American Heritage College Dictionary, 3rd Ed.

1. (used with singular verb) The mathematics of the collection, organization, and interpretation of numerical data, esp. the analysis of population characteristics by inference from sampling.

2. (used with plural verb) Numerical data.

Utts and Heckard

Statistics is a collection of procedures and principles for gathering data and analyzing information in order to help people make decisions when faced with uncertainty.

Why Study Statistics?

Several possible answers to this question:

- "I have to because it is a requirement for my major"
- "I need to in order to advance at work"
- "I want to because I see more and more uses of statistics and I don't really understand the methods",…

Why Study Statistics?

What you should recognize is:

- statistics is now an essential communications tool
- studying statistics will
  - give you access to powerful problem solving and analysis methods
  - the power to make informed decisions about things you hear or read

Types of Statistical Applications

- ***Descriptive Statistics*** uses numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present that information in a convenient form
- ***Inferential Statistics*** utilizes sample data to make estimates, decisions, predictions, or other generalizations about a larger set of data

Basic Terms

Raw data --- numbers and category labels that are collected, but not yet processed

Variable --- a characteristic that differs from one individual to the next

Population data --- measurements taken from all individuals in a population

Sample data --- measurements taken from a subset of a population

Statistic --- a summary measure computed from sample data

Parameter --- a summary measure computed for an entire population

Descriptive Statistics --- summary numbers for either a population or a sample

Reliability---how good is the statistical inference?

- Inferences based on a complete census of the population is "certain"
- Inferences made from samples contain an element of uncertainty—want to be able to make statements about the degree of uncertainty in estimates based on sample data

Types of Data

Qualitative variables---cannot be measured on a natural numerical scale; data classified into categories

- ***Nominal variables***---group or category names that have no inherent ordering
  - Type of transportation used to get to school (walk, bus, bike, drive)
  - Student lives on/off campus (on, off)
- ***Ordinal variables***---group or category names where one response is greater than or less than another
  - Year in school (freshman, sophomore, junior, senior)
  - Exam grades recorded as letter grades (A – F)
  - Clothing sizes (XS, S, M, L, XL)

Quantitative variables---recorded numerical values; the data are either measurements or counts taken on each **individual**

Variables are classified as either continuous or discrete

- ***Continuous variable*** --- A variable where every value within some interval is a possible result;
    - height
    - weight
- ***Discrete variable*** --- A variable which may take on only one of a certain number of possible values
    - number of children in a family
    - number of emergency room admittances each night
    - up-face on the roll of a die

Collecting Data

- ***Published source***---data already collected; published in a book, journal, newspaper
- ***Designed experiment***---sets up strict controls over units in the study; often includes one or more treatment and control groups
- ***Survey***---conducted via phone, mail, email, or in-person
- ***Observational study***---observation of experimental units in their natural setting

Representative Sample---exhibits characteristics typical of those possessed by the target population; required to apply inferential statistical methods, regardless of data collection method

Simple Random sample---use of a selection method that ensures that every subset of fixed size in the population has the same chance of being included in the sample

Explanatory and Response Variables

Many questions are about the **relationship** between ***two variables***.

It is useful to identify one variable as the **independent variable (explanatory variable, predictor, covariate)** and the other variable as the **dependent variable (response variable).**

Generally, the *value of the independent variable* for an individual is thought to **partially explain** the *value of the dependent variable* for that individual.

Example

Age (continuous) + smoking (yes/no) ➔ cancer (yes/no)

Age and smoking are explanatory or independent variables; and cancer is the response

NOTE: unless data are from a randomized experiment, an observed relationship between exploratory and response variables ***does not*** imply a causal relationship.

