Introduction to Biostatistics (171:161)
Breheny

Lab #7

In Lab #7, we are going to use R and SAS to calculate factorials, binomial coefficients, and probabilities from the binomial distribution.

In this regard, working in R is more convenient than working in SAS. Because SAS only works with data sets, not individual numbers, it is rather bulky and awkward as a calculator: you have to create a data set, create a variable that contains the value you are interested in, and then print out the data set. So, to do something simple like multiply two numbers together, we have to submit:

```
DATA tmp;
  x = 5*4;
  PUT x;
RUN;
```

The PUT statement tells SAS to output the value of "x" to the Log window (you should see the value 20 displayed there). Alternatively, you could leave the PUT statement out and use the Table Editor to look at the data set tmp. For what follows, we will refrain from writing out the entire data step, just the part that replaces the "5*4" above.

1 The binomial coefficients

In previous lectures, we discussed factorials and binomial coefficients. Factorials can be calculated with:

SAS:

```
FACT(5);
```

R:

```
factorial(5)
```

which computes 5!. So, you can compute the binomial coefficients $5!/(3!(5-3)!)$ with

SAS:

```
FACT(5)/(FACT(3)*FACT(5-3));
```

R:

```
factorial(5)/(factorial(3)*factorial(5-3))
```

Calculating binomial coefficients is a common task, and there are SAS and R commands specifically for doing so, named COMB (for "combinations") and choose (because it gives you the number of ways of choosing 3 items, given 5 choices):

SAS:

```
COMB(5, 3);
```

R:

```
choose(5, 3)
```

These do the exact same thing as the longer way listed above.

2 The binomial distribution

SAS and R also have functions for calculating probabilities coming from a number of distributions. SAS calculates probabilities for distributions using the PDF function (which stands for "probability distribution function"). R uses the function dbinom for calculating probabilities from the binomial distribution, dnorm for calculating probabilities from the normal distribution, and so on.

For example, as we discussed in class, the CDC estimates that 22% of adults in the U.S. smoke. We can get the probability that 5 people in a random 10-person sample would smoke using:

SAS:

```
PDF('Binomial', 5, .22, 10);
```

R:

```
dbinom(5, size = 10, prob = .22)
```

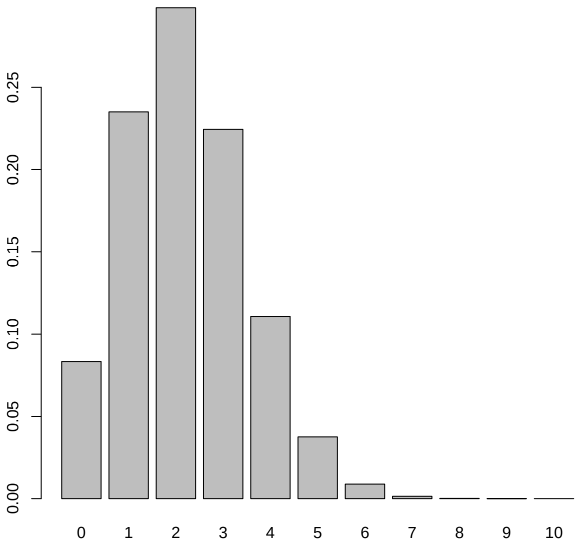
This returns 3.7%, the same result we obtained in class.

In R, you can leave out the `size =` and `prob =` , but if you do so, you have to have everything in a particular order. For example:

```
> dbinom(5, prob = .22, size = 10)
[1] 0.03749617
> dbinom(5, .22, 10)
[1] NaN
Warning message:
In dbinom(x, size, prob, log) : NaNs produced
```

Continuing in R, we can get the entire distribution with a single command:

```
> d <- dbinom(0:10, size = 10, prob = .22)
> d
 [1] 8.335776e-02 2.351116e-01 2.984109e-01 2.244458e-01 1.107842e-01
 [6] 3.749617e-02 8.813203e-03 1.420443e-03 1.502392e-04 9.416700e-06
[11] 2.655992e-07
> sum(d)
[1] 1
> 100*round(d, digits = 3)
 [1]  8.3 23.5 29.8 22.4 11.1  3.7  0.9  0.1  0.0  0.0  0.0
> barplot(d, names = 0:10)
```



The plot gives us a visual idea of the probability that we will see 0, 1, 2, and so on smokers in our sample. One can do all of these things in SAS, but they require some programming with loops that is a bit beyond the scope of this course.

