Notes 1: Descriptive Statistics

1. Descriptive and Inferential Statistics

- descriptive statistics are used to describe data
- inferential statistics are used to draw inferences from sample to population
- statistic vs. parameter (M vs. μ), sample vs. census

2. Variables

Anything that varies or takes different values is a variable. Anything that does not vary is a constant.

3. Scales of measurement

Scales	Criteria
Nominal	categories
Ordinal	categories, rank
Interval	categories, rank, equal interval
Ratio	categories, rank, equal interval, true zero point

Scales	Examples
Nominal	Types of flowers, sex, dropout/stay-in, vote/abstain
Ordinal	SES, Likert scales responses, class rank
Interval	time, temperature (in the abstract with no beginning or ending)
Ratio	age, weight, height, time to complete task

- researchers usually do not make a distinction between interval and ratio variables, and it is seldom necessary to do so
- the majority of variables in education are nominal or ordinal; very few interval or ratio variables exist in the social sciences

4. Types of Variables

- independent (IV, cause, predictor) and dependent (DV, effect, criterion), the IV and DV can be identified by noting the temporal order of the variables, the IV will be that which is first in the time sequence (e.g., there will be difference in mathematics scores between males and females; IV = sex, DV = mathematics scores; sex occurs before mathematics scores)
- qualitative, which is also referred to as categorical and nominal (with the simplest case being dichotomous or binary)
- quantitative, which is usually any variable with an underlying continuum (variables measured either at the ordinal, interval, or ratio scale)
- continuous; the measurement of such variables that could theoretically take more refined values (height, weight, IQ); with continuous variables, one only has reported values, not the exact values, to work with (e.g., IQ = 107), but one can establish the "limits of the exact value" by adding and subtracting one-half the unit of measurement from the reported value (e.g., IQ [106.5, 107.5])
- discrete; the measurement of such variables takes only separate values or whole numbers, such as counts of something

5. Hypotheses

Researcher's expectations about research outcomes; either directional, non-directional, or null. These three types of hypotheses are explained below in the matrix.

Type of Hypothesis	Type of Independent Variable	
	Qualitative (Categorical)	Quantitative (Continuous)
Directional	Group differences exist; one group expected to perform better than the other group(s). Example: Group A will do better than group B.	Either a positive or negative relationship will exist. Example: Higher scores on A are associated with higher scores on B. Example: Higher scores on A are associated with lower scores on B
Non-directional	Group differences exist, but it is not clear which group will do better. Example: There will be a difference between groups A and B.	Relationship will exist, but it is not clear if it will be positive or negative. Example: Variable A is associated with variable B.
Null	No difference expected; groups will do the same. Example: There is no difference between groups A and B.	No relationship expected. Example: Variable A is not associated with variable B.

6. Frequency Distributions

A frequency distribution typically displays raw scores in a distribution of scores in rank order and indicates the number of times a given raw score occurred in the distribution. Two types of frequency distributions exist:

- un-grouped frequency: indicates how often each raw score occurred (e.g., each IQ score, each age)
- grouped frequency: with larger range of values, it is often better to group scores into classes or intervals and count frequency by class or interval; determine number of classes by dividing the range by an appropriate number, such as 10 (this will give class widths)
- relative frequency: proportion or percentage of occurrence; appropriate for either un-grouped or grouped frequency
- cumulative relative frequency: running total of relative frequency; shows total of all scores in proportion or percentage as a cumulative for a given score

Examples

Below are ages for a group of students in an introductory statistics course. The un-grouped frequency is displayed for the ages. A grouped frequency is also displayed. Note the relative frequency for both examples.

Scores: 21, 22, 29, 22, 22, 31, 35, 43, 44, 51, 51, and 55

Un-grouped Frequency Distribution

```
                                                    Valid     Cum
Value Label             Value  Frequency  Percent  Percent  Percent
                        21.00         1      8.3      8.3      8.3
                        22.00         3     25.0     25.0     33.3
                        29.00         1      8.3      8.3     41.7
                        31.00         1      8.3      8.3     50.0
                        35.00         1      8.3      8.3     58.3
                        43.00         1      8.3      8.3     66.7
                        44.00         1      8.3      8.3     75.0
                        51.00         2     16.7     16.7     91.7
                        55.00         1      8.3      8.3    100.0
                                 -------  -------  -------
                        Total        12    100.0    100.0
```

