About this

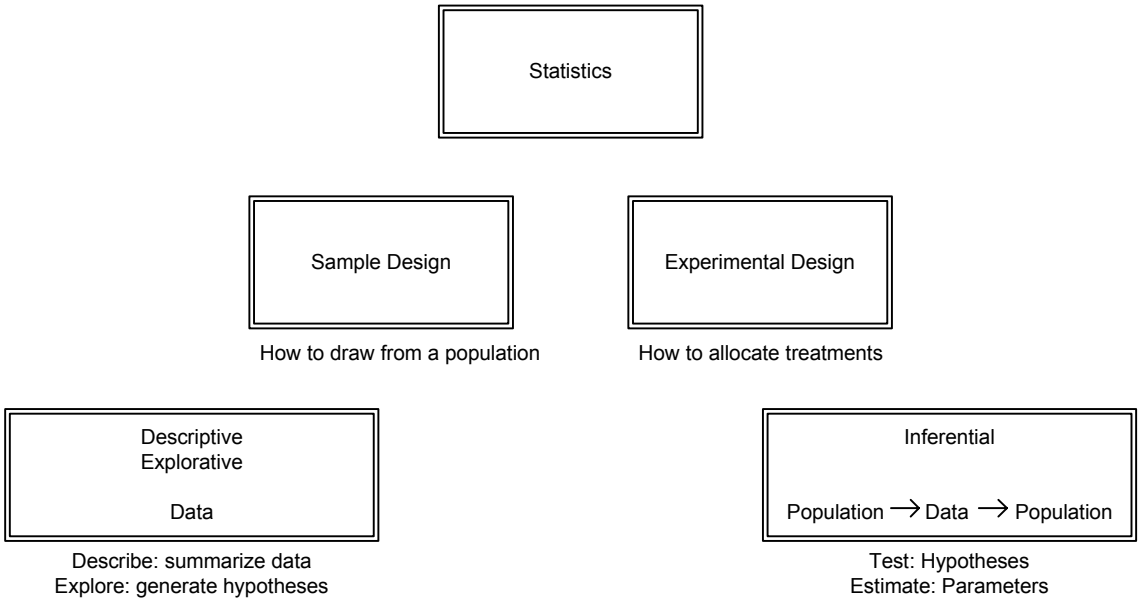
Descriptive Statistics

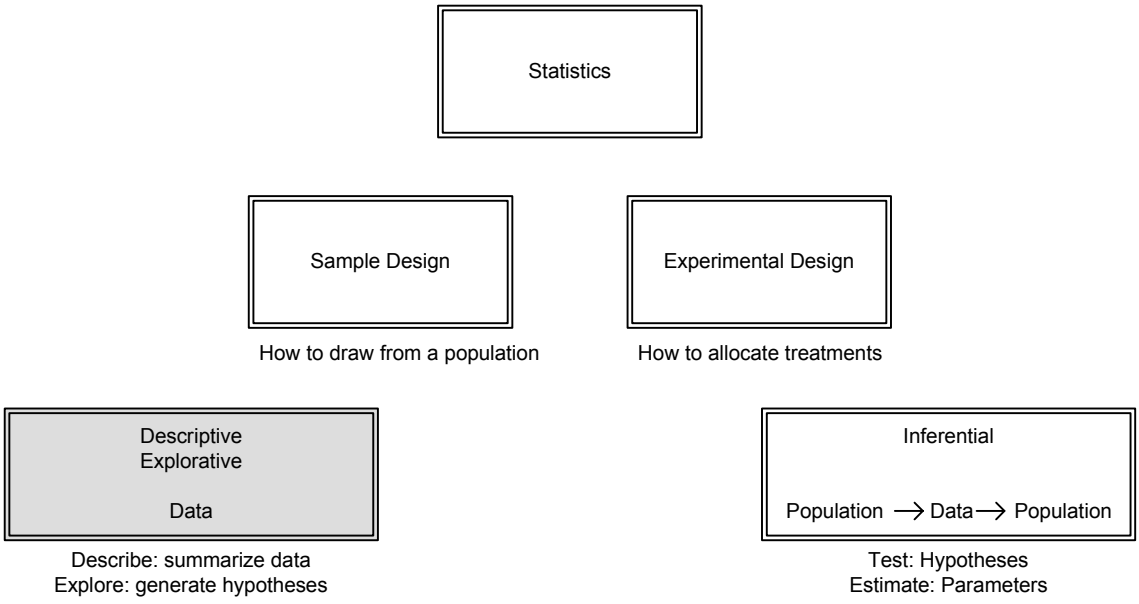
This overview summarizes mainly the first three chapters of Moore (2006).

Added are respective contents from van Belle (2008) and Good (2006).

You will also be informed about the state of the arts and controversial questions.

The text is not intended to substitute textbooks.





Levels of measurement		
Categorical		
Univariate	Bivariate	Multivariate
Freq. tables Bar Pie	Crosstable Stacked Bar / Pie	Crosstables
Quantitative		
Univariate	Bivariate	Multivariate
Mean, Std, SEM, Median, Q1, Q3, Skewness, Kurtosis Five number summary Histogram, Boxplot	Correlation (Pearson, Spearman) Partial Correlation Regression (OLS, WLS,Quantile) Scatterplot	Factor Analysis Cluster Analysis Multidim. scaling (MDS) Discriminant Analysis Data Mining Online Analytical Processing (OLAP) others e.g.Sun ray

Grayed topics are covered in this paper

Definitions

Individual	Object described by data
Variable X	Any characteristic of an individual (e.g. Sex; Height; Income …)
Levels of measurement	Nominal / Ordinal / Quantitative (discrete/continuous)

Levels of measurement

Levels of measurement and appropriate statistics

Levels of measurement		Operations & relations	Appropriate statistics				Examples
	Sublevel		Proportion	Mode	Median	Moments*	
Categorical Nominal			✓	✓			sex; colors; blood groups
Ordinal		$\leq$	✓	✓	✓	**?**	grades; "low medium high"
Quantitative Metric Continuous Scale	**Interval**	$\leq \pm$	✓	✓	✓	✓	IQ; Date
	Ratio	$\leq \pm \cdot \div$	✓	✓	✓	✓	Weight; Age; Income; Counts

