2 Descriptive Statistics

Reading: SW Chapter 2, Sections 1-6

A natural first step towards answering a research question is for the experimenter to design a study or experiment to collect data from the **population** (i.e. collection of individuals) of interest. In most studies, data are collected on only a subset or **sample** from the **population**. Typically, a number of different characteristics or **variables** are measured on each selected individual.

Once the data are collected, we should summarize the information **graphically** and **numerically**. The actual methods used to summarize data depend on the types of variables that were recorded.

Quantitative versus Qualitative

Simply, a quantitative variable is a variable expressed by a quantity, while a qualitative variable is expressed by a quality (i.e. **categorical**).

Examples:

- number of pregnancies (quantitative)
- eye color (qualitative or categorical)
- age (quantitative)
- ethnic group (qualitative or categorical)

Discrete versus Continuous:

Variables that are expressed numerically can be further subdivided into **discrete** and **continuous** variables. A discrete or counting variable is a variable that takes on a finite or countably infinite number of values, while a continuous variable is a variable that assumes any of the values in at least one interval of the real number line.

Examples:

- number of pregnancies (discrete)
- age (continuous)
- city population size (discrete)
- proportion of population who are HIV+ (continuous)

Nominal versus Ordinal:

Categorical variables are **ordinal** if the order of the categories is meaningful and are **nominal** if the order is unimportant.

Examples:

- stage of cancer: in situ, local, regional, distant (ordinal)
- ethnic group (nominal)

Notes:

1. Continuous variables often have a well-defined measurement scale. For example, time in seconds, temperature in degrees Celsius. However, the scale is often not unique. With continuous variables you should always define the unit of measurement.
2. Discrete variables can be constructed from continuous variables. For example, age is a continuous variable, but the variable X defined by $X = 1$ if age is less than 40, otherwise $X = 2$ is a discrete variable that has been created by categorizing age. Note X is ordinal.
3. A qualitative variable can be **coded** to have numerical values. For example, if the variable is eye color, we might define $X = 1$ if person has blue eyes, $X = 0$ otherwise.
4. A discrete variable that has only two possible values is called **binary**. The variable X above is binary.
5. We are limited in our ability to measure continuous variables. Furthermore, many discrete variables can be analyzed with methods for continuous variables, provided the discrete variables are "close-enough" to being continuous. For example, if scores on a psychological test can take on integer values from 1 to 50, then the score variable is discrete. However, if a sample distribution of the scores contains many of the possible values, then it may be possible to use methods for continuous data for analyzing the discrete data.

REMARK: We commonly use capital letters, say X and Y, to identify variables. This is useful mathematical shorthand that is not intended to confuse you.

Summarizing and Displaying Numerical Data

Suppose we have a sample of n individuals, and we measure each individual's response on one quantitative characteristic, say height, weight, or systolic blood pressure. For notational simplicity, the collected measurements are denoted by $Y_1,\ Y_2, ..., Y_n$, where n is the **sample size**. The order in which the measurements are assigned to the place-holders $Y_1,\ Y_2, ..., Y_n$ is irrelevant.

Two standard numerical summary measures are the **sample mean** $\overline{Y}$ and the **sample standard deviation** s. A numerical summary measure is called a **statistic**, so both the sample mean and standard deviation are statistics.

The sample mean is a measure of **central location**, or a measure of a **typical value** for the data set. The standard deviation is a measure of **spread** in the data set. These summary statistics might be familiar to you. Let us consider a simple example to show you how to compute them. Suppose we have a sample of $n = 8$ children with weights (in pounds): 5, 9, 12, 30, 14, 18, 32, 40. Then

$$\overline{Y} = \frac{\sum_i Y_i}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$
$$= \frac{5 + 9 + 12 + 30 + 14 + 18 + 32 + 40}{8} = \frac{160}{8} = 20.$$

The sample standard deviation is the square root of the sample variance given by the formula:

$$s^2 = \frac{\sum_i (Y_i - \overline{Y})^2}{n-1} = \frac{(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + \cdots + (Y_k - \overline{Y})^2}{n-1}.$$

For hand calculations, it is common to create a **table** from which s is computed, as below:

Data	Deviation	Squared Deviation
5	5-20 = -15	(-15)^2 = 225
9	9-20 = -11	(-11)^2 = 121
12	12-20 = - 8	(- 8)^2 = 64
14	14-20 = - 6	(- 6)^2 = 36
18	18-20 = - 2	(- 2)^2 = 4
30	30-20 = 10	10^2 = 100
32	32-20 = 12	12^2 = 144
40	40-20 = 20	20^2 = 400

