Fundamentals of Business Statistics – PT IMBA 2007/8

Prof. Dr. Stijn Viaene

Vlerick Leuven Gent Management School

Course Description

- To make informed business decisions, today's managers must be confident and comfortable in their use and interpretation of data.
- This course introduces you to fundamental statistical tools and methods for data analysis and statistical inference that are applicable to a wide variety of business problems.
- The textbook that is used for this course is "Data Analysis for Managers with Microsoft Excel, 2nd ed." by Albright, Winston and Zappe.
- The course/book illustrates in a practical way a variety of statistical methods, from simple to complex, to help you analyze data sets and uncover important information, and emphasizes the usefulness of concepts and theory to effective managerial decision making. Mathematical analysis will be kept to a minimum.

Critical Success Factor

Positive Attitude!

Stijn, I have a ‘non-math mind’

- At least I hope your mind is open, critical
- Maximize your ‘learning experience’
- Exercises (Excel & small manual calculations)
- **Always make a drawing**
- 80% sound judgment – 20% computer
- Check relevance
- **Start with an intelligent ‘guess’**

There are 3 Kinds of Lies: Lies, Damned Lies, and Statistics

- "How to lie with statistics" (Darrel Huff, 1954)
- Refers to the persuasive power of numbers and succinctly describes how statistics, even accurate one, can be used to bolster an inaccurate argument
- Statistics' big foes:
  - Selectively choosing data sets (GIGO)
  - Small data sets
  - Very skewed value distributions
  - Ignoring bad results (or overemphasizing good results)

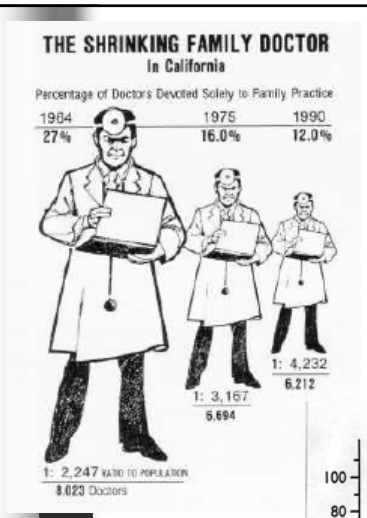
Statistics' Biggest Foe

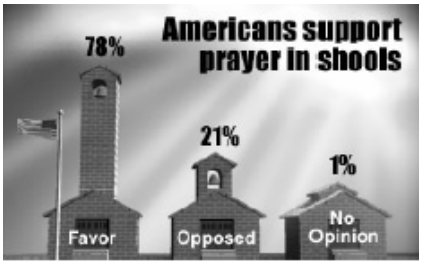
Blockheads, fools, morons, idiots, prigs and authoritarian personalities

Graphical Excellence
= substance + statistics + design

- Principles (E.R. Tufte)
  - Show the *data*
  - Focus on the *substance* of the graph
  - Avoid distortion
  - Encourage *comparison*
  - Serve a clear *purpose*
  - Integrate the statistical and *verbal* descriptions of the graph
- Graphical integrity + design aesthetics
  - Lie factor (misrepresentation: the size of the effect data vs. effect in graph)
  - Data ink ratio (design aesthetics: ink used for data / ink used for graph)

http://www.edwardtufte.com





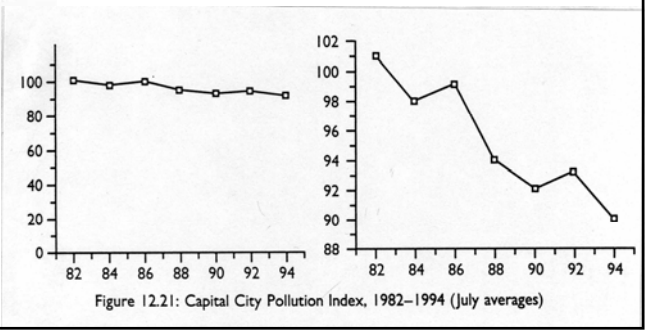


Figure 12.21: Capital City Pollution Index, 1982–1994 (July averages)

Course Outline

Graphs and tables	Chapter 2	Part A
Summary measures	Chapter 3	
Probability and probability distributions	Chapter 5	Part B
Normal, binomial, Poisson, and exponential distributions	Chapter 6	
Sampling & statistical inference	Chapters 8-11	Part C

Part A: Describe (Summarize) Data in Samples
Part B: Language of Probability & Random Variables
Part C: Sampling & Statistical Inference

"We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression,"
R.A. Fisher in *The Design of Experiments*

Course Assessment

The final assessment for this course consists of **a critical paper** with the following elements:

- A description of a data set with at least three variables and fifty observations; Take care that you choose the database in such a way that the following assignments are statistically meaningful;
- A description of the variables (discrete, continuous, etc.);
- An application of the most relevant methods in descriptive statistics for at least one variable, with comments on the relevance of those techniques and comments on the meaning of the results;
- One box-plot diagram with interpretation;
- A calculation of the correlation between all variables with interpretation of the results;
- An estimation of a confidence interval for a mean and/or a proportion for one variable, with interpretation;
- The formulation of two hypotheses you test with your data; Please comment the results;
- One linear regression analysis (simple regression is sufficient; multiple regression is, of course, also possible) with interpretation of the results;
- Describe (outside your data set) one meaningful application in your company of the Poisson distribution; illustrate with some relevant calculations.

