Lab 8: Break free from tables!
STT 421: Summer, 2004
Vince Melfi

For many years published tables of probabilities, like Tables A–F of Normal, Binomial, etc., probabilites, were indispensible to statisticians. But now computers can reproduce the values in these tables very quickly. Today we'll look at some of the SAS capabilities for computing probabilities.

Binomial Probabilities

We have learned that the binomial distribution often provides a good model for choosing a sample at random from a population in the case where we are interested in a variable that has only two values. For example, if we choose $n = 20$ registered voters at random and ask them whether they plan to vote for George Bush in 2004, then the number who say "yes" is modeled by a binomial distribution with parameters $n = 20$ and p, where p is the true proportion of registered voters who plan to vote for George Bush in 2004.

In SAS it's easy to compute binomial and other probabilities via the **pdf** function. The following program shows how to compute the probability that $X = 3$, where X has a binomial distribution with parameters $n = 20$ and $p = 0.1$. (This would be the model for the numbre of George Bush supporters in a sample of size $n = 20$ if the population proportion of George Bush supporters is 0.1.)

```
data binom1;
  x = pdf('binomial', 3, 0.1, 20);

proc print data=binom1;

run;
```

When you run this program, you should see that $x = 0.19012$. So there is a probability of 0.19012 that exactly 3 of the 20 people are supporters of George Bush, assuming that $p = 0.1$ of the population are George Bush supporters. In terms of proportions, this tells us that there is a 0.19012 probability that our estimate $\hat{p} = X/n$ of the proportion of George Bush supporters is 3/20.

It's almost as easy to compute a whole binomial table of probabilities. For example, suppose we want to know all the binomial probabilities when $n = 20$ and $p = 0.1$. The following program does the trick.

```
data binom2;
  do i = 0 to 20 by 1;
  prob = pdf('binomial', i, 0.1, 20);
  output binom2;
  end;

proc print data=binom2;
```

```
run;
```

Almost immediately you should see something like the table below, giving all the probabilities. We can see, for example, that there is a very small (zero to 5 decimal places) probability that $X = 17$.

```
Obs     i        prob

  1     0     0.12158
  2     1     0.27017
  3     2     0.28518
  4     3     0.19012
  5     4     0.08978
  6     5     0.03192
  7     6     0.00887
  8     7     0.00197
  9     8     0.00036
 10     9     0.00005
 11    10     0.00001
 12    11     0.00000
 13    12     0.00000
 14    13     0.00000
 15    14     0.00000
 16    15     0.00000
 17    16     0.00000
 18    17     0.00000
 19    18     0.00000
 20    19     0.00000
 21    20     0.00000
```

The normal density

The **pdf** function also gives an easy way to draw a picture of a density function. For example, the following program should draw a picture of the normal density with $\mu = 3$ and $\sigma = 5$. We know that the density is very close to zero when we're more than 3σ from the mean, so we'll concentrate on x values between -12 and 18.

```
data normal1;
  do x = -12 to 18 by 0.05;
  density = pdf('normal', x, 3, 5);
  output normal1;
  end;

proc gplot data = normal1;
  plot density*x;
  symbol interpol=join;

run;
```

Explanation

Hopefully you got a plot of the appropriate normal density.

1. The statement **do x = -12 to 18 by 0.05** tells SAS to loop, starting with **x=-12**, increasing by **0.05** each time, and stopping when **x = 18**.

2. The statement **density = pdf('normal', x, 3, 5)** computes the appropriate normal density at **x** and stores the value in the variable `density`.

3. There's a new twist in **proc gplot**: The statement **symbol interpol=join** tells SAS that we want to join the plotted points, which is appropriate when we're plotting a function.

Cumulative distribution functions

With continuous random variables, and often with discrete random variables, we want to compute probabilities like $P(3 < X \leq 10)$ or $P(.02 < hatp < .075)$. For such problems, a cumulative distribution function (CDF) is much more useful than a PDF. Let's return to the binomial case, this time with $n = 25$ and $p = 0.45$. We'll be interested in computing $P(10 < X \leq 20)$. One way to do this is to add $P(X = 11) + P(X = 12) + \cdots + P(X = 20)$.[1] We can do this via the **pdf** function.