* mean; variance; skewness; kurtosis

Descriptive Statistics

Categorical

(Nominal / Qualitative)

	Univariate	Bivariate	Multivariate
Numerical	Table Items and frequencies and percentages	Crosstab = Contingeny table Chi^2 Odds ratio & Relative risk (2 x 2) Polychoric (if not 2x2 then ordinal) Lambda, Tau Gamma (ordinal) Kappa (n x n)	Nested Crosstab Mantel Haenszel (m x n) x (2 x 2)
Graphical	Bar Chart Pie Diagram	Bar Chart Pie Diagram clustered or stacked	Correspondence analysis

Quantitative

(Metric / Scale / Quantitative / Continuous)

	Univariate	Bivariate	Multivariate
Numerical	Min Q1 Q2 (Median) Q3 Max Skewness Kurtosis If symmetric: Mean Standard deviation Standard error of mean SEM	Correlation matrix Pearson or Spearman Correlation Zero order or partial	Principal component analysis or Factor analysis
Graphical	Boxplot Histogram	Scattergram	

Quantitative x Categorical

Boxplot (bivariate)

ROC-curves (m x bivariate)
Receiver operating characteristic

Cluster analysis (multivariate)

Discriminant analysis:
substituted by Logistic Regression

Guido Lüchters 10/2009

Descriptive Statistics

Categorical

(Nominal / Qualitative)

Univariate

Favoured Political Party

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Conservative	497	50%	50%	50%
	Liberal	153	15%	15%	65%
	Social demo.	216	22%	22%	87%
	Green	35	4%	4%	90%
	Communist	52	5%	5%	95%
	Right wing	47	5%	5%	100%
	Total	1000	100%	100%	

Tab

Bar

Pie

Bivariate

		Favoured Political Orientation (derived)		
		Left	Right	Total
Sex	female	131	369	500
		26%	74%	100%
	male	172	328	500
		34%	66%	100%
Total		303	697	1000
		30%	70%	100%

Chi²
Odds ratio & Relative risk (2 x 2)
Polychoric (if not 2x2 then ordinal)
Lambda, Tau
Gamma (ordinal)
Kappa (n x n)

Crosstab

clustered

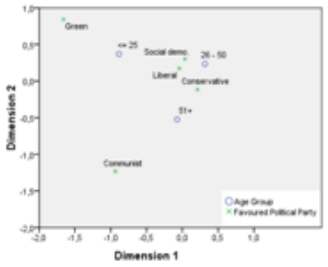
stacked

Multivariate

			Favoured Political Orientation (derived)		
Former State of Germany			Left	Right	Total
west	Sex	female	79	171	250
			32%	68%	100%
		male	118	132	250
			47%	53%	100%
	Total		197	303	500
			39%	61%	100%
east	Sex	female	52	198	250
			21%	79%	100%
		male	54	196	250
			22%	78%	100%
	Total		106	394	500
			21%	79%	100%

Mantel Haenszel (m x n) x (2 x 2)

Nested Crosstab



Correspondence analysis

Quantitative

(Metric / Scale / Quantitative / Continuous)

Univariate

Descriptives

			Statistic	Std. Error
age	Mean		45,2	,4
	95% Confidence Interval for Mean	Lower Bound	44,5	
		Upper Bound	45,9	
	5% Trimmed Mean		45,3	
	Median		44,9	
	Variance		25,5	
	Std. Deviation		5,0	
	Minimum		28,2	
	Maximum		58,9	
	Range		30,8	
	Interquartile Range		6,6	
	Skewness		-,2	,2
	Kurtosis		,4	,3

(Five) number summaries:

Boxplot

Histogram

Scattergram

Bivariate Multivariate

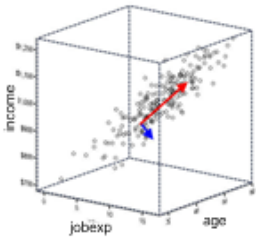
Correlations

		income	age	jobexp
income	Pearson Correlation	1		
	Sig. (2-tailed)			
	N	200		
age	Pearson Correlation	,518**	1	
	Sig. (2-tailed)	,000		
	N	200	200	
jobexp	Pearson Correlation	,894**	,517**	1
	Sig. (2-tailed)	,000	,000	
	N	200	200	200

**. Correlation is significant at the 0.01 level (2-tailed).

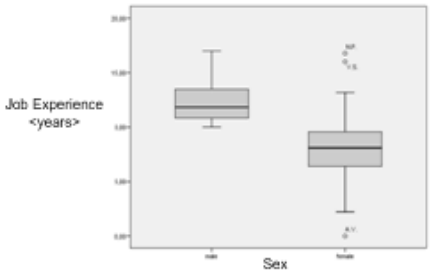
Pearson or Spearman Correlation
(zero or partial order)

Factor analysis

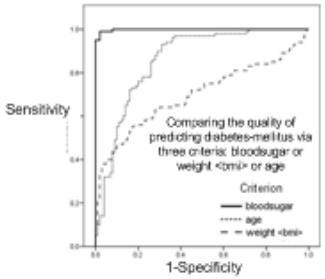


Quantitative x Categorical

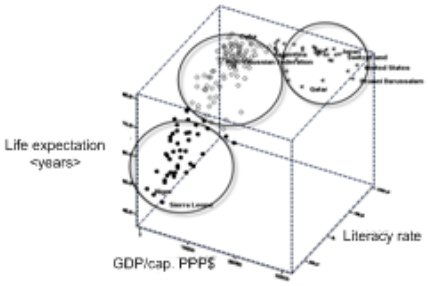
Boxplot (bivariate)



ROC-curves (m x bivariate)



Cluster analysis (multivariate)



