Introduction; Descriptive & Univariate Statistics

I. KEY CONCEPTS

A. Population. Definitions:

1. The entire set of members in a group. EXAMPLES: All U.S. citizens; all Notre Dame Students.
2. All values of a variable in a definable group (e.g. Catholic, Protestant, Jewish)
3. The set of all values of interest

B. Sample - A subset of a population. We usually analyze samples - samples are supposed to tell us about the population.

1. Probability sample - a subset of the population for which all members had a known, non-zero probability of inclusion in the sample.

2. Random sample - a subset of the population for which all members had an equal probability of inclusion in the sample.

C. Population parameters - numerical characteristics of a population (the mean, median, mode are some of the simpler examples of population parameters)

D. Sample statistics, or sample parameters - sample estimates of population parameters.

II. LEVELS OF MEASUREMENT

There are different levels of measurement. That is, the way we interpret the numbers we assign to our measurements depends upon the level of measurement that is used.

NOMINAL. Nominal measurement is a classification system; we use numbers instead of names to identify things. For example, if we wanted to code religion, we might say 1 = Catholic, 2 = Protestant, 3 = Jewish, etc. That does not mean that a protestant is more religious than a Catholic and less religious than a Jew. The numbers we use our arbitrary, and you can't perform mathematical operations (i.e. add a Catholic and a protestant and get a Jew). Categories should be mutually exclusive and exhaustive. That is, you should only be able to classify something one way, and you should have a category for every possible value.

ORDINAL. With ordinal measurement, categories are ranked in order of their values on some property. Class ranks are an example (highest score, second highest score, etc.) However, the distances between ranks do not have to be the same. For example, the highest scoring person may have scored one more point than the 2nd highest, she may have scored 5 more points than the third highest, etc.

INTERVAL. With interval level measurement, the distance between each number is the same. For example, the distance between 1 and 2 is the same as the distance between 15 and 16.

With interval measurement, we can determine not only that a person ranks higher but how much higher they rank. You can do addition and subtraction with interval level measures, but not multiplication and division.

RATIO. With ratio level measures, you can do addition, subtraction, and multiplication and division. With ratio measures, you have an absolute, fixed, and nonarbitrary zero point.

EX: Fahrenheit and centigrade scales of temperatures are interval-level measures. They are not ratio-level because the zero point is arbitrary. For example, in the F scale 32 degrees happens to be the point where water freezes. There is no reason you couldn't shift everything down by 32 degrees, and have 0 be the point where water freezes. Or, add 68, and have 100 be the freezing point. The zero point is arbitrary. It is not correct to say that, if it is 70 degrees outside, that it is twice as warm as it would be if it were 35 degrees outside.

Such things as age and income, however, have nonarbitrary zero points. If you have zero dollars, that literally means that you have no income. If you are 20 years old, that literally means you have been around for 20 years. Further, $10,000 is exactly twice as much as $5,000. If you are 20, you are half as old as someone who is 40.

III. SUMMARY MEASURES OF CENTRAL TENDENCY.

We often hear the terms "average" or "typical" used. But, what the speaker means may vary.

A. Mode - value that occurs most often
B. Median - middle value in a set of numbers arranged in order of magnitude
C. Mean - Arithmetic average of all numbers; the sum of all values divided by the total number of values.
D. Appropriateness of each measure:

1. Mode is the only appropriate measure for nominal data. Other measures make no sense.

2. Median is most appropriate for ordinal data - uses only the rank order, ignores distance. Is also sometimes good for interval and ratio level data that have some extreme values - for example, income figures could be misleading if the sample or population includes a few multi-millionaires.

3. Mean uses both rank order and distance between ranks. As noted above, outliers are potentially problematic.

NOTE: In a normal distribution, the mean, median and mode are all the same.

IV. MEASURES OF DISPERSION.

A. Rationale. We often want to know how much variability, or spread, there is in the numbers. For example, suppose the average income is $25,000. It could be that most people have incomes ranging from $24,000 - $26,000, or the range of values could be from $1,000 to $100,000. Hence, we would like to have some sort of measure of dispersion.

