Lab 3: Distributions of Random Variables

In this lab, we investigate the probability distribution that is most central to statistics: the normal distribution. If we are confident that our data are nearly normal, that opens the door to many powerful statistical methods. Here we'll use the graphical tools of SAS to assess the normality of our data and also learn how to generate random numbers from a normal distribution.[1]

The Data

We'll be working with measurements of body dimensions. This data set contains measurements from 247 men and 260 women, most of whom were considered healthy young adults.

If you are using SAS University Edition, you need to ensure that interactive mode is turned off. To do this, click the button to the right of **Sign Out** in the upper right corner of the window and then click **Preferences**. In the Preferences window, on the General tab, the bottom check box (located next to the text **Start new programs in interactive mode**) should ***not*** be selected. If the box is selected, you need to clear it and save your change.

```
filename bdims url 'http://www.openintro.org/stat/data/bdims.csv';

proc import datafile=bdims
            out=bdims
            dbms=csv
            replace;
   getnames=yes;
run;
```

Let's take a peek at the first few rows of the data.

```
proc print data=bdims (obs=6);
run;
```

You see that for every observation, we have 25 measurements, many of which are either diameters or girths. A key to the variable names can be found at **http:// www.openintro.org/ stat/ data/ bdims.php**, but we'll focus on only three columns to get started: weight in kg (**wgt**), height in cm (**hgt**), and **sex** (*1* indicates male, *0* indicates female).

Because males and females tend to have different body dimensions, it will be useful to create two additional data sets: one with only men and another with only women.

```
data mdims fdims;
   set bdims;
   if sex=1 then output mdims;
   if sex=0 then output fdims;
run;
```

Exercise 1: Make a histogram of men's heights and a histogram of women's heights. How would you compare the various aspects of the two distributions?

---

[1]

The Normal Distribution

In your description of the distributions, did you use words such as “bell-shaped” or “normal?” It’s tempting to say so when faced with a unimodal symmetric distribution.

To see how accurate that description is, we can plot a normal distribution curve on top of a histogram to see how closely the data follow a normal distribution. This normal curve should have the same mean and standard deviation as the data.

Next we make a density histogram with a normal probability curve overlaid. The difference between a frequency histogram and a density histogram is that in a frequency histogram, the ***heights*** of the bars add up to the total number of observations. In a density histogram, the ***areas*** of the bars add up to 1. The area of each bar can be calculated as simply the height × the width of the bar. Using a density histogram enables us to properly overlay a normal distribution curve over the histogram because the curve is a normal probability density function. Frequency and density histograms both display the same exact shape. They differ only in their y-axis. The commands below create a density histogram. To request a frequency histogram, you can simply add the VSCALE=COUNT option to the HISTOGRAM statement after the / in the code below.

```
title 'Histogram of hgt for females';
ods select histogram;
proc univariate data=fdims;
   var hgt;
   histogram / normal;
   output out=estimates n=n mean=mean std=std;
run;
```

We used the ODS SELECT statement to request that only the histogram be printed. The DATA statement specifies **fdims** as our data set. The VAR statement specifies **hgt** as the variable of interest. The HISTOGRAM statement requests a histogram for each variable listed in the VAR statement (in this case, only **hgt**), and the NORMAL option requests that the normal distribution curve be overlaid. The OUTPUT statement sends the estimates of the sample size, the mean, and the standard deviation to a new data set, **estimates**. To save time later, we can save the estimates generated from the UNIVARIATE procedure as macro variables.

```
data _null_;
   set estimates;
   call symputx('n',n);
   call symputx('mean',mean);
   call symputx('std',std);
run;
```

The SYMPUTX function saves each variable as a macro variable. These macro variables will hold the sample size, mean, and standard deviation until you exit this SAS session. Because the data set is specified as _NULL_ (a SAS keyword), SAS saves the values to the macro variables without creating a data set.

Exercise 2: Based on the plot, does it appear that the data follow a nearly normal distribution?

