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 both the binomial and the normal distributions. We will also analyze data from one-sample studies in which the outcome is categorical.

For the first goal, 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 coefficient

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

SAS:	`R`:
`FACT(5);`	`factorial(5)`

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

SAS:	`R`:
`FACT(5)/(FACT(3)*FACT(5-3));`	`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)
```

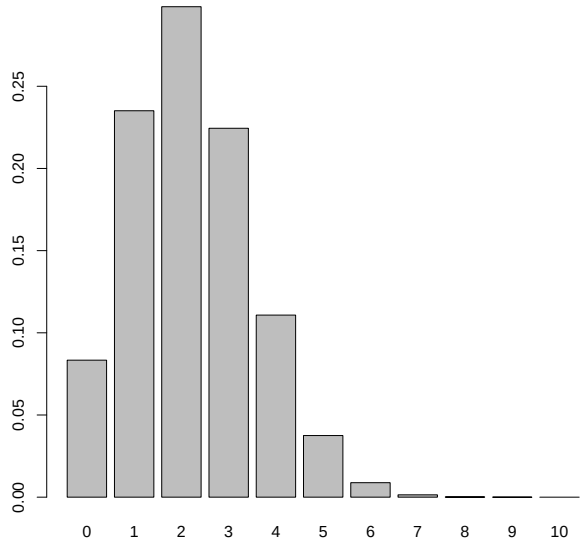
This returns 3.7%, the same result we obtained in class.

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 The normal distribution

The syntax for calculating probabilities from the normal distribution is very similar to the syntax for the binomial distribution. The `pnorm` function in R will calculate the area under the normal curve to the left of any number; using `CDF` with the `'Normal'` does the same thing in SAS. For example:

SAS:

```
DATA _NULL_;
  a = CDF('Normal', -1);
  PUT a;
  b = CDF('Normal', 0);
  PUT b;
  c = CDF('Normal', 2);
  PUT c;
RUN;
```

R:

```
pnorm(-1)
pnorm(0)
pnorm(2)
```

You can use this to calculate the area between, say, 1 and 2, or outside ± 1:

SAS:

```
CDF('Normal', 2) - CDF('Normal', 1)
CDF('Normal', -1) + 1-CDF('Normal', 1)
2*CDF('Normal', -1)
```

R:

```
pnorm(2) - pnorm(1)
pnorm(-1) + 1 - pnorm(1)
2*pnorm(-1)
```

Another helpful function is `qnorm`, which calculates the quantiles of the normal distribution; the SAS equivalent is `QUANTILE`. So, for example, what is the value for which 10% of the area lies to the right of it?

SAS:

```
QUANTILE('Normal', 0.1)
```

R:

```
qnorm(.1)
```

Or, what is the value z for which 10% area lies outside $\pm z$?

SAS:

```
QUANTILE('Normal', 0.1/2)
```

R:

```
qnorm(.1/2)
```

As with binomial distributions, using R (or SAS) is a great way to check your work, but be sure you also know how to perform these calculations using the table, as you will need this skill on quizzes.

4 Cystic fibrosis crossover study data

Download and import the data set `cysticfibrosis.txt`. Then create a variable that indicates whether or not each patient did better on drug or not. We named our data set "cf" and our indicator variable "DrugBetter".

You can obtain confidence intervals and hypothesis tests all in one bundle. The first step is to make a table (in this case, a one-variable table) of variables that we are interested in.

In SAS, we make tables using `PROC FREQ`, as we have already covered. However, we are now going to add an `EXACT` statement with a `BINOMIAL` option, specifying that we want exact tests and confidence intervals for the table based on the binomial distribution. In R, you can use the function `binom.test` to accomplish the same thing.

SAS:

```
PROC FREQ DATA = cf;
  TABLES DrugBetter;
  EXACT BINOMIAL;
RUN;
```

R:

```
binom.test(sum(DrugBetter), length(DrugBetter))
# or
binom.test(11, 14)
```

Note that you can just enter the data directly into the R code (that 11 out of 14 patients did better on the drug); you cannot do anything like this in SAS.

In the SAS Results Viewer window, two sets (approximate and exact) of confidence intervals are reported, along with two sets (approximate and exact) of hypothesis tests of the null hypothesis that $p = 0.5$. You want the exact, two-sided results. R only gives you what you asked for: the exact results.

5 Premature infant survival data

Often (in this class and in real life), you will not have access to an entire data set, or have it in a SAS-friendly format. You may simply know the summary statistics of how many individuals fell into the two categories. For example, in a previous lecture we discussed a study in which, out of a sample of 39 infants born at 25 weeks gestation, 31 survived. This is all the information we need in order to calculate confidence intervals and perform hypothesis tests.

As we saw above, we can just enter the 31 and 39 directly into `R` to obtain these results. However, in SAS everything has to be a data set, so to use SAS, we are going to have to create a data set first.

The survival data can be represented in the following manner:

Outcome	N
Survived (1)	31
Died (0)	8

We can easily use `DATALINES` to create this data set.

Now, if you try the following:

