BU609: Making Business Decisions: Descriptive Statistics

BU609-1

"If you can't measure it,
you can't manage it"

19-4-08

Course S/W

- Excel
  - Data Analysis : ensure **Analysis ToolPak** is enabled
  - PhStat (Evans text)
    - » Another Excel add-in; use is optional
- Other s/w (later in course)
  - TreePlan (Decision Tree analysis)
  - Premium Solver (linear programming)
  - Crystal Ball (Monte Carlo simulation)

Objectives for this Session

- Understand terminology & concepts
- Understand numerical data properties
  - Summary measures of central tendency, variation, & shape
- Be able to
  - Organize, tabulate, & graph numerical data
  - Build tables & charts for categorical data
  - Complete 1-way & 2-way analysis of data
  - Use Excel
- Understand good/bad data presentation

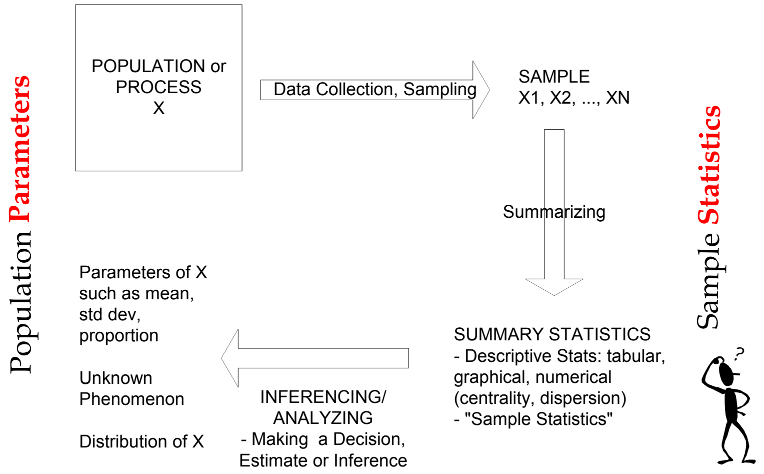
What is Statistics?

- Webster's: branch of math dealing with collection, analysis, interpretation & presentation of masses of numerical data
- Way to get **information** from **data**

Major Areas Within Statistics

- Sampling
  - collecting data from a population
  - why sample?
- Descriptive Statistics
  - summarizing & presenting data
- Inferential Statistics
  - using statistics and sample data to make statements about parent population
  - estimation & hypothesis testing

Key Statistical Concepts



Population **Parameters**

POPULATION or PROCESS X

Data Collection, Sampling

SAMPLE X1, X2, ..., XN

Summarizing

SUMMARY STATISTICS
- Descriptive Stats: tabular, graphical, numerical (centrality, dispersion)
- "Sample Statistics"

INFERENCING/ ANALYZING
- Making a Decision, Estimate or Inference

Parameters of X such as mean, std dev, proportion

Unknown Phenomenon

Distribution of X

Sample **Statistics**

Terminology

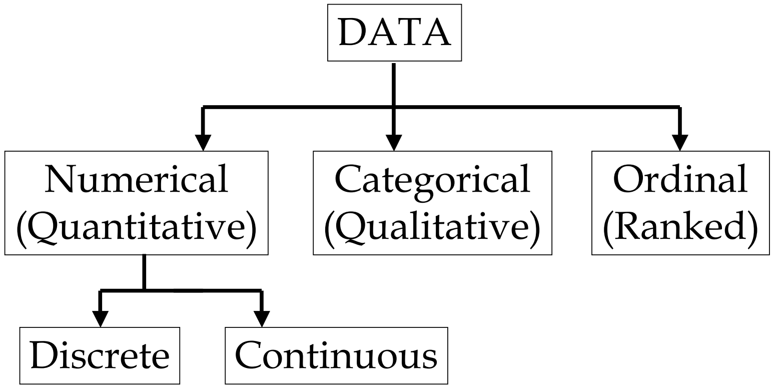
- Variables - characteristics of population or sample which are of interest
- Data - actual values of variables
  - Quantitative: numerical observations
  - Qualitative: categorical observations
  - Knowing type of data is necessary to properly select technique to be used
- Population vs Sample
  - Population parameter
  - Sample statistic

(cont.)