Guido Lüchters 10/2009

Categorical

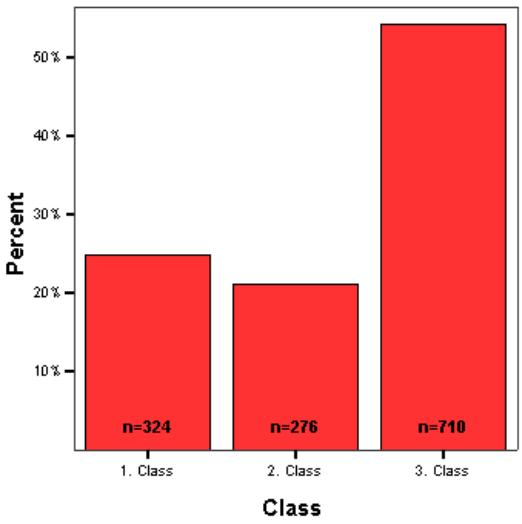
Univariate

Frequency table

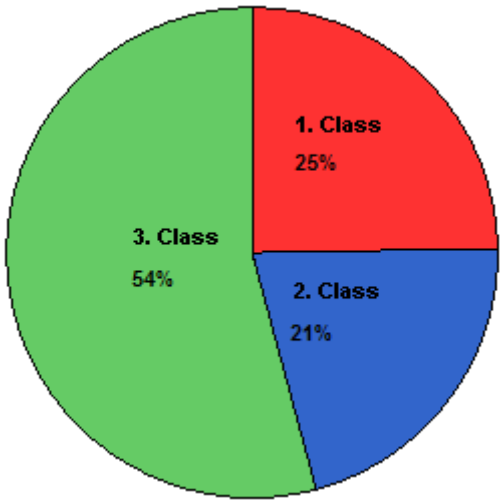
Passengers on the Tlitanic by travel-class

		Frequency	Percent
Valid	1. Class	324	24,7
	2. Class	276	21,1
	3. Class	710	54,2
	Total	1310	100,0

Bar graph



Pie graph



Categorical

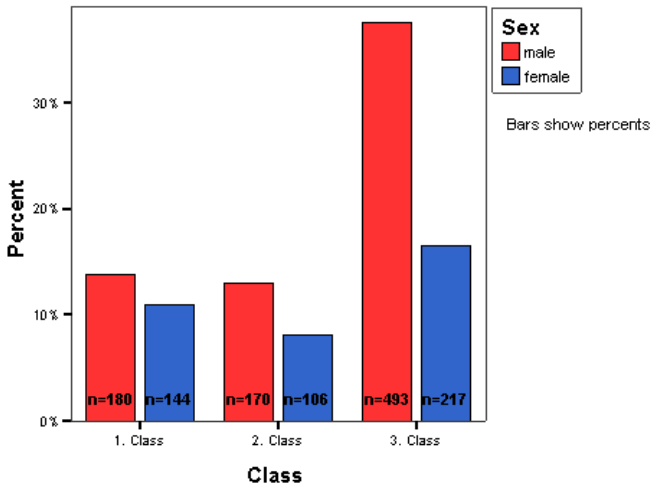
Bivariate

Cross table

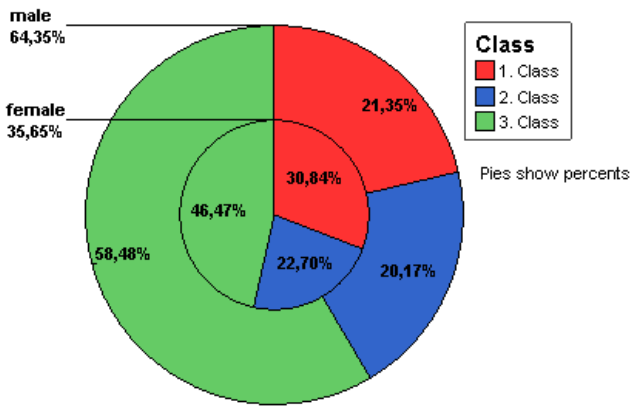
Passengers on the Titanic by Sex and Class of travel

		Sex		
		male	female	Total
Class	1. Class	180	144	324
		56%	44%	100%
	2. Class	170	106	276
		62%	38%	100%
	3. Class	493	217	710
		69%	31%	100%
Total		843	467	1310
		64%	36%	100%

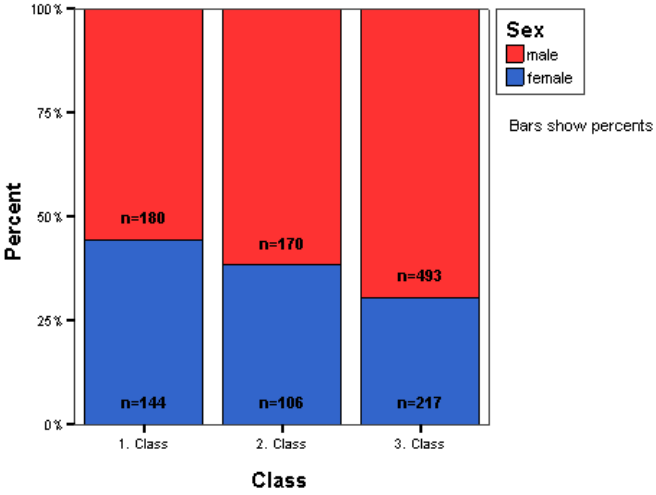
Bar graph (clustered)



Pie graph (stacked)



Bar graph (stacked)



Remark
Moore (2006) suggests bar- and pie-graphs, though with some restrictions (Moore, 2006, pp. 7-9).

