3. Descriptive Statistics: Tabular, Graphical and Numerical

Readings: ASW, Chapters 2 & 3

1

Answers to "Why Are You Taking This Course?"

Interesting	Need	Fit	Fit	Need	Need
Need	Interesting	Need	Need	Need	Fit
Need	Need	Need	Need	Need	Interesting
Other	Easy	Need	Interesting	Interesting	Fit
Easy	Easy	Need	–	Need	Need
Need	Interesting	Fit	Interesting	Fit	Need
Interesting	Need	Need	Need	Need	Need
Need	Need	Interesting	Interesting		

2

Answers to "How Much Income Did You Earn?"

0	1200	0	8000	0	–	0	–
6400	0	2500	3000	6000	5000	8000	4000
11000	25000	4000	8800	5000	7000	8000	0
0	18000	5400	15000	3500	24000	1000	8000
0	0	2100	8000	0	4000	0	0
1000	0	3000	0	2000	4800		

3

A) Summarizing **Qualitative** Data Visually

- First: classify data into nonoverlapping groups.
- Frequency distribution: tabular summary of the groups.

4

Frequency of "Why are You Taking This Course?"

Answer	Frequency
Needed it	25
It fit timetable	6
Looked Easy	3
Looked Interesting	10
Liked Prof	0
Other/Blank	2
Total	46

5

Frequency of "Why are You Taking This Course?"

Answer	Frequency	Relative Frequency	Percent Frequency
Needed it	25	0.543478	54.3%
It fit timetable	6	0.130435	13.0%
Looked Easy	3	0.065217	6.5%
Looked Interesting	9	0.195652	19.6%
Liked Prof	0	0	0.0%
Other/blank	3	0.065217	6.5%
Total	46	1	100%

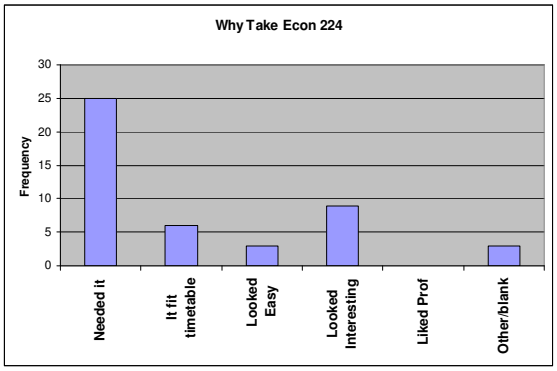
=25/46

6

Bar Graphs

- Put your answers into Excel.
  - Construct relative frequency.
  - Construct Bar graphs, Pie graphs.
- See Appendix 2.2 of the text.

7



B) Summarizing **Quantitative** Data Visually

- First: classify data into nonoverlapping groups.
  - Need a bin width and number of bins.

$$\text{Bin Width} = \frac{\text{max value - min value}}{\text{no. of bins}}$$

- Second, construct the frequency table, histogram.
  - Use Excel's Histogram data analysis function.

9

Answers to "How Much Income Did You Earn?"

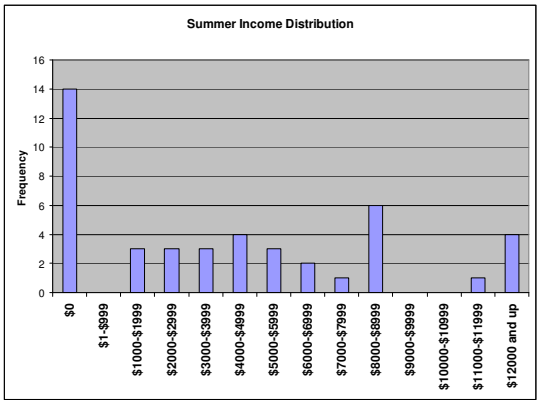
0	1200	0	8000	0	–	0	–
6400	0	2500	3000	6000	5000	8000	4000
11000	25000	4000	8800	5000	7000	8000	0
0	18000	5400	15000	3500	24000	1000	8000
0	0	2100	8000	0	4000	0	0
1000	0	3000	0	2000	4800		

