STATISTICS REVIEW

Statistics and the Scientific Method

- In general, the scientific method includes:
    1. A review of facts, theories, and proposals,
    2. Formulation of a logical hypothesis that can be evaluated by experimental methods, and
    3. Objective evaluation of the hypothesis on the basis of experimental results.

- Objective evaluation of a hypothesis is difficult because:
    1. It is not possible to observe all conceivable events (populations vs. samples) and
    2. Inherent variability exists because the exact laws of cause and effect are generally unknown.

- Scientists must reason from particular cases to wider generalities.
- This is a process of uncertain inference.
- This process allows us to disprove hypotheses that are incorrect, but it does not allow us to prove hypotheses that are correct.

- Statistics is defined as the science, pure and applied, of creating, developing, and applying techniques by which the uncertainty of inductive inferences may be evaluated.

- Statistics enables a researcher to draw meaningful conclusions from masses of data.

- Statistics is a tool applicable in scientific measurement.

- Its application lies in many aspects of the design of the experiment. This includes:
    1. Initial planning of the experiment,
    2. Collection of data,
    3. Analysis of results,
    4. Summarizing of the data, and
    5. Evaluation of the uncertainty of any statistical inference drawn from the results.

Types of variables

1. Qualitative variable:
    - One in which numerical measurement is not possible.
    - An observation is made when an individual is assigned to one of several mutually exclusive categories (i.e. cannot be assigned to more than one category).
    - Non-numerical data.
    - Observations can be neither meaningfully ordered nor measured (e.g. hair color, resistance vs. susceptibility to a pathogen, etc.).

2. Quantitative variable:
    1. One in which observations can be measured.
    2. Observations have a natural order of ranking.
    3. Observations have a numerical value (e.g. yield, height, enzyme activity, etc.)

- Quantitative variables can be subdivided into two classes:
    1. Continuous: One in which all values in a range are possible (e.g. yield, height, weight, etc.).
    2. Discrete: One in which all values in a range are not possible, often counting data (number of insects, lesions, etc.).

Steven's Classification of Variables

- Stevens (1966)[1] developed a commonly accepted method of classifying variables.

    1. *Nominal variable*:

        - Each observation belongs to one of several distinct categories.

        - The categories don't have to be numerical.

        - Examples are sex, hair color, race, etc.

    2. *Ordinal variable*:
        - Observations can be placed into categories that can be ranked.
        - An example would be rating for disease resistance using a 1-10 scale, where 1=very resistant and 10=very susceptible.

        - The interval between each value in the scale is not certain.

    3. *Interval variables*:
        - Differences between successive values are always the same.

        - Examples would be temperature and date.

    4. *Ratio variables*:
        - A type of interval variable where there is a natural zero point or origin of measurement.

        - Examples would be height and weight.

        - The difference between two interval variables is a ratio variable.

[1] Stevens, S.S. 1966. Mathematics, measurement and psychophysics. pp. 1-49. *In* S.S. Stevens (ed.) Handbook of experimental psychology. Wiley, New York.

Descriptive measures depending on Steven's scale†

Classification	Graphical measures	Measures of central tendency	Measures of dispersion
Nominal	Bar graphs Pie charts	Mode	Binomial or multinomial variance
Ordinal	Bar graphs Histogram	Median	Range
Interval	Histogram areas are measurable	Mean	Standard deviation
Ratio	Histogram areas are measurable	Geometric mean Harmonic mean	Coefficient of variation

†Table adapted from Afifi, A., S. May, and V.A. Clark. 2012. Practical multivariate analysis 5th edition. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Presenting variables

1. Y_i notation

   - In this course, we are going to use the letter Y to signify a variable using the Y_i notation.
   - Y_i is the i^{th} observation of the data set Y. (Y_1, Y_2, Y_3 . . . Y_n).
   - If Y=1, 3, 5, 9, then Y_1=___ and Y_3=___.