Other authors find both graphs generally not suitable:
"Never use a pie chart … Bar graphs waste ink; they don't illuminate complex relationships … Stacked bar graphs are worse than bar graphs … Three-dimensional bar graphs constitute misdirected artistry … Always think of alternatives to bar graphs"
(van Belle, 2008, pp. 203-210.) (Good, 2006, pp.125-127)

Categorical

Multivariate

Passengers on the Titanic by Sex, Class of travel, and Survived status

Sex			Survived		
			survived	died	Total
male	Class	1. Class	62	118	180
			34%	66%	100%
		2. Class	24	146	170
			14%	86%	100%
		3. Class	77	416	493
			16%	84%	100%
	Total		163	680	843
			19%	81%	100%
female	Class	1. Class	139	5	144
			97%	3%	100%
		2. Class	94	12	106
			89%	11%	100%
		3. Class	106	111	217
			49%	51%	100%
	Total		339	128	467
			73%	27%	100%

Quantitative Univariate

Statistics: Mean; variance; standard deviation; median; quartiles (Q1; Q2; Q3); IQR

Definitions

Distribution of a variable X	Tells us what values it takes and how often it takes these values
Characteristics of distributions	
Center	Mean / Median also (trimmed and weighted means)
Spread	Min Max / Range / Variance / Standard deviation / Quartiles (Q1, Q2, Q3)
Shape parameters	Modes (number of peaks) / Skewness / Kurtosis

Statistic	A quantity calculated from the data (e.g. mode, min, max, mean, median, standard deviation, Q1, Q3, skewness, kurtosis, …)

Center

Mean $\bar{x}$

$$\frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Median $\tilde{x}$

$$\begin{cases} x_{[n+1]/2} & n \text{ odd} \\ \dfrac{x_{[n]/2} + x_{[n/2+1]}}{2} & n \text{ even} \end{cases}$$

The median is the midpoint of the distribution: 50% of the observations are smaller and the other 50% are larger

where **X[i]** denotes the ordered ith value of x_1, …, x_n

Spread

Variance s^2

$$\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Standard deviation s $\sqrt{s^2}$

Q1	First quartile 25% of the data are smaller than Q1 Median of values left from the median (excluded)
Q2	Second quartile = Median
Q3	Third quartile 25% of the data are larger than Q3 Median of values right from the median (excluded)
Interquartile Range IQR	Q3-Q1

A resistant measure	is relatively unaffected by changes in the numerical value of a small portion of observations

Quantitative Univariate

How to calculate statistics (mean, variance, standard deviation, median)

Examples: Mean, Variance, Standard deviation, Median

x_1,x_2,x_3,x_4 = 5,3,1,7

Mean $\bar{x}$ = (5+3+1+7) / 4 = 16 / 4 = **4**
Variance s^2 = [(5 - **4**)² + (3 - **4**)² + (1 - **4**)² + (7 - **4**)²] / (4-1) = 20 / 3 = **6.7**
Standard deviation s = $\sqrt{6.7}$ = **2.6**

To determine the Median have to order the four numbers
$x_{[1]},x_{[2]},x_{[3]},x_{[4]}$ = 1,3,5,7

Then - because these are 4 numbers (even) - the median is the middle between the second and third number:(3+5)/2 = 4
Median $\tilde{x}$ = **4**

---

x_1,x_2,x_3,x_4 = 5,3,1,1000

Mean $\bar{x}$ = (5+3+1+1000) / 4 = 1009 / 4 = **252.3**
Variance s^2 = [(5 - **252.3**)² + (3 - **252.3**)² + (1 - **252.3**)² + (1000 - **252.3**)²] / (4-1) = 745514.8 / 3 = **248504.9**
Standard deviation s = $\sqrt{248504.9}$ = **498.5**

To determine the Median have to order the four numbers
$x_{[1]},x_{[2]},x_{[3]},x_{[4]}$ = 1,3,5,1000

Then - because these are 4 numbers (even) - the median is the middle between the second and third number:(3+5)/2 = 4
Median $\tilde{x}$ = **4**

As you can see in this example: the median is not changing even if extreme values appear, while mean, variance and standard deviation are changing substantially.

The median is resistant.

The mean, the variance and the standard deviation are not resistant.

Quantitative
Univariate

How to calculate statistics (median, Q1, Q3)

Examples: Median, Q1, Q3 (n odd)

Data already ordered (n=13)

19 22 23 23 23 26 26 27 28 29 29 31 32

19 22 23 23 23 26 26 27 28 29 29 31 32
Because n is odd the median is the center observation (n+1)/2 = 7th observation

Median = 26

50% of the values are smaller equal than 26 (Median = Q2)

19 22 23 | 23 23 26 26 27 28 29 29 31 32

Q1 is the median of 6 values left from the median (excluded)
Therefore is the middle between the 6/2 = third and the 6/2+1 = 4th value:

Q1 = (23+23)/2 = 23

25% of the values are smaller equal than 23 (Q1)

19 22 23 23 23 26 26 27 28 29 | 29 31 32
Q3 is the median of 6 values right from the median (excluded)
Therefore is the middle between the 6/2 = third and the 6/2+1 = 4th value counting from the median

Q3 = (29+29)/2 = 29

75% of the values are smaller equal than 29 (Q3)

Examples: Median, Q1, Q3 (n even)

Data already ordered (n=20)

13 15 16 16 17 19 20 22 23 23 23 24 25 25 26 28 28 28 29 32

13 15 16 16 17 19 20 22 23 23 | 23 24 25 25 26 28 28 28 29 32
Because n is even the median is the middle between the 20/2 = 10th and the 20/2+1 = 11th value:

Median = (23+23)/2 = 23

---

13 15 16 16 17 | 19 20 22 23 23 | 23 24 25 25 26 28 28 28 29 32
Q1 is the median of values left from the median
These are 10 values so the median is the middle between the 10/2 = 5th and the 10/2+1 = 6th value:

Q1 = (17+19)/2 = 18

---

13 15 16 16 17 19 20 22 23 23 | 23 24 25 25 26 | 28 28 28 29 32
Q3 is the median of values right from the median
These are 10 values so the median is the middle between the 10/2 = 5th and the 10/2+1 = 6th value counting from the median

Q3 = (26+28)/2 = 27

Q1, Median and Q3 will not change if we exchange some extreme values:
-1000 -500 -300 -100 17 | 19 20 22 23 23 | 23 24 25 25 26 | 28 500 1000 1000 100000
Still
Q1 = 18
Median = 23
Q3 = 27

Remark

The median, Q1 and Q3 are resistant..

As you can see in this example:

The median, Q1 and Q3 are not changing even if a lot (8 of 20) of extreme values appear,
So that Q1 and Q3 correctly and resistantly describe the spread of the mid 50% of the data 18-27 but not the spread of all data.

To describe the full spread, minimum and maximum are also necessary.