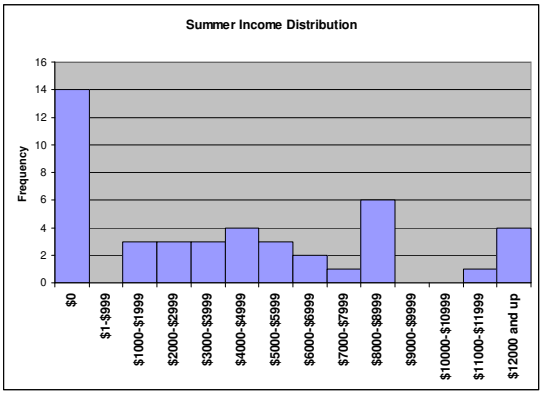
10

Frequency Table for Income Earned

Income Level	Frequency	Rel. Frequency	Percent
$0	14	0.318182	31.8%
$1-$999	0	0	0.0%
$1000-$1999	3	0.068182	6.8%
$2000-$2999	3	0.068182	6.8%
$3000-$3999	3	0.068182	6.8%
$4000-$4999	4	0.090909	9.1%
$5000-$5999	3	0.068182	6.8%
$6000-$6999	2	0.045455	4.5%
$7000-$7999	1	0.022727	2.3%
$8000-$8999	6	0.136364	13.6%
$9000-$9999	0	0	0.0%
$10000-$10999	0	0	0.0%
$11000-$11999	1	0.022727	2.3%
$12000 and up	4	0.090909	9.1%
Total	44	1	100%

11



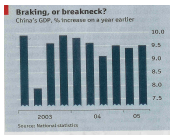


Presentation and Bin Widths

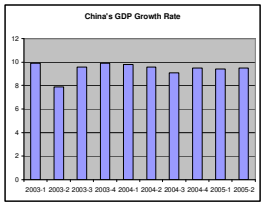
Bin Widths matter: http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html.

14

Examples of "Poor" Presentation of Graphs

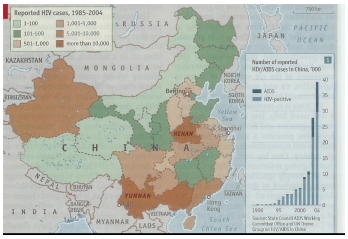


Source: Economist, July 23, 2005



15

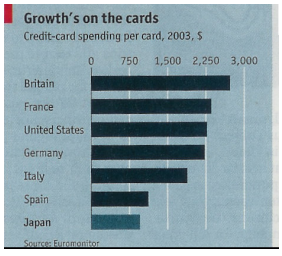
Examples of "Poor" Presentation of Graphs II



Source: Economist, July 30, 2005

16

Sideways Bar Graphs



Source: Economist, July 23, 2005

17

C) Quantitative Data: Mean, Median, etc.

- Measures of location.
- If it is a population: population parameters.
- If it is a sample: sample statistics (estimates of population value).

18

Answers to "How Much Income Did You Earn?"

0	1200	0	8000	0	–	0	–
6400	0	2500	3000	6000	5000	8000	4000
11000	25000	4000	8800	5000	7000	8000	0
0	18000	5400	15000	3500	24000	1000	8000
0	0	2100	8000	0	4000	0	0
1000	0	3000	0	2000	4800		

19

Definitions

- Observations: $x_i = x_1, x_2, \ldots x_n$ for n observations.
- Sample mean = average = xbar = $\bar{x}$
  - $\bar{x}$ = point estimate of pop'n mean μ.

$$\bar{x} = \frac{\text{sum of all sample observations}}{\text{number of sample observations}} = \frac{\sum x_i}{n}$$

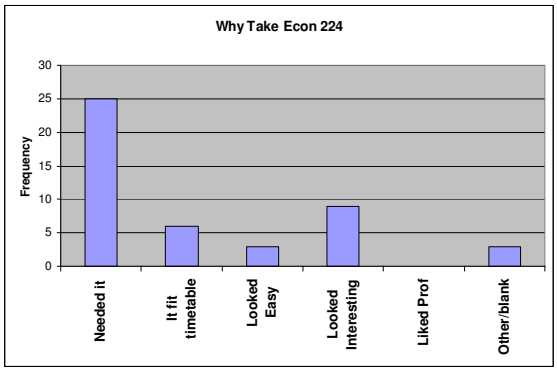
$$\mu = \frac{\text{sum of all pop'n observations}}{\text{number of pop'n observations}} = \frac{\sum x_i}{N}$$

20

Definitions cont'd

- Median = middle value for the data set when in ascending order.
- Mode: value with greatest frequency.