2. Vector notation
   - The modern approach to presenting data uses vectors.
   - Specifically, a vector is an ordered set of n elements enclosed by a pair of brackets.

$$Y=\begin{vmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_n \end{vmatrix}$$

Using numbers from the previous example,

$$Y=\begin{vmatrix}1\\3\\5\\9\end{vmatrix}$$

- Y' is called the transpose of Y.
- The transpose of a column vector is a row vector.
- Using the previous example, $Y'=\begin{vmatrix}1 & 3 & 5 & 9\end{vmatrix}$

Vector math

- A row and column vector can be multiplied if each vector has the same number of elements.
- The product of vector multiplication is the sum of the cross products of the corresponding entries.
- Multiplication between two column vectors requires taking the transpose of one of the vectors.
- For example, if

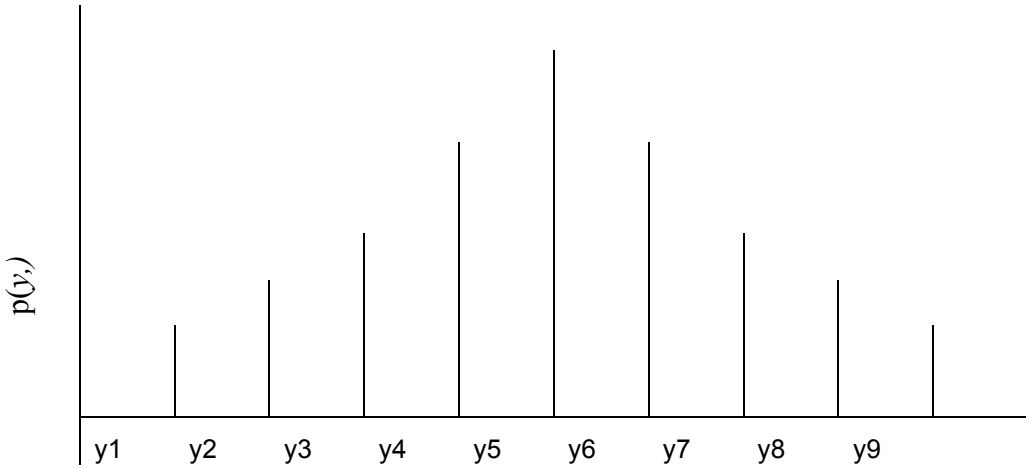
$$X=\begin{vmatrix}1\\3\\4\end{vmatrix} \text{ and } Y=\begin{vmatrix}2\\4\\5\end{vmatrix}$$

$$X'Y=\begin{vmatrix}1 & 3 & 4\end{vmatrix} \times \begin{vmatrix}2\\4\\5\end{vmatrix}$$

