FAQ and Rules

Not all first day handout information is repeated in the course packet, **so don't lose it**.

Be sure to use my correct email address: James P. "Jay" Geaghan = **jgeaghan@lsu.edu**

"Common" courtesies: I expect you to refrain from talking in class and to SILENCE YOUR PHONE.

This is not a distance ed class – attendance is expected.
Absence from quizzes is excused ONLY with an email from your major professor;
you still must provide answers to the daily quiz via email

Attendance at review sessions, when offered, is optional. The intent is to answer your questions; I will have no material to present and no new material will be covered.
There will be no review for the final exam.

How to get SAS on your computer: http://www.stat.lsu.edu/html/sas_information.html

Daily Design

I will be placing a description of an analysis problem on the Internet for each class. You should plan on examining this design before class. At the beginning of each class I will randomly determine whether we have a quiz on that design or not.

I do not intend to spend much time on this daily activity. If there is a quiz, I will allow 5 minutes for you to answer and turn in quiz. If not, I will give you the answers.

We will address these analyses during the course. However, there will be quizzes on material that we have not covered in depth.

Notes on the Labs

Assignments are due the week after they are made unless you have made arrangements with the TA. One point is lost each week an assignment is late up to 50%.

Notes on Exams

Exam examples are posted for you to see my style of questions.

All exams and the final are in the regular classroom.

Corrections of Exam scores must be made within 48 hours of the return of the exam.

Final exams not returned. Grades for final exam%, lab%, and daily quizzes will be posted only with signed permission. Grades will not be provided by email.

On the exam you will be allowed to bring a **calculator.** I do not expect to have a lot of calculations on exams, but there will be some.

You may bring one "cheat sheet" to class; an 8.5 by 11 page written on both sides. I are not collect. The number of pages you may have for the 3 exams and final are 1, 1, 1 & 4 respectively.

I will provide you with t-tables for an exam when I feel they are needed. You will need to understand MY tables. Tables are available in the course packet and in the internet.

Cheating and Plagiarism: Neither will be tolerated, of course. This course is conducted in accordance with university policy concerning cheating and plagiarism.

More fine print: *This syllabus is meant to be suggestive, not absolute. Any and all of the information on this syllabus is subject to change at any time, including exam dates, grading policies, office hours, etc. Changes will be announced in class and via email. We may cover more than what is listed on the syllabus, or less.*

Daily quizzes!!!

First, is there really going to be a daily quiz every day? No, probably not. On the day before each class I will post the description of an experiment or research question. These will be available on the Internet. You will examine the questions and determine the answer. I do not mind if you discuss these questions with your friends under two conditions. One, discussion stops when you get to the classroom, and two, the discussion consists of more than just "What is the answer to today's quiz." Talk it over and try to understand the questions, because similar questions will be on the exams.

Once in the classroom, at the beginning of class I will roll a dice. If it comes up a 6, there will be a quiz. You will have 5 minutes to write your name on a slip of paper and pass it to the front of the room. As soon as I have everyone's answers I will discuss the analysis. (*If you have an excused absence, confirmed by your major professor, you may email me your answer.*)

Once a number has been rolled on a die, that number will not be counted again until after a 6 occurs. For example, if I roll a 3 on the first day; no quiz. If I roll a 3 on the second day it will not count; I will roll again. If it then comes up a 5; no quiz. On the third day 3's and 5's will be skipped, etc. Once a 6 is rolled, all numbers are back in contention.

If things go as planned we will have 5 or 6 quizzes during the semester. I will drop your lowest quiz and grade on the basis of the remaining quizzes.

Instructions: Examine each of the design descriptions below prior to coming to class and determine the answers to the questions. On randomly selected dates a quiz will be given over these designs.

- Z test (one-sample) – tests a mean against an hypothesized value (variance known)
- Z test (two-sample) – tests two means for equality (variance known)
- t test (one-sample) – tests a mean against an hypothesized value (variance unknown)
- t test (two-sample) – tests two means for equality (variance unknown)
- Chi square test variance against an hypothesized value
- Chi square test of independence
- Chi square test of goodness of fit
- F test (one-sample) – tests a variance against an hypothesized value (Chi square test is better)
- F test (two-sample) – tests two variances against each other
- Analysis of Variance – tests two **or more** means for equality
- Regression

Analyses for the daily quizzes

Disclaimer: The analyses described below are intended to cover a broad range of introductory level experiment and analytical types. Many of the designs were inspired by material found on the internet. Where tables or other materials were used I have tried to cite the original internet link and recognize the authors. However, the experiments described below are intended only for teaching purposes, so many have been simplified or modified to accomplish teaching objectives. As a result, the described analyses are not necessarily faithful to the original experiment. It should also be noted that the analyses included were not evaluated nor were they chosen because they are either particularly well done or poorly done.