Exploratory Data Analysis

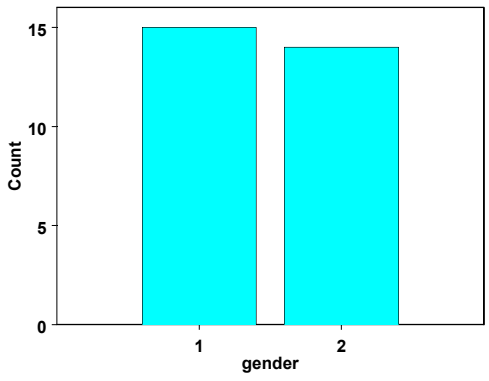
Raw data:

(taken from http://abacus.bates.edu/acad/depts/psychology/SPSSPC/spsspc.html#raw)

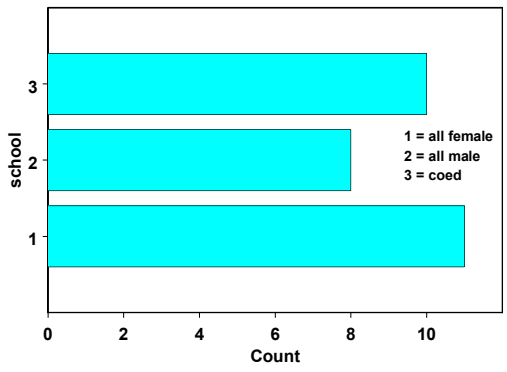
case	gender	school	write	math	esteem	conf
1	1	1	286	279	1	5
2	2	2	281	321	3	6
3	1	1	306	341	3	7
4	1	1	300	298	3	8
5	1	3	277	303	3	5
6	2	3	290	312	4	6

Variable name	Description and coding
case	participant identification number
gender	1 = female, 2 = male
school	type of high school: 1 = all female, 2 = all male, 3 = coed
write	writing score on the National Assessment of Educational Progress test
math	math score on the National Assessment of Educational Progress test
esteem	response to the statement, "On the whole, I am satisfied with myself."; 1 = strongly disagree, 2 = somewhat disagree, 3 = somewhat agree, 4 = strongly agree
conf	response to the question "How do you feel about participating in class discussions." 1 - 10 Likert scale: 1= not at all confident, 10 = very confident