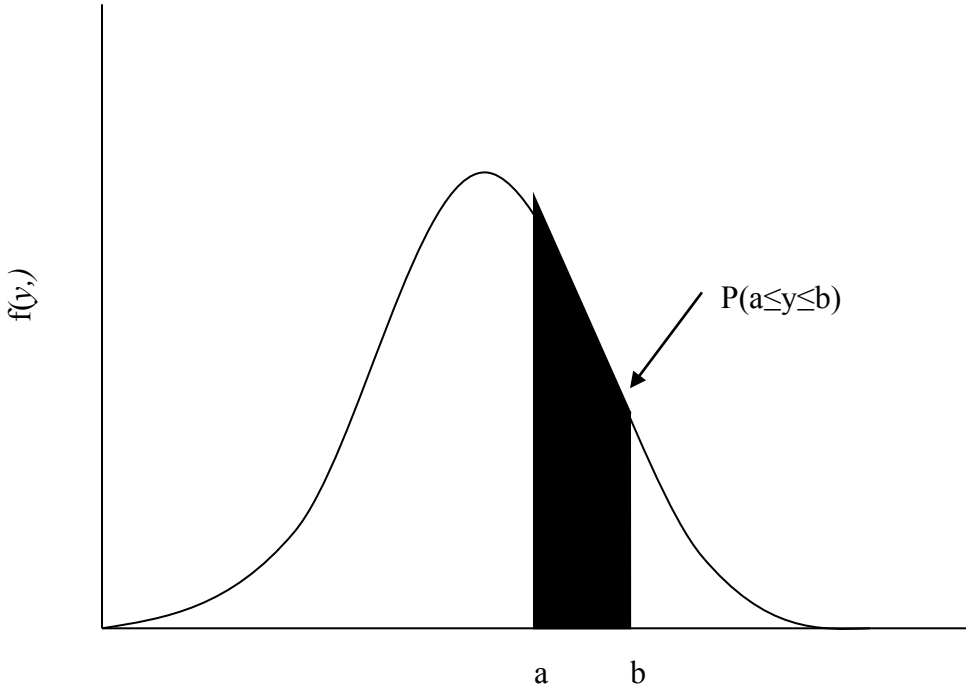
X'Y=(1*2) + (3*4) + (4*5) = 34

Probability Distributions

- The probability structure of the random variable *y* is described by its probability distribution.
- If the variable *y* represents a discrete quantitative variable, we can call the probability of the distribution of *y*, p(y), the probability function of y.
- If the variable *y* represents a continuous quantitative variable, we can call the probability distribution of *y*, f(*y*), the probability density function of *y*.
- In the graph of a discrete probability function, the probability is represented by the height of the function p(y_i).
- In the graph of a continuous probability function, the probability is represented by the area under the curve, f(*y*), for a specified interval.



Distribution of a discrete variable



Distribution of a continuous variable

Properties of Probability Distributions

y discrete: $0 \le p(y_j) \le 1$ all values of y_j

$P(y = y_j) = p(y_j)$ all values of y_j

$\sum p(y_j) = 1$

y continuous: $0 \le f(y)$

$$P(a \le y \le b) = \int_a^b f(y)dy$$

$$\int_{-\infty}^{\infty} f(y)dy = 1$$

Populations vs. Samples

Population:

- Consists of all possible values of a variable.
- Are defined by the experimenter.
- Are characterized by parameters.
- Parameters usually are specified using Greek letters.

Sample:

- Consists of part of the population.
- We often make inferences about populations using information gathered from samples.
- It is extremely important that samples be representative of the population.
- Samples are characterized by statistics.

Item	Parameter	Statistic
Mean	μ	$\bar{Y}$
Variance	σ^2	s^2
Standard deviation	σ	s

Three Measures of Central Tendency

1. Mean (arithmetic)

2. Median – Central value

3. Mode – Most widely observed value

Mean

$$\bar{Y} = (\sum_{i=1}^{n} Y_i)/n = (\sum Y_i)/n$$

- Where: $\bar{Y}$ = mean,
  Y_i is the i^{th} observation of the variable Y, and
  n = number of observations

- We can also use "Y dot" notation to indicate arithmetic functions.

- For example $\sum Y_i$ = Y. where the "." means summed across all i.

- Fact: The sum of the deviations from the mean equals zero.

$$\sum (Y_i - \bar{Y}) = 0$$

Example of Calculating the Mean

	Y_i	$Y_i - \bar{Y}$
	9	9-10=-1
	5	5-10=-5
	14	14-10=4
	12	12-10=2
$\sum Y_i =$	40	$\sum (Y_i - \bar{Y}) = 0$

Step 1. Calculate $\sum Y_i = (9+5+14+12) = 40$

Step 2. Divide $\sum Y_i$ by n
(40/4)=10

- Use of the mean to calculate the sum of squares (SS) allows the calculation of the minimum sum of square.

SS= $\sum (Y_i - \bar{Y})^2$

- The mathematical technique that we will be using in this course for most of our statistical analyses is called the **Least Squares Method**.

Weighted Mean

- An arithmetic technique used to calculate the average of a group of means that do not have the same number of observations.

- Weighted mean= $\bar{Y}_w = \sum n_i \bar{Y}_i / \sum n_i$

Example of Calculating a Weighted Mean

Mean grain protein of barley grown in three North Dakota Counties.

County	n_i	$\bar{Y}_i$
Cass	5	13.5%
Traill	7	13.0%
Grand Forks	3	13.9%

$$\bar{Y}_w = [5(13.5) + 7(13.0) + 3(13.9)]/(5 + 7 + 3)$$
$$= 13.34\%$$

- Non-weighted mean would be (13.5+13.0+13.9)/3 = 13.47%

Median

- Value for which 50% of the observations, when arranged in order from low to high, lie on each side.

- For data with a more or less normal distribution, the mean and median are similar.

- For skewed data, data with a pronounced tail to the right or left, the median often is a "better" measure of central tendency.