- Sometimes, especially to perform **nonparametric** techniques, it is also important to know if data can be ranked
- Type of analysis allowed for each type of data
  - Quantitative data - arithmetic calculations (e.g., mean, variance, range)
  - Qualitative data - counting the number of observation in each category; proportions
  - Ranked data - computations based on an ordering process

Data Types

- Determines statistical technique(s)



DATA

Numerical (Quantitative)

Categorical (Qualitative)

Ordinal (Ranked)

Discrete

Continuous

Cross-Sectional vs Time-Series

- **Cross sectional** data collected at point in time
  - Marketing survey (observe preferences by gender, age, etc - demographics, psychographics)
  - Test score in a statistics course
  - Starting salaries of MBA graduates
- **Time series** data is collected over successive points in time
  - Weekly closing price of gold
  - Amount of crude oil imported monthly
- Usually dealing with quantitative data
- May want to forecast

Descriptive Statistics

- Involves arrangement, summary, & presentation of data, to enable meaningful interpretation, and to support decision making
- Methods make use of
  - graphical techniques
  - numerical measures.
- Methods presented apply to both
  - entire population
  - population sample
- Use in Country project, case reports, …

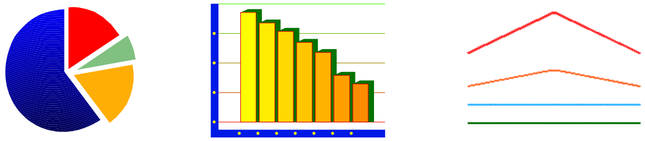
Graphical Descriptive Statistics

Viewer discretion advised, graphic content!

A picture is worth a thousand words

Displaying Data with Charts

- Pie, Bar & Line Charts

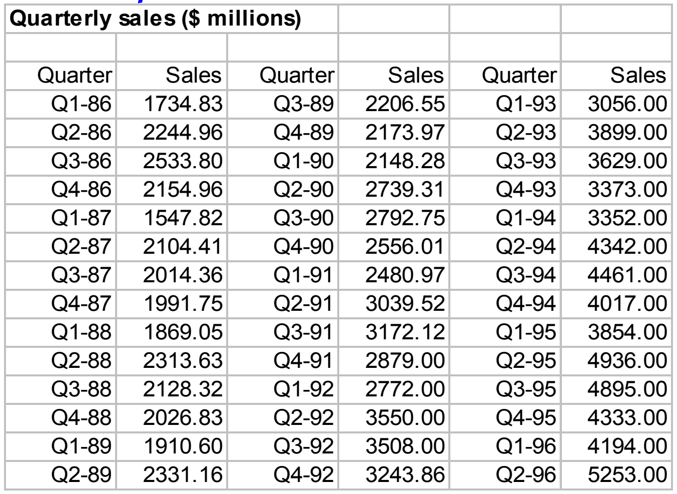


- For qualitative and/or quantitative data
- Most appropriate when raw data can be naturally **categorized** in a meaningful manner
- XY plot for quantitative data (2 variables)

Characteristics Of Good Graph

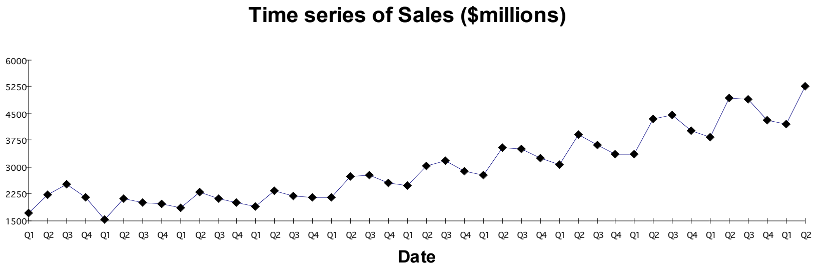
- Several variables (reasonable #)
- Shows association or causal relationship
- Text + graphics; use appropriately
- Efficient (viewer doesn't have to work at deciphering)
- Effective (gets point across; unbiased)

Example: Raw Data



Quarterly sales ($ millions)					
Quarter	Sales	Quarter	Sales	Quarter	Sales
Q1-86	1734.83	Q3-89	2206.55	Q1-93	3056.00
Q2-86	2244.96	Q4-89	2173.97	Q2-93	3899.00
Q3-86	2533.80	Q1-90	2148.28	Q3-93	3629.00
Q4-86	2154.96	Q2-90	2739.31	Q4-93	3373.00
Q1-87	1547.82	Q3-90	2792.75	Q1-94	3352.00
Q2-87	2104.41	Q4-90	2556.01	Q2-94	4342.00
Q3-87	2014.36	Q1-91	2480.97	Q3-94	4461.00
Q4-87	1991.75	Q2-91	3039.52	Q4-94	4017.00
Q1-88	1869.05	Q3-91	3172.12	Q1-95	3854.00
Q2-88	2313.63	Q4-91	2879.00	Q2-95	4936.00
Q3-88	2128.32	Q1-92	2772.00	Q3-95	4895.00
Q4-88	2026.83	Q2-92	3550.00	Q4-95	4333.00
Q1-89	1910.60	Q3-92	3508.00	Q1-96	4194.00
Q2-89	2331.16	Q4-92	3243.86	Q2-96	5253.00

(cont.) - Processed Data



Data > Information > Knowledge > Decisions & Business Outcomes

Displaying Data

- Frequency Distribution Table
  - tabular numeric report of interval or category frequencies
- Histogram
  - graphical display of frequencies
  - relative frequencies often used
- Graphs
  - scatter plot (for 2 variables)
  - other types of graphs
  - useful for **exploratory analysis**

Frequency Distribution Table

- Determine range (quantitative data)
  - Max(DataRange) - Min(DataRange)
- Select # of classes (depends on # data points)
  - usually 5-10 (may go higher); for categories, this is easy

# of Observations	**# of Classes**
Less then 50	5 - 7
50 - 200	7 - 9
200 - 500	9 - 10
500 - 1,000	10 - 11

(cont.)