Quantitative Univariate

Five- number summary and Boxplot

Definition

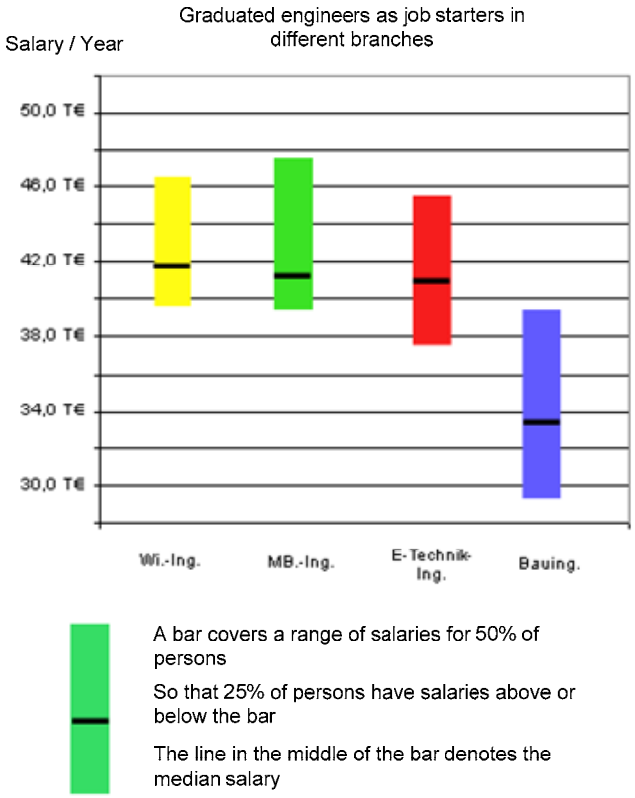
Five-number summary[1]	**Min Q1 Median Q 3 Max**

"These five numbers offer a reasonably complete description of center and spread. … The median describes the center of the distribution; the quartiles show the spread of the center half of the data; the minimum and maximum show the full spread of the data." (Moore, 2006, p. 46)

Definitions

Boxplot[2] **= Box-Whisker Plot**	"A graph of the five-number summary - A central box spans the quartiles Q1 and Q3 - A line in the box marks the median - Lines extend from the box out to the smallest and largest observations - Call an observation a suspected outlier if it falls more than 1.5 X IQR above the third quartile or below the first quartile" (Moore, 2006)
Stem leaf Histogram	old form of histogram
Histogram	shows counts or percentage of observations that fall into classes

Example Boxplot (no whiskers) plots in the German journal "Der Spiegel"

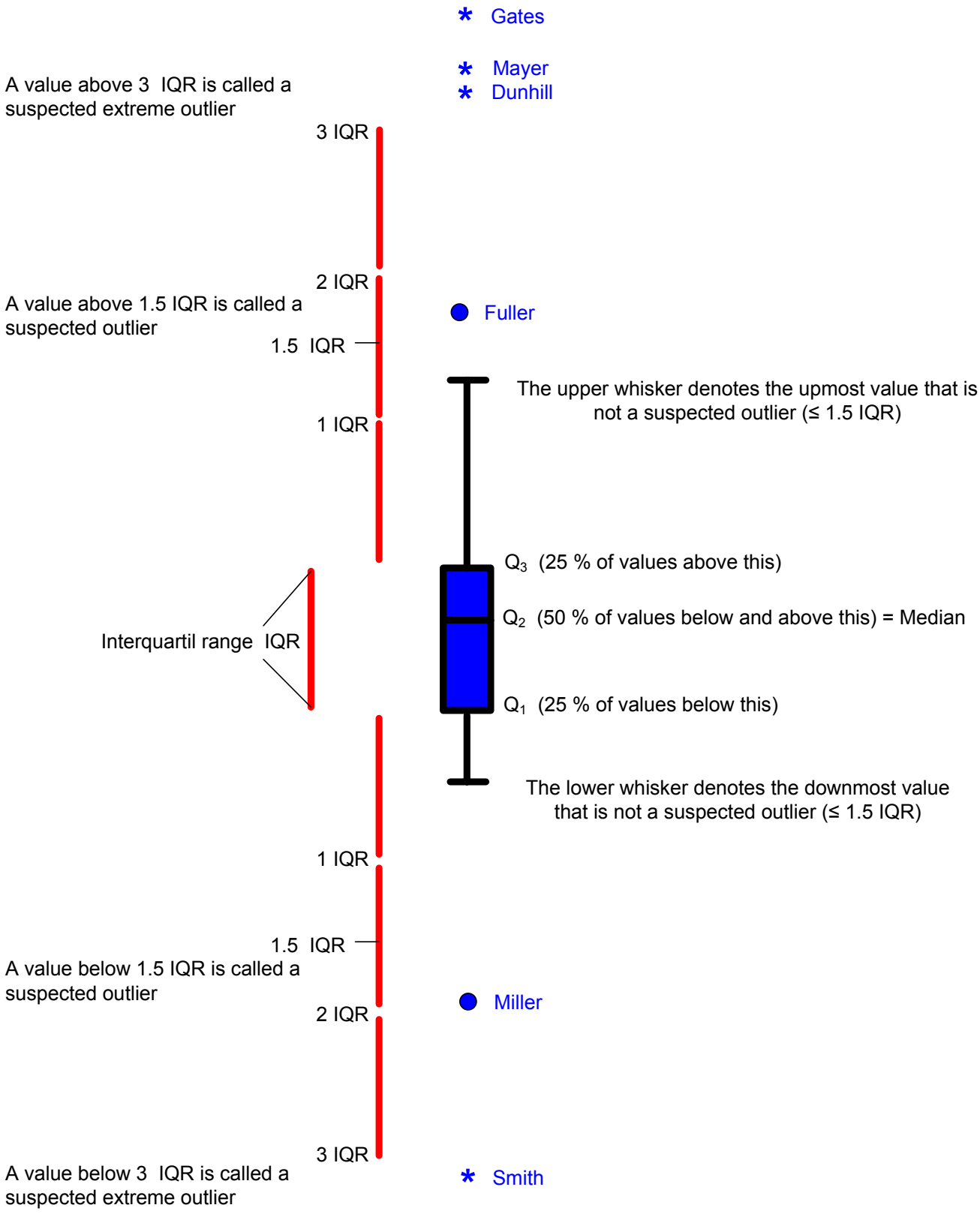


[1] By Tuckey (1977). Bowley (1901) suggested seven-figure summary which included also percentiles (5% and 95%)
[2] Quartile box plots by Tuckey (1977), Percentile box plots by Bowley (1901)