21



Definition's cont'd

- Percentile:
  The *p*th percentile of a data set is a value such that at least *p* percent of the items take on this value or less and at least (100 - *p*) percent of the items are greater than or equal to this value.
- Decile: 10, 20, 30, …., 90 percentile.
- Quartile: 25, 50, 75, 100.
- Quintile: 20, 40, 60, 80, 100.

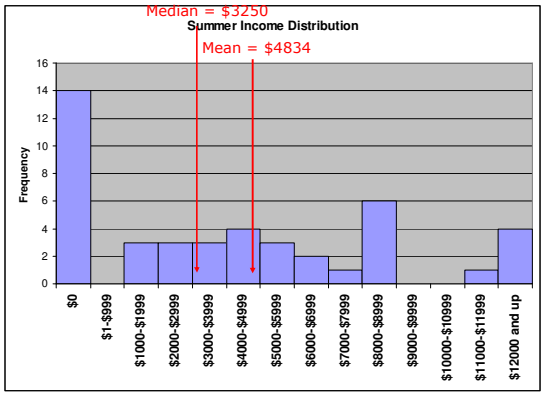
23

Summer Income and Excel's Descriptive Stats

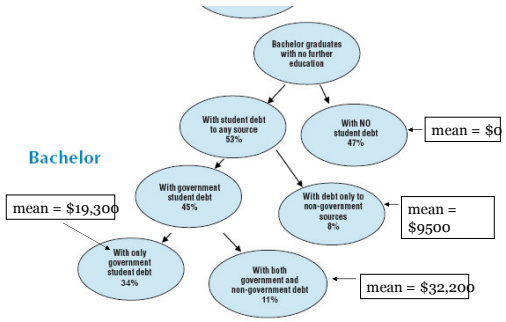
- Excel's descriptive statistics Data Analysis yields:

Summer Income	
Mean	4834.091
Standard Error	909.4113
Median	3250
Mode	0
Standard Deviation	6032.352
Sample Variance	36389276
Kurtosis	3.880683
Skewness	1.912365
Range	25000
Minimum	0
Maximum	25000
Sum	212700
Count	44

24



Student Loans



Source: Statistics Canada, Class of 2000: Profile of postsecondary graduates and student debt

26

Student Loans Cont'd

- Mode: $0
- Mean = (0.47 x $0) + (0.34 x $19,300) + (0.08 x $9,500) + (0.11 x $32,200) = $10,864.
- Mean (with debt) = $20,498.

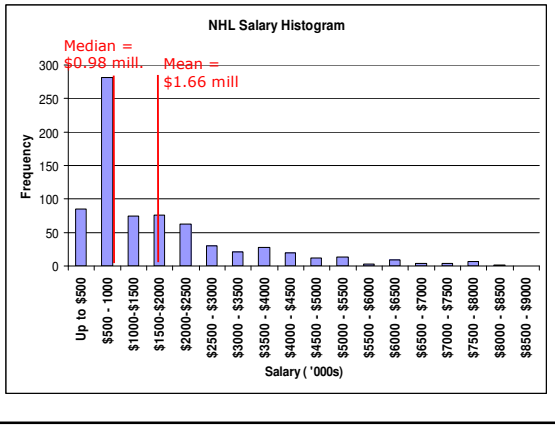
27

Mean vs. Median – Hockey Salaries

- Rodney Fort's webpage, 2006-07 NHL salaries.

	Total Salary
Mean	1661542.494
Standard Error	56159.67404
Median	984200
Mode	450000
Standard Deviation	1514229.783
Sample Variance	2.29289E+12
Kurtosis	3.295901175
Skewness	1.843760272
Range	7910000
Minimum	450000
Maximum	8360000
Sum	1207941393
Count	727

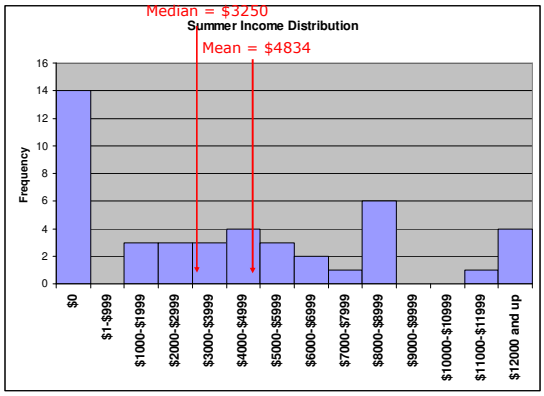
28



D) Summarizing Quantitative Data: Variability

- Why variability matters:
  - The average student in this class is:
    59% male,
    23.8 years old,
    has a GPA of 71.8%,
    expected grade of 78.3%,
    earned $4834 last year.