- Calculate class width
  - Range / # Classes
- Calculate class boundaries (lower & upper limits)
  - Excel only wants upper boundary (“bin table”)
  - For future calculations, may want class midpoints
- Assign observations to classes & count
- From Frequency Table, draw histogram

Relative Frequency

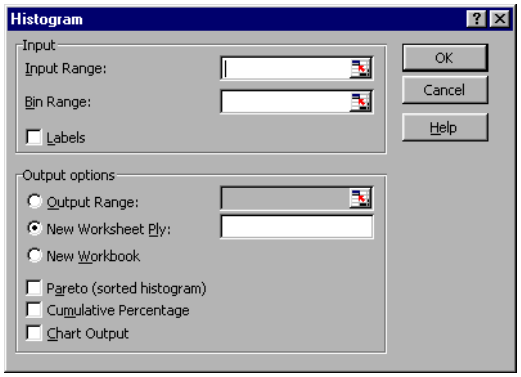
- Often preferable to show relative frequency (proportion) of observations falling into classes, rather than actual frequency

$$\text{Class relative frequency} = \frac{\text{Class frequency}}{\text{Total number of observations}}$$

- Use relative frequencies when
  - population relative frequencies are studied
  - comparing two or more histograms
  - sample sizes are different

Excel

- Tools > Data Analysis > Histogram
  - Can customize Chart Output afterwards



Histogram

Input

Input Range:

Bin Range:

□ Labels

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

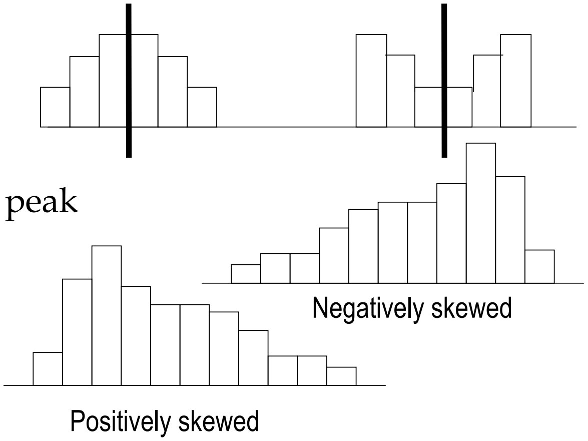
□ Pareto (sorted histogram)

□ Cumulative Percentage

□ Chart Output

Shapes of Histograms

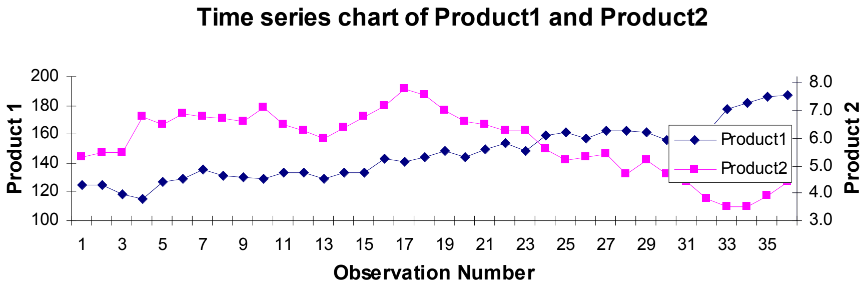
- Symmetry
- Skewness
  - Tail, not peak
- # Modes



Graphing with Excel

- Start with Chart Wizard
  - hilite data
  - click on Chart Wizard
  - select chart type (can ‘View Sample’)
  - continue thru prompts
- After Finish, customize
  - right click on area for options
  - double left click on areas for options
- Example next slide

Time Series Example



Graphical Excellence

- **Graphical excellence** deals with the effective use of graphical techniques
- Effective graphical techniques are
  - informative
  - concise
  - provide clear presentation of data to viewer

How can we achieve graphical excellence?