A good measure of dispersion should:

1. Be independent of the mean. You could add or subtract the same value to all cases, and the measure would not change value.
2. Should take into account all observations, rather than just a few selected values.
3. Should be convenient to manipulate mathematically.

B. Measures of dispersion.

1. The Range = absolute difference between the highest and lowest values. Meets criteria 1 and 3, but not 2.

2. Variance = average squared deviation about the mean. The variance meets all three criteria.

NOTE:

a. If you don't square, the sum will equal 0.
b. You could use the absolute value of the difference, but this is hard to work with mathematically.

3. Standard Deviation = square root of the variance. The standard deviation is usually more convenient for interpreting variability, since σ is in the same units as the original data.

C. For a normal distribution (bell curve) about 68% of all values fall within one standard deviation to either side of the mean, and about 95% fall within 2 standard deviations.

D. Sample variance uses N - 1 instead of N in the denominator. Estimating the mean eats up 1 degree of freedom - 1 number cannot vary. With large Ns, this is trivial.

V. SHAPES OF DISTRIBUTIONS.

A. Being able to describe shape is often more helpful than just being able to describe the central location or spread of a set of numbers.

B. Properties of shapes:

1. Unimodal, Bimodal, or Multimodal - can have one or more modes.
2. Symmetry - distribution has same shape on both sides.
3. Skewed - Not symmetric. If most of the values fall to the right of the mode, distribution is skewed positively. With positive skew, Mean > Median > Mode. We usually only worry about skewness in extreme cases.

VI. ADDITIONAL COMMENTS RELATED TO THE FOLLOWING FORMULAS FOR *UNIVARIATE STATISTICS*.

A. We sometimes add subscripts to the mean, variance, or s.d. (e.g. μ_x) because we may be looking at more than one variable, e.g. we might have both X and Y and we want to distinguish between them.

B. With summation, Σ, we often leave off the i=1 to N; when we do this it is assumed summing is done across all cases. Or, we might just put an index letter (e.g. i) underneath Σ, meaning summation is done across all values of that index.

C. The alternative formulas for the population variance have the advantage of computational simplicity. You only have to make one pass through the data. With the original formulas, you have to make two passes through the data, since you must first compute the mean.

D. When intervals are used in a frequency distribution, the interval actually starts one-half unit before the first point and ends one-half unit after the last point. For example, the interval 100-199 actually stretches from 99.50 to 199.50.

Univariate statistics

Population mean:

$$\frac{1}{N}\sum_{i=1}^{N} X_i = \mu = \mu_X$$

NOTE: X_i = value of the ith case on variable X (incidentally, $\Sigma X_i = N\mu$)
N = population size
Σ = Summation. It means add the values for all N cases.

Population mean for a frequency distribution:

$$\frac{1}{N}\sum_{i=1}^{C} X_i f_i = \mu$$

NOTE: X_i = value of the ith category
f_i = number of times that value occurs (incidentally, $\Sigma f_i = N$)
c = number of categories or groups

Population variance:

$$\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2 = \sigma^2 = \sigma_{XX} = \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N} X_i^2 - \mu^2$$

Population variance for a frequency distribution:

$$\frac{1}{N}\sum_{i=1}^{c}(X_i - \mu)^2 f_i = \sigma^2 = \sum_{i=1}^{c} X_i^2 \frac{f_i}{N} - \mu^2$$

Population standard deviation:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2} = \sigma = \sigma_x$$

HYPOTHETICAL EXAMPLE: Consider the following hypothetical values for weekly income from a population of 10 people:

X_i	$(X_i - \mu)^2$	X_i^2
\$100	22,500	10,000
\$150	10,000	22,500
\$200	2,500	40,000
\$250	0	62,500
\$250	0	62,500
\$250	0	62,500
\$250	0	62,500
\$325	5,625	105,625
\$325	5,625	105,625
\$400	22,500	160,000
Σ 2,500	68,750	693,750

So, $\mu = 2{,}500/10 = 250$, $\sigma^2 = 68{,}750/10 = 6{,}875$, $\sigma = 82.9156$. Or, $\sigma^2 = 693{,}750/10 - 250^2 = 69{,}375 - 62{,}500 = 6{,}875$. (Note too that, in this case, the median and mode also equal \$250.)