This critical paper has to be handed in on hardcopy (paper) to the *program coordinator*

by Thursday, October 30, 2007.

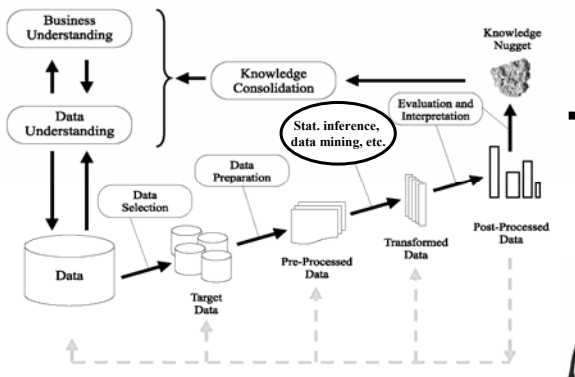
Fundamentals of Business Statistics – PT IMBA 2007/8

PART A

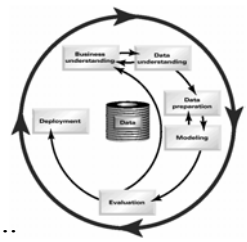
A Toolkit for Describing (Summarizing) Data in Samples (Data Sets)

Chapters 2 & 3

A word of caution – KDD Process



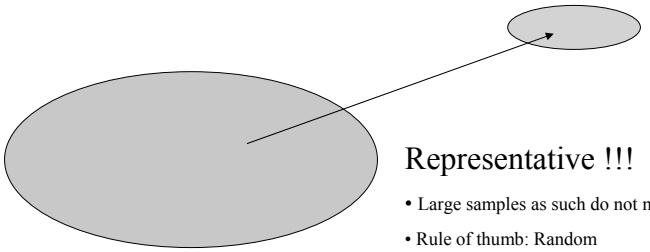
Standardized: CRISP-DM



Data, information, knowledge, learning, ...

Population vs. Sample

- A **population** includes all of the objects of interest in a study. A **sample** is a subset of the population of interest, often randomly *chosen* and preferably **representative** of the population as a whole.



Representative !!!

- Large samples as such do not mean much
- Rule of thumb: Random
- Rule of thumb: Avoid too much human judgment

Sample Format: Spreadsheet

- An **observation** is a member of the population or sample. Alternative terms for observations are **cases** and **records**.
- Each row corresponds to an observation.
- Each column represents a simple (single, univariate) **variable,** i.e. quantified or qualified characteristic or attribute with different values among observations. An alternative term for variable that is commonly used in database packages is **field**.

	A	B	C	D	E	F
1	**Data from a questionnaire on environmental policy**					
2						
3	Age	Gender	State	Children	Salary	Opinion
4	35	Male	Minnesota	1	$65,400	5
5	61	Female	Texas	2	$62,000	1
6	35	Male	Ohio	0	$63,200	3
7	37	Male	Florida	2	$52,000	5
8	32	Female	California	3	$81,400	1
9	33	Female	New York	3	$46,300	5
10	65	Female	Minnesota	2	$49,600	1
11	45	Male	New York	1	$45,900	5
12	40	Male	Texas	3	$47,700	4
13	32	Female	Texas	1	$59,900	4
14	57	Male	New York	1	$48,100	4
15	38	Female	Virginia	0	$58,100	3
16	37	Female	Illinois	2	$56,000	1
17	42	Female	Virginia	2	$53,400	1
18	38	Female	New York	2	$39,000	2
19	48	Male	Michigan	1	$61,500	2
20	40	Male	Ohio	0	$37,700	1
21	57	Female	Michigan	2	$36,700	4
22	44	Male	Florida	2	$45,200	3
23	40	Male	Michigan	0	$59,000	4
24	21	Female	Minnesota	2	$54,300	2
25	49	Male	New York	1	$62,100	4
26	34	Male	New York	0	$78,000	3
27	49	Male	Arizona	0	$43,200	5
28	40	Male	Arizona	1	$44,500	3
29	38	Male	Ohio	1	$43,300	1
30	27	Male	Illinois	3	$45,400	2
31	63	Male	Michigan	2	$53,900	1
32	52	Male	California	1	$44,100	3
33	48	Female	New York	2	$31,000	4

coding.xls

Types of Data

- There are several ways to categorize data
  - **Numerical** vs. **categorical**
    (Do we intend to do arithmetic on the data? Choice?)
  - **Cross-sectional** vs. **time series** (vs. panel data?)
- Numerical: **discrete** vs. **continuous**
  (Can we count values? Choice? Implication?)
- Categorical: **nominal** vs. **ordinal**
  (Is ordering meaningful?)
- Coding of a variable can be misleading!

