Economics 40/Statistics M11 Lab 2: Normal Distribution and Z-Scores

Due Friday February 9, 2001

Purpose: The purpose of this lab is to use Stata to examine distributions.

Data: We will use data on assorted variables, some normal, some not. Issue the Stata command:

use "http://www.stat.ucla.edu/~vlew/stat11/labs/dists.dta"

If you cannot access the data with the "use" command, go to the course web page, find the labs link and read the instructions on downloading data there.

Introduction:

You were introduced to normal distributions in Chapter 1.3. The goal here is to give you practice in working with them. Once you have the data set above loaded into Stata, if you issue a

describe

You should see this:

```
Contains data from dists.dta
  obs:         2,000
 vars:             6                          22 Jan 2001 22:57
 size:        44,000 (100.0% of memory free)
-------------------------------------------------------------------------------
   1. standard  float  %9.0g                  Standard Normal Distribution
   2. students  float  %9.0g                  Age distribution of 2000
                                                college students
   3. heights   float  %9.0g                  Heights of 2,000 American Women
   4. prichg    float  %9.0g                  Percentage Change in Price for
                                                2,000 stocks
   5. dice      byte   %8.0g                  Results from 2,000 rolls of a
                                                fair 6 sided die
   6. yesno     byte   %8.0g                  Survey where 1=yes and 0=no
-------------------------------------------------------------------------------
Sorted by:
```

And a summarize will yield:

```
. summarize

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+------------------------------------------------------
standard |    2000   -.0243922   1.005488  -4.009344   3.611203
students |    2000    19.46102   2.010673   13.00286   25.69446
 heights |    2000    63.95163   2.473171   56.23328   74.02336
  prichg |    2000    1.950179   1.720361  -3.741254   7.669019
    dice |    2000      3.5175   1.705036          1          6
   yesno |    2000        .486    .499929          0          1
```

So there are 6 variables. The first variable, called standard, has 2,000 values created to look very much like a perfect, standard, normal distribution. What can we see from the above summarize command, is that the variable standard has a mean of nearly zero and a standard deviation of 1.005 or very close to what a standard normal should have (a standard normal has a mean of zero and a standard deviation of 1). We can take a look a histogram of the values for standard, try:

```
graph standard, bin(15) normal ylabel xlabel
```

What this does is create a histogram of values with 15 bins, draws a normal curve over it (so you can see how it compares with a normal) and fixes the vertical and horizontal axes so they look rather nice. It looks pretty close to normal. Here is something else you can examine, it's called pnorm:

```
pnorm standard
```

All pnorm does is graph the distribution of a variable against a standardized normal probability table. If the variable is distributed close to normal, all of the values will lie on the line going from the lower left hand corner of the screen to the upper right hand corner of the screen. If a variable deviates substantially from normal, it will not lie on this line. For example, try a pnorm of the variable dice or the variable yesno as a contrast:

```
pnorm dice
```

Or

```
pnorm yesno
```

For part 1 of your assignment, I would like you to construct histograms of the variables students, heights, prichg, dice, and yesno. For part 2, I would like you to issue the pnorm command for the same five variables. Print out the graphs. Which variables appear to be normal? Which ones are not normally distributed?

Z-scores

Remember, Z scores are also called standard units or standard scores or standardized scores. I like to think of the as standard "deviation" units because that is simply what they are. A Z-score of 2.41 is interpreted as the value that has a Z score of 2.41 is 2.41 standard deviations above average. A Z-score of -1.68 is interpreted as some value that has a Z score of -1.68 is 1.68 standard deviations BELOW average.

This dataset has 2,000 observations and 6 variables. Don't calculate 2,000 Z scores by hand, instead, use a Stata function called "egen" and let's use the variable heights:

```
egen zheight=std(heights)
```

all egen does is create a new variable, we'll call it zheight (to stand for the Z scores of height) and assign it the value of its standardized height. Issue a list command to see the results:

list heights zheight

Your first 10 values should look like this:

```
        heights    zheight
   1.  64.91271    .388603
   2.   65.9706   .8163501
   3.  61.24952  -1.092569
   4.  66.89182   1.188835
   5.  65.26472   .5309355
   6.  59.42522  -1.830205
   7.  63.14916  -.3244694
   8.  61.89093  -.8332235
   9.  62.19126  -.7117863
  10.  63.72664  -.0909714
```

And if you hand check zheight, for example, the first one:

$$Z = \frac{x - m}{s} = \frac{64.91271 - 63.95163}{2.473171} = .3886$$

Is very close to what Stata is generating for you. Note that x stands for the value of interest, Greek letter mu (ì) is the mean of the population and Greek letter sigma (ó) is the standard deviation of the population. We got the mean and standard deviation from the summarize command above.

Now that we have Z scores, we could ask…well, how well does the variable on women's heights fit a normal? Let's look at those Z scores, from your Table A, about .1587 or 15.87% of all heights will fall at or below a Z=-1.0. We can examine our 2,000 to see if this is true. Issue the command:

. *count if zheight <= -1.0*

and it should return the value

319

which means that 319/2000 or .1595 or 15.95% fell at or below a Z=-1.0, not bad. By the way, Stata has a built-in calculator. If you issue the command:

. *display 319/2000*

You will get this value back:

.1595

To find out what percentage of heights fell at or above +1.0, just issue the command

```
count if zheight >= 1.0
```

you will get 299, and then to find the percentage:

```
display 299/2000
```

and you'll get .1495, again, pretty close to what we would expect.

How would you find out what percentage fell between Z= -1 and Z=+1? Do this:

```
count if zheight >= -1 & zheight <=+1
```

this will return a value of 1382. Note the & stands for the word "and" all it means is that for a value to count, it must be greater than -1 and less than +1 at the same time. To find out what percentage 1,382 is of 2000, type:

```
display 1382/2000
```

you'll get .691 or about 69.1%, again very close to the 68% we were expecting. How would you figure out what percentage was either greater than +1 or less than -1? Frankly, I'd just subtract .691 from 1.0 or 69.1% from 100% but you could issue the command

```
. count if zheight <= -1  |  zheight >= +1
```

and you will get 618. Note that line is a vertical bar. It means "OR" to Stata. All this means is that you want the count of observations with a zheight of less than or equal to 1 or a zheight of greater than or equal to 1. You can't use an "&" (this stands for "and") because a value can't satisfy both conditions (compare it to the previous statement).

I would like you to generate the z-scores for variables prichg and dice and then calculate the following percentages using the method shown above: (a) between z = -1 and z = +1 (b) between z = -2 and z= +2 and finally (c) between z = -3 and z = +3. Include these percentages in your assignment, you don't need to show me the actual calculated z scores.

ASSIGNMENT RECAP

Do the 4 things listed below, staple the necessary output together, put your name and your TA's name on it and turn it all in on February 9, 2001.

1. I would like you to construct histograms of the variables students, heights, prichg, dice, and yesno and print out the graphs.
2. I would like you to issue the pnorm command for the same five variables and print out the graphs.
3. Answer the questions: which variables appear to be normal? Which ones are not normally distributed?
4. I would like you to generate the z-scores for variables prichg and dice and then calculate the following percentages using the method shown above: (a) between z = -1 and z = +1 (b) between z = -2 and z= +2 and finally (c) between z = -3 and z = +3. Include these percentages in your assignment, you don't need to show me the actual z scores generated by Stata.