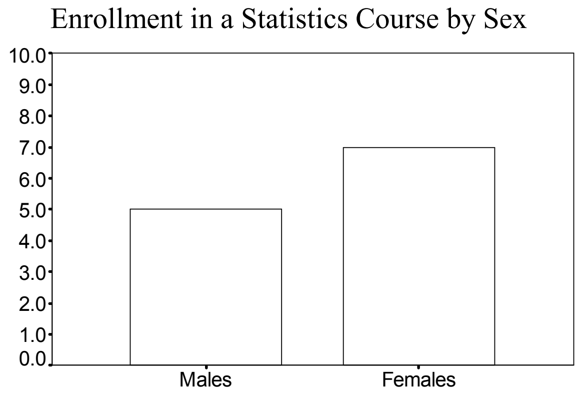
The column denoted as "Valid Percent" is the relative frequency, and the column "Cum Percent" is the cumulative relative frequency.

Grouped Frequency Distribution

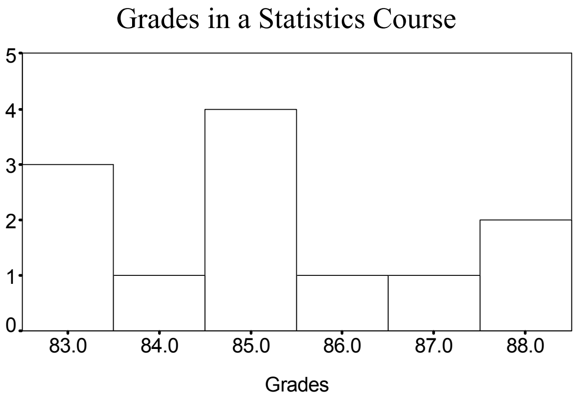
```
                                                  Valid     Cum
Value Label          Values    Frequency  Percent  Percent  Percent
                 20 thur 29         5      41.7    41.7     41.7
                 30 thru 39         2      16.6    16.6     58.3
                 40 thru 49         2      16.6    16.6     75.0
                 50 thur 59         3      25.0    25.0     100.0
                               -------  -------  -------
                      Total        12    100.0    100.0
```

7. Graphs

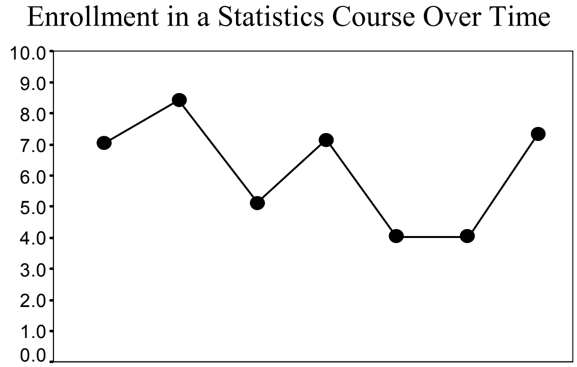
- bar graph: bars represent frequency of score on X; bars don't touch; good for categorical or qualitative variables
- histogram: similar to bar graph, but bars touch for successive scores (such as 84 and 85); good for quantitative variables
- frequency (percentage) polygon: line connects the midpoint dots of histogram; bars of histogram may or may not be present
- times-series graph: a polygon in which X-axis is time and Y-axis (ordinate) is variable of interest
- stem-and-leaf display: first digit(s) base, rest are leaf; this is a refined grouped frequency distribution
- box-and-whisker: shows distribution of scores in quartiles and extreme scores (see below)



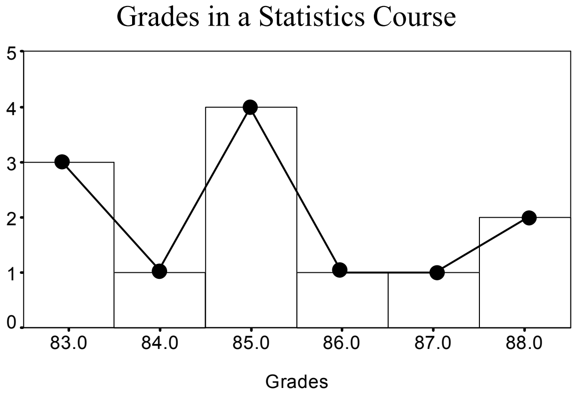
Bar Chart



Histogram



Time-Series



Frequency Polygon

Grades in a Statistics Course

```
Frequency     Stem &  Leaf
     3.00        8 .  333
     1.00        8 .  4
     4.00        8 .  5555
     1.00        8 .  6
     1.00        8 .  7
     2.00        8 .  88

 Stem width:      1.00
 Each leaf:       1 case(s)
```