30



Source: *AP Photo*, retrieved from http://wcco.com

What is Variability?

- What is the chance of being near the average of the distribution, or near one or both tails?

33

Measures of Variability

- **Range**: largest value – smallest value.
- **Interquartile range**: 75th percentile – 25th percentile.
- **Population variance** is

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

- **Sample variance** is

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

34

Measures of Variability cont'd

- More usual to use the standard deviations: sample **standard deviation**: $s = \sqrt{s^2}$ pop'n **standard deviation**: $\sigma = \sqrt{\sigma^2}$.
- **Coefficient of variation**: (standard deviation/mean)x100% = (σ/μ)x100% or $(s/\bar{x})$x100%.
  - Relative size of standard deviation.

35

Standard Deviations and the Distribution

Chebyshev's Theorem (all distributions)	Empirical Rule (normal distribution only)
–	≈ 68% of distribution within ONE standard deviation of the mean
At least 75% of distribution within TWO standard deviations of the mean	≈ 95% of distribution within TWO standard deviations of the mean
At Least 89% of distribution within THREE standard deviations of the mean	≈ 99% of distribution within THREE standard deviations of the mean

36

What is Variability Again

Standard deviation (left) > standard deviation (right)

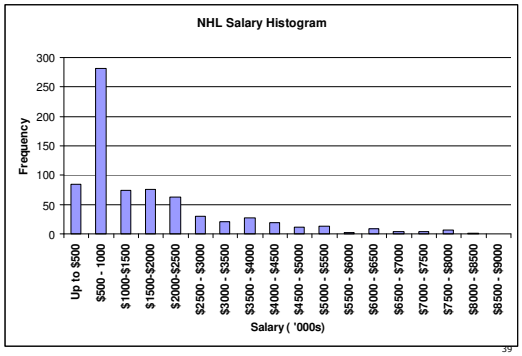
37

Summer Income and GPAs

True Econ 224 Summer Income	"Adjusted" Econ 224 Summer Income
Mean = $4834.09	Mean = $ 4834.09
Sample Variance = $36,3892,246	Sample Variance = 9,091,310.40
Standard Deviation = $6032.35	Standard Deviation = $3015.18
Coeff. of Variation = 124.7%	Coeff. of Variation = 62.4%
± 1 St. Dev. Range: –$1198.26 – $10866.44	± 1 St. Dev. Range: $1818.91 – $7849.27

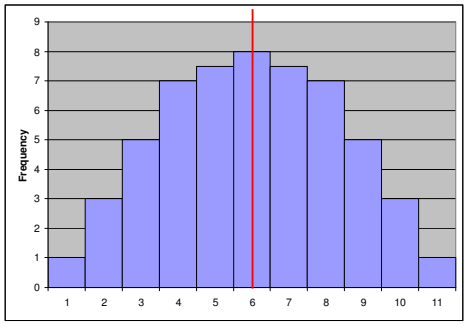
38

E) Skewness, Kurtosis, Outliers



39

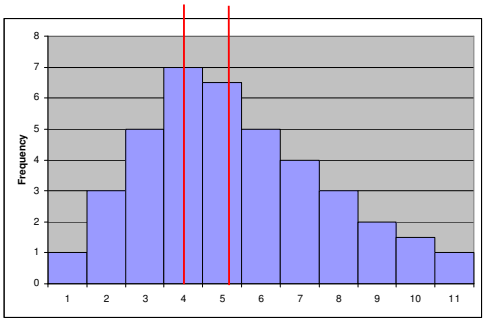
Symmetric Distribution – Skewness measure = 0



Mean = Median = 6

40

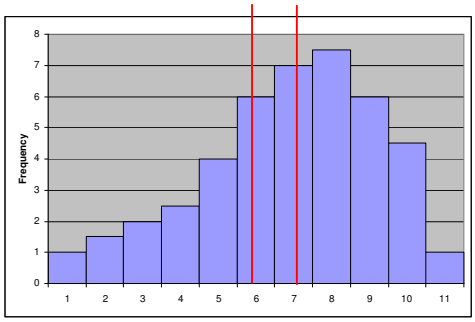
Skewed Right – Skewness measure > 0



Median = 4.5 Mean = 5.3

41

Skewed Left – Skewness measure < 0



Median = 6.5 Mean = 6.9

42

Kurtosis

- How fat are the tails.
  - How likely are extreme values?

