STAT 541

Creating Samples in SAS

Creating a Systematic Sample from a Known Number of Observations

- Observations are chosen from data set at regular intervals

SET *data-set-name* POINT= *point-variable;*

- *point-variable* names a temporary numeric variable whose value is the observation number of the observation to be read, must be given a value before SET statement execution, and must be a variable and not a constant value

Creating a Systematic Sample from a Known Number of Observations (continued)

- *point-variable* values should be positive integers less than or equal to the number of observations in the SAS data set
- Assign the value of *point-variable* within the program so that it has a value when the SET statement begins execution.
- The value of *point-variable* must change during DATA step execution so that another observation is selected.

Creating a Systematic Sample from a Known Number of Observations (continued)

- Use the STOP statement to stop processing the current DATA step immediately and resume processing statements after the end of the current DATA step.

```
data everyevenrecord;
do obsnum=2 to 136 by 2;
   set original point=obsnum;
   output;
end;
stop;
run;
```

Creating a Systematic Sample from an Unknown Number of Observations

- When you don't know the number of observations in the data set, use the NOBS= option in the SET statement to determine how many observations there are in a SAS data set.

SET *data-set-name* NOBS= *variable;*

- *variable* is a temporary numeric variable whose value is the number of observations in the input data set

Creating a Systematic Sample from an Unknown Number of Observations (continued)

```
data everyevenrecord;
do obsnum=2 to totobs by 2;
   set original point=obsnum nobs=totobs;
   output;
end;
stop;
run;
```

Creating a Random Sample with Replacement

```
data subset (drop=i totobs);
samplesize=20;
do i =1 to samplesize;
   obsnum=ceil(ranuni(0)*totobs);
   set original point=obsnum nobs=totobs;
   output;
end;
stop;
run;
```

Creating a Random Sample with Replacement (continued)

The RANUNI function generates a number between 0 and 1.

RANUNI (*seed*)

where *seed* is a nonnegative integer less than 2,147,483,647

- If 0 is the seed, the computer clock initializes the stream and the stream of random numbers is NOT replicable. Using a specific positive seed will produce replicable results.

Creating a Random Sample with Replacement (continued)

- ranuni(0)*totobs

Using a multiplier (positive integer) with the RANUNI function changes the outcome's range to a number between 0 and the multiplier

- obsnum=ceil(ranuni(0)*totobs);

obsnum will have a value that ranges from 1 to totobs (total number of observations) because the CEIL function returns the smallest integer that is **greater than or equal** to the argument

Creating a Random Sample without Replacement

```
data subset (drop=obsleft samplesize);
samplesize=20;
obsleft=totobs;
do while (samplesize>0);
  obsnum+1;
 if ranuni(0)<samplesize/obsleft then do;
    set original point=obsnum nobs=totobs;
    output;
    samplesize=samplesize-1;
  end;
  obsleft=obsleft-1;
end;
stop;
run;
```

Creating a Random Sample without Replacement (continued)

- Each observation in the original data set is considered for selection only once.
- *samplesize* is the number of observations to read into the sample and decreases by 1 per DO loop iteration
- *obsleft* is the number of observations in the original data set that have not yet been considered for selection and decreases by 1 per DO loop iteration
- *totobs* is the total number of observations in the original data set
- *obsnum* is the number of the observation considered for selection (starting value is 0 and increments by 1 per DO loop iteration)
- When the IF-condition is true, the observation (as per *obsnum* value) is selected, and not selected otherwise.