Quantitative Univariate

Boxplot

Box-whisker plot definition by graph



Quantitative Univariate

Boxplot

Example Boxplot

Data:

Income x 1000 / year:	11 12 11 21 22 23 24 25 22 22 22 23 30 31 41
ordered:	11 11 12 **21** 22 22 22 **22** 23 23 24 **25** 30 31 41
Min	11
Q1	21
Median	22
Q3	25
Max	41
IQR	25-21 = 4

Limits:

Q3 + 1.5 IQR	25 + 1.5*4	= 31
Q1 - 1.5 IQR	21 - 1.5*4	= 15
Q3 + 3 IQR	25 + 3*4	= 37
Q1 - 3 IQR	21 - 3*4	= 9

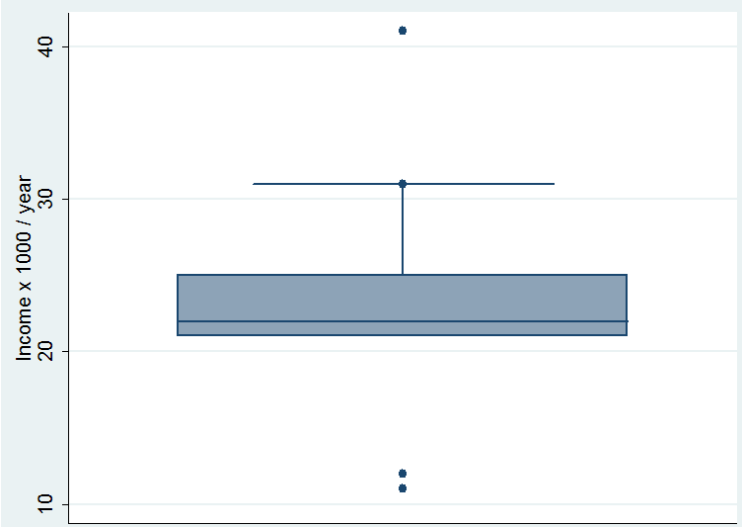
The data point 31 could be seen as the upper whisker-end or as a suspected upper outlier (border case)
The lower whisker-end is 21 which is equal to Q1 therefore there is no low whisker

41 is a an upper suspected extreme outlier (> 37)

11, 11, 12 are three suspected lower outlier (<15)

STATA ***boxplot***

boxplot x



Quantitative Univariate

Stem-leaf

Examples STEM-LEAF Histogram

The stem-leaf[3] histogram is a simple-graphical old form of a histogram which has the advantage that individual numbers can be identified (only for small data sets)

Data:
Income x 1000 / year: 11 12 11 21 22 23 24 25 22 22 22 23 30 31 41
ordered: 11 11 12 21 22 22 22 22 23 23 24 25 30 31 41

SPSS Explore

```
Income x 1000 / year Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     3,00 Extremes    (=<12,0)
     1,00       21 .  0
     4,00       22 .  0000
     2,00       23 .  00
     1,00       24 .  0
     1,00       25 .  0
     3,00 Extremes    (>=30,0)

 Stem width:      1,00
 Each leaf:       1 case(s)
```

STATA stem

```
stem x, lines(1)

Stem-and-leaf plot for (Income x 1000 / year)

  1* | 112
  2* | 122223345
  3* | 01
  4* | 1


stem x, lines(2)

Stem-and-leaf plot for (Income x 1000 / year)

  1* | 112
  1. |
  2* | 12222334
  2. | 5
  3* | 01
  3. |
  4* | 1
```

Remark

The form of stem-leaf for small data sets depends on how classes are defined

[3] By Arthur Lion Bowley (1901)

Quantitative Univariate

Histogram

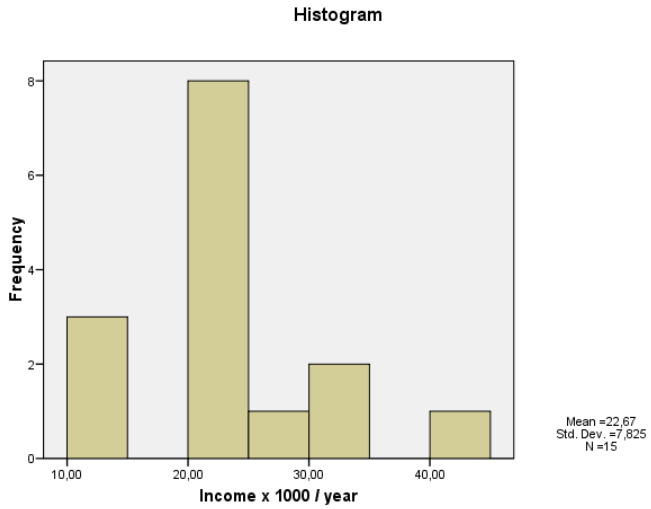
Examples Histogram

Histogram shows counts or percentage of observations that fall into **classes**

Data:

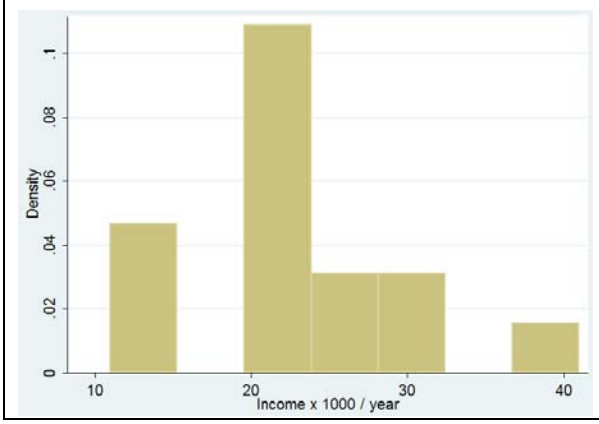
Income x 1000 / year:	11 12 11 21 22 23 24 25 22 22 22 23 30 31 41
ordered:	11 11 12 21 22 22 22 22 23 23 24 25 30 31 41