43

F) Measuring Relationships Between Variables

- NHL Salary Data – what might we be interested in correlating this with?
- Summer Income?

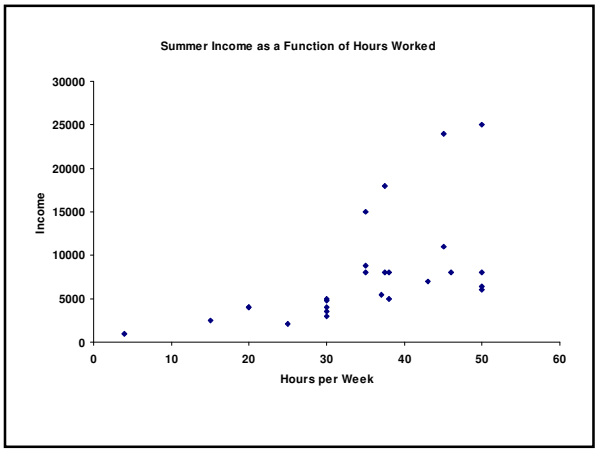
44

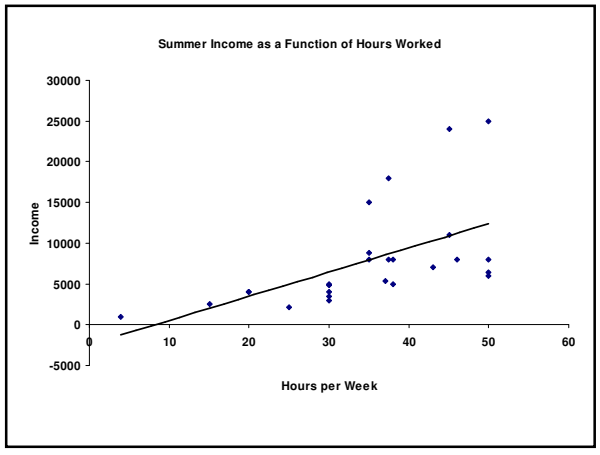
Methods

- Cross-tabulation tables – read on your own.
- Our focus:
  - scatter diagrams
  - covariance
  - correlation
  - regression analysis (eventually).

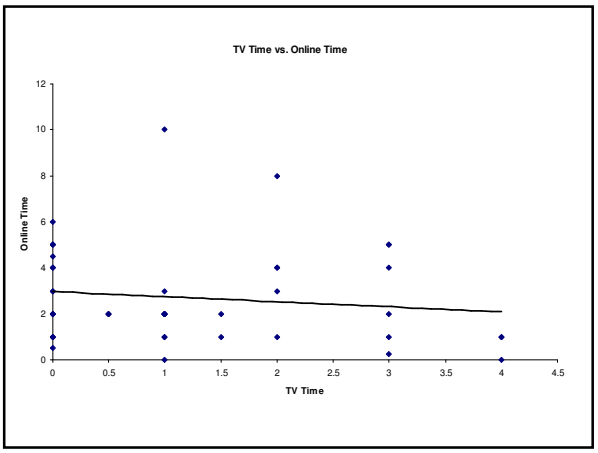
45

Income	hrs/week	Income	hrs/week
8000	38	8000	35
6400	50	18000	37.5
2500	15	5400	37
3000	30	15000	35
6000	50	3500	30
5000	38	24000	45
8000	50	1000	4
4000	20	8000	37.5
11000	45	2100	25
25000	50	8000	46
4000	20	4000	30
8800	35	1000	200
5000	30	2000	200
7000	43	4800	30