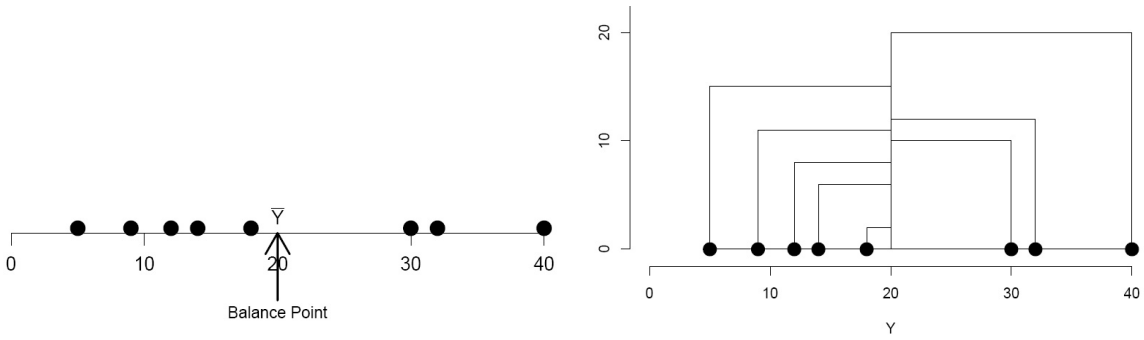
The sample variance is obtained by adding the entries in the last column and dividing by $n-1$:

$$s^2 = \frac{225 + 121 + 64 + 36 + 4 + 100 + 144 + 400}{8-1} = \frac{1094}{7} = 156.3.$$

Thus, $s = \sqrt{s^2} = 12.5$. Summary statistics have well-defined units of measurement, for example, $\overline{Y} = 20lb$, $s^2 = 156.3lb^2$, and $s = 12.5lb$. The standard deviation is often used instead of s^2 as a measure of spread because s is measured in the same units as the data.

REMARK: If the divisor for s^2 was n instead of $n-1$, then the variance would be the average squared deviation observations are from the center of the data as measured by the mean.

The following graphs should help you to see some physical meaning of the sample mean and variance. If the data values were placed on a "massless" ruler, the balance point would be the mean (20). The variance is basically the "average" (remember n-1 instead of n) of the total areas of all the squares obtained when squares are formed by joining each value to the mean. In both cases think about the implication of unusual values (**outliers**). What happens to the balance point if the 40 were a 400 instead of a 40? What happens to the squares?



The **sample median** M is an alternative measure of central location. The measure of spread reported along with M is the **interquartile range**, $IQR = Q_3 - Q_1$, where Q_1 and Q_3 are the **first** and **third** quartiles of the data set, respectively. To calculate the median and interquartile range, order the data from lowest to highest values, all repeated values included. The ordered weights are

$$5 \;\; 9 \;\; 12 \;\; 14 \;\; 18 \;\; 30 \;\; 32 \;\; 40.$$

The median M is the value located at the half-way point of the ordered string. There is an even number of observations, so M is defined to be half-way between the two middle values, 14 and 18. That is, $M = .5(14 + 18) = 16lb$. To get the quartiles, break the data into the lower half: 5 9 12 14, and the upper half: 18 30 32 and 40. Then

$$Q_1 = \text{first quartile} = \text{median of lower half of data} = \tfrac{9+12}{2} = 10.5lb,$$

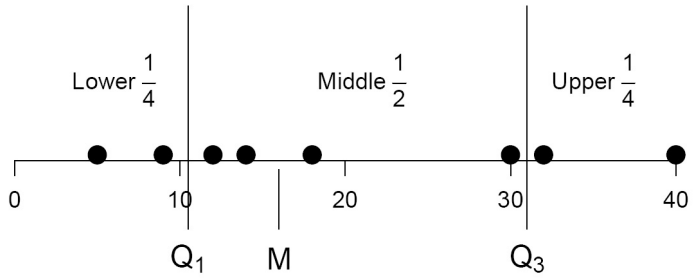
and

$$Q_3 = \text{third quartile} = \text{median of upper half of data} = .5(30{+}32) = 31lb.$$

The interquartile range is

$$IQR = Q_3 - Q_1 = 31 - 10.5 = 20.5lb.$$

The quartiles, with M being the second quartile, break the data set roughly into fourths. The first quartile is also called the 25^{th} percentile, whereas the median and third quartiles are the 50^{th} and 75^{th} percentiles, respectively.. The IQR is the **range** for the middle half of the data.



Suppose we omit the largest observation from the weight data:

5 9 12 14 18 30 32.

