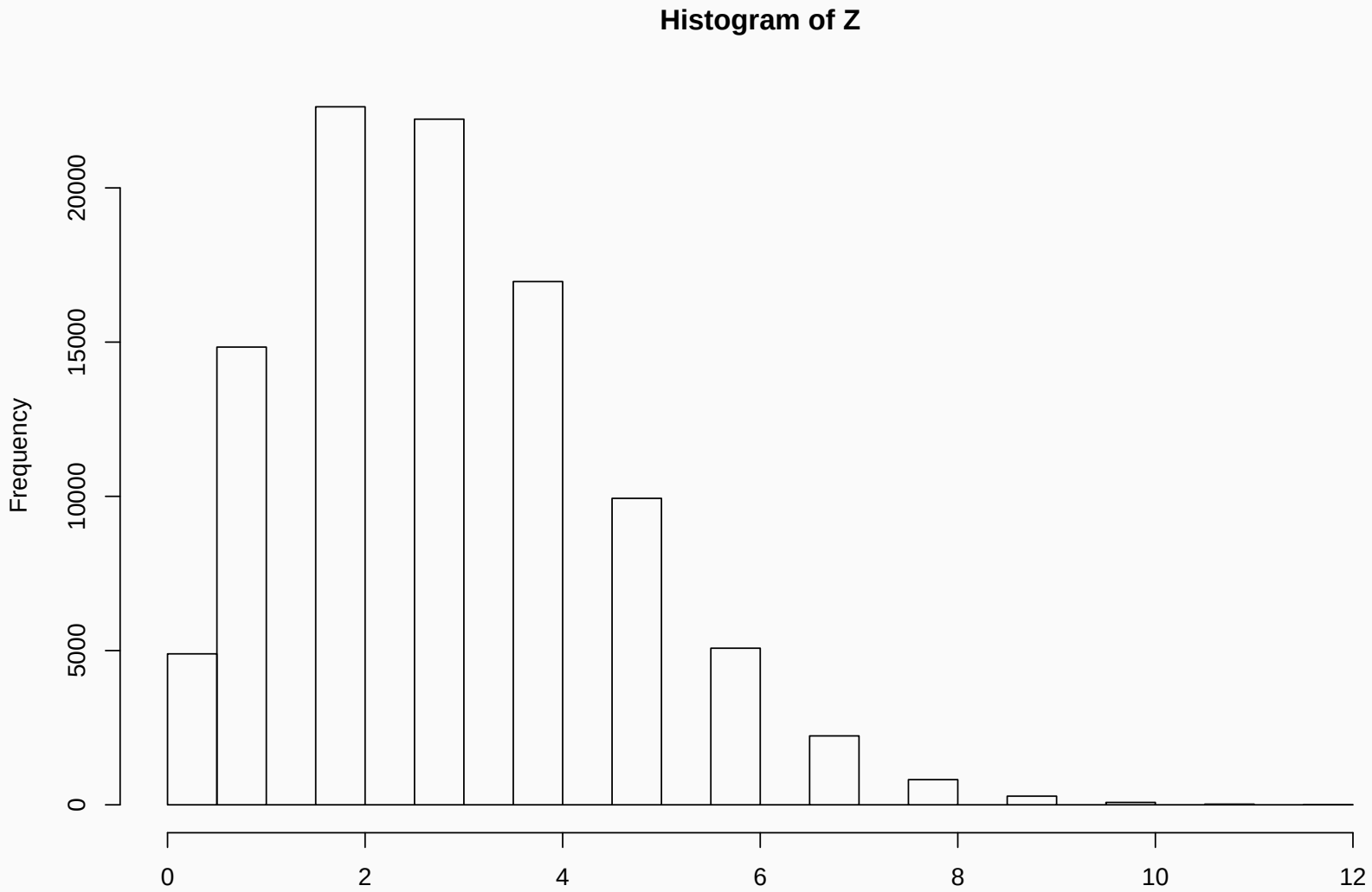
```
## [1] 0.13416
```

##Poisson distribution

- One of the useful properties of the Poisson distribution is that when you add multiple Poisson distributions together, the result is also a Poisson distribution.

Poisson distribution

```
# Use compare_histograms to compare Z to the Poisson(3)
hist(Z)
```



Geometric Distribution

- Simulating waiting for heads.

$$E[X] = 1/p - 1$$

- Think of this as the number of tails before the first heads (the number of non-event before the first event).

```
# Simulate 100 instances of flipping a 20% coin
flips <- rbinom(100, 1, .2)

# Use which to find the first case of 1 ("heads")
which(flips==1)[1]
```

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

```
# Existing code for finding the first instance of heads
which(rbinom(100, 1, .2) == 1)[1]
```

Geometric Distribution

- Use the `replicate()` function to simulate 100,000 trials of waiting for the first heads after flipping coins with 20% chance of heads. Plot a histogram of this simulation by calling `qplot()`

```
library(ggplot2)

# Replicate this 100,000 times using replicate
replications <- replicate(100000, which(rbinom(100, 1, .2)

# Histogram the replications with qplot
qplot(replications)

## `stat_bin()` using `bins = 30`. Pick better value with `
```