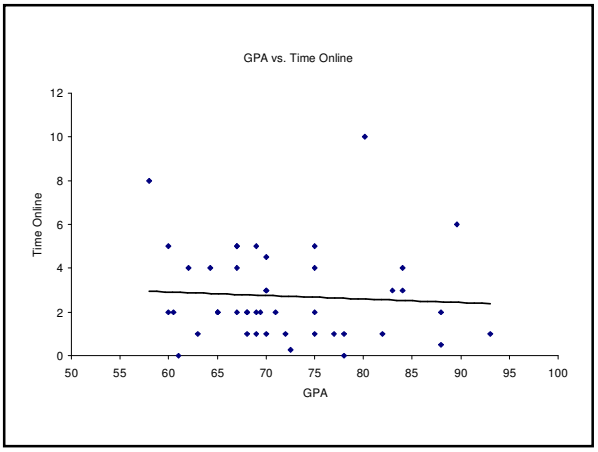




TV	Online	TV	Online	TV	Online
0	5	1.5	1	1	2
0	1	3	5	0.5	2
0	3	0	6	3	2
0	2	4	1	1	2
0.5	2	0	3	2	1
2	8	0	2	0	1
0	5	0	4	1	3
1	10	4	1	1	2
3	4	0	0.5	4	0
2	4	1	1	2	4
0	4	0	5	2	3
3	0.25	1	2	1	2
0	4.5	3	5	0	2
1.5	2	1	1	0	1
3	1	0.5	2		
1	0	1	2		



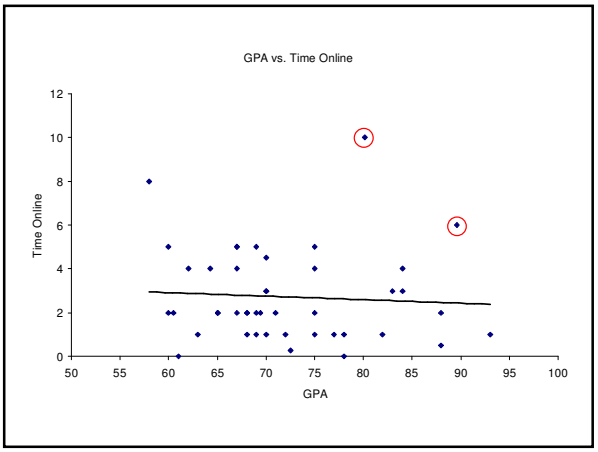
GPA	Online	GPA	Online	GPA	Online
75	5	61	0	88	0.5
63	1	69	1	72	1
70	3	84	3	67	5
68	2	71	2	69	2
60.5	2	78	0	69	5
58	8	75	4	68	1
67	5	70	3	65	2
80.12	10	82	1	65	2
64.3	4	60	5	88	2
84	4	89.6	6	68	2
62	4	93	1	67	2
72.5	0.25	83	3	60	2
70	4.5	65	2	70	1
75	2	67	4	75	1
77	1	78	1	69.4	2

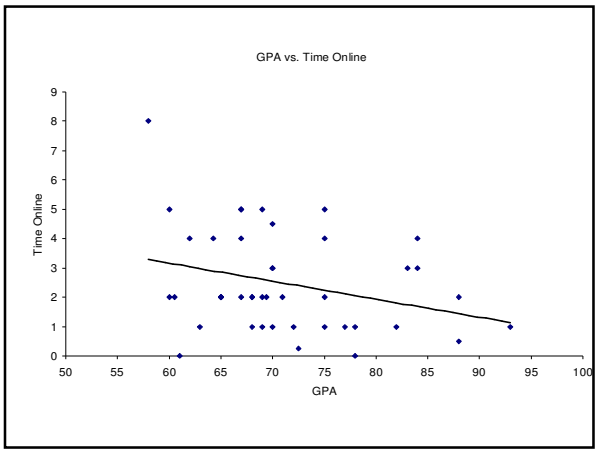


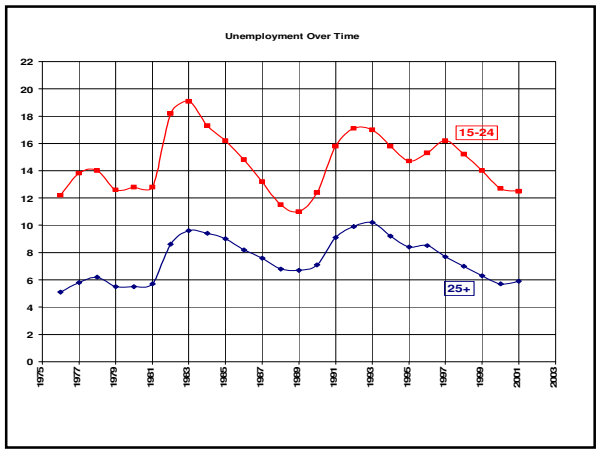
Outliers

- Rare, extreme values may distort the outcome.
  - Could be an error.
  - Could be a very important observation.
- Outlier: more than 3 standard deviations from the mean.