Stem-and-Leaf #1

Another Set of Grades in a Statistics Course

```
Frequency     Stem &  Leaf
      .00        6 *
     2.00        6 .  56
     3.00        7 *  012
     3.00        7 .  589
     2.00        8 *  23
     6.00        8 .  666779
     3.00        9 *  012
     2.00        9 .  59

 Stem width:     10.00
 Each leaf:        1 case(s)
```

Stem-and-Leaf #2

8. <u>Distributions</u>

- displayed often with smoothed histrograms
- normal
- bi-modal; multi-modal
- rectangular

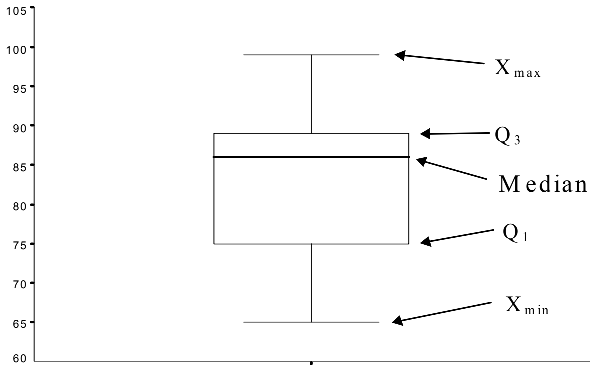
- skewed (positive and negative)

9. Percentiles and Percentile Ranks

- percentiles: points (or scores) in distribution below which a given percent, P, of cases lie (e.g., P_{75} = 110 for IQ)
- percentile ranks: percentage of a distribution which lies below a score (e.g., IQ = 110, PR_{110} = 75)
- quartiles: divides a distribution at three points—P_{25}, P_{50}, and P_{75}—thus creating four quarters which are called quartiles; the first quartile, Q_1, represents the bottom 25% of scores, the second quartile, Q_2, the next 25%, and so on to Q_4.

10. Box-and-whisker display

Another graph that displays various bits of information for a distribution. See example below.



11. Univariate Summary Measures

Univariate refers to one variable. Summary measures refer to indices that provide concise descriptions of distributions of scores, like the mean (average). There are two common types of summary measures:

- Central tendency: a typical score or average
- Variability: the spread or dispersion of scores

12. Notation and Symbols

Before discussing measures of central tendency and variability, it is first important to understand various mathematical symbols used in calculating univariate summary measures.

- X_i: this represents the i^{th} raw score in the distribution; often presented as just X
- n: the sample size; the number of scores in a distribution; sometimes it has a subscript, such as n_2 or n_b, to denote to which group it is referring (e.g., n_g, for sample size of girls)
- $\sum$: just means to sum or add scores; sometimes it is displayed as $\sum_{i=1}^{n=3}$ which means that the total number of observations to sum is three, and one starts the summation at observation number 1
- $\sum X$: sum all the X's (e.g., 1, 2, 3, 4, 5; $\sum N = 1 + 2 + 3 + 4 + 5= 15$)
- $\sum X^2$: sum the square of the X's (e.g., 1^2, 2^2, $3^2 = 1 + 4 + 9 = 14$)
- $(\sum X)^2$: square the sum of the X's (e.g., $1 + 2 + 3 = 6$, then square 6, $6^2 = 36$)

13. Measures of Central Tendency
Central tendency refers to typical or average scores in distribution. There are three commonly used measures of central tendency:

- mode (Mo): most frequent score in distribution; uni-, bi-, tri-, multi-modal distributions; good for nominal or qualitative data, but can also be used with ordinal, interval, and ratio variables
- median (Md, Mdn, X_{50}): score in middle is scores are rank ordered; 50% above, 50% below; good for ordinal data, or interval/ratio data when the distribution is highly skewed (ex. income in US is positively skewed, so use Mdn); not appropriate for nominal data
- mean ($\overline{X}, M$): what one usually thinks of as the average; $\sum X_i / n$ = mean = $\overline{X}$; the balancing point—that point at which the sum of squares [SS] is minimum, and the Sum of Deviations scores will equal exactly zero (both discussed later); best used with ratio or interval data, but may be used with some ordinal data; not appropriate for nominal data.