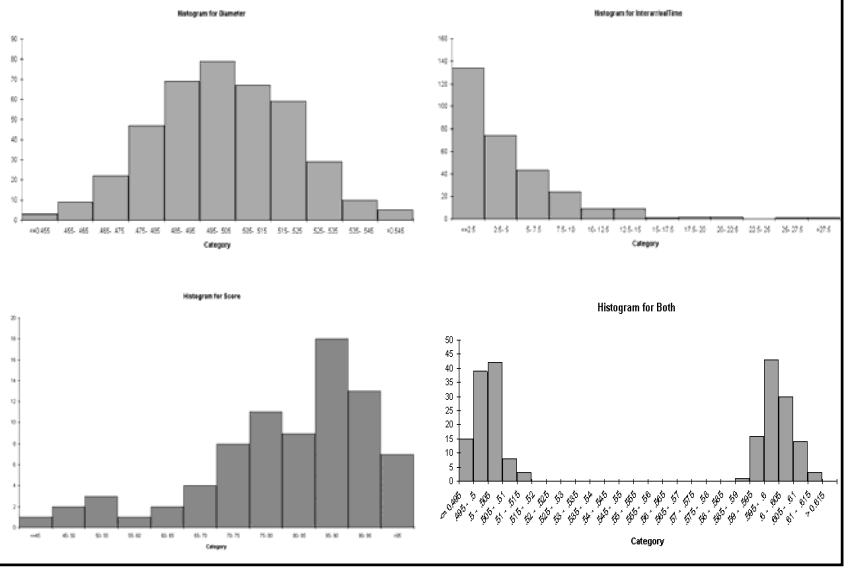

coding.xls

Frequency Tables & Histograms

- A **frequency table** indicates how many observations fall in various categories of a single variable. (**~distribution) (~likelihood**)
- To obtain a frequency table for a variable that is truly continuous we must first choose '**appropriate**' **categories**.
  - We want to have enough categories so that we can see a meaningful distribution, but we don't want so many categories that there are only a few observations per category.
  - A good rule of thumb is to divide the range of values into 8 to 15 equally spaced categories, plus a possible open-ended category at either end of the range.
- A **histogram** is the graphical analog of a frequency table.

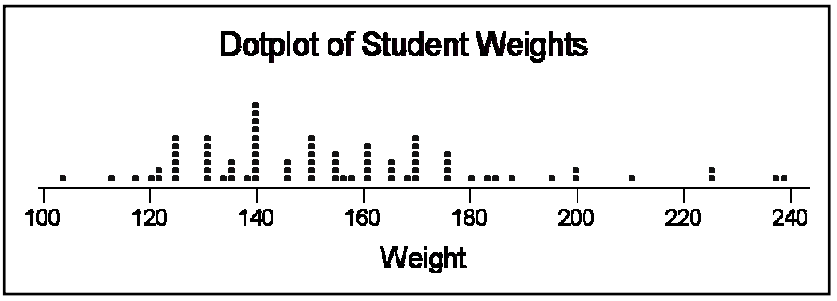

actor_salary
.xls

Histogram Shapes



Dot Plot (cf. SPC)

- Each dot represents one observation that takes the indicated value.



Summary Measures (Univariate)

- Maximum, minimum, and range
  = What is value range?
- **Measures of central location**:
  Mean (average), median, mode
  = What is 'typical value'?
- Quartiles (IQR) and percentiles
  = How are values distributed?
  (shape of distribution)
- Most popular measures of variability: Variance and standard deviation

	A	B	C
1	*Summary measures for selected variables*		
2			Salary
3		Count	190.000
4		Mean	29762.105
5		Median	29850.000
6		Standard deviation	3707.212
7		Minimum	17100.000
8		Maximum	38200.000
9		Range	21100.000
10		Variance	13743424.116
11		First quartile	27325.000
12		Third quartile	32300.000
13		Interquartile range	4975.000
14		Mean absolute deviation	2967.767
15		Skewness	-0.166
16		Kurtosis	-0.071
17		5th percentile	23690.000
18		95th percentile	35810.000

salary.xls

Mean (Average)

- The **sample mean** (or average value) $\overline{X}$ of a variable X is the sum of all observations' value for this variable (X_i) divided by the number of observations (n).

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Alternatively: a weighted average of the distinct values with weights equal to relative frequencies.

Value	3	4	5	6	7	100
Frequency	1	1	2	3	2	1
Rel. Freq.	0.1	0.1	0.2	0.3	0.2	0.1
Rel.Freq.*Value	0.3	0.4	1	1.8	1.4	10
	Mean?					

- The mean can be misleading due to outliers (extreme values)
- Property: Sum of deviations around the mean is zero

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

Median

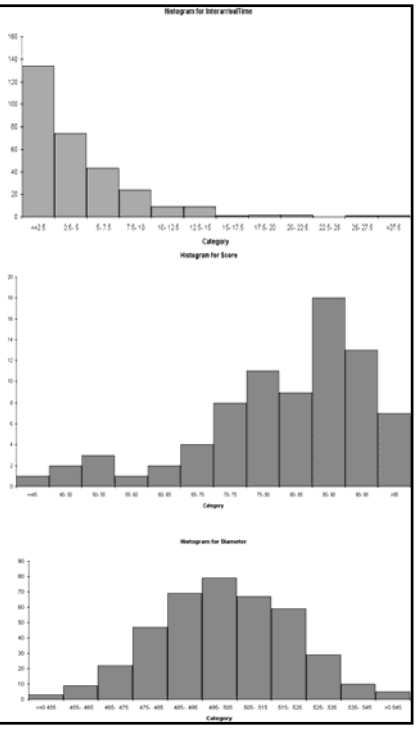
- The **median** is the 'middle observation' value when the data are listed from smallest to largest.
  - If odd number of observations, then middle observation
  - If even number of observations, then average of two middle observations
- About 50% of the observations have a lower/higher value.