How do M and IQR change? With an odd number of observations, there is a unique middle observation in the ordered string which is M. Here $M = 14lb$. It is unclear which half the median should fall into, so M is placed into both the lower and upper halves of the data. The lower half is 5 9 12 14, and the upper half is 14 18 30 32. With this convention, $Q_1 = .5(9 + 12) = 10.5$ and $Q_3 = .5(18 + 30) = 24$, giving $IQR = 24 - 10.5 = 13.5(lb)$.

If you look at the data set with all eight observations, there actually are many numbers that split the data set in half, so the median is not uniquely defined, although "everybody" agrees to use the average of the two middle values. With quartiles there is the same ambiguity but no such universal agreement on what to do about it, however, so Minitab will give slightly different values for Q_1 and Q_3 than we just calculated, and other packages will report even different values. This has no practical implication (all the values are "correct") but it can appear confusing.

Minitab Implementation

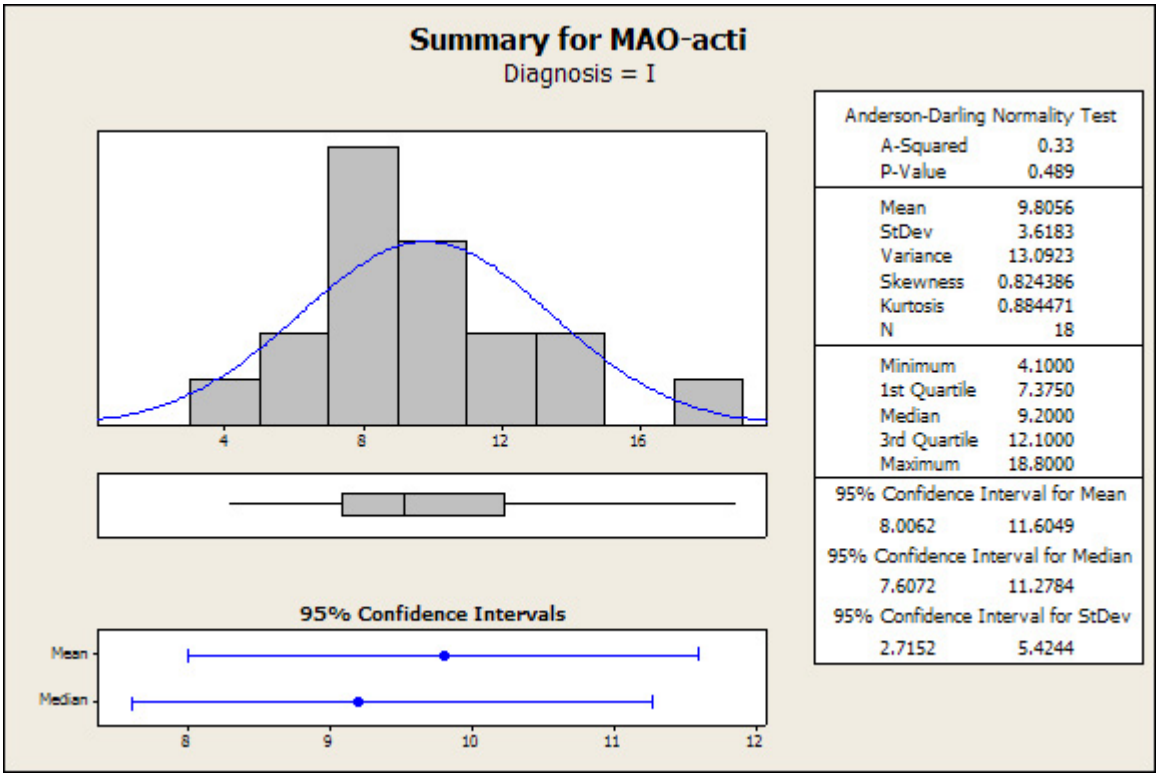
Minitab will automatically compute the summaries we have discussed, and others. Erik will show you how to do this in LAB. Following are numerical and graphical summaries for the data in Example 1.4 pages 3-4 of SW. Monoamine oxidase (MAO) activity expressed as nmol benzylaldehyde product per 108 platelets per hour was measured on schizophrenic patients of three different diagnoses. The data are on the CD in the back of SW.