- The median often is used as the measure of central tendency when discussing data associated with income.

Example of Determining the Median

Let Y=14, 7, 4, 17, 19, 10, 11

Step 1. Arrange the data from low to high
4, 7, 10, 11, 14, 17, 19

Step 2. Identify the middle observation. This value is the median
11

- If the number of observations in a data set is an even number, the median is equal to the average of the two middle observations.

Mode

- Value that appears most often in a set of data.

- There can be more than one value for the mode.

Three Measures of Dispersion

1. Variance
2. Standard deviation
3. Range

- Measures of dispersion provide information on the variability present in the data set.

- For example, given you have two data sets with the same mean:

Y_1=1, 3, 5, 7 $\bar{Y}_1 = 4$

Y_2=1, 2, 3, 10 $\bar{Y}_2 = 4$

- By looking at only the means, we cannot determine anything about the variability present in the data.

- However, by determining the variance, standard deviation, or range for each data set, we can learn more about the variability present.

Range

- The range is the difference between the maximum and minimum value in the data set.

- For example the range for Y_1=7-1=6 and the range for Y_2=10-1.

Variance of a Sample

Definition formula: $s^2 = \sum (Y_i - \bar{Y})^2 \Big/ (n-1)$

Working formula: $s^2 = \sum Y_i^2 - \dfrac{(\sum Y_i)^2}{n} \Big/ (n-1)$

- The numerator of the formula is called the sum of squares (SS).

- The denominator of the formula is called the degrees of freedom (df).

- df=number of observations – number of independent parameters being estimated
    - A data set contains *n* observations, the observations can be used either to estimate parameters or variability.
    - Each item to be estimated uses one df.
    - In calculating the sample variance, one estimate of a parameter is used, $\bar{Y}$.
    - This leaves *n-1* degrees of freedom for estimating the sample variance.
- The parameter being estimated by $\bar{Y}$ is the population mean μ.

Example of Calculating the Variance Using the Definition Formula

Using the previous data of: Y_1= 1, 3, 5, 7 and $\bar{Y}_1 = 4$
Y_2= 1, 2, 3, 10 and $\bar{Y}_2 = 4$

$$s^2 = \sum (Y_i - \bar{Y})^2 \Big/ (n-1)$$

$$s_1^2 = [(1-4)^2 + (3-4)^2 + (5-4)^2 + (7-4)^2] \Big/ (4-1) = 6.67$$

$$s_2^2 = [(1-4)^2 + (2-4)^2 + (3-4)^2 + (10-4)^2] \Big/ (n-1) = 16.67$$

- It would be very difficult and time consuming to calculate the variance using the definition formula; thus, the working formula generally is used to calculate the variance.

Example of Calculating the Variance Using the Working Formula

$$s^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} \Big/ (n-1)$$

$$s_1^2 = (1^2 + 3^2 + 5^2 + 7^2) - \tfrac{(1+3+5+7)^2}{4} \Big/ (n-1)$$

$$= 84 - \frac{256}{4} \Big/ 3$$

$$= 6.67$$

Variance of a Population

- The variance of a population has a slightly different formula than the variance of a sample.

$$\sigma^2 = \sum (Y_i - \mu)^2 / N$$

Where Y_i= the i^{th} observation in the population,
μ =Population mean, and
N=number of observations in the population.

- The denominator is different in that we do not calculate the df.

- Calculation of the df is not needed since there is no parameter being estimated.

Standard Deviation

- Equal to the square root of the variance.

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

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

Variance of the Mean and the Standard Deviation of the Mean

Given the following means, each with a different numbers of individuals:

County	n_i	$\overline{Y}_i$
Cass	5	13.5%
Traill	7	13.0%
Grand Forks	3	13.9%

First look at the thoughts of the variance of the mean

- In the table above, which mean do you think is the best? Why?

- Would you expect the variability of the mean based on 7 seven observations to be less than one based on 5 or 3 observations?

- In fact, we should not be surprised to learn that sample means are less variable than single observations because means tend to cluster closer about some central value than do single observations.

A different approach to looking at the variance of the mean

- From a population containing N observations, we can collect any number of random samples of size n.
- We can plot the distribution of these n means.
- The distribution of all these n means is called "*The distribution of the sample mean."*
- The distribution of the sample mean, like any other distribution, has a mean, a variance, and a standard deviation.