This semester we will be generally concerned with testing hypotheses. There are many types of hypotheses that may be of interest. Those listed below will be introduced this semester. Note that although our tests are generally applied to samples, we are always testing hypotheses about population parameters and the Null Hypothesis statement should reflect this fact.

Z test (one and two sample)	H_0: $\mu = \mu_0$ where variance (σ^2) is known
One-sample t-test	H_0: $\mu = \mu_0$
Paired t-test	H_0: $\mu = \mu_0$
Two-sample t-test	H_0: $\mu_1 = \mu_2$ (Two-sample Z-test when σ^2 is known)
Chi square test of variance	H_0: $\sigma^2 = \sigma_0^2$
Chi square test of independence	H_0: the two sets of categories are independent
Chi square goodness of fit test	H_0: the pattern of counts follow a given pattern
F test (two-tailed of variance)	H_0: $\sigma_1^2 = \sigma_2^2$ (also one-tailed test but Chi square is better)
Analysis of Variance (one-way)	H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots$
Analysis of Variance (two-way) also Randomized Block Designs	H_{01}: $\mu_{11} = \mu_{12} = \mu_{13} = \ldots$ & H_{02}: $\mu_{21} = \mu_{22} = \mu_{23} = \ldots$
Regression	H_0: $\beta_i = 0$

1) To test a mean for equality to a hypothesized value, use the one sample test. To test for the equality of two means use a two sample test.
   - a) If the variance is known, use a Z test.
   - b) If the variance is not known and must be estimated from a sample, use a t-test.
   - c) The paired t-test is a special case where observations are paired. In this case you can take the difference between the two paired observations and test the mean of the pairs as a single sample of differences.

2) Test a mean against another mean.
   - a) A two-sample t test is used to test one mean against another mean.
   - b) An Analysis of Variance is used to test between means when there are three or more means. It can also be to test only two means, in which case it gives the same result as the t-test most cases. The exception is when the variances for the two means are not equal, in which case the two-sample t-test is more easily modified. Some computer algorithms used to analyze Analysis of Variance will not handle the unequal variance case.
   - c) Analysis of Variance can also be used when means from more than two separate groups are to be compared. For example, if you wish to compare test results for class (freshman, sophomores, juniors and seniors) and for gender (male and female). This is a “two-way” or factorial ANOVA.

3) Testing variances.

a) A variance can be tested against a hypothesized value with a Chi-square test.

b) Two variances can be tested for equality with the F test. These tests can be one-tailed or two-tailed tests.

c) There are some tests available for testing for equality among more than two means. We will only see this test in the context of Analysis of Variance.

4) Chi square tests can also be used to test certain proportional patterns of count data.

a) They can be used to determine if numbers of observations in some categories occur equally frequently or not. Patterns other than "equally distributed" can also be tested.

b) This test can also be used to determine if counts in a two-way table are dependent on the table categories or not.

5) Regression is a different kind of analysis used to relate (actually correlate) two variables. It is used to fit the linear equation $Y_i = b_0 + b_1X_i$

Examples and solutions:

1) According to published reports the production of soybeans in northern Louisiana should average 38 bushels per acre. A researcher has produced an average of 36 bushels lbs per acre on 24 experimental plots (variance = 9 lb per acre squared). He wants to know if his production differs significantly from the published mean.

2) Packing crates used for tomatoes are designed to accommodate a mean size of 3 inches with a standard deviation of no more than 0.3 inches (variance = 0.09 inches squared). An agronomy student has developed a new variety. A sample of 100 tomatoes has a mean of 3.00 inches and a variance of 0.110 inches squared. He wants to determine if the variance for his variety exceeds the required standard by a statistically significant amount.

3) Mendelian genetics dictate that the phenotypic expression of a gene with incomplete dominance should follow a 9:6:1 ratio when heterozygous individuals are crossed. The results of a genetics experiment on butterfly color patterns show a 127:86:17 ratio. The investigator wants to determine if this pattern departs significantly from the expected proportions.