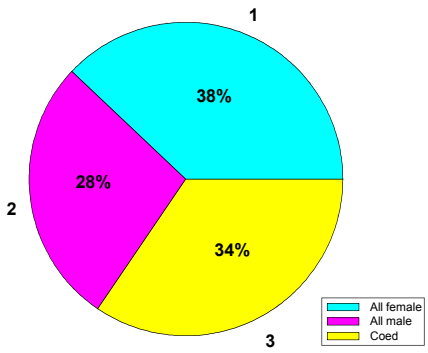
Data Visualization—Vertical bar plot



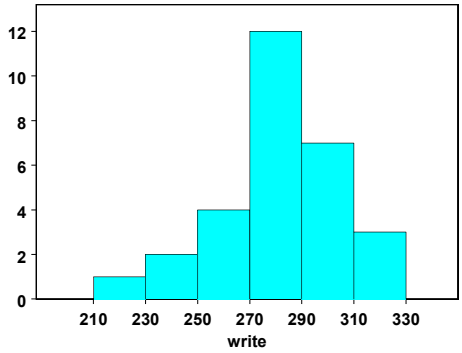
Data Visualization—Horizontal bar plot



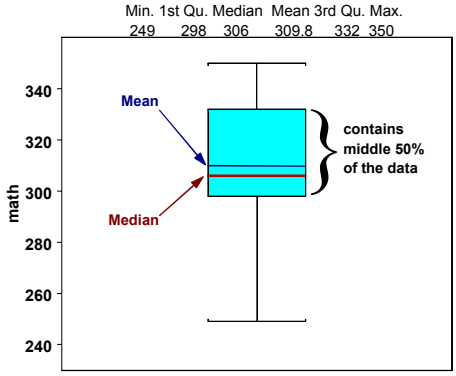
Data visualization—Pie chart



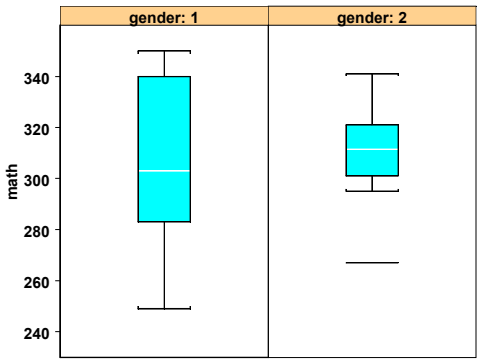
Data Visualization—Histogram



Data Visualization—Box plot 1

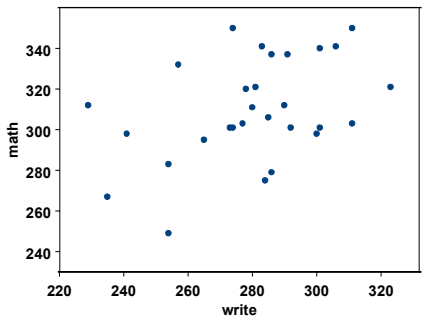


Data Visualization—Box plot 2



Data Visualization—Scatter plot

We will revisit this topic



Summation Notation

Let $y_1, y_2, y_3, \ldots, y_n$ be n measurements from the quantitative data set Y.

Then define:

- ***The sum of all the measurements in the data set***

$$y_1 + y_2 + y_3 + \ldots + y_n = \sum_{i=1}^{n} y_i$$

If $Y = \{286, 281, 306, 300, 277, 290\}$ is a sample of writing scores, then

$$\sum_{i=1}^{6} y_i = 286 + 281 + 306 + 300 + 277 + 290 = 1{,}740 \text{ points}$$

- ***The sum of the squares of the measurements***

$$\sum_{i=1}^{n} y_i^2 = y_1^2 + y_2^2 + \ldots + y_n^2$$

For the writing scores,

$$\sum_{i=1}^{6} y_i^2 = 286^2 + 281^2 + 306^2 + 300^2 + 277^2 + 290^2 =$$

$$505{,}222 \text{ points}^2$$

- ***The sum of the measurements, quantity squared***

$$\left(\sum_{i=1}^{n} y_i\right)^2 = \left(y_1 + y_2 + \ldots + y_n\right)^2$$

$$= (1{,}740)^2$$

$$= 3{,}027{,}600 \text{ points}^2$$

Numerical Descriptive Measures

- ***Measures of central tendency*** **(location)**---tendency of data to cluster or center around certain numerical values
  - ***Mean***---sum of the measurements divided by the number of measurements in the data set

Sample mean, $\bar{y}$, is calculated as

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

For the writing scores sample

$$\bar{y} = \frac{\sum_{i=1}^{6} y_i}{6} = \frac{286+281+306+300+277+290}{6} = \frac{1740}{6} = 290 \text{ points}$$

Population mean, μ (mu), is calculated as

$$\mu = \frac{\sum_{i=1}^{N} y_i}{N}$$

where N is the size of the population

- ***Median***---the middle number when the measurements are arranged in order

What is the middle measurement for the sample of writing scores?

286, 281, 306, 300, 277, 290

277, 281, 286, 290, 300, 306

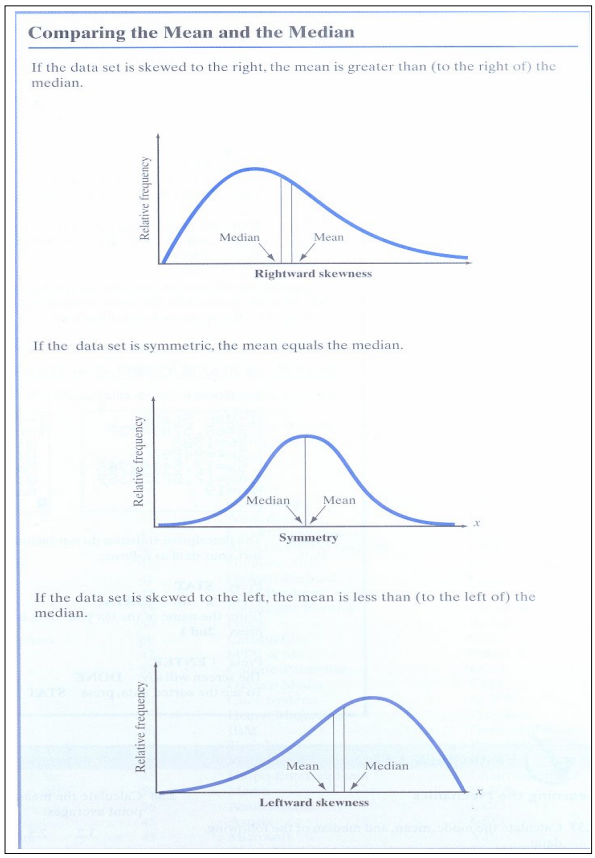
If the number of observations is even, the median is the mean of the middle two numbers---

