PRINCIPLES OF DATA ANALYSIS

Peter Avery

MiraCosta College

pavery@miracosta.edu

American Mathematical Association of Two-Year Colleges

31 October 2013

ABSTRACT

Analyzing numerical data can be simplified to the same few guiding principles, whether working with data in one variable or two. Discover these principles, see related student learning outcomes, and complete two extended authentic assessments using real data. Bring your graphing calculator. A complete handout will be provided.

THE PRINCIPLES OF DATA ANALYSIS

PRINCIPLE 1

Start with a graph.

PRINCIPLE 2

Look at the graph for the overall pattern and striking deviations from the overall pattern.

PRINCIPLE 3

Add numerical descriptions of the data.

PRINCIPLE 4

Sometimes the overall pattern is so regular that it can be described mathematically.

STUDENT LEARNING OUTCOME

Analyze the distribution of a quantitative variable.

PRINCIPLE 1

Stemplot

Histogram

PRINCIPLE 2

Shape of distribution - symmetric, skewed, other

Outliers

PRINCIPLE 3

Center - mean, median

Spread - five-number summary, standard deviation

PRINCIPLE 4

Sometimes the overall pattern of a large number of observations is so regular that it can be described by the smooth curve of the Normal distribution.

68-95-99.7 rule

Standard scores

Percentiles of Normal distributions

ASSESSMENT - Women's Literacy Rates in Islamic Nations

Here are data showing the percent of women at least 15 years old who were literate in 2002 in the major Islamic nations. Countries with populations of less than 3 million are omitted. Data for a few nations, such as Afghanistan, Indonesia and Iraq, are not available.

Country	Percent of Women Literate
Algeria	60
Bangladesh	31
Egypt	46
Iran	71
Jordan	86
Kazakhstan	99
Lebanon	82
Libya	71
Malaysia	85
Morocco	38
Saudi Arabia	70
Syria	63
Tajikistan	99
Tunisia	63
Turkey	78
Uzbekistan	99
Yemen	29

A. Give a suitable graphical representation of the distribution. Draw your graph carefully.

B. Comment briefly on the shape of the distribution and list any outliers.

Shape:

Outliers:

C. You have two choices for a numerical summary of the distribution. Circle your choice (1 or 2 below) and give a reason for your choice.

1. Mean and Standard Deviation

2. Five-number summary

Reason:

D. For your choice 1 or 2 in C above, calculate the appropriate numerical summary. Give appropriate units. You may use a calculator.

Mean:

Standard Deviation:

Minimum:

First Quartile:

Median:

Third Quartile:

Maximum:

STUDENT LEARNING OUTCOME

Analyze the relationship between two quantitative variables.

PRINCIPLE 1

Scatterplot

PRINCIPLE 2

Association - direction, form, strength

Outliers

PRINCIPLE 3

Correlation

PRINCIPLE 4

Sometimes the overall pattern is so linear that it can be described by a linear equation.

Regression line

Use of r^2

Prediction

ASSESSMENT - Wine and Heart Disease

Is wine good for your heart?

Drinking moderate amounts of wine may help prevent heart attacks. The data below give yearly wine consumption (liters of alcohol from drinking wine, per person) and yearly deaths from heart disease (deaths per 100,000 people) in 19 developed nations in 2001.

Country	Alcohol from wine	Heart disease deaths
Australia	2.5	211
Austria	3.9	167
Belgium	2.9	131
Canada	2.4	191
Denmark	2.9	220
Finland	0.8	297
France	9.1	71
Iceland	0.8	211
Ireland	0.7	300
Italy	7.9	107
Netherlands	1.8	167
New Zealand	1.9	266
Norway	0.8	227
Spain	6.5	86
Sweden	1.6	207
Switzerland	5.8	115
United Kingdom	1.3	285
United States	1.2	199
West Germany	2.7	172

A. State the explanatory variable, if any.

B. State the response variable, if any.

C. Make a scatterplot that shows how national wine consumption helps explain heart disease death rates. Use a graphing calculator. You do not need to draw the scatterplot on this page.

D. Describe the direction, form and strength of the relationship.

E. Which points, if any, are outliers?

F. Explain in simple language (one sentence) what the direction of the association says about the relationship between wine consumption and heart disease deaths.

G. Calculate the correlation coefficient. Use a graphing calculator.

H. Calculate the equation of the least squares regression line. Use a graphing calculator. Draw this line on your scatterplot.

I. What percentage of the variation in heart disease death rates is explained by the least squares regression line?

J. What percentage of the variation in heart disease death rates is explained by factors other than the least squares regression line?

K. Predict the heart disease death rate for a country where wine consumption is 5 liters of alcohol per person.

L. Predict the heart disease death rate for a country where wine consumption is 20 liters of alcohol per person.

M. Do these data give good evidence that increased wine drinking causes a reduction in the heart disease death rate?

YES / NO (circle one)

N. Justify your choice in M above.

O. Suggest some differences among nations that may be confounded with wine drinking habits.

(Moreover, data about nations may tell us little about individual people. So these data are not evidence that you can lower your risk of heart disease by drinking more wine.)

THE QUESTION OF CAUSATION

THE FLOW OF IDEAS

IDEA 1

A strong relationship between two quantitative variables does not always mean that changes in one variable cause changes in the other - "causation".

IDEA 2

An observed relationship between two quantitative variables may be due to

Causation
Common response
Confounding

or a combination of two or more of these factors.

IDEA 3

The best evidence for causation comes from randomized comparative experiments.

IDEA 4

An observed relationship that may not be due to causation can still be used for prediction.

STUDENT LEARNING OUTCOME

Design an experiment to investigate causation.

ASSESSMENT - Testing Raloxifene

How are new medical drugs tested? Consider the following example of an actual field test of a new drug.

A. How were the subjects chosen to take part in this experiment?

RANDOMLY SELECTED / VOLUNTEERS / OTHER (circle one)

B. State the explanatory and response variables (be precise). For each one, also choose whether it is quantitative (numerical) or categorical.

1. State the explanatory variable:

QUANTITATIVE / CATEGORICAL (circle one)

2. State the response variable:

QUANTITATIVE / CATEGORICAL (circle one)

C. Explain the meanings of the following terms as used in the study.

1. Randomized:

2. Placebo-controlled:

3. Blind:

D. State all the treatments (be precise):

E. State the number of subjects assigned to take the placebo.

F. Draw a diagram that outlines the design of the experiment. Be sure to show the treatments and the response variable.

G. Choose the type of design used (circle one):

Completely randomized design

Matched pairs design

Block design

Other

H. State how you would label the subjects in order to carry out the randomization.

I. Use the following line of random digits, starting at the beginning, to select only the first 5 subjects to be assigned to the first treatment group. Write down the 5 labels.

13873 81598 95052 90908 73592 75186 87136 95761

J. Name the three principles of experimental design:

1.

2.

3.

Did this experiment follow all three principles?

YES / NO (circle one)

K. State the advantages of a well-designed experiment over an observational study in this situation.

L. Name the three principles of basic data ethics:

1.

2.

3.

Did this experiment follow all three principles?

YES / NO (circle one)