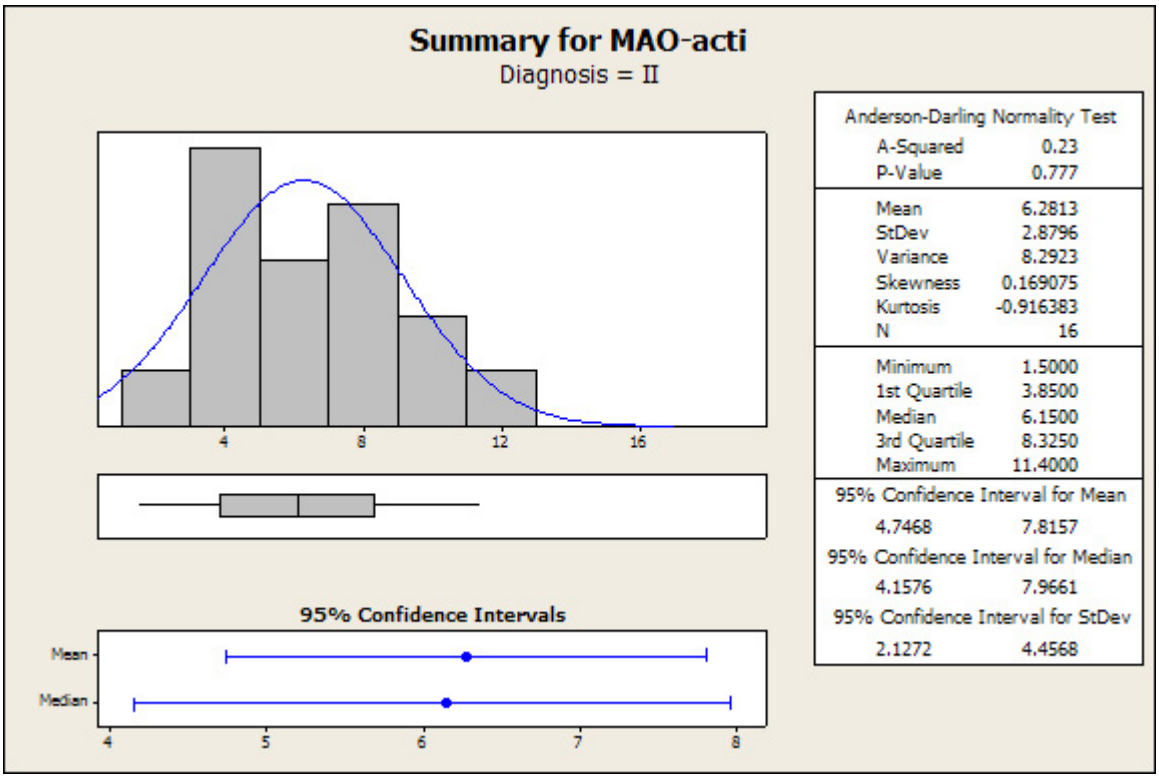
The first display is simple descriptive statistics, the graphs are an enhancement of the simple descriptive statistics. Eric will show you how to obtain both, and how to import into a program like WORD. Let us discuss the output.