Geometric Distribution

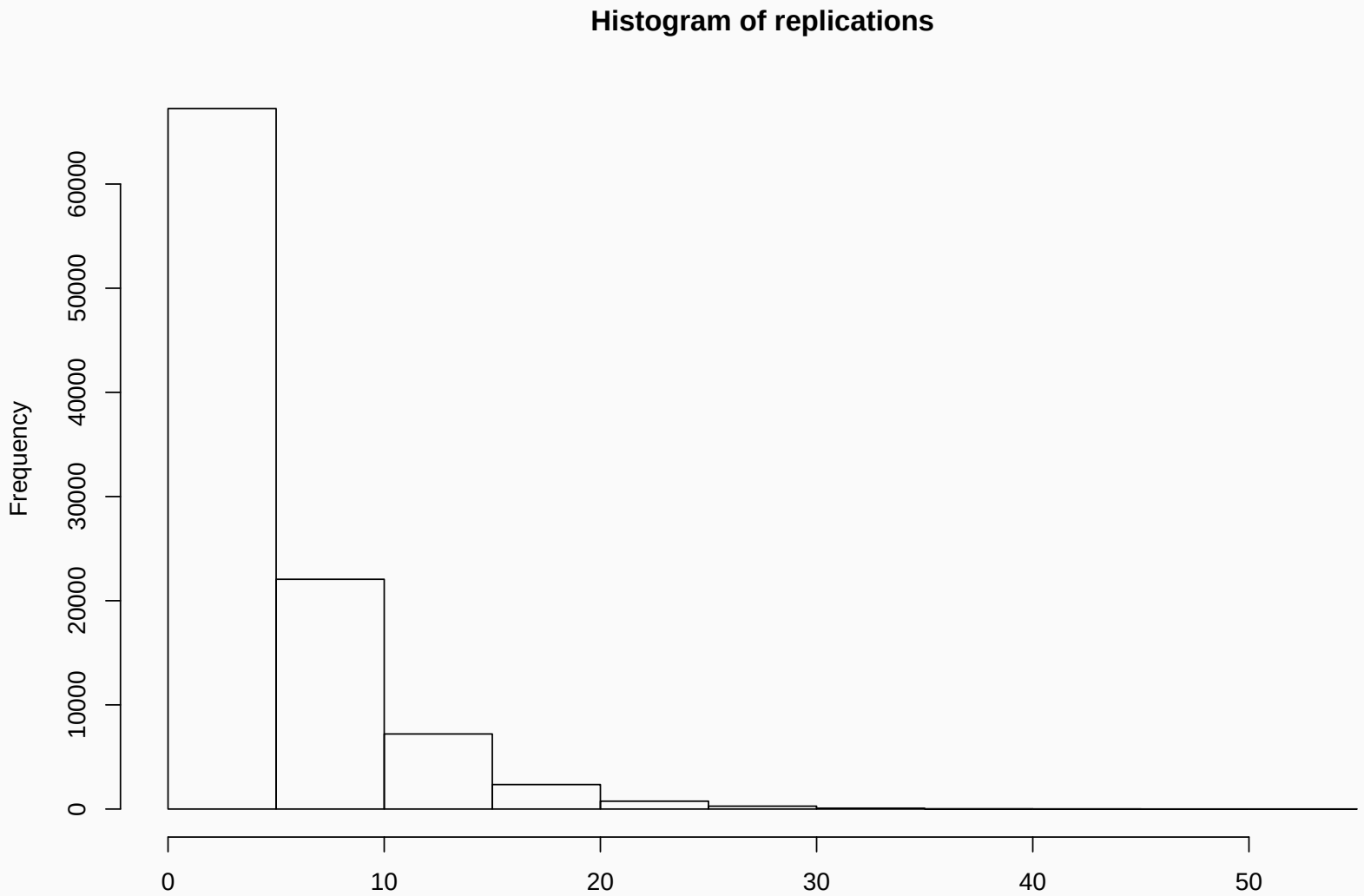
- Compare your replications with the output of rgeom().

```
# Replications from the last exercise
replications <- replicate(100000, which(rbinom(100, 1, .2)

# Generate 100,000 draws from the corresponding geometric
geom_sample <- rgeom(100000, .2)
```

Geometric Distribution

```
# Compare the two distributions with compare_histograms
hist(replications)
```



Geometric Distribution

- A new machine arrives in a factory. This type of machine is very unreliable: every day, it has a 10% chance of breaking permanently. How long would you expect it to last?
- Notice that this is described by the cumulative distribution of the geometric distribution, and therefore the `pgeom()` function. `pgeom(X, .1)` would describe the probability that there are X working days before the day it breaks (that is, that it breaks on day X + 1).

```
# Find the probability the machine breaks on 5th day or ea
pgeom(4, .1)
```

```
## [1] 0.40951
```

```
# Find the probability the machine is still working on 20t
1 - pgeom(19, .1)
```

Geometric Distribution

- If you were a supervisor at the factory with the unreliable machine, you might want to understand how likely the machine is to keep working over time. You'll plot the probability that the machine is still working across the first 30 days.

```
# Calculate the probability of machine working on day 1-30
still_working <- 1 - pgeom(0:29, .1)

# Plot the probability for days 1 to 30
qplot(1:30, still_working)
```

Hypothesis Testing

- The main purpose of statistics is to test a hypothesis. For example, you might run an experiment and find that a certain drug is effective at treating headaches.
- A good hypothesis statement should:
  - Include an "if" and "then" statement.
  - Include both the independent and dependent variables.
  - Be testable by experiment, survey or other scientifically sound technique.