4) A forest products student is developing a new binder for laminated wood products. He wants to compare the mean strength of laminated wood prepared with his new formulation against the commercial standard. He prepares 10 pieces of laminated wood with each binder and measures their strength. The mean values of shear (a measure of strength) for the new binder was 7.33 megapascals (MPa) (s.d. = 0.33) and for the old standard was 8.15 MPa (s.d.=1.15).

5) A sociology student studying college student spending habits. She has interviewed several hundred students and wants to compare the mean credit card balance for freshmen, sophomores, juniors and seniors to determine if there are statistically significant differences in the mean monthly value dollars spent.

Quiz 1)

A researcher investigating water treatment systems in Haiti found “that 14.74% of the sample population without filters suffered a diarrheal episode [56 of 380] compared with 6.51% of the sample population with filters [30 of 461]”. How would we determine if the proportions who suffered illness differ between those with and without filters?

	No filter	**Filter**
No illness	**324**	**431**
Illness	**56**	**30**

Quiz 2)

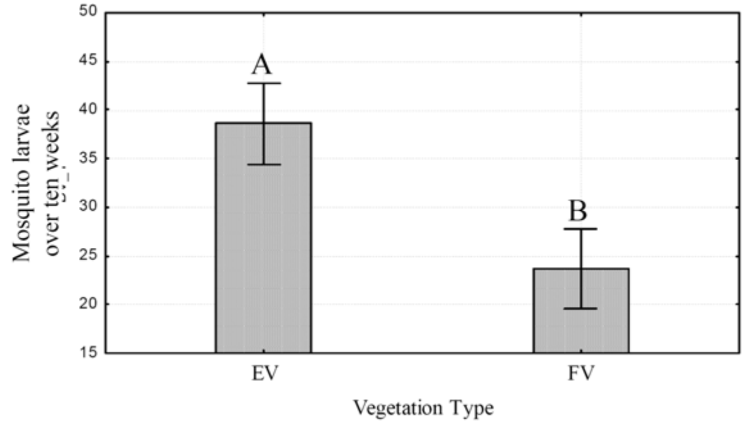


Figure 11. Mosquito density in emergent vegetation (EV) vs floating vegetation(FV).

Part of a large study of mosquito populations examined summer mosquito numbers from two experimental flow-through wetlands at the Olentangy River Wetland Research Park (ORWRP) in Columbus, Ohio. A single sample of larval-stage mosquitoes consisted of 10 dips taken randomly with a white plastic container (11 cm diameter and 350 ml capacity) and an adjustable plastic handle. The larva in the 10 dips were combined into a single composite sample. The objective of the study is to compare the mean of twenty samples from emergent vegetation (from two sites over 10 dates) to the mean of twenty samples from floating vegetation (from the same sites and dates). Mosquito density from outflow regions of two experimental wetlands were measured and normalized by log transformation.

Introduction

Course Objectives

The objectives of this course are to provide the student with an understanding of elemental statistics and to develop the ability to understand and apply basic statistical procedures. Initially we will develop some fundamental concepts of statistics and express data using some basic descriptive methods.

As part of the course we will develop a framework of statistical notation and terminology. This is necessary to understand advanced statistical methodology in the published literature and for communication with statisticians providing statistical support.

We will further develop an understanding and appreciation of statistical inference, particularly hypothesis testing concepts, in order to understand the role of statistics in the decision making processes. The early topics covered involve concepts that will form the foundation for understanding hypothesis testing. We will develop procedures for basic statistical tests of hypothesis and estimation in order to understand and utilize basic statistical methods.

Finally, we will use modern statistical software to apply statistics. Although simple software, such as spreadsheets, are useful for some basic statistical analyses, eventually the user will be severely limited if these are the only applications available. This course is intended to take the user through introductory statistics and into more advanced statistical analysis typically needed for scientific research. The software we currently use is SAS® 9.3, 2004.

The Scientific Method

Understanding the importance of statistics to a myriad of scientific disciplines depends, in part, on recognizing its role in the scientific method. The scientific method is a way of approaching an investigation, and statistics is an integral part of the method. The scientific method is described below as a 5 step process.