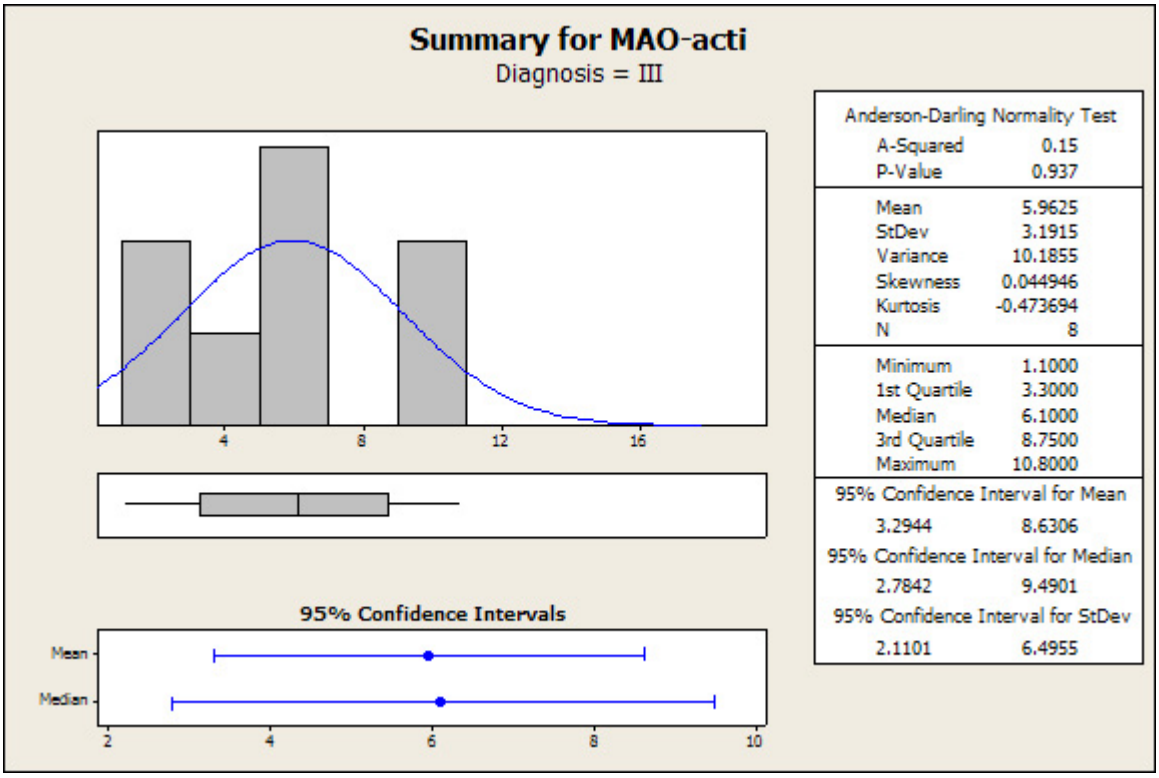
```
Descriptive Statistics: MAO-acti

Variable  Diagnosis   N  N*   Mean  SE Mean  StDev  Minimum     Q1 Median
MAO-acti  I          18   0  9.806    0.853  3.618    4.100  7.375  9.200
          II         16   2  6.281    0.720  2.880    1.500  3.850  6.150
          III         8  10   5.96     1.13   3.19     1.10   3.30   6.10

Variable  Diagnosis      Q3  Maximum
MAO-acti  I          12.100   18.800
          II          8.325   11.400
          III          8.75    10.80
```


Summary for MAO-acti
Diagnosis = I
4
8
12
16
95% Confidence Intervals
Mean
Median
8
9
10
11
12
Anderson-Darling Normality Test
A-Squared 0.33
P-Value 0.489
Mean 9.8056
StDev 3.6183
Variance 13.0923
Skewness 0.824386
Kurtosis 0.884471
N 18
Minimum 4.1000
1st Quartile 7.3750
Median 9.2000
3rd Quartile 12.1000
Maximum 18.8000
95% Confidence Interval for Mean
8.0062 11.6049
95% Confidence Interval for Median
7.6072 11.2784
95% Confidence Interval for StDev
2.7152 5.4244


Summary for MAO-acti
Diagnosis = II
4
8
12
16
95% Confidence Intervals
Mean
Median
4
5
6
7
8
Anderson-Darling Normality Test
A-Squared 0.23
P-Value 0.777
Mean 6.2813
StDev 2.8796
Variance 8.2923
Skewness 0.169075
Kurtosis -0.916383
N 16
Minimum 1.5000
1st Quartile 3.8500
Median 6.1500
3rd Quartile 8.3250
Maximum 11.4000
95% Confidence Interval for Mean
4.7468 7.8157
95% Confidence Interval for Median
4.1576 7.9661
95% Confidence Interval for StDev
2.1272 4.4568



Mean versus Median

Although the mean is the most commonly used measure of central location, it (and the standard deviation) is very sensitive to the presence of extreme observations, sometimes called **outliers**. The median and interquartile range are more **robust** (less sensitive) to the presence of outliers.

For example, the following data are the incomes in 1000 dollar units for a sample of 12 retired couples: 7, 1110, 7, 5, 8, 12, 0, 5, 2, 2, 46, 7. The sample has two extreme outliers at 46 and 1110. For these data $\overline{Y} = 100.9$ and $s = 318$, whereas $M = 7$ and $IQR = 8.3$. If we hold out the two outliers, then $\overline{Y} = 5.5$ and $s = 3.8$, whereas $M = 6$ and $IQR = 5.25$.

The mean and median often have similar values in data sets without outliers, so in such a case it does not matter much which one is used as the **typical value**. This issue is important, however, in data sets with extreme outliers. In such instances, the median is often more reasonable. For example, is $\overline{Y} = 100.9$ a reasonable measure for a typical income in this sample, given that the second largest income is only 46?

Further Points That Will Emphasized in Class:

1. I will mention another summary measure, the **coefficient of variation**: $CV = 100\% * s/\overline{Y}$.

2. I will briefly discuss how the mean and standard deviation change if the units are changed. For example, what happens in the weight problem if I change units from pounds to ounces?

3. The size of the standard deviation depends on the units of measure. We often use s to compare spreads from different samples measured on the same attribute.