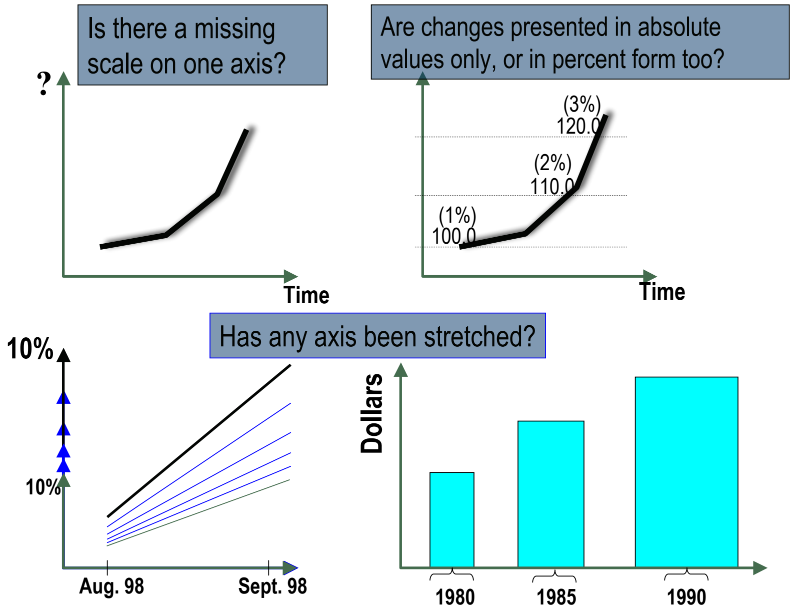
(cont.)

- Graphical excellence is achieved when
  - Graph presents large data sets **concisely** and **coherently**;
  - Ideas & concepts to be delivered are **clearly understood** to viewer
  - Graph encourages viewer to compare variables
  - Display induces viewer to address **substance** of data rather than **form** of graph
  - No **distortion** of what data reveal
- Avoid **chartjunk**

http://www.uh.edu/~tech132/3360grx.doc

Graphical Deception

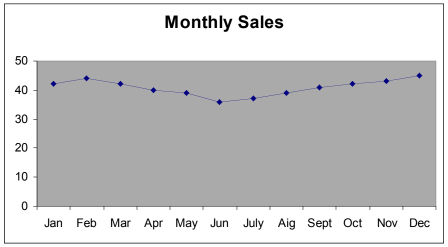
- Important to be able to critically **evaluate** information presented by graphical techniques
- Things to be **cautious** about when observing a graph:
  - Is there a missing scale on one axis, or no zero point?
  - Do not be influenced by a graph's caption
  - Are changes presented in absolute values only, or in percent form too?


Is there a missing scale on one axis?

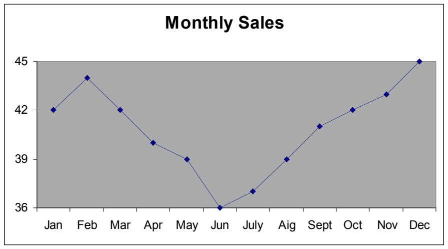
Are changes presented in absolute values only, or in percent form too?

Has any axis been stretched?

Is there no zero point on the vertical axis?

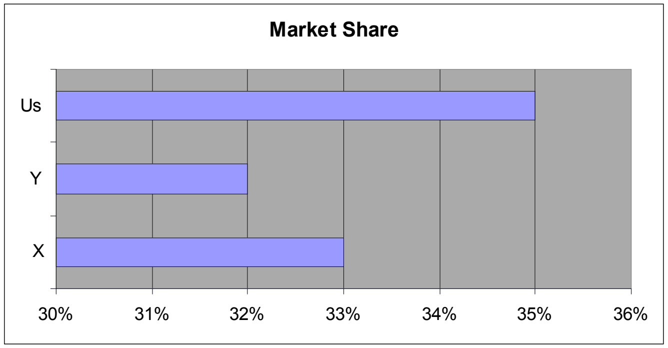


Marketing: Demand is almost level throughout the year



Production: We have to deal with severe seasonality!

Sales Manager: Our market share **far exceed** all competitors!



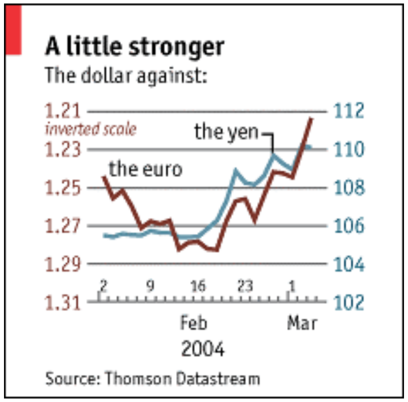
Cheating with Statistics

- "Figures don't lie, but liars sure figure"
- "There are three kinds of lies: lies, damned lies, & statistics" (Disraeli)
- "He uses statistics as a drunken man uses lampposts - for support rather than for illumination" (Lang)
- "Statistics is like a bikini; what it reveals is interesting, what it hides is essential" (BU609 grad)
- Survivorship bias, sampling & non-sampling errors

Good Graph?

"The **dollar** reversed some of the past year's declines against the euro and yen. The greenback hit its highest level in 2004 against those currencies as investors were increasingly heartened by America's growth prospects."

