1 Creating a simulated dataset

The following code will create a simulated dataset with 1000 observations drawn from a normal distribution with mean μ = 2 and standard deviation σ = 1.

data sym:
  seed = 35242;
  retain seed;
  do i = 1 to 1000;
    y = 2 + rannor(seed);
  output;
  end;
  drop seed;
run;

We are going to treat this simulated dataset as a population. This is the complete set of items which we are interested. To display just the first 10 records of this dataset,

proc print data = sym(Cha=10):
  run;

2 Proc means

We can use the proc means to get various summary statistics in a more compact format than proc summary provides. The default statistic provided are

- n = number of observations
- mean
- std dev = standard deviation
- minimum
- maximum

proc means data = sym:
  var y;
  run;

Output:

Analysis Variable: Y

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>2.04229E+0</td>
<td>0.98612E+0</td>
<td>1.34028E+0</td>
<td>3.02318E+0</td>
</tr>
</tbody>
</table>

3 Drawing simple random samples from our population

We will use proc plan to draw a simple random sample of size 10 from our “population” of 1000 values. We can then use proc means to get summary statistics for our simple random sample by copying the selected record id’s from the proc plan output into the code for proc means.

call plan sampling size = 10 of 1000:
  seed = 35242;
  factors a = 10 of 1000:
run;

call plan sampling size = 50 of 1000:
  seed = 35242;
  factors a = 50 of 1000:
run;

Copy the selected record id’s into the code for proc means and delete the separating lines.

4 Confidence intervals

We can request other descriptive statistics by specifying them as part of the proc means statement. One that you will need sooner in the confidence interval for the mean. Putting “clm alpha = .05” on the end of the proc means statement provides a 95% confidence interval.

call plan sampling size = 10 of 1000:
  seed = 35242;
  factors a = 10 of 1000:
  clm alpha = .05:
run;

call plan sampling size = 50 of 1000:
  seed = 35242;
  factors a = 50 of 1000:
  clm alpha = .05:
run;
Analysis Variables: y

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>1.8596915</td>
<td>1.0462922</td>
<td>1.660717</td>
<td>1.147113</td>
</tr>
</tbody>
</table>

5 Assignment for the rest of the lab period

(You may work with another person to do this, but I need at least 2 sets of results (preferably more than 2 sets) for each of 4 kinds of samples from each person.)

1. Do the following 2 times:
   (a) Draw a simple random sample of size 10 from our simulated "population." Use a different seed each time so you get different samples.
   (b) Calculate the sample mean and sample standard deviation from the sample and record it.

2. Do the same for 2 different simple random samples of size 90.

3. Sample from a skewed distribution.
   (a) Use the code below to simulate a dataset from a skewed distribution.

   ```plaintext
   data skewed;
   seed = 325;
   retain seed;
   do i = 1 to 1000;
     y = 2 * rannor(seed, 2);
   output;
   end;
   drop seed;
   run;
   ```

   (b) Use PROC UNIVARIATE to verify that you got a skewed distribution.

   ```plaintext
   proc univariate plot data = skewed;
   var y;
   run;
   ```

   (c) Draw 2 simple random samples of size 10 from the skewed population. Calculate and record the sample mean and standard deviation from each one.

   (d) Do the same for 2 simple random samples of size 90 from the skewed distribution.