Evaluating the Normal Distribution

Eyeballing the shape of the histogram is one way to determine whether the data appear to be nearly normally distributed, but it can be frustrating to decide just how close the histogram is to the curve. An alternative approach involves constructing a normal probability plot, also called a normal Q-Q plot (for quantile- quantile).

```
title 'Q-Q plot of hgt for females';
ods select qqplot;
proc univariate data=fdims;
   var hgt;
   qqplot / normal(mu=est sigma=est);
run;
```

We used the ODS SELECT statement to request that only the Q-Q plot be printed. The DATA statement specifies **fdims** as the working data set. The VAR statement specifies **hgt** as the variable of interest. The QQPLOT statement requests a Q-Q plot for each variable in the VAR statement (in this case only **hgt**). The NORMAL option requests that the reference line be from a normal distribution with mean and standard deviation estimated from the data. A data set that is nearly normal will result in a probability plot where the points closely follow the line.

Any deviations from normality lead to deviations of these points from the line. The plot for female heights shows points that tend to follow the line but with some errant points toward the tails. We're left with the same problem that we encountered with the histogram above: how close is close enough?

A useful way to address this question is to rephrase it as follows: what do probability plots look like for data that I ***know*** came from a normal distribution? We can answer this by simulating data from a normal distribution using a DATA step.

```
data sim_norm;
   do i=1 to &n;
      x1=rand('NORMAL',&mean,&std);
      output;
   end;
run;
```

The DO loop tells SAS that we want to perform the actions in the loop *n* times, where *n* is the number of females in the data set. The RAND function draws a single value from a normal distribution with mean and standard deviation equal to those that we estimated previously. The OUTPUT statement adds the newly drawn random value to the data set.

We can then generate a Q-Q plot for the simulated variable.

```
title 'Q-Q plot for simulate x1';
ods select qqplot;
proc univariate data=sim_norm;
   var x1;
   qqplot x1/ normal(mu=&mean sigma=&std);
run;
```

Exercise 3: Make a normal probability plot for variable **x1** in the data set **sim_norm**. Do all of the points fall on the line? How does this plot compare to the probability plot for the real data?

Even better than comparing the original plot to a single plot generated from a normal distribution is to compare it to many more plots. We modify the previous data simulation syntax to generate nine simulated variables (although we could generate as many as we want).

```
data simulated;
   array x {9} x1-x9;
   do i=1 to &n;
      do j=1 to 9;
         x[j]=rand('NORMAL',&mean,&std);
      end;
      output;
   end;
run;
```

The Q-Q plots for the variables can then be generated as follows:

```
title 'Q-Q plots for simulate x1-x9';
ods select qqplot;
proc univariate data=simulated;
   var x1-x9;
   qqplot x1-x9 / normal(mu=&mean sigma=&std);
run;
```

Exercise 4: Does the normal probability plot for **hgt** for females look similar to the plots created for the simulated data? That is, do plots provide evidence that the female heights are nearly normal?

Exercise 5: Using the same technique, determine whether female weights appear to come from a normal distribution.

Normal Probabilities

Okay, so now you have a slew of tools to judge whether a variable is normally distributed. Why should we care?

It turns out that statisticians know a lot about the normal distribution. When we decide that a random variable is approximately normal, we can answer all sorts of questions about that variable related to probability. Take, for example, this question: "What is the probability that a randomly chosen young adult female is taller than 6 feet (about 182 cm)?"[2]

If we assume that female heights are normally distributed (a very close approximation is also okay), we can find this probability by calculating a Z score and consulting a Z table (also called a *normal probability table*).

In SAS, this is done in two steps.

```
data temp;
   result=1 - cdf('NORMAL',182,&mean,&std);
run;

title 'Probability of a female being taller than 182cm';
proc print data=temp;
   var result;
run;
```

We make the necessary calculation within the DATA step, saving the calculation to a DATA step named **temp**. We then use the PRINT procedure to display the contents of the data set.

[2] The study that published this data set is clear to point out that the sample was not random and therefore inference to a general population is not suggested. We do so here only as an exercise.

Note that the function CDF gives the area under the normal curve below a given value with a given mean and standard deviation. Because we're interested in the probability that someone is taller than 182 cm, we have to take 1 minus that probability.