Economist.com, 4-Mar-04



Source: Thomson Datastream

FRB Exchange rates (3-Mar-04):
Euro 1.2088, Yen 110.25
(Rates in currency units per U.S. dollar)

Questions to Ask

- Who says?
- How do they know?
- What's missing?
- Did somebody change the subject?
- Does it make sense?

(Huff, "How to Lie with Statistics")

Thinking and caution is necessary! Don't blindly accept data & conclusions simply because they are available.

2-Way Analysis of Data

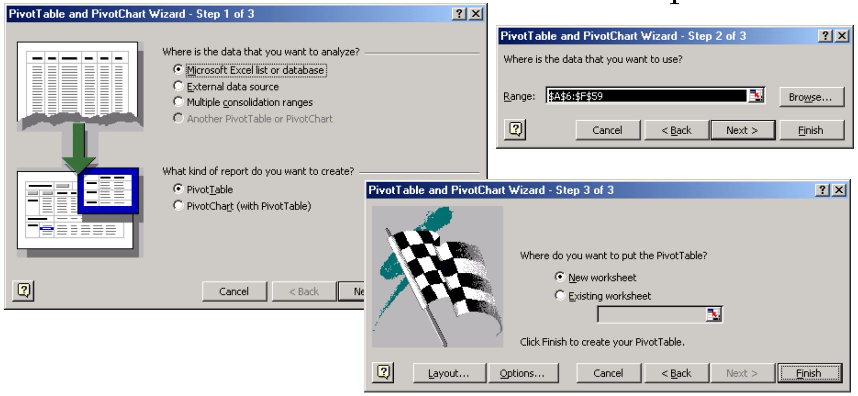
Pivot (Contingency) Tables,
Scatter Plots
Useful for Exploratory Data Analysis

Contingency Tables

- Shows # observations jointly for 2 variables
- May include row & column proportions, or total %
- Helps find relationships
- Used to calculate "joint probabilities"
- Widely used
- Excel: Pivot Tables (2-way or 3-way)

Exploring Data: Pivot Tables

- "Slice & Dice" data
  - contingency tables, crosstabulation
- Data > PivotTable and Pivot Chart Report...



PivotTable and PivotChart Wizard - Step 1 of 3

Where is the data that you want to analyze?

● Microsoft Excel list or database
○ External data source
○ Multiple consolidation ranges
○ Another PivotTable or PivotChart

What kind of report do you want to create?

● PivotTable
○ PivotChart (with PivotTable)

PivotTable and PivotChart Wizard - Step 2 of 3

Where is the data that you want to use?

Range: $A$6:$F$59

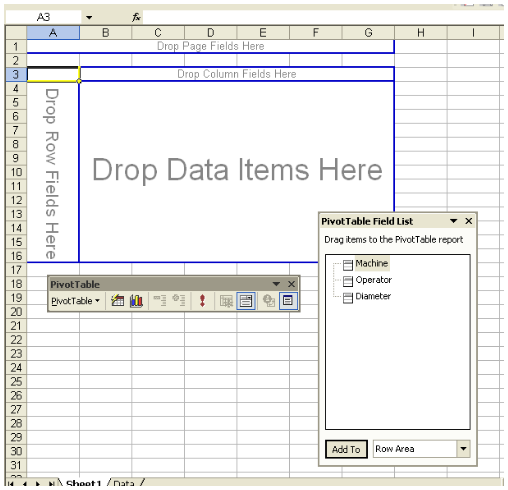
PivotTable and PivotChart Wizard - Step 3 of 3

Where do you want to put the PivotTable?

● New worksheet
○ Existing worksheet

Click Finish to create your PivotTable.

(cont.) Pivot Tables



Scatter Diagrams

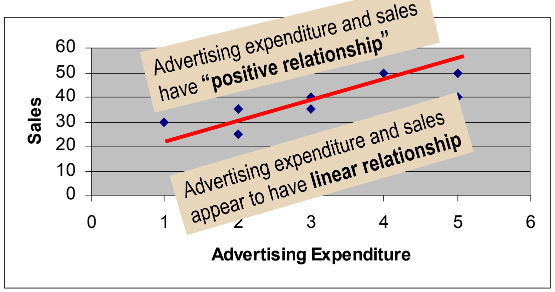
- Often interested in relationships between **two** numerical variables
- Use for **exploratory analysis**
- Example
  - Small business owner wants to assess effects of advertising on sales
  - Paired observation data collected. Each pair consisted of monthly advertising expenditure and monthly sales.

(cont.)