3 4 5 5 6 6 6 7 7 100 (Median?)

- Median is 'robust' with respect to outliers.
- Thus, the median is a better measure of central location than the mean for (very) skew distributions.

Median (cont.)

- (Mean-Median) is a measure of **skewness**:
  - If (Mean-Median) > 0 then positively skewed (= to the right)
  - If (Mean-Median) < 0 then negatively skewed (= to the left)
  - If (Mean-Median) = 0 then symmetric ('bell shaped'???)
- Question: What is the range of this measure of skewness?



Mode

- The **mode** is the most frequently *occurring value*.
- If the values are essentially continuous, as with the salaries in Example 3.1, then the mode is essentially irrelevant. There is typically no single value that occurs more than once.

  *(Option: group values as in histogram → **modal class**)*

- Note:
  - A distribution can have more than one mode.
  - Mode = Median for an 'indicator' or 'dummy' variable (red flag).
  - 'Missing value' is often replaced by mode for categorical variable. (Mode is often 'default' value.)
  - Mode is typically not a good measure of central location.

Summary Measures (Univariate)

- Maximum, minimum, and range
  = What is value range?
- Measures of central location:
  Mode, mean (average), median
  = What is 'typical value'?
- **Quartiles (IQR) and percentiles**
  = How are values distributed?
  (shape of distribution)
- Most popular measures of variability: Variance and standard deviation

	A	B	C
1	**Summary measures using Excel functions**		
2			
3	Count	190	
4	Minimum	$17,100	
5	Maximum	$38,200	
6	Average	$29,762	
7	Median	$29,850	
8	Lower quartile	$27,325	
9	Upper quartile	$32,300	
10	5-percentile	$23,690	
11	95-percentile	$35,810	
12	Range	$21,100	
13	Standard deviation	$3,707	
14	Variance	13743424	

salary.xls

Quartiles and Percentiles

- How are the variable values for the observations **spread** over the range from smallest to largest value? **(~variability**)
- **Range** is only meaningful for numerical data, sensitive to outliers and does not take into account all values.
- Median is a.k.a. 50th percentile or 2nd quartile.
- The ***p*th percentile** is the value such that at least *p*% of the observations take this value or less AND at least (100 - *p*)% of the observations take this value or more.
  - Arrange the data (i.e. for *n* observations) in ascending order of variable value.
  - Compute index *i*, the position of the *p*th percentile.

$$i = (p/100)n$$

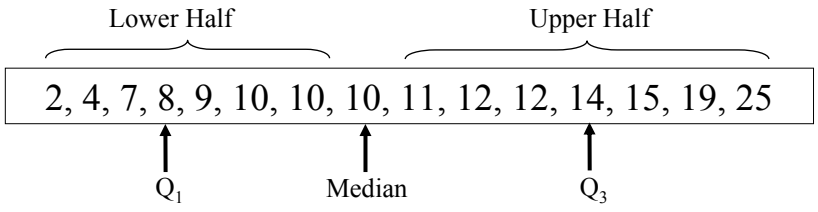
  - If *i* is not an integer, round up. The *p*th percentile is the value in the *i*th position.
  - If *i* is an integer, the *p*th percentile is the average of the values in positions *i* and *i*+1.

15th, 90th percentile?

5	7	9	11	13	15	17	19	21	23
25	27	29	31	33	35	37	39	41	43
45	47	49	51	53	55	57	59	61	63
65	67	69	71	73	75	77	79	81	83
85	87	89	91	93	95	97	99	101	103

Quartiles and Percentiles (cont.)

- Quartiles are specific percentiles.
  - Q1: 1st quartile = 25th percentile (= Median of lower half)
  - Q2: 2nd quartile = 50th percentile = Median
  - Q3: 3rd quartile = 75th percentile (= Median of upper half)
- Inter-Quartile-Range (IQR) is [Q1,Q3] with length (Q3-Q1)
- IQR contains about 50% of the observed variable values (which?)



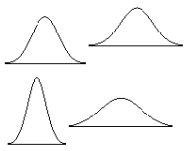
Summary Measures (Uni-variate)

- Maximum, minimum, and range
  = What is value range?
- Measures of central location:
  Mean (average), median, mode
  = What is 'typical value'?
- Quartiles (IQR) and percentiles
  = How are values distributed?
  (shape of distribution)
- **Most popular measures of variability:** Variance and standard deviation

	A	B	C
1	**Summary measures using Excel functions**		
2			
3	Count	190	
4	Minimum	$17,100	
5	Maximum	$38,200	
6	Average	$29,762	
7	Median	$29,850	
8	Lower quartile	$27,325	
9	Upper quartile	$32,300	
10	5-percentile	$23,690	
11	95-percentile	$35,810	
12	Range	$21,100	
13	Standard deviation	$3,707	
14	Variance	13743424	

Part diameter.xls

Variance



- The **sample variance** s^2 of a variable X is the sum of all observations' squared deviation from the sample mean $\overline{X}$ for this variable divided by the number of observations (n) - 1.