SPSS ***Explore***

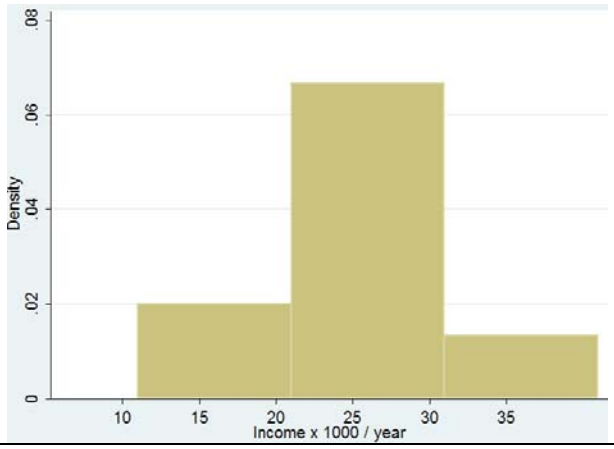


STATA ***histogram***

histogram x, **bin(7)**



histogram x, **bin(3)**



Remark

The form of histogram depends on how classes (bins[4]) are defined

[4] "bin" here has the meaning of "basket" or "container" ~ class

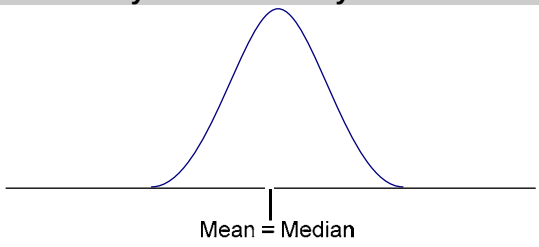
Quantitative
Univariate

Characteristics of density functions

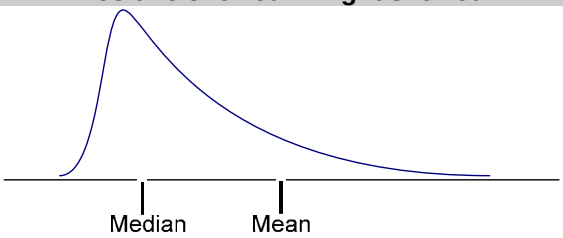
Definitions

Density function	**Provisional definition** (precise next chapter) "A smooth approximation to the irregular bars of the histogram. … is always on or above the horizontal axis and has area exactly 1 underneath it" (Moore, 2006, pp.64-67)
Shape parameters	
Mode	Modes (unimodal, bimodal, …, multimodal)
Skewness	Right skewed = positive skewed : more small than large value Left skewed = negative skewed : more large than small values

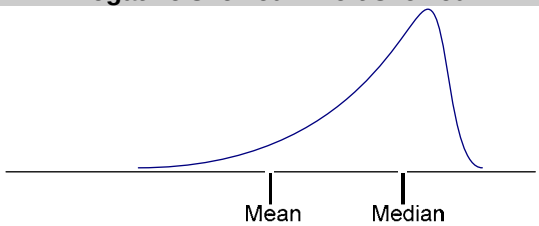
Unimodal	**Symmetric density curve**



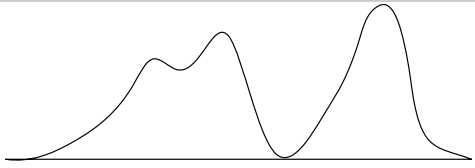
Unimodal	**Positive skewed = Right skewed**
More small values	



Unimodal	**Negative skewed = Left skewed**
More high values	



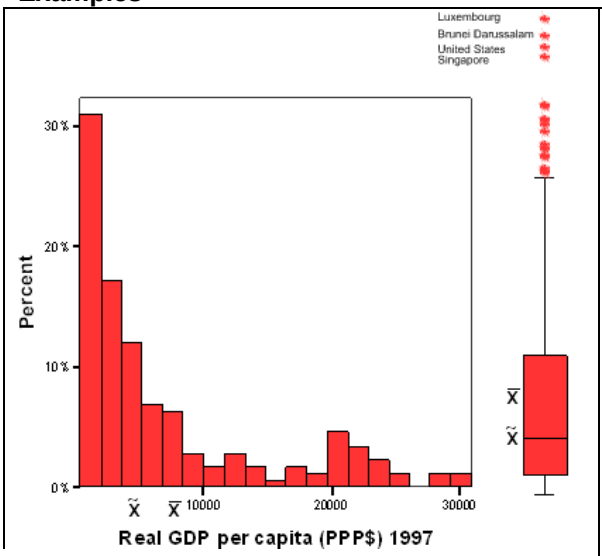
Multimodal

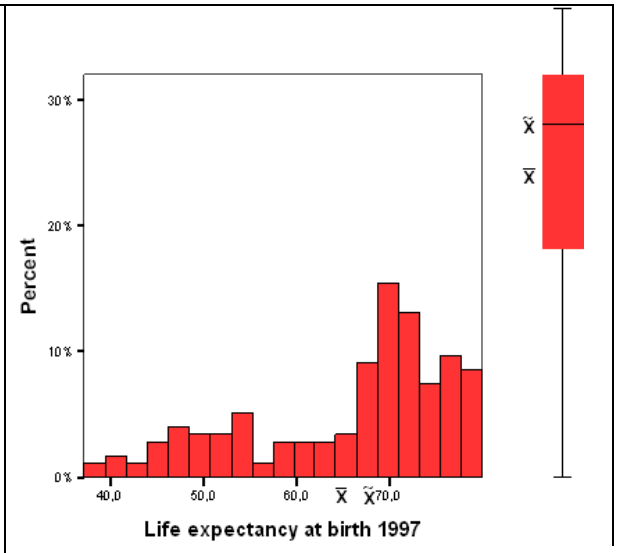


Quantitative Univariate

Characteristics of density functions

Examples





	Statistic	SE		Statistic	SE
Mean	7,200.0	580.0	Mean	66.0	0.8
95% CI for Mean	[6,100 ; 8,400]		95% CI for Mean	[64 ; 67]	
Trimmed Mean	6,500.0		Trimmed Mean	66.0	
Variance	58 Mio.		Variance	120.0	
Std. Deviation	7,700.0		Std. Deviation	11.0	
Minimum	410.0		Minimum	37.2	
Q1	1700.0		Q1	58.0	
Median (Q2)	3,900.0		Median (Q2)	70.0	
Q3	9300.0		Q3	74.0	
Maximum	30863.0		Maximum	80.0	
Range	30453.0		Range	43.0	
Skewness	1.4	0.2	Skewness	-0.9	0.2
Kurtosis	0.9	0.4	Kurtosis	-0.4	0.4