- Scatter diagram can describe relationship between advertising expenditures & sales

Advert	Sales
1	30
3	40
5	40
4	50
2	35
5	50
3	35
2	25

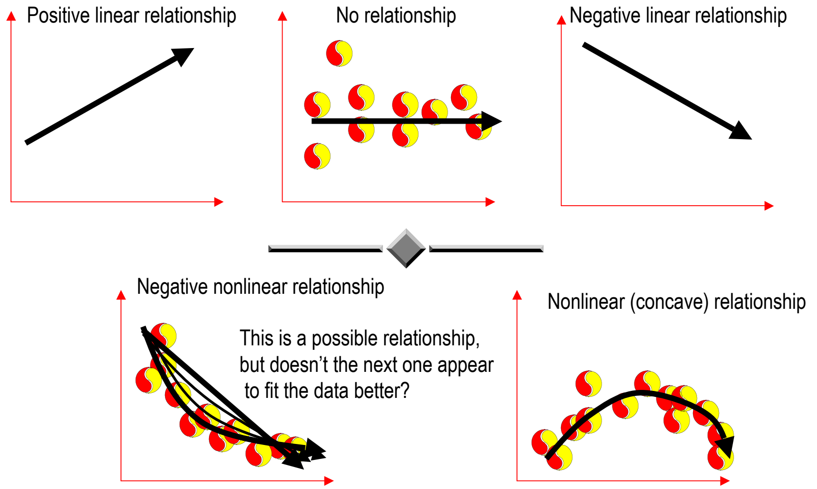
Excel scatter diagram



Advertising expenditure and sales have **"positive relationship"**

Advertising expenditure and sales appear to have **linear relationship**

Typical Patterns


Positive linear relationship
No relationship
Negative linear relationship
Negative nonlinear relationship
Nonlinear (concave) relationship

This is a possible relationship, but doesn't the next one appear to fit the data better?

Numerical Descriptive Measures

Quantitative Data

Standard Notation

Measure	Sample	Population
Mean	$\bar{x}$	μ
Std. Dev.	s	σ
Variance	s^2	σ^2
Size	n	N

Measures of Central Location

- Usually focus our attention on two aspects of measures of central location:
  - Measure of central data point (average).
  - Measure of dispersion of data about average/mean

Central data point reflects locations of all actual data points

Arithmetic Mean (Average)

- Most popular & useful measure of central location

$$\text{Mean} = \frac{\text{Sum of the measurements}}{\text{Number of measurements}}$$

Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample size

Population mean

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

Population size

Cautions re Mean

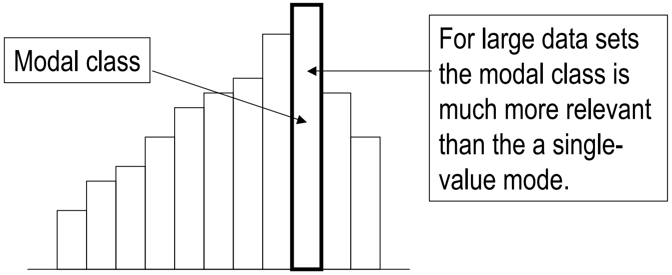
- Mean Reversion
  - Tendency of things to tend to equilibrium (long term average/mean)
  - Is change (system shock) permanent or temporary?
- Self-attribution bias
  - Tendency for people to think of themselves as "better than average"; attribute positive events to their own abilities & negative events to external forces
  - Can everyone in 609 do better than class average?
- Flaw of Averages

Median

- **Median** of set of measurements is value that falls in middle when measurements are arranged in increasing order of magnitude
  - odd # of measurements > middle value
  - even # of measurements > take average of two middle values
- Sometimes better representation of "midpoint"

Mode

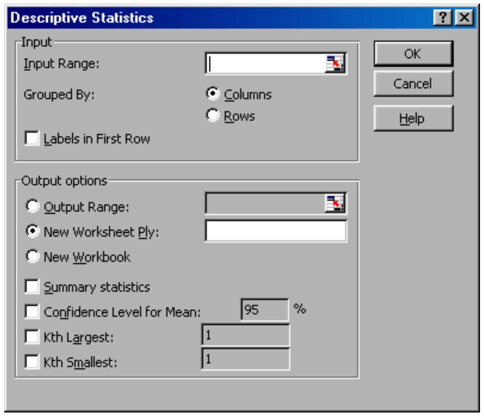
- Mode of a set of measurements is value that occurs most frequently
- Set of data may have one mode (or modal class), or two or more modes



For large data sets the modal class is much more relevant than the a single-value mode.

Excel & Descriptive Statistics

- Tools > Data Analysis > Descriptive Statistics



Input

Input Range:

Grouped By: ● Columns ○ Rows

□ Labels in First Row

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

□ Summary statistics

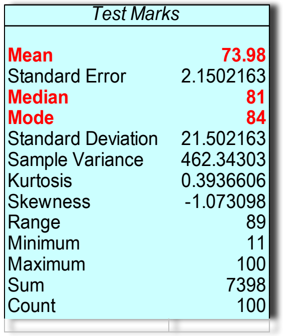
□ Confidence Level for Mean: 95 %

□ Kth Largest: 1

□ Kth Smallest: 1

Example

- HR department has administered an aptitude test to 100 job applicants. Find mean, median, and mode, and describe information they provide.