These numbers can also be written as a frequency distribution,

X_i	f_i	$X_i * f_i$	$(X_i - \mu)^2$	$(X_i - \mu)^2 f_i$	X_i^2	$X_i^2 f_i$
\$100	1	\$100	22,500	22,500	10,000	10,000
\$150	1	\$150	10,000	10,000	22,500	22,500
\$200	1	\$200	2,500	2,500	40,000	40,000
\$250	4	\$1,000	0	0	62,500	250,000
\$325	2	\$650	5,625	11,250	105,625	211,250
\$400	1	\$400	22,500	22,500	160,000	160,000
Σ	10	2,500		68,750		693,750

So, $\mu = 2{,}500/10 = 250$, $\sigma^2 = 68{,}750/10 = 6{,}875$, $\sigma = 82.9156$. Or, $\sigma^2 = 693{,}750/10 - 250^2 = 69{,}375 - 62{,}500 = 6{,}875$.

NOTE: Often frequency distributions report an *interval* rather than a specific value for X_i (e.g. \$100 - \$199). When intervals are used in a frequency distribution, the interval actually starts one-half unit before the first point and ends one-half unit after the last point. For example, the interval 100-199 actually stretches from 99.50 to 199.50. The *midpoint* of the interval (e.g. \$149.50) is typically used in the calculations.

SPSS Solution. Here is how we could (almost) do the above in SPSS. As noted in the comments, some results are slightly different because SPSS assumes we are analyzing a sample rather than the entire population.

```
* Univar.sps.
* Sample SPSS descriptive statistics example.  Replicates examples in handout.

* This program is really quite short, but these painstakingly detailed
* comment lines stretch it out.  Comment lines are very handy though
* if you are ever trying to figure out why you did something the way you did.

* Also, while I am giving you this program, this could all easily be done
* interactively using SPSS Menus.  In effect, SPSS will generate most
* of this syntax for you.

* First, enter the data.  Normally I would create a separate data file, but
for
* now I will enter the data directly into the program using the
* data list, begin data and end data commands.

data list free / X.
begin data.
100
150
200
250
250
250
250
325
325
400
end data.

* The formats command tells SPSS that X is measured in dollars.
* Not essential, but it helps make the display easier to read.  This could
* also be done using the SPSS Data Editor.  The Var Labels Command
* will also make the output easier to read.

Formats X (dollar8).
Var Labels X "Weekly Income".

* Next, run the frequencies command, indicating what stats I want.
* I used SPSS menus to generate the syntax for this command, but it
* could also be typed in directly.

FREQUENCIES
  VARIABLES=x
  /STATISTICS=STDDEV VARIANCE MEAN MEDIAN MODE SUM
  /ORDER=  ANALYSIS .
```

Frequencies

Statistics

X Weekly Income

N	Valid	10
	Missing	0
Mean		$250.00
Median		$250.00
Mode		$250
Std. Deviation		$87.40
Variance		$7,638.89
Sum		$2,500

X Weekly Income

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	$100	1	10.0	10.0	10.0
	$150	1	10.0	10.0	20.0
	$200	1	10.0	10.0	30.0
	$250	4	40.0	40.0	70.0
	$325	2	20.0	20.0	90.0
	$400	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

```
* In the output, notice that the variance and SD are bigger than what I had in
* my example.
* This is because SPSS assumes we are analyzing a sample, whereas
* in the the example I stated we had the entire population,
* hence SPSS uses a slightly different formula for its calculations.  The
* difference between sample estimates and population parameters
* will be discussed in class.


* Now, here is how to run the problem when the data are already grouped
* in a frequency distribution.
* The variable WGT indicates how often the value occurs in the data.

data list free / X WGT.
begin data.
100    1
150    1
200    1
250    4
325    2
400    1
end data.

Formats X (dollar8).
Var Labels X "Weekly Income"/ Wgt "Weighting Var".

* The Weight command causes cases to be weighted by the # of times
* the value occurs.

Weight by Wgt.
```

```
* Now just run the frequencies again.

FREQUENCIES
  VARIABLES=x
  /STATISTICS=STDDEV VARIANCE MEAN MEDIAN MODE SUM
  /ORDER=  ANALYSIS .
```