Note that I used the `sum` command to add up all the probabilities. Of course, they have to add up to 1. We can use this to quickly get probabilities like the probability of getting two or fewer smokers:

```
> sum(dbinom(0:2, size = 10, prob = .22))
[1] 0.6168803
```

This matches up with the 61.7% that we got in class.

Adding up these probabilities cumulatively is another common task, and both SAS and `R` have dedicated functions for doing so. SAS uses `CDF` (which stands for “cumulative distribution function”) which returns the total probability that the random variable will equal any of number up to and including the number you specify. `R` has the same function, but calls it `pbinom` (or `pnorm` for the normal distribution, and so on). So, to get the probability that our sample will contain two or fewer smokers:

SAS:

```
CDF('Binomial', 2, .22, 10);
```

`R`:

```
pbinom(2,size = 10, prob = .22)
```

Again, we get 61.7%, which is equivalent to the `PDF` of 0 plus the `PDF` of 1 plus `PDF` of 2.

Note that both of these functions calculate the probability of two or fewer smokers by default. Thus, if we want to get the probability of something like “7 or more”, we have to subtract the probability of “6 or fewer” from 1:

SAS:

```
1 - CDF('Binomial', 6, .22, 10);
```

R:

```
1 - pbinom(6, size = 10, prob = .22)
```

In R, We could also get the same answer directly with:

```
> sum(dbinom(7:10, size = 10, prob = .22))
[1] 0.001580364
```

We could also achieve the same answer by specifying `lower.tail = F`:

```
> pbinom(6, size = 10, prob = .22, lower.tail = F)
[1] 0.001580364
```

Feel free to use R/SAS to check your answers, or to try to get the same answer as the computer in order to get extra practice working with the binomial distribution. However, keep in mind that you have to know how to calculate binomial probabilities by hand for quizzes, so don't use SAS/R exclusively unless you are sure you don't need the practice.

3 Practice problems

1. Suppose a group of 20 men, all unrelated, received a flu vaccine. Assume each man in this group has a 0.05 chance of dying in the next year. How likely it is that at least 2 of these men will die in the following year?

$$\begin{aligned}
P[X \geq 2] &= 1 - P[X < 2] \\
&= 1 - (P[X = 0] + P[X = 1]) \\
&= \texttt{1 - sum(dbinom(0:1, size = 20, prob = 0.05))} \\
&= \boxed{0.264}
\end{aligned}$$

2. Suppose 67% of Americans watch TV on a daily basis. Answer the following question based on repeated samples of size 19 from the U.S. population.

   (a) What is the mean number of individuals per sample who watch TV on a daily basis?

$$\begin{aligned}
\bar{x} &= np \\
&= (19)(0.67) \\
&= \boxed{12.73 \text{ people}}
\end{aligned}$$

   (b) What is the standard deviation?

$$\begin{aligned}
sd &= np(1-p) \\
&= \sqrt{(19)(0.67)(1-0.67)} \\
&= \boxed{2.04961 \text{ people}}
\end{aligned}$$

   (c) What is the probability that at least 3 of the randomly selected individuals watch TV on a daily basis?

$$\begin{aligned}
P[X \geq 3] &= 1 - (P[X = 0] + P[X = 1] + P[X = 2]) \\
&= 1 - \left(\binom{19}{0}(0.67)^0(0.33)^{19-0} + \binom{19}{1}(0.67)^1(0.33)^{19-1} + \binom{19}{2}(0.67)^0(0.33)^{19-2} \right) \\
&= \boxed{0.9999995}
\end{aligned}$$