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Note: the variable X is often mentioned in the subscript of s^2, i.e. s_X^2 or between brackets, i.e. $s^2(X)$

- The variance tends to increase when there is more **variability around the mean**. (implicitly assumes symmetry!)
- Large deviations from the mean contribute heavily to the variance because they are **squared**. (positive! squared units!)
- **Population variance**:

(N is population size)
(μ is population mean)

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

Part diameter.xls

Standard Deviation

- The **standard deviation** s , defined as the square root of the variance s^2, is a more intuitive measure of variability.

$$s = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

- The standard deviation is measured in **original units**, and it is much easier to interpret.
- **Population standard deviation**:

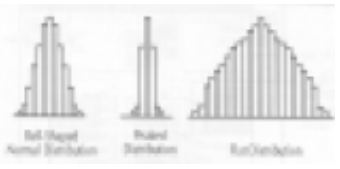
(N is population size)
(μ is population mean)

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

Part diameter.xls

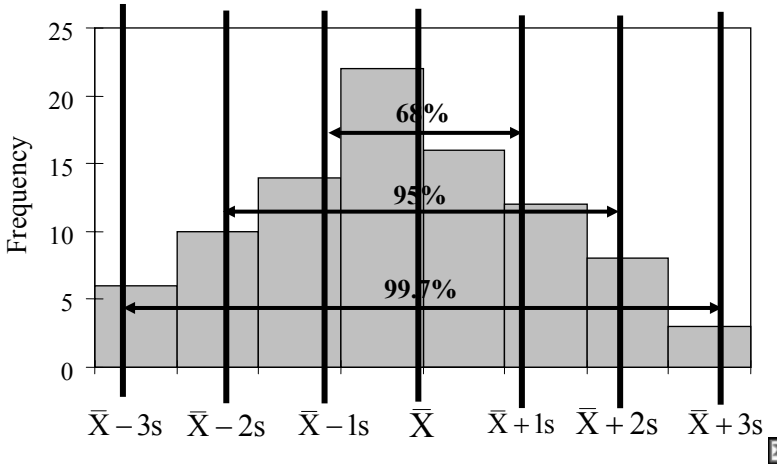
Empirical Rules: Interpretation of Standard Deviation

- **If** histogram is **unimodal and relatively symmetric** ('bell shaped'):
  - Approximately 68% of the observations are within 1 standard deviation of the mean.
  - Approximately 95% of the observations are within 2 standard deviations of the mean.
  - Approximately 99.7% - almost all - of the observations are within 3 standard deviations of the mean.



Part diameter.xls

Empirical Rules: Interpretation of Standard Deviation (cont.)



Dow returns.xls

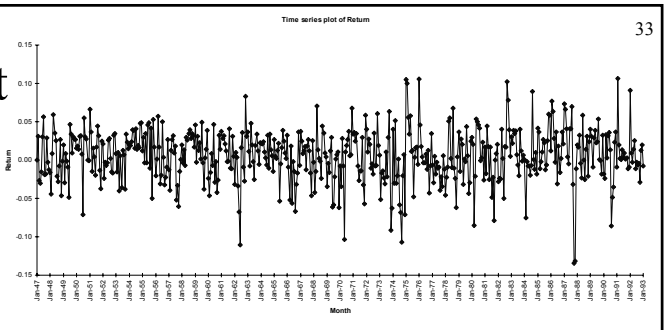
Boxplot

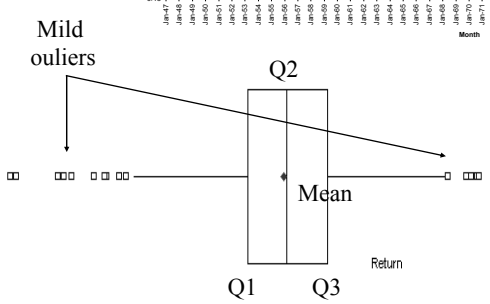
- In one drawing: **information on central location and variability**
- Boxplots can be used in two ways: (1) to describe a single variable in a data set, or (2) to compare two (or more) variables
- Understanding StatPro conventions:
  - **Box** (with reference to underlying, horizontally drawn unit scale axes)
    - Hinges at Q1 and Q3
    - Median as vertical line inside box
    - Mean as dot
  - **Whiskers** (or horizontal lines beyond box hinges)
    - Extend to furthest observations within 'inner fence length' (= 1.5 IQR) from box hinges. These boundary observations are called upper and lower 'adjacent values'.
  - Mild and extreme **outliers**
    - Mild: observations beyond 1.5 IQR but within 3 IQR (= 'outer fence length') from box hinges (drawn as open dots)
    - Extreme: observations beyond 3 IQR from box hinges (drawn as solid dots)

Dow returns.xls

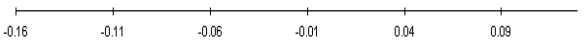
Boxplot (cont.)