Another way to write this problem is

$$P(10 < X \leq 20) = P(X \leq 20) - P(X \leq 10).$$

(We first compute $P(X \leq 20)$, but this is too much because it includes $X = 0, X = 1, \ldots, X = 10$. So we then subtract $P(X \leq 10)$.)

The CDF at x returns $P(X \leq x)$, so we can solve the above problem by computing the CDF at 20 and the CDF at 10 and subtracting. Here's the appropriate SAS code to compute the PDF and the CDF.

```
data binom3;
  do i = 0 to 25 by 1;
  binompdf = pdf('binomial', i, 0.45, 25);
  binomcdf = cdf('binomial', i, 0.45, 25);
  output binom3;
  end;

proc print data = binom3;

run;
```

[1] Note that $X = 10$ is not included, since the inequality has $10 < X$.

Here is an abbreviated version of the output.

```
Obs    i    binompdf    binomcdf
 9     8     0.07013     0.13398
10     9     0.10839     0.24237
11    10     0.14189     0.38426
12    11     0.15831     0.54257
13    12     0.15111     0.69368
14    13     0.12364     0.81731
15    14     0.08671     0.90402
16    15     0.05202     0.95604
17    16     0.02660     0.98264
18    17     0.01152     0.99417
19    18     0.00419     0.99836
20    19     0.00126     0.99962
21    20     0.00031     0.99993
22    21     0.00006     0.99999
```

To answer our problem we can either use

$$0.14189 + 0.15831 + \cdots + 0.00126 \quad \text{or} \quad 0.99993 - 0.38426.$$

Clearly the second is easier.

Central Limit Theorem, Take I

As the course progesses we'll learn that the normal distribution provides a very good approximation for probabilities in a wide variety of problems. Here we'll see that we can approximate binomial distributions with large values of n by appropriate normal distributions.

First we'll plot the pdf for a binomial distribution with $n = 75$ and $p = 0.4$.

```
data binom4;
  do i = 0 to 75 by 1;
  binompdf = pdf('binomial', i, 0.4, 75);
  output binom4;
  end;

proc gplot;
  plot binompdf * i;
  symbol value=circle;
  symbol interpol=none;

run;
```

Note that the shape of the binomial pdf is close to the shape of a normal density. (We've set **symbol interpol = none** to avoid connecting the points, since this is a discrete pdf, not a density.)

Now we'll see how well we can approximate the binomial probabilties with normal probabilities. We'll use the fact that the mean of a binomial distribution is np and the standard deviation is $\sqrt{np(1-p)}$. In our case this yields

$$\mu = (75)(0.4) = 30 \quad \text{and} \quad \sigma = \sqrt{75(0.4)(0.6)} = 4.24.$$

We'll compute the normal CDF at some values between 20 and 40 (this is where most of the probability is for the binomial) and compare these to the binomial CDF.

```
data normbinom;
  do i = 20 to 40 by 0.5;
  binomcdf = cdf(’binomial’, floor(i), 0.4, 75);
  normcdf = cdf(’normal’, i, 30, 4.24);
  output normbinom;
  end;

proc gplot data=normbinom;
  plot binomcdf * i normcdf * i /overlay;
  symbol interpol=join;

proc print data = normbinom;

run;
```

Explanation

1. In the line beginning **binomcdf** we specified **floor(i)** as the second argument to **cdf** to make sure that this argument was an integer, because the binomial distribution only takes integer values. The **floor** function returns the greatest integer less than or equal to a value. For example, **floor(34.5) = 34**; **floor(34) = 34**.

2. We placed both plot requests on the same line and specified **overlay** to tell SAS to place both plots on the same graph.

Use the output from the **proc print** statement to answer the following questions.

1. Compute $P(24.5 < X < 30.5)$ using the binomial cdf and the normal cdf. What is the difference between your answers?

2. Compute $P(30 < X < 40)$ using the binomial cdf and the normal cdf. What is the difference between your answers?

3. Compute $P(X \leq 25.5)$ using the binomial CDF and the normal CDF. What is the difference between your answers?

4. Compute $P(X \leq 25)$ using the binomial CDF and the normal CDF. What is the difference between your answers?