1. REVIEW and OBSERVATION: This includes all mechanisms by which a scientist becomes knowledgeable and formulates concepts in his discipline including literature searches, formal course work, laboratory and field observations, and communication with other scientists.
2. HYPOTHESIS: The researcher develops a testable contention about the functioning of some aspect of his discipline. The hypothesis often involves the comparison of the performance of two or more categories, such as comparisons between environments, plant varieties, educational alternatives, pharmaceutical applications or agricultural practices. *Statistical concepts are involved here*
3. EXPERIMENT: The researcher plans and executes an experiment designed to test the hypothesis that has been developed. *Statistical techniques are involved here*
4. EVALUATE the HYPOTHESIS: This involves the analysis of the data gathered in the experiment, and should result in the confirmation or rejection of the hypothesis. *This is a statistical application.*
5. DRAW CONCLUSIONS: Based on the initial understanding of the situation, and on the results of the experimental procedure conducted, the researcher will state a conclusion. Conclusions and interpretation of the results should be stated in the context of the original field of study and may not appear to be inherently statistical.

Some Areas of Statistics

- Descriptive Statistics – graphs and charts
- Exploratory Statistics – a group varying from descriptive statistics to multivariate analyses
- Designed research studies – a variety of scientific experiments
    - Experimental Design – investigator controls individuals (the objective is usually a comparison). These studies often involve statistical testing of differences
    - Sample Survey – individuals are not controlled (e.g. find out "How many" or "How much"). These studies may test for differences or estimate values with confidence intervals.

Organizing, Tabulating and Summarizing Data

The first order of business in many scientific endevours is to find some expression of the data. This may be included in a report as an end in itself, or be used to better understand the results of a study to guide further analysis.

- Descriptive Statistics or Exploratory Statistics
- Frequency distributions
- Graphs and histograms
- Pie charts and star charts
- Drawing Conclusions and Assessing reliability – This will be our main concern

Definitions (you do not have to know these terms verbatim)

- STATISTICAL INFERENCE is the drawing of a conclusion from incomplete information
    - DEDUCTIVE Inference – conclude about a part from knowledge of the whole population
    - INDUCTIVE Inference – conclude about whole from a part
- CONSTANT – a quantity or characteristic whose value remains constant from one individual to another
- VARIABLE – a quantity or characteristic whose value changes from one individual to another
- OBSERVATION – the measurement of some characteristic or variable on an individual
- DATA – a set of observations taken from a group of individuals being studied
- POPULATION – all possible individuals on which a variable may be measured. The total group (as defined by the investigator) about which inferences are to be made.
- SAMPLE – a finite number (subset) of individuals selected from a population for study in a given experiment.
- SAMPLE SIZE – the number of observations or measurements in the sample, usually designated *n*.
- RANDOM SAMPLE – a sample drawn in such a way that every individual in the population has an equal chance of being included in the sample.
- PARAMETER – a summary number that describes a population. It is a constant since it involves measurement of every individual in the population (e.g. μ or σ^2 or β).

- STATISTIC – a summary number that describes a sample. It is a variable since many different samples can be drawn from a population. A statistic is used to estimate a parameter (e.g. $\overline{Y}$ or S^2 or b).
- EXPERIMENT – a planned inquiry to obtain new knowledge or confirm or deny results of previous experiments.
- TREATMENT – a procedure whose effect is to be measured or compared with other experiments.
- EXPERIMENTAL UNIT – the unit to which one application of the treatment is applied.
- SAMPLING UNIT – the unit on which the effect of the treatment is measured. This may be the same as the experimental unit, or smaller than the experimental unit.

CLASSIFICATION OF VARIABLES

- QUALITATIVE – each individual belongs to one or several mutually exclusive categories.
  - Ordinal scale – ranked category variables; e.g. small, medium, large
  - Nominal scale – a classification or group; e.g. male, female
- QUANTITATIVE – an observation resulting from a true numerical measurement.
  - CONTINUOUS – a quantitative variable for which all values within some range are possible (e.g. height, weight, depth). These variables are often grouped in intervals.
  - DISCRETE – a quantitative variable which does not take on all values in a continuum; often the variable can assume integer values only (e.g. number of objects or individuals).