Assuming a normal distribution has enabled us to calculate a theoretical probability. If we want to calculate the probability empirically, we simply need to determine how many observations fall above 182 and then divide this number by the total sample size.

```
title 'Proportion of females taller than 182cm';
proc sql;
select b.n_tall / a.n as proportion
   from (select count(hgt) as n from fdims) as a,
        (select count(hgt) as n_tall from fdims
           where hgt>182) as b;
quit;

title;
```

The SQL procedure uses the Structured Query Language to count the number of females (labeled **a.n**) and the number of females with heights above 182 cm (labeled **b.n_tall**). SQL then calculates the proportion of females above 182 cm as **b.n_tall** divided by **a.n**.

Although the probabilities are not exactly the same, they are reasonably close. The closer that your distribution is to being normal, the more accurate the theoretical probabilities will be.

The TITLE statement at the end of the syntax is used to clear the previous title. This ensures that if we forget to include or choose not to include a title for our next procedure, the output will not automatically be given the title **Proportion of females taller than 182cm**.

Exercise 6: Write out two probability questions that you would like to answer; one regarding female heights and one regarding female weights. Calculate the probabilities using both the theoretical normal distribution as well as the empirical distribution (four probabilities in all). Which variable, height or weight, had a closer agreement between the two methods?

On Your Own

1. Now let's consider some of the other variables in the body dimensions data set. Using the figures on the next page, match the histogram to its normal probability plot. All of the variables have been standardized (first subtract the mean and then divide by the standard deviation), so the units won't be of any help. If you are uncertain based on these figures, generate the plots in SAS to check.

   a. The histogram for female biiliac (pelvic) diameter (**bii_di**) belongs to normal probability plot letter ____.

   b. The histogram for female elbow diameter (**elb_di**) belongs to normal probability plot letter ____.

   c. The histogram for general age (**age**) belongs to normal probability plot letter ____.

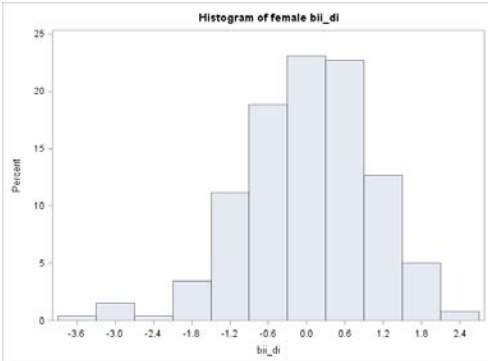
   d. The histogram for female chest depth (**che_de**) belongs to normal probability plot letter ____.

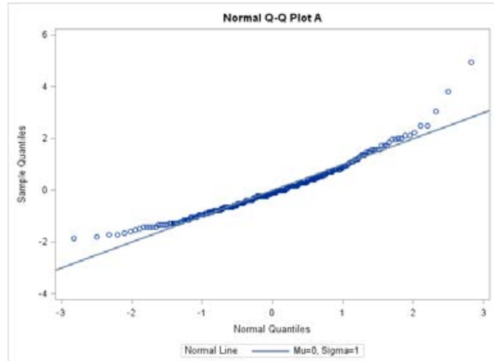
2. Note that normal probability plots C and D have a slight stepwise pattern. Why do you think this is the case?

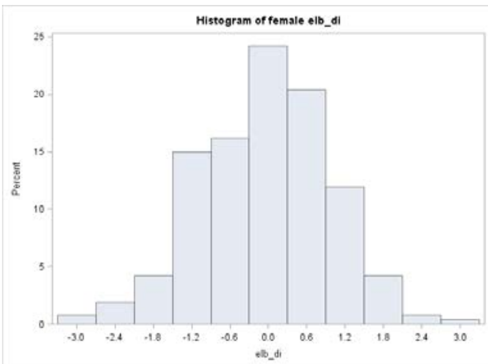
3. As you can see, normal probability plots can be used both to assess normality and visualize skewness.