14. Central Tendency and Distributions
Placement of mean, median, and mode for normal, bi-modal, rectangular, and skewed distributions

15. Mean of Groups
To find the mean of two (or more) groups, use the following formula (or a simply adaptation of it)

$$\overline{X}. = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}$$

where $\overline{X}_1$ is the mean of group 1, $\overline{X}_2$ is the mean of group 2, n_1 is the sample size for group 1 and n_2 is the sample size for group 2, and $\overline{X}.$ is the grand mean.

16. Inferences and Sampling Error
- population mean: if one uses a sample, then the mean is denoted by the symbol $\overline{X}$ or *M*, if one refers to the population (i.e., census) mean, the symbol μ is used (μ is the Greek letter for small m)
- sampling error: the measures of central tendency can be used for inferences from sample to population; any randomly formed error (chance error) between sample statistic (e.g., $\overline{X}$) and population parameter (e.g., μ) is called sampling error:

  sampling error = $statistic - parameter$ (e.g. $\overline{X} - \mu$)

17. Components of Variance
- deviation score ($X - \overline{X}$)= x_i: the mean subtracted from the raw score
- sum of deviation scores $\sum X - \overline{X} = \sum x_i$ sum of all deviation scores
- sums of squares (SS) = $\sum (X - \overline{X})^2 = \sum x_i^2$: square each deviation score and sum
- least squares criterion: the mean, $\overline{X}$, is know as the least-squares criterion because the mean will provide the smallest sum of squares, and since it provides the smallest SS, it is know as the least squares criterion

18. Variability

Variability is the spread or dispersion of the scores (also may be viewed as the tendency for scores to differ from one another or to depart from the typical or average score). There are a number of indices of variability. The most commonly used are:

- exclusive range (R): difference between largest and smallest scores, i.e., $X_{max} - X_{min}$ (e.g., X_i = 3, 6, 2, 9: $X_{max} = 9$, $X_{min} = 2$, R = 9 – 2 = 7)
- inclusive range: we won't use this range, but some statisticians use the inclusive range which is defined as the difference between largest and smallest score plus 1, i.e., $X_{max} - X_{min} + 1$ (e.g., X_i = 3, 6, 2, 9: $X_{max} = 9$, $X_{min} = 2$, R = 9 – 2 + 1 = 8)

- sample variance (VAR or s^2): find SS and divide it by n - 1, i.e.,

$$s^2 = \frac{\sum(X-\overline{X})^2}{n-1} = \frac{\sum x^2}{n-1} = \frac{SS}{n-1}$$

the computational formula is

$$s^2 = \frac{n\sum X^2 - (\sum X)^2}{n(n-1)}$$

- population variance (σ^2):

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

- sample standard deviation (SD or s):

$$s = \sqrt{s^2} = \sqrt{\frac{\sum(X-\overline{X})^2}{n-1}} = \sqrt{\frac{\sum x_i^2}{n-1}}$$

so s is just like the variance except in the original scale of measurement

- population standard deviation (σ):

$$\sigma = \sqrt{\sigma^2}$$

Why isn't the range as good a measure of variability as s?

Note: Always calculate to three decimal places

Example Calculation of Central Tendency and Variability

Find the three measures of central tendency, s^2, and s for the following:

9, 10, 5, 6, 5, 7, 8, and 5.

M = (5 + 6 + 5 + 7 + 8 + 9 + 10 + 5)/8 = (55)/8 = 6.875

Mo = 5

Mdn = 5 5 5 6 7 8 9 10, 6 and 7 are the middle values, so (6 + 7)/2 = 6.5

Raw Scores	**Mean**	**Deviation Scores**	**Deviation Scores Squared**
X_i	M	$x = X - M$	$x^2 = (X - M)^2$
5	6.875	-1.875	3.516
6	6.875	-0.875	0.766
5	6.875	-1.875	3.516
7	6.875	0.125	0.016
8	6.875	1.125	1.266
9	6.875	2.125	4.516
10	6.875	3.125	9.766
5	6.875	-1.875	3.516

$$\text{SS} = \sum\left(X-\overline{X}\right)^2 \text{ or } \sum\left(X-M\right)^2 \text{ or } \sum x^2 = 26.878$$

$$s^2 = \frac{\sum\left(X-\overline{X}\right)^2}{n-1} = \frac{\sum x^2}{n-1} = \frac{SS}{n-1} = \frac{26.878}{8 - 1} = \frac{26.878}{7} = 3.839$$

$$s = \sqrt{s^2} = \sqrt{3.839} = 1.959$$