- There is a known relationship between the variance among individuals and that among means of individuals. This relation and the one for the standard deviation are:

$$\sigma^2_{\bar{Y}} = \frac{\sigma^2}{n} \qquad \text{And} \qquad \sigma_{\bar{Y}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Thus:
$$s^2_{\bar{Y}} = \frac{s^2}{n} \qquad \text{And} \qquad s_{\bar{Y}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$$

- The standard deviation of the mean is more commonly referred to as the **Standard Error**.

Coefficient of Variation (CV)

- The CV is a relative measure of the variability present in your data.
- To know if the CV for your data is large or small takes experience with similar data.
- The size of the CV is related to the experimental material used.
- Data collected using physical measurements (e.g. height, yield, enzyme activity, etc.) generally have a lower CV than data collected using a subjective scale (e.g. lodging, herbicide injury, etc.)

$$\%CV = (s/\bar{Y}) * 100$$

Example of use of the CV for checking data for problems

- Typically, a single CV value is calculated for the experiment; however, a CV can also be calculated for each treatment.

- If the experiment CV is larger than expected, one can look at the CV values for each treatment to see if one or more of them are causal of the experiment CV.

- The individual treatment CVs can be calculated using a program such as Excel.

Treatment	Replicate	Yield	Treatment CV
A	1	89.3	
A	2	78.2	
A	3	97.6	11.0
B	1	76.3	
B	2	107.9	
B	3	70.7	23.6
C	1	56.2	
C	2	52.9	
C	3	120.3	49.7
D	1	78.2	
D	2	83.5	
D	3	80.3	3.3
E	1	79.2	
E	2	69.0	
E	3	75.3	6.9
Expt mean		81.0	
Expt. Standard deviation		17.6	
Expt. CV		21.8	

Linear Additive Model

- We try to explain things we observe in science using models.

- In statistics, each observation is comprised of at least two components:

1. Mean (μ)
2. Random error (ε_i).

Thus, the linear model for any observation can be written as:

$$Y_i = \mu + \varepsilon_i$$

Mean, Variance, and Expected Values

- The mean, μ, of a probability distribution is a measure of its central tendency. Mathematically, the mean can be defined as:

$$\mu = \begin{cases} \sum_{all\ y} yp(y) & \text{y discrete} \\ \int_{-\infty}^{\infty} f(y)dy & \text{y continuous} \end{cases}$$

- The mean also can be expressed as the expected value, where μ = E(y).

$$\mu = E(y) = \begin{cases} \sum_{all\ y} yp(y) & \text{y discrete} \\ \int_{-\infty}^{\infty} f(y)dy & \text{y continuous} \end{cases}$$

- The variability or dispersion of a probability distribution can be measured by the variance, defined as:

$$\sigma^2 = \begin{cases} \sum_{all\ y} (y-\mu)^2 p(y) & \text{y discrete} \\ \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy & \text{y continuous} \end{cases}$$

- The variance also can be expressed as an expectation as:

$\sigma^2 = E[(y-\mu)^2] = V(y)$

Summary of Concepts of Expectation

1. E(c) = c
2. E(y) = μ
3. E(cy) = cE(y) = cμ
4. V(c) = 0
5. V(y) = σ^2
6. V(cy) = $c^2V(y) = c^2\sigma^2$

If there are two random variables y_1, with $E(y_1) = \mu_1$ and $V(y_1) = \sigma_1^2$, and y_2, with $E(y_2) = \mu_2$ and $V(y_2) = \sigma_2^2$, then

7. $E(y_1 + y_2) = E(y_1) + E(y_2) = \mu_1 + \mu_2$

8. $V(y_1 + y_2) = V(y_1) + V(y_2) + 2Cov(y_1, y_2)$

   Where $Cov(y_1, y_2) = E[(y_1 - \mu_1)(y_2 - \mu_2)]$ and this is the covariance of the random variables y_1 and y_2. If y_1 and y_2 are independent, the $Cov(y_1, y_2) = 0$.

