James Scruggs

M01255052

5 Distribution

Today we will explore the Binomial and Normal Distributions.

The Binomial Distribution

Suppose we have a yes/no variable. For example, "Do you prefer Coke over Pepsi?" or "Are you right-handed?" or "Did the medical test come back positive?" These questions yield answers that are one of two outcomes. These outcomes can be considered yes/no or success/failure or 1/0.

Suppose further that we ask these questions to *n* people in an independent way. That is, the people we ask are not related to each other in any way. Finally, suppose we want to estimate the population proportion of people who say "yes." Let's call that proportion *p*. In order to do the estimation properly, we will need to understand how the distribution of likely answers behaves.

So, to take a look at that distribution, let's review what the distribution is counting:

> Counting the number of "successes" when
> 1. There are only two outcomes.
> 2. There are *n* trials per experiment.
> 3. The trials are independent.
> 4. The probability of success on each trial is a fixed value, *p*.

This distribution is called the binomial distribution.

1. To explore the binomial distribution in R, we need to talk about four commands: pbinom; qbinom; dbinom; and rbinom. Let's start with rbinom. The r stands for random. rbinom generates random variables with the specified distribution. For example, to simulate how many heads came up with 20 flips of a fair coin, you'd type

   rbinom(n=1, size=20, p=0.5) #Note: R calls the number of trials per experiment size. The variable n is how many observations we want.

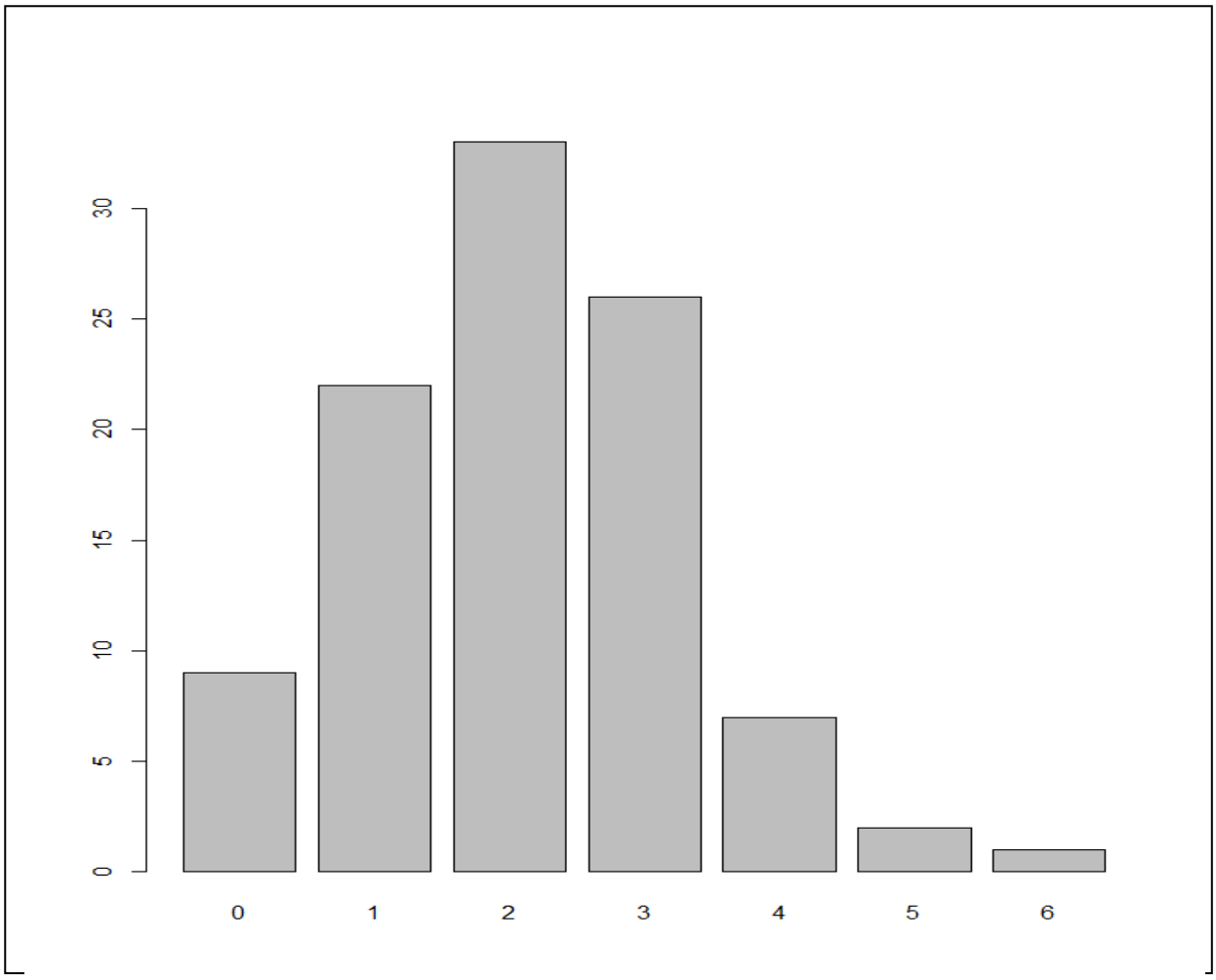
   To simulate flipping the coin 20 times, then repeating 10 times, you'd type
   rbinom(n=10, size=20, p=0.5)