All numbers - except minimum and maximum - are rounded to first relevant digit

SE = Standard Error of the statistic
Trimmed Mean = Mean after excluding the highest 5% and the 5% lowest values (exclude suspect values first)
Q1, Q2, Q3 = first, second and third quartile
CI = Confidence Interval
Kurtosis = positive = leptokurtic = peakedness / negative = platikurtic = flat / for a normal distribution = 0 = mesokurtic

Data source: Human Development Report (HDR), UNDP, 1999

Ordinal / Quantitative Bivariate

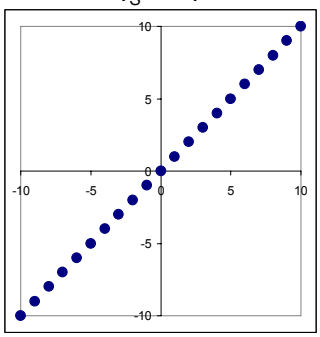
Correlation

Definitions

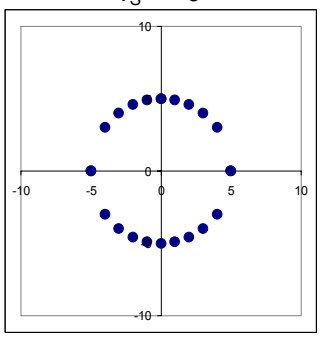
Correlation	Describes the association between variables
Pearson correlation r	Direction and strength of linear association $-1 \leqslant r \leq +1$
Spearman correlation r_S	Direction and strength of monotone association $-1 \leqslant r_S \leq +1$

Correlation is a symmetric property **r(y;x)=r(x;y)**

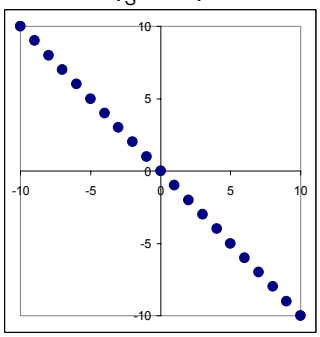
Linear with perfect strength
r = 1
r_S = 1



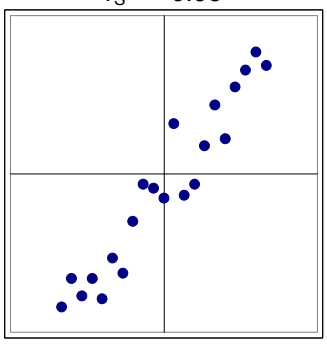
uncorrelated
r = 0
r_S = 0



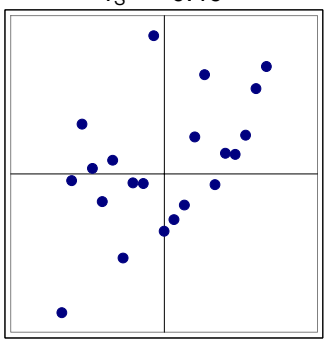
Linear with perfect strength
r = -1
r_S = -1



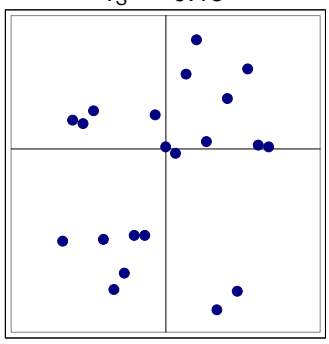
r = 0.95
r_S = 0.95



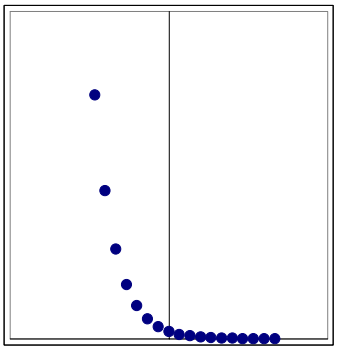
Linear with decreasing strength
r = 0.50
r_S = 0.46



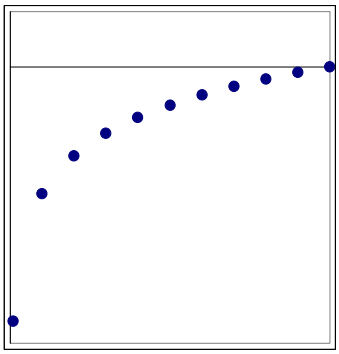
r = 0.20
r_S = 0.15



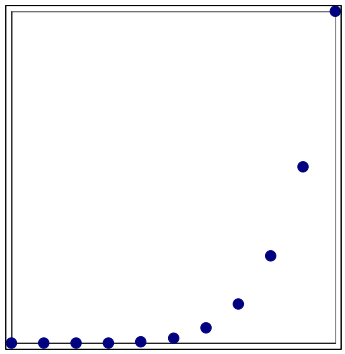
r = - 0.7
r_S = - 1



Monotone
r = 0.7
r_S = 1



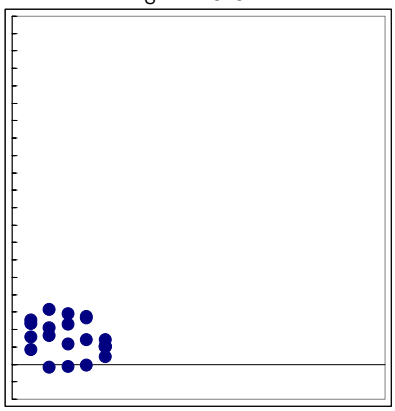
r = 0.7
r_S = 1