P03.37





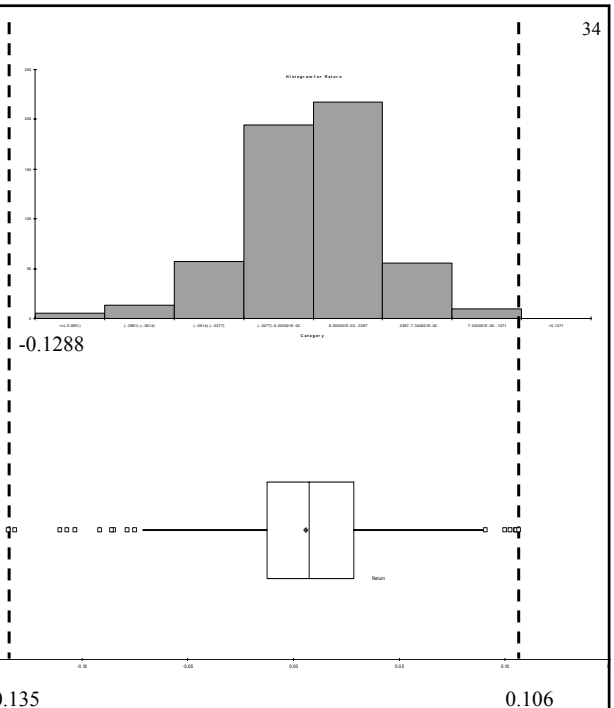
Note: whisker length!



	A	B
1	**Summary measures for boxplots**	
2		Return
3	Mean	0.00588
4	Median	0.007417
5	Q1	-0.012566
6	Q3	0.02836
7	IQR	0.040926
8		
9	Outer lower fence	-0.135342
10	Outer upper fence	0.151137
11		
12	Inner lower fence	-0.073954
13	Inner upper fence	0.089748
14		
15	Lower adjacent value	-0.071125
16	Upper adjacent value	0.089421
17		
18	# of extreme outliers	0
19	# of mild outliers	16
20		
21	# of low outliers	10
22	# of high outliers	6

Boxplot (cont.)

Use a boxplot to summarize the distribution of these returns.

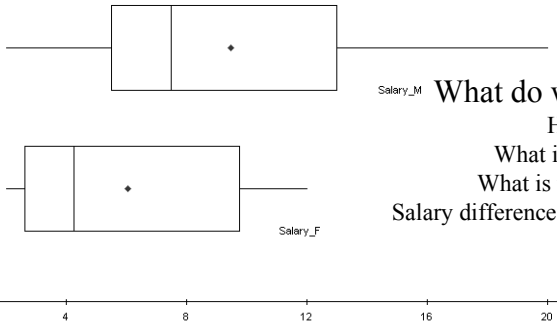


Boxplots Probably Most Useful for Comparing Populations

	A	B	C	D	E
1	**Famous actors and actresses**				
2					
3	Note: All monetary values are in $ millions.				
4					
5	Name	Gender	DomesticGross	ForeignGross	Salary
6	Angela Bassett	F	32	17	2.5
7	Jessica Lange	F	21	27	2.5
8	Winona Ryder	F	36	30	4
9	Michelle Pfeiffer	F	66	31	10
10	Whoopi Goldberg	F	32	33	10
11	Emma Thompson	F	26	44	3
12	Julia Roberts	F	57	47	12
13	Sharon Stone	F	32	47	6
14	Meryl Streep	F	34	47	4.5
15	Susan Sarandon	F	38	49	3
16	Nicole Kidman	F	55	51	4
17	Holly Hunter	F	51	53	2.5
18	Meg Ryan	F	43	55	8.5
19	Andie Macdowell	F	26	75	2
20	Jodie Foster	F	62	85	9
21	Rene Russo	F	69	85	2.5
22	Sandra Bullock	F	64	104	11
23	Demi Moore	F	65	125	12
24	Danny Glover	M	42	4	2
25	Billy Crystal	M	52	14	7

P03.39 (P02.50)

'stacked' (categorical) variable identifies two subpopulations → drawing boxplots conditional on values of this code variable can identify differences among the subpopulations



What do we observe?

Hint:
What is typical?
What is variability?
Salary difference *because of* Gender?

salary crimination.