Test Marks	
Mean	**73.98**
Standard Error	2.1502163
Median	**81**
Mode	**84**
Standard Deviation	21.502163
Sample Variance	462.34303
Kurtosis	0.3936606
Skewness	-1.073098
Range	89
Minimum	11
Maximum	100
Sum	7398
Count	100

Measures of variability

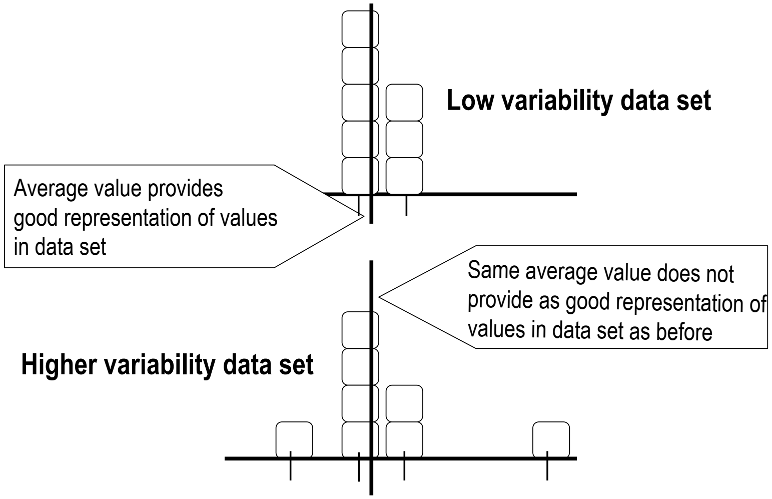
- Measures of central location fail to tell whole story about distribution!
- A question of interest still remains unanswered:

How typical is average value of all measurements in data set?

or

How much spread out are measurements about average value?

Observe two hypothetical data sets



Low variability data set

Average value provides good representation of values in data set

Same average value does not provide as good representation of values in data set as before

Higher variability data set

Range

- Range of a set of measurements is difference between largest & smallest measurements
  - Major advantage is ease with which it can be calculated
  - Major shortcoming is failure to provide information on **dispersion** of values between two end points

Percentiles

- p^{th} percentile of a set of measurements is value for which
  - at most p% of measurements are less than that value; at most 100(1-p)% of all measurements are greater than that value
- Commonly used percentiles
  - First (lower) decile = 10th percentile
  - First (lower) quartile (**Q_1**) = 25th percentile
  - Second (middle)quartile (**Q_2**) = 50th percentile
  - Third quartile (**Q_3**) = 75th percentile
  - Ninth (upper) decile = 90th percentile

Variance

- Measure of dispersion reflecting values of ***all*** measurements
- **Variance** of **population** of 'N' measurements $x_1, x_2, \ldots, x_N$ having mean μ is defined as

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

Variance of **sample** of 'n' measurements $x_1, x_2, \ldots, x_n$ having mean $\bar{x}$ is defined as

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

Standard Deviation

- **Standard Deviation** of set of measurements is square root of **variance** (same units as mean)

$$\text{Sample standard deviation: } s = \sqrt{s^2}$$

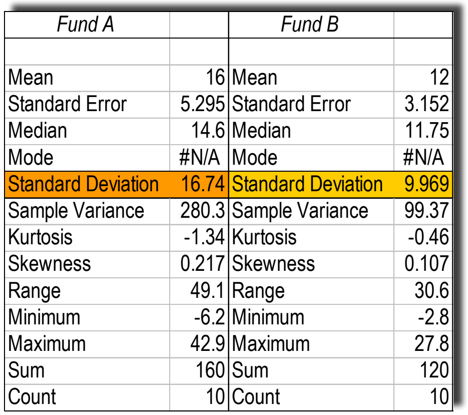
$$\text{Population standard deviation: } \sigma = \sqrt{\sigma^2}$$

- **Example**
  - Rates of return over past 10 years for two mutual funds are shown. Which one has higher level of risk (greater variation)?
  - Fund A: 8.3, -6.2, 20.9, -2.7, 33.6, 42.9, 24.4, 5.2, 3.1, 30.05
  - Fund B: 12.1, -2.8, 6.4, 12.2, 27.8, 25.3, 18.2, 10.7, -1.3, 11.4

Solution

- Use Excel printout (run from "Descriptive Statistics" sub-menu)

Fund A should be considered riskier because its standard deviation is larger



Fund A		*Fund B*	
Mean	16	Mean	12
Standard Error	5.295	Standard Error	3.152
Median	14.6	Median	11.75
Mode	#N/A	Mode	#N/A
Standard Deviation	16.74	Standard Deviation	9.969
Sample Variance	280.3	Sample Variance	99.37
Kurtosis	-1.34	Kurtosis	-0.46
Skewness	0.217	Skewness	0.107
Range	49.1	Range	30.6
Minimum	-6.2	Minimum	-2.8
Maximum	42.9	Maximum	27.8
Sum	160	Sum	120
Count	10	Count	10