Frequencies

Statistics

X Weekly Income

N	Valid	10
	Missing	0
Mean		$250.00
Median		$250.00
Mode		$250
Std. Deviation		$87.40
Variance		$7,638.89
Sum		$2,500

X Weekly Income

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	$100	1	10.0	10.0	10.0
	$150	1	10.0	10.0	20.0
	$200	1	10.0	10.0	30.0
	$250	4	40.0	40.0	70.0
	$325	2	20.0	20.0	90.0
	$400	1	10.0	10.0	100.0
	Total	10	100.0	100.0	

Stata Solution. Stata has a variety of ways of doing similar things. Here I make use of the `summarize, tabstat` and `tab1` commands. Like SPSS, Stata assumes we are analyzing a sample rather than the entire population which causes some of the numbers to differ slightly from our hand calculations. Among other things, note that SPSS uses a separate `weight` command, whereas in Stata weighting is done via a parameter on the statistical procedure command.

```
. * univar.do.
. * It is easier to enter data into the data editor, but for now
. * I will enter it using the input and end commands.
. clear

. input x

            x
  1. 100
  2. 150
  3. 200
  4. 250
  5. 250
  6. 250
  7. 250
  8. 325
  9. 325
 10. 400
 11. end

. label variable x "Weekly Income"

.
. * The summarize command gives basic summary stats.
. * The tabstat command offers more summary stats
. * The tab1 command gives frequencies.
.
. summarize x

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |        10         250    87.40074        100        400


. tabstat x, statistics( count mean median min max sd var sum )

    variable |         N      mean       p50       min       max        sd  variance       sum
-------------+--------------------------------------------------------------------------------
           x |        10       250       250       100       400  87.40074  7638.889      2500
----------------------------------------------------------------------------------------------

. tab1 x

-> tabulation of x

     Weekly |
     Income |      Freq.     Percent        Cum.
------------+-----------------------------------
        100 |          1       10.00       10.00
        150 |          1       10.00       20.00
        200 |          1       10.00       30.00
        250 |          4       40.00       70.00
        325 |          2       20.00       90.00
        400 |          1       10.00      100.00
------------+-----------------------------------
      Total |         10      100.00
```

```
. * The clear command will clear out the data.
. clear

. * Now, we will redo the problem using grouped data and the weight parameter.
. input x wgt

            x        wgt
  1. 100 1
  2. 150 1
  3. 200 1
  4. 250 4
  5. 325 2
  6. 400 1
  7. end

. label variable x "Weekly Income"

. * [fw = wgt] tells Stata to weight each case by the value of wgt.
. * Stata has several other weighting options which can be more powerful
. * and easier to use than SPSS.
. * fw stands for frequency weights.
.
. summarize x [fw = wgt]

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |        10         250    87.40074        100        400


. tabstat x [fweight=wgt ], statistics( count mean median min max sd var sum )

    variable |         N      mean       p50       min       max        sd  variance       sum
-------------+-----------------------------------------------------------------------------------
           x |        10       250       250       100       400  87.40074  7638.889      2500
-----------------------------------------------------------------------------------------------

. tab1 x [fw=wgt]

-> tabulation of x

     Weekly |
     Income |      Freq.     Percent        Cum.
------------+-----------------------------------
        100 |          1       10.00       10.00
        150 |          1       10.00       20.00
        200 |          1       10.00       30.00
        250 |          4       40.00       70.00
        325 |          2       20.00       90.00
        400 |          1       10.00      100.00
------------+-----------------------------------
      Total |         10      100.00
```