Symbolic Notation

Greek letters are used to indicate PARAMETERS	Arabic (English) letters are used to indicate STATISTICS
μ (means)	$\overline{X}$, $\overline{Y}$ (means)
σ (standard deviations)	S (standard deviations)
β (slopes)	b (slopes)
ρ (correlation)	r (correlation)
τ (experimental treatments)	t (experimental treatments)

Other Symbolic Notation

- Letters at the beginning of the alphabet are used for CONSTANTS (a, b, c)
- Letters at the end of the alphabet are used for VARIABLES (X, Y, Z)
- Letters in the middle of the alphabet (i, j, k, l) are used as subscripts, often italicized (e.g. X_i and Y_{ijk})

Means and measures of central tendency

Means, or the arithmetic "average", are important statistics in characterizing data. It is a measure that provides an indication of where the center of the distributions lies and is an important reference point. In hypothesis testing it is usually the means that are compared to determine of two samples are potentially drawn from the same population or not. Variances and other parameters can also be tested to compare populations, but the test of the means is more common.

Summation Operations

The symbol Σ is used to represent summation. Given a variable, Y_i, representing a series of observations from Y_1 (the first observation) to Y_n (the last observation out of "n" observations), the notation ΣY_i represents the sum of all of the Y_i values from the first to the last. Since the summation is for values of i from 1 to n the summation sign is often subscripted with "i = 1" and superscripted with an n (e.g. $\sum_{i=1}^{n} Y_i$)

Example of Summation: A variable "length of Bluegill in centimeters" is measured for individuals captured in a seine. This quantitative variable will be called "Y", and the number of individuals captured will be represented by "n".

For this example let $n = 4$

The variable Y_i is subscripted in order to distinguish between the individual fish (i)

$$Y_1 = 3,\ Y_2 = 4,\ Y_3 = 1,\ Y_4 = 2$$

Summation operation: To indicate that the sum all individuals in the sample (size n) write.

$$\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + Y_3 + Y_4 = 3 + 4 + 1 + 2 = 10$$, and where, $n = 4$

the mean is given by $\sum_{i=1}^{n} Y_i \Big/ n = 10\big/4 = 2.5$

Sum of Squares

Two other values that will have to be calculated are the "sum of the squares" and the "square of the sums". To indicate the sum of squared numbers, simply indicate the square of the variable after the summation notation.

$$\sum_{i=1}^{n} Y_i^2 = Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2 = 3^2 + 4^2 + 1^2 + 2^2 = 9 + 16 + 1 + 4 = 30$$

where $n = 4$

This is called the Sum of Squares, and should not be confused with the ...

Square of the Sum: The sum was $\sum_{i=1}^{n} Y_i = 10$

The square of the sum is given by simply squaring the sum, $\left(\sum_{i=1}^{n} Y_i\right)^2 = 10^2 = 100$.

Both of these calculations will be needed in calculating the variance.

Measures of Central Tendency

These measures provide an indication of location on a scale. The most common measure is called the arithmetic mean or the “average”. It is the sum of all observations of the variable of interest ($\sum_{i=1}^{n} Y_i$) divided by the number of values summed (n).

Calculation of the Mean

Example: The calculation of the mean for 4 fish lengths.

It was previously determined that

$$\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + Y_3 + Y_4 = 3 + 4 + 1 + 2 = 10$$

where, $n = 4$

The mean is given by $\sum_{i=1}^{n} Y_i \Big/ n = 10/4 = 2.5$

For a larger sample of fish

Y_i = 7, 9, 9, 3, 6, 5, 0, 7, 0, 7

n = 10

ΣY_i = (7 + 9 + 9 + 3 + 6 + 5 + 0 + 7 + 0 + 7) = 53

The mean is then $\sum_{i=1}^{n} Y_i \Big/ n = (7+9+9+3+6+5+0+7+0+7) \Big/ 10 = 53/10 = 5.3$.

Other measures of central tendency

MEDIAN – the central-most observation in a ranked (ordered or sorted) set of observations. If the number of observations is even, take the mean of the center most 2 observations

Example: for the fish sample used earlier, rank the observations

Y_i = 0, 0, 3, 5, 6, 7, 7, 7, 9, 9

If a single observation was in the center it would be used as the median. In this case the number of observations is even and the center falls between two numbers, 6 and 7, so calculate the mean of those two numbers.

MEDIAN = (6 + 7) / 2 = 6.5

MODE – the value of the most frequently occurring observation

Example: For the fish sample,

Y = 0, 0, 3, 5, 6, 7, 7, 7, 9, 9

The most frequently occurring value was “7”.

Therefore, the MODE = 7

MIDRANGE – average of the largest and smallest observation.

Example: The smallest observation in the fish sample was 0 and the largest was 9. The midrange is calculated as the midpoint between these values

MIDRANGE = (0 + 9) / 2 = 4.5

Do not make the mistake of subtracting the lower value from the higher value and dividing by 2. That would be half of the RANGE, not the MIDRANGE.