   Suppose we wanted to simulate the number of heads that came up in 20 flips of a coin that had probability 2/3 of coming up heads, and we wanted to repeat that simulation 15 times. Type the correct command into R, and put the command and the output in the following box.

```
rbinom(n=15, size=20, p=0.66)

[1] 11 11 14 16 12  7 14 14 12 13 12 13 17  9 14
```

we will assign 100 independent binomial numbers with parameters n = 7 and p = 0.3 to a vector object called V. generate numbers using **rbinom** function and tabulate data and plot using **barplot**



2. pbinom and dbinom are commands that tell you the probability of certain output. dbinom(x, size, prob) gives the probability of getting exactly x heads. pbinom(x, size, prob) gives the probability of getting x or fewer heads. Try the following:

dbinom(0:10, size=10, prob=0.5)

pbinom(0:10, size=10, prob=0.5)

Notice that the output of pbinom is the sum of dbinom up to that x value.

What is the probability of getting less than or equal to 12 heads, if you flip a fair coin 20 times?

87%

What is the probability of getting less than or equal to 7 heads, if you flip a fair coin 20 times?

13%

What is the probability of getting between 8 and 12 heads, if you flip a fair coin 20 times? (There are a couple of ways to get this number. You can subtract the first two numbers, or you can add up output from dbinom for x for 8 to 12.)

74%

3. Finally, qbinom(p, size, prob) gives the smallest x value so that the probability of being less than or equal to x is at least p. Try the following:

pbinom(0:10, size=10, prob=0.5)

qbinom(0.75, size=10, prob=0.5)

Notice that the value of pbinom at x=6 is the first value bigger than 0.75 in that list.

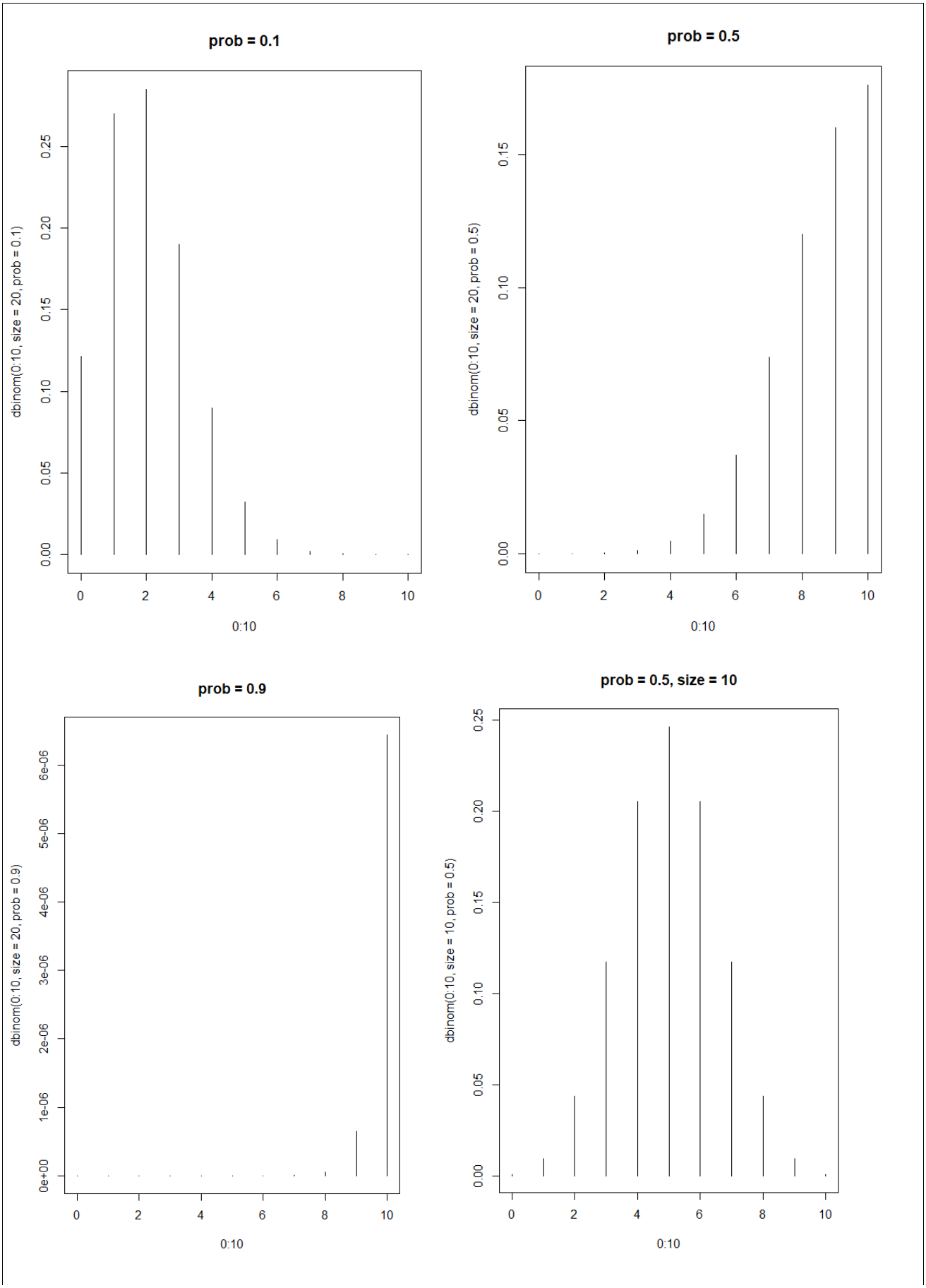
What is the value for which 90 percent of the probability is smaller than it, when you flip a fair coin 20 times?