53







G) Covariance and Correlations

- Covariance measures relationship between 2 variables.
- Covariance > 0, positive relationship.
- Covariance < 0, negative relationship.
- Covariance = 0, no relationship

- Sample covariance:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

57

Covariance cont'd

- Population covariance:

$$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

58

Correlation Coefficient

- Controls for units, and is between -1 and +1.
- Sample correlation coefficient:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- Population correlation coefficient:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

59

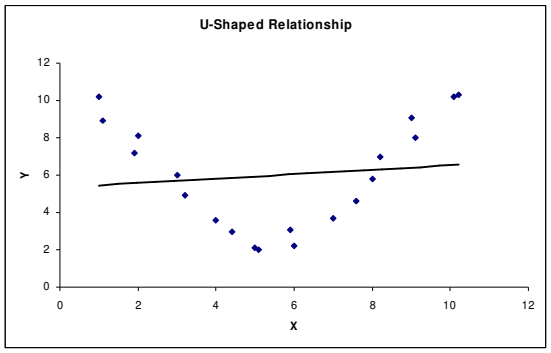
Excel's Correlation Tables

	Age	GPA	E(grade)	Courses	Econ	Stat	TV	Online	Income	hrs/wk
Age	1.0000									
GPA	-0.1173	1.0000								
E(grade)	0.0004	0.3621	1.0000							
Courses	-0.2362	0.1773	0.2770	1.0000						
Econ	0.0079	0.0697	0.2468	0.4995	1.0000					
Stat	-0.0840	0.3247	-0.0108	0.1372	0.2113	1.0000				
TV	0.0241	0.1676	0.0113	0.2419	0.2142	0.1533	1.0000			
Online	-0.2000	-0.0848	-0.0157	-0.0477	0.0308	0.2131	-0.2320	1.0000		
Income	-0.1137	0.3479	0.2534	0.3797	0.1530	-0.0618	0.3821	-0.2910	1.0000	
hrs/wk	-0.0845	0.0328	0.1347	-0.0058	-0.0791	-0.1020	-0.0324	-0.2312	0.2129	1.0000

≈ Zero Correlation

Positive Correlation

Negative Correlation



Correlation = +0.12.

61

Correlation ≠ Causation

- Cross-Tab all.
- Russian doctors and the plague.

62

H) Index Numbers and Linear Transformations

- **NFL Quarterback Rating Formula**

  a = (((Comp/Att) * 100) -30) / 20
  b = ((TDs/Att) * 100) / 5
  c = (9.5 - ((Int/Att) * 100)) / 4
  d = ((Yards/Att) - 3) / 4

  a, b, c and d can not be greater than 2.375 or less than zero.

  QB Rating = (a + b + c + d) / .06

Source: www.primetimecomputing.com

63

U.N.'s Human Development Index

HDI =

$$\frac{1}{3}\times\left(\frac{\text{life expectancy}}{100}\right)+$$
$$\frac{1}{3}\times\left(\frac{2}{3}\times\frac{\text{Adult Literacy Rate}}{100}+\frac{1}{3}\times\frac{\text{School Enrolment Rate}}{100}\right)+$$
$$\frac{1}{3}\times\left(\frac{\text{Log(GDP per capita)} - \log(100)}{\text{Log}(40000) - \log(100)}\right)$$

Source: Wikipedia

64

How Valid are Index Numbers

- Where do the weights come from?
  - Quarterbacks: why are completions so heavily weighted?
  - HDI: why is real GDP capped at 40,000? Why is it 1/3?
- The underlying components are surveys with means and standard deviations.
  - What is the mean and standard deviation of the overall index?

65

Mean, Variance of an Index

- Suppose X = a + bY
- Mean (X) = a + b*mean(Y).
  - Simple for linear functions.
- Variance (X) = b^2*variance(Y)
- Suppose X = a + bY + cZ.
- Variance (X) = b^2*variance(Y) + c^2*variance(Z) + 2bc*covariance(YZ)
- Indexes inherit the properties of their components.

66