Analyzing Relationships with Scatterplots

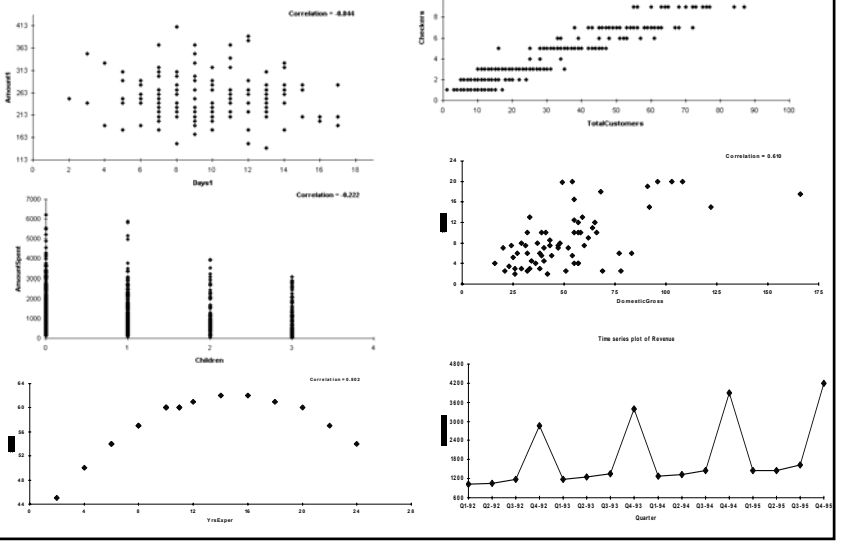
P02.14 (P02.07) P02.17 P02.18

- We are often interested in the **relationship** between two (*numerical*) variables. (How do values of two different variables tend to 'move together'?)
- A useful way to picture this relationship is to **plot a point for each observation** (= natural pairing of values), where the coordinates of the point represent the values of the two variables.
- The resulting graph is a **scatterplot**.
- Two basic questions:
  - Is there a relationship?
  - What type of relationship?

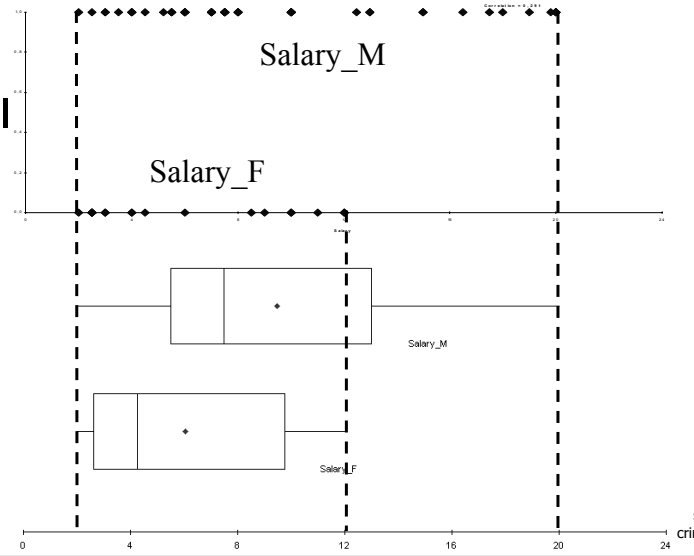
Examples:

Salary_DomticGrossRev.. SalesProduciv_YrsExp.xls Revenue_Quarter.xls

Analyzing Relationships with Scatterplots (Cont.)



Side-by-Side Boxplots Revisited



Covariance and Correlation

- Both summarize the type of behavior (~ relationship linear?) observed in a scatterplot.
- Each measures the direction (= *sign of measure*) and strength (= *absolute value of measure*) of a **linear** relationship between two numerical (numerically encoded) variables.
- Informally:
  - If this line rises form left to right then the relationship is **positive**. If it falls from left to right then the relationship is **negative**.
  - The relationship is **strong** if the points in a scatterplot cluster tightly around some straight line (*i.e. it makes sense to 'simplify the world' in the form of a simple line; how perfect is the fit?*) .

Covariance

- Formally for **sample**:

$$Cov(X,Y) = s(X,Y) = s_{XY} = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

- Formally for **population**:

$$Covp(X,Y) = \sigma(X,Y) = \sigma_{XY} = \frac{1}{(N)}\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)$$

- Interpretation via the **geometrics** of the scatter of points
- Note:
  - $s_{XX} = s_X^2$ and $s_{YY} = s_Y^2$
  - $s_{XY} = s_{YX}$

spending habits.xls

P03.30
(P02.03)

Correlation

- Formally for **sample**:

$$Corr(X,Y) = r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

- Formally for **population**:

$$Corrp(X,Y) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- It is difficult to interpret/compare the magnitudes of covariances. These depend on the fact that the data are measured in dollars rather than, say, thousands of dollars. It is much easier to interpret the magnitudes of the correlations because they are scaled to be **between -1 and +1**. (unitless!)
- **Table of correlations for more than 2 variables**

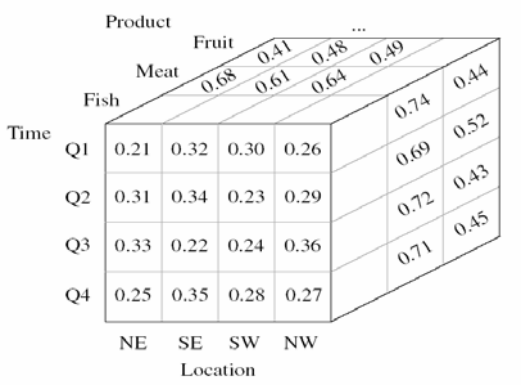
spending habits.xls

Exploring Data with Pivot Tables

- Statistics: crosstabs or contingency tables
- MDA: multidimensional data analysis
- **Most basic use = like a histogram, i.e. to look at value distributions, but now in a multidimensional way!**
- Excel's spreadsheet version of the 'data cube', the analytical (navigation) tool of data warehousing environments