13

4. Note that we can draw the distribution by using *plot* with *dbinom*.

plot(0:10,dbinom(0:10, size=20, prob=0.5), type='h') #type = “h” makes the plot give “histo gram-like” bars rather than dots or connected lines.

Plot several different graphs with different values of prob and size. Then summarize what changing those values does to the graph.


prob = 0.1
dbinom(0:10, size = 20, prob = 0.1)
0:10
prob = 0.5
dbinom(0:10, size = 20, prob = 0.5)
0:10
prob = 0.9
dbinom(0:10, size = 20, prob = 0.9)
0:10
prob = 0.5, size = 10
dbinom(0:10, size = 10, prob = 0.5)
0:10

By changing the probability, we make the occurrence of a certain value less likely/ more likely. By changing the size from 20 to 10, we make the number of flips equal to 10 so with a probability of 50%, 5 occurrences of heads is going to be highest value. If the size is 20 then 10 occurrences is going to be the highest value with p=.5.

The Normal Distribution

The Normal Distribution is used for quantitative data that has a mound-shape. In R, it also has four commands: pnorm; qnorm; dnorm; and rnorm. Notice that the prefixes p, q, d, and r are the same as the binom commands. The suffix norm is what changes. This pattern of commands holds for other distributions as well. If the distribution family has a name, then there are four commands for the distribution in R.

5. *rnorm(n, mean, sd)* creates n observations of normal random variables with the given mean and standard deviation. Create 10 normal random variables with mean 100 and standard deviation 15, and paste them here.

[1] 75.00419 69.26612 96.53335 92.04615 130.04785 95.14320 77.33493

[8] 87.62895 109.95548 111.93271

6. *pnorm(x, mean, sd)* gives the probability of being less than or equal to x. This is similar to *pbinom*. But *dnorm(x, mean, sd)* doesn't give the probability of being exactly x, which *dbinom* does. This is because normal random variables, like all continuous random variables, have zero probability of being an exact value. I'll go into this further in a couple of minutes.

In the mean time, what is the probability of a normal random variable being less than 120 if the mean is 100 and the standard deviation is 15? What is the probability of being more than 120? (Subtract the previous answer from 1.)

90% for being less than 120

9% for being more than 120

What is the probability of being between 80 and 120 if the mean is 100 and the standard deviation is 15? (*pnorm(120, mean=100, sd=15)–pnorm(80, mean=100, sd=15)*)

82%

7. *qnorm(p, mean, sd)* gives the x value for which the probability of being less than that x value is p. What value is the 90^{th} percentile (ie. 90 percent of the population is less than this number) for a normal random variable with mean 100 and standard deviation 15?

119.2233

8. I mentioned that *dnorm(x, mean, sd)* doesn't give the probability of being exactly equal to x. So what does it give? It gives the height of the normal curve at that x.

xvalues <- seq(from=70,to=130, by=1)

plot(xvalues, dnorm(xvalues, mean=100, sd=15), type="l")

For continuous random variables, the probability of exactly one value is 0, and the probability of being between two values is the area under the curve between these two values.

Plot the graph of a normal random variable with mean 50 and standard deviation 10. Note you will have to change the xvalues to the correct values.