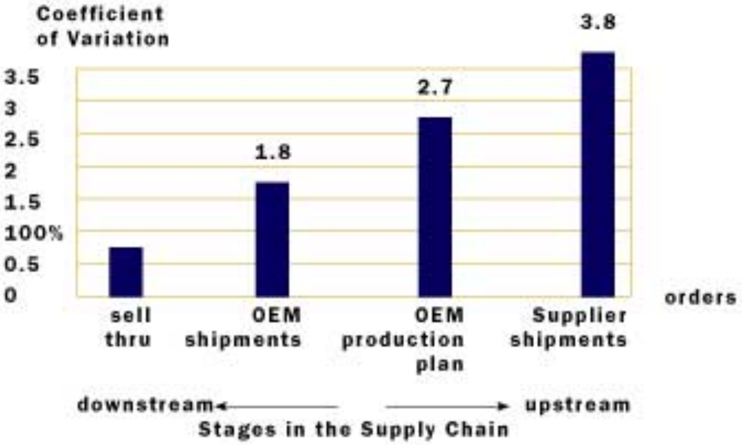
Excel Functions

- AVERAGE
  - Calculates average (arithmetic mean)
- VAR
  - Calculates variance based on a sample (n - 1 in denominator)
- VARP
  - For a population (n in denominator)
- STDEV, STDEVP

Coefficient of Variation

- CV = σ/μ (pop), cv = s/x-bar (sample)
  - Consider impact of smaller/larger values of mean and variance
  - Can be reported as % (*100)
- Sharpe ratio (inverse of CV)
  - measure of the mean excess return per unit of risk in an investment asset or a trading strategy
  - $(R – R_f)/\sigma$

SCM Example



Callioni & Billington ('01), "Effective Collaboration", ORMS Today, 28(5)

Measures of Association

- Scatter plot provides visual measure of association between two variables
- Two common numeric measures used for describing linear association
  - **Covariance** - any pattern to way two variables move together?
  - **Correlation Coefficient** - how strong is linear relationship between two variables?

*Covariance**

$$\text{Population covariance} = \text{COV(X,Y)} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

μ_x (μ_y) is the population mean of the variable X (Y)

N is population size; n is sample size

$$\text{Sample covariance} = \text{cov(X,Y)} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{n-1}$$

Coefficient of Correlation

Population coeff. of correlation

$$\rho = \frac{\text{COV(X,Y)}}{\sigma_x \sigma_y}$$

Sample coeff. of correlation

$$r = \frac{\text{cov(X,Y)}}{s_x s_y}$$

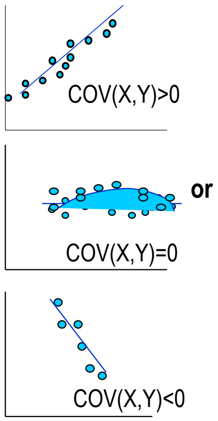
- Answers how **strong** association is between X & Y, & **direction**
- Between –1 & +1

(cont.)

+1 Perfect positive linear relationship

ρ or r = 0 No linear relationship

-1 Perfect negative linear relationship



(cont.)

- If two variables are very strongly positively related, coefficient value is close to +1 (strong positive linear relationship)
- If two variables are very strongly negatively related, coefficient value is close to -1 (strong negative linear relationship)
- No straight line relationship is indicated by a coefficient close to zero

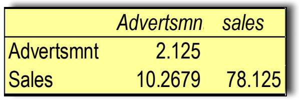
Example

- Compute covariance and coefficient of correlation to measure how advertising expenditure and sales level are related to one another

Advert	Sales
1	30
3	40
5	40
4	50
2	35
5	50
3	35
2	25

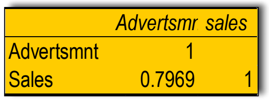
(cont.)

- Excel: Tools > Data Analysis > Covariance *or* Correlation



	Advertsmn	*sales*
Advertsmnt	2.125	
Sales	10.2679	78.125

Covariance matrix



	Advertsmr	*sales*
Advertsmnt	1	
Sales	0.7969	1

Correlation matrix

- Interpretation
  - Covariance (10.2679) indicates advertisement expenditure & sales are positively related
  - Coefficient of correlation (.797) indicates a strong positive linear relationship between advertisement expenditure & sales

Causality vs Correlation

- A → B
- B → A

Correlation doesn't show direction (causality)

- A → C → B Intervening Variable

Don't fall into the 'causality trap'

Excel Workshop Exercises

- Online file: TSX-Data_Nortel_Royal.xls
  - plot time series of each, calculate descriptive statistics for each; correlation & covariance for Nortel/Royal, and each vs TSX300