9. $V(y_1 - y_2) = V(y_1) + V(y_2) - 2Cov(y_1, y_2)$

   If y_1 and y_2 are independent, then:

10. $V(y_1 \pm y_2) = V(y_1) + V(y_2) = \sigma_1^2 + \sigma_2^2$.

11. $E(y_1 \cdot y_2) = E(y_1) \cdot E(y_2) = \mu_1 \cdot \mu_2$.

However, note that

12. $E\left(\frac{y_1}{y_2}\right) \neq \frac{E(y_1)}{E(y_2)}$

Mean and Variance of a Linear Function

- Given a single random variable, y1, and the constant a1:

  Then $E(a_1y_1) = a_1E(y_1) = a_1\mu_1$

  and

  $V(a_1y_1) = a_1^2V(y_1) = a_1^2\sigma_1^2$

- Given the y_1 and y_2 are random variables and a_1 and a_2 are constants:

  Then $E(a_1y_1 + a_2y_2) = a_1E(y_1) + a_2E(_{y2}) = a_1\mu_1 + a_2\mu_2$

  and

  $V(a_1y_1 + a_2y_2) = a_1^2V(y_1) + a_2^2V(y_2) + 2a_1a_2Cov(y_1, y_2)$

  $= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1a_2Cov(y_1, y_2)$

- If there is no correlation between y1 and y2, then

  $V(a_1y_1 + a_2y_2) = a_1^2V(y_1) + a_2^2V(y_2)$

Expectation of the Sample Mean and the Sample Variance

Sample mean

$$E(\bar{y}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

This relationship works because each y_i is an unbiased estimator of μ.

Sample variance

$$E(S^2) = E\left[E\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] = \frac{1}{n-1}E(SS)$$

where

$$E(SS) = E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] =$$

$$= E\left[\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} 2\bar{y} \bullet y_i + \sum_{i=1}^{n} \bar{y}^2\right]$$

$$= E\left[\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} 2\bar{y}^2 + \sum_{i=1}^{n} \bar{y}^2\right]$$

$$= E\left[\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y}^2\right]$$

$$= E\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right] \qquad \text{because } \sigma^2 = E(y_i^2) - \mu^2 \text{ and } E(\bar{y}^2) = \left(\mu^2 + \frac{\sigma^2}{n}\right)$$

we get

$$= \sum_{i=1}^{n}(\mu^2 + \sigma^2) - n(\mu^2 + \frac{\sigma^2}{n})$$

$$= n(\mu^2 + \sigma^2) - n\mu^2 - \sigma^2$$

$$= n\mu^2 + n\sigma^2 - n\mu^2 - \sigma^2$$

$$= n\sigma^2 - \sigma^2$$

$$= (n-1)\sigma^2$$

Therefore,

$$E(S^2) = \frac{1}{n-1}E(SS) = \sigma^2$$

Thus, we can see that S^2 is an unbiased estimator of σ^2.

Central Limit Theorem

Definition: The distribution of an average tends to be Normal, even if the distribution from which the average is computed is non-Normal.

SAS COMMANDS FOR UNIVARIATE ANALYSIS

```
options pageno=1;
data example;
input y1 y2;
datalines;
3 8
4 10
11 8
5 13
5 8
12 12
8 13
7 10
;;
ods rtf file='example.rtf';
proc print;
*comment this procedure statement will print the data.;
title 'Output of Proc Print Statement';
*comment a title statement following the procedure (proc) statement
will give a title for the output related to the proc statement.  the
title must be within single quotes.;
proc means mean var std min max range stderr cv;
*comment mean=mean var=variance std=standard deviation min=minimum
value max
=maximum value range=range stderr=standard error cv=coefficient of
variation.;
var y1 y2;
title 'Output of Proc Means Statement';
run;
ods rtf close;
Run;
```

Output of Proc Print Statement

Obs	y1	y2
1	3	8
2	4	10
3	11	8
4	5	13
5	5	8
6	12	12
7	8	13
8	7	10

Output of Proc Means Statement

Variable	Mean	Variance	Std Dev	Minimum	Maximum	Range	Std Error	Coeff of Variation
y1	6.87500	10.696	3.270	3.000	12.000	9.000	1.156	47.575
y2	10.2500	4.785	2.187	8.000	13.000	5.000	0.773	21.342