Percentiles and Quartiles

Percentiles – the value of an observation that has a given percent of the observations below that value and the remaining observations above that value.

The 50th percentile is the value where 50% of the sample observations would have values below it and 50% would be above it. This is also known as the median.

It is often useful to know what value has 5% of the observations below it and 95% above it. This is the 5th percentile. Conversely the 95th percentile is the observation whose value exceeds 95% of the observations in the data set and is exceeded by 5% of the values.

Likewise, the value of the 75th percentile would have 75% of the observations below the value and 25% above.

Quartiles – observations that have one, two or three quarters of the observations above and below their value.

The first quartile is the value of the observation that has one quarter of the observations below it and three quarters above the value. It is the 25th percentile.

The second quartile is the value of the observation that has half (two quarters) of the observations below and above. This value is the same as the MEDIAN or 50th percentile.

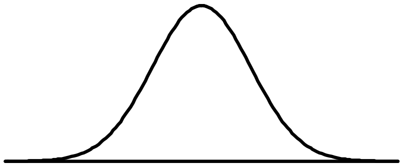
The third quartile is the value of the observation that has three quarters of the observations below and one quarter of the observations above the value. It is the 75th percentile.

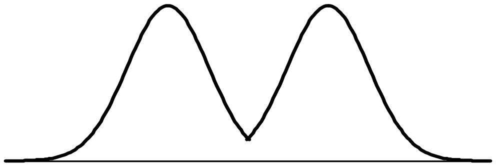
Which measure of Central Tendency is best?

This depends on the distribution. If the distribution is monomodal and symmetric then the

MEAN = MEDIAN = MODE = MIDRANGE

This is true for the NORMAL bell-shaped curve.



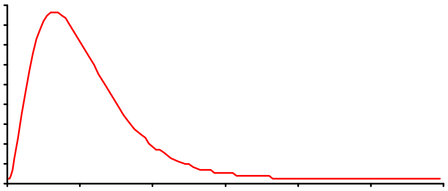


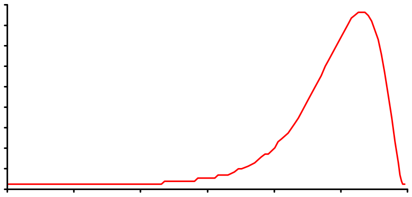
Bimodal distributions are not well described by any measure of central tendency, particularly a single MODE.

Asymmetrical distributions may be best described by the MEDIAN or MODE, depending on the objectives.

POSITIVE SKEW (the long tail of the distribution to the right)

Mode < MEDIAN < Mean





NEGATIVE SKEW

Mean < MEDIAN < Mode

Relative positions of the MEAN, MEDIAN and MODE for asymmetric distributions

- The MEAN is closest to the drawn out tail of the distribution.
- The MODE is farthest from the tail.
- The MEDIAN is intermediate.

Statistic	Negatively Skewed	Symmetric	Positively Skewed
MODE	Largest	Middle	Smallest
MEDIAN	Middle	Middle	Middle
MEAN	Smallest	Middle	Largest

Selecting a measure of central tendency

The MEAN is generally preferred because:

- it utilizes all information in the data set
- it is widely recognized and is easy to work with
- the distribution of the means tends to be normally distributed even if the original observations are not.
- it is generally more sensitive to changes in the form of the distribution (e.g. asymmetry) though this is not always an advantage.

The MEDIAN or MODE may be desirable for asymmetric data sets and may give a more representative measure of location of the center of the data.

Example: Find a "typical" salary for a business employing 5 individuals.

The salaries are: $100,000 $30,000 $20,000 $15,000 $10,000

MEAN = $175000 / 5 = $35,000

MEDIAN = $20,000

MODE – there is no mode unless the data are arbitrarily grouped, and if grouped, two different groupings may not give the same mode

Salary Interval	Frequency	
1 – 10000	1	
10001 – 20000	2	<=== MODE here
20001 – 30000	1	
30001 and over	1	

Since the MEDIAN and the MODE do not use all of the information in the data set for calculation, so they are less sensitive to change.

For example if the top person above gets a raise to $200,000 the MEDIAN and MODE do not change. This can be either an advantage or a disadvantage, depending on the objectives. However, the MEAN would increase from $35,000 to $75, 000.