$$M = \frac{286+290}{2} = 288 \text{ points}$$

If the number of observations is odd, the median is the middle number

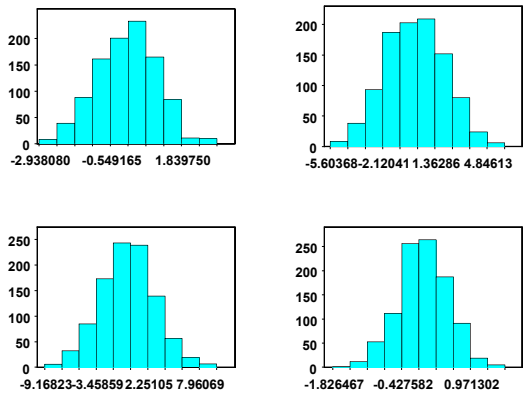
Use $\hat{M}$ for a sample median and use M for a population median

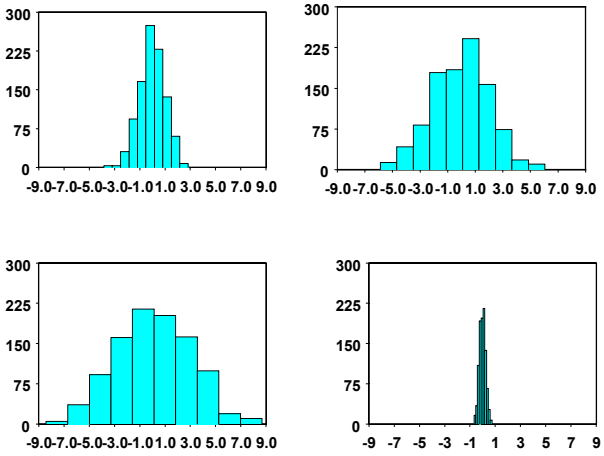
Shape



Numerical Measures of Variability (Spread)

Used in conjunction with measures of central tendency to more fully summarize a data set





- ***Range***---largest measurement (max) minus the smallest (min) measurement

  Note: easy to compute, but beware that two data sets can have the same range, but quite different variability

  Example:

  Let $X = \{10, 50, 50, 50, 50, 50, 50, 90\}$

  versus

  Let $Y = \{10, 20, 30, 45, 50, 70, 85, 90\}$

  The ranges for both sets of data are $90 - 10 = 80$.

  The second data set (Y), however, is clearly more variable.

- ***Sample variance***---denoted as s^2; equals the sum of the squared distance from the mean divided by $(n-1)$

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

- ***Sample standard deviation***---is the positive square root of the sample variance

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

For the data in the range examples, the standard deviations are:

$$s_x = \sqrt{s_x^2} = \sqrt{457.1} = 21.4$$

$$s_y = \sqrt{s_y^2} = \sqrt{878.6} = 29.6$$

Try to use the given data to calculate the standard deviations and see if you get the correct answers.

- ***Population variance***---denoted as σ^2 (sigma squared); calculated using population data; sample mean is replace with μ; $n-1$ is replace with N

- ***Population standard deviation***---denoted as σ; calculated as the positive square root of σ^2