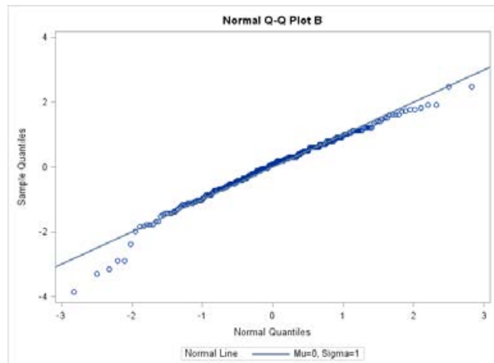
   Make a normal probability plot for female knee diameter (**kne_di**). Based on this normal probability plot is this variable left skewed, symmetric, or right skewed? Use a histogram to confirm your findings.

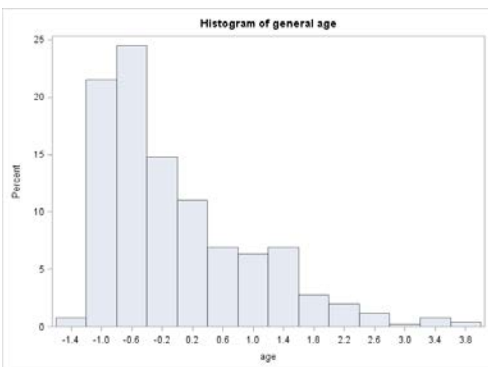
4. What concepts from the textbook are covered in this lab? What concepts, if any, are not covered in the textbook? Have you seen these concepts elsewhere (for example, lecture, discussion section, previous labs, or homework problems)? Be specific in your answer.

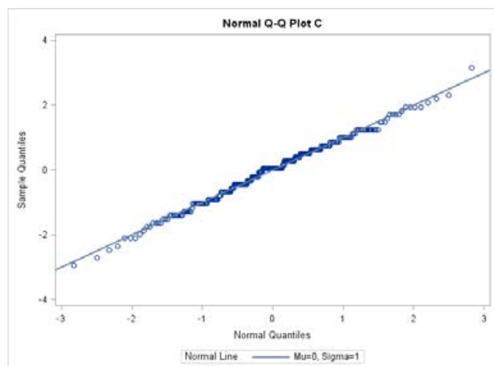

Histogram of female bii_di
Percent
bii_di

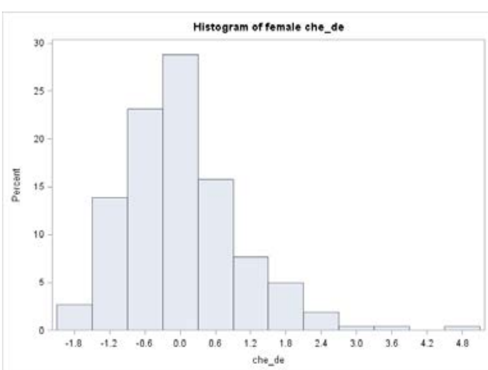

Normal Q-Q Plot A
Sample Quantiles
Normal Quantiles
Normal Line
Mu=0, Sigma=1

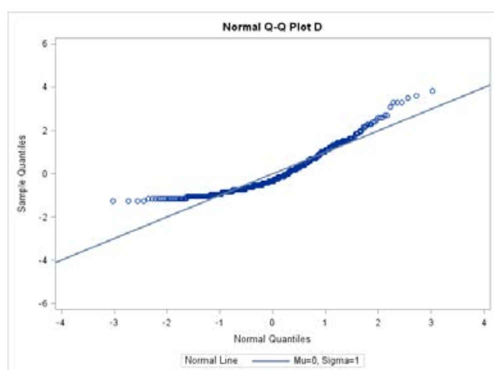

Histogram of female elb_di
Percent
elb_di


Normal Q-Q Plot B
Sample Quantiles
Normal Quantiles
Normal Line
Mu=0, Sigma=1


Histogram of general age
Percent
age


Normal Q-Q Plot C
Sample Quantiles
Normal Quantiles
Normal Line
Mu=0, Sigma=1


Histogram of female che_de
Percent
che_de


Normal Q-Q Plot D
Sample Quantiles
Normal Quantiles
Normal Line
Mu=0, Sigma=1