Count of Gender	Age			
Gender	0-29	30-59	60-89	Grand Total
0	19.20%	36.20%	5.40%	60.80%
1	12.60%	22.90%	3.70%	39.20%
Grand Total	31.80%	59.10%	9.10%	100.00%

Multidimensional Data Analysis



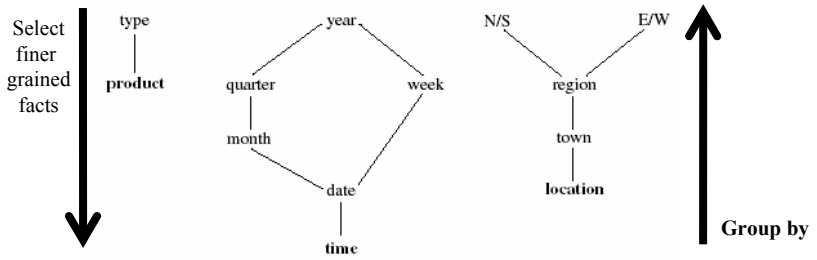
3D cube representation of sales data

Facts: numerical measures of the subject area of interest (i.e. sales)

Dimensions: perspectives along which one wants to view facts (i.e. time, product, location)

Used for
- selection
- grouping

Multidimensional Data Analysis (cont.)



Concept hierarchies for dimensions product, time and location

Multidimensional Data Analysis (cont.)

- Cube operations:
  - Roll-up (→ drill down): performs an aggregation, either by ascending one or more concept hierarchies or eliminating one or more dimension
  - Slice: defines a subcube by performing a selection on one dimension of a given cube
  - Dice: defines a subcube by performing a selection on two or more dimensions of a given cube
  - Pivot: rotates a cube along one or more dimensions to provide an alternative presentation of the data
  - Express facts as counts, percentages, sums, …

P02.41 (P02.1) P03.29 P02.43

Male vs. female famous actor salaries: Discrimination??? 46

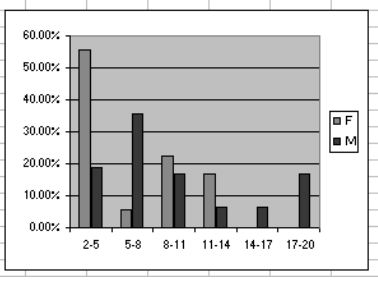
Example

1) Regroup salaries
2) Re-express facts

	A	B	C	D
1	Drop Page Fields Here			
2				
3	Count of Gender	Gender		
4	Salary	F	M	Grand Total
5	2	1	1	2
6	2.5	4	1	5
7	3	2	2	4
8	3.5		1	1
9	4	2	3	5
24	15		2	2
25	16.5		1	1
26	17.5		1	1
27	18		1	1
28	19		1	1
29	19.8		1	1
30	20		4	4
31	Grand Total	18	48	66

	A	B	C	D
3	Count of Gender	Gender		
4	Salary	F	M	Grand Total
5	2-5	55.56%	18.75%	28.79%
6	5-8	5.56%	35.42%	27.27%
7	8-11	22.22%	16.67%	18.18%
8	11-14	16.67%	6.25%	9.09%
9	14-17	0.00%	6.25%	4.55%
10	17-20	0.00%	16.67%	12.12%
11	Grand Total	100.00%	100.00%	100.00%

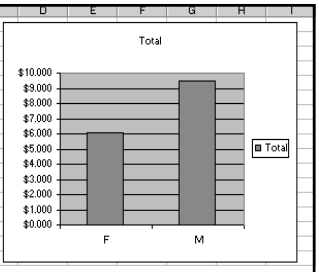
Draw pivot chart



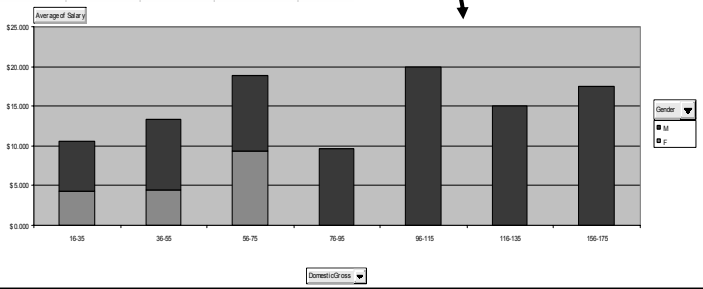
Example (Cont.)

Average of Salary	Gender		
DomesticGross	F	M	Grand Total
16-35	$4.357	$6.155	$5.456
36-55	$4.400	$8.990	$8.072
56-75	$9.417	$9.500	$9.462
76-95		$9.700	$9.700
96-115		$20.000	$20.000
116-135		$15.000	$15.000
156-175		$17.500	$17.500
Grand Total	$6.056	$9.479	$8.545

Average of Salary	
Gender	Total
F	$6.056
M	$9.479
Grand Total	$8.545



Is it <u>because</u>...???



Assignment Questions

- Look at all your charts, tables, and measures. Assess their use for each of the different types of variables we've covered. (= '**sensibility analysis**')
- What is the **effect of outliers** on each of the measures of typicality, variability and relationship we've covered?
- Play around with Pivot Tables in Excel!!!!