```
PROC FREQ DATA = gestsurv;
  TABLES surv;
  EXACT BINOMIAL;
  WEIGHT count;
RUN;
```

You may notice that by default, SAS gives you information about the probability of dying (0) instead of the probability of surviving (1) (this is because 0 is less than 1); the same would have happened if we had used "Survived" and "Died" because "Died" occurs before "Survived" alphabetically. To get the results that we obtained in class, you can use a `LEVEL` option to specify that you want a different level of the categorical variable. To use it, submit:

```
PROC FREQ DATA = gestsurv;
  TABLES surv / binomial (level = 2);
  WEIGHT count;
RUN;
```

which tells SAS to use "level 2" of the categorical variable as the category of interest (*i.e.* the one that comes second in alphabetical order). You should now have the results we got in class.

Note: we could also just subtract everything from 1 to get the other estimates and confidence intervals; you can compare these results to the results above.

Finally, note that we left out the `EXACT BINOMIAL` statement above; as a result, SAS does not provide the results of the binomial test for whether or not $p = 0.5$. In this case, that test is not meaningful, which brings up an important point: don't be distracted by superfluous SAS output. SAS will often output far more than you want to know, and much of it might be meaningless for your particular analysis. If, however, you were in a situation where you wanted to conduct a hypothesis test, you can add the `EXACT BINOMIAL` line back in.

6 Binomial 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. Suppose repeated samples of size 19 are drawn from the U.S. population.

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

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

7 Normal practice problems

1. Find the area under the normal curve...

   (a) below 0.3.

$$\begin{aligned} P[X \leq 0.3] &= \texttt{pnorm(.3)} \\ &= \boxed{0.6179114} \end{aligned}$$

   Using the table, we find the probability to be 0.618.

   (b) above 0.65.

$$\begin{aligned} P[X \geq 0.65] &= \texttt{1 - pnorm(0.65)} \\ &= \boxed{0.2578461} \end{aligned}$$

$$OR$$

$$\begin{aligned} P[X \geq 0.65] &= \texttt{pnorm(-0.65)} \\ &= \boxed{0.2578461} \end{aligned}$$

   Using the table, we find the probability to be $1 - 0.742 = 0.258$.

   (c) between 0.3 and 0.65.

$$\begin{aligned} P[0.3 \leq X \leq 0.65] &= \texttt{pnorm(0.65) - pnorm(0.3)} \\ &= \boxed{0.1242425} \end{aligned}$$

$$OR$$

$$\begin{aligned} P[0.3 \leq X \leq 0.65] &= \texttt{1 - (pnorm(0.3) + pnorm(-0.65))} \\ &= \boxed{0.1242425} \end{aligned}$$

   Using the table, we find the probability to be $0.742 - 0.618 = 0.124$.

(d) below -0.45.

$$P[X \leq -0.45] = \texttt{pnorm(-0.45)}$$
$$= \boxed{0.3263552}$$

Using the table, we find the probability to be 0.326.

2. Find the following percentiles of the normal curve.

(a) 20^{th}

$$P[Z \leq z] = 0.1$$
$$z = \texttt{qnorm(0.2)}$$
$$= \boxed{-0.8416212}$$

Using the table, we find the percentile to be -0.84.

(b) 80^{th}

$$P[Z \leq z] = 0.80$$
$$z = \texttt{qnorm(0.80)}$$
$$= \boxed{0.8416212}$$

Using the table, we find the percentile to be 0.84.

(c) 95^{th}

$$P[Z \leq z] = 0.95$$
$$z = \texttt{qnorm(.95)}$$
$$= \boxed{1.644854}$$

Using the table, we find that the percentile is between 1.64 and 1.65. From the `R` code above, we see that the percentile is actually rounded to 1.645; this is value commonly used for the 95^{th} percentile.

(d) 90^{th}

$$P[Z \leq z] = 0.90$$
$$z = \texttt{qnorm(.90)}$$
$$= \boxed{1.281552}$$

Using the table, we find the percentile to be 1.28.

8 Categorical practice problems

1. Use the table below summarizing the survival data at gestational age 22 weeks to answer the following questions.

Outcome	Count
Survived	0
Died	29

(a) What are the exact 95% Confidence Limits for probability of surviving?

$$\begin{aligned}\text{Lower Bound} &= 1 - 95\% \text{ Upper Conf Limit} \\ &= 1 - 1.0000 \\ &= \boxed{0.0000} \\ \text{Upper Bound} &= 1 - 95\% \text{ Lower Conf Limit} \\ &= 1 - 0.8806 \\ &= \boxed{0.1194}\end{aligned}$$

(b) What is the p-value for the approximate test and exact test?

$$\begin{aligned}\text{approx} &= \boxed{< .0001} \\ \text{exact} &= \boxed{3.725\text{E-}09}\end{aligned}$$

(c) What test does the p-value correspond to?

$$\begin{aligned}H_0 &: p = 0.50 \\ &vs. \\ H_1 &: p \neq 0.50\end{aligned}$$

2. Use the smoking data set to answer the following questions.

(a) What proportion of the observations survived?

$$\frac{\#\text{survived}}{\#\text{observations}} = \boxed{0.7191781}$$

(b) What is the exact confidence interval for survival?

$$\text{CI} = \boxed{(0.6940270,\ 7433448)}$$

(c) What is the exact p-value testing that the proportion of survival is equal to 0.5?

$$P[p \geq 0.7191781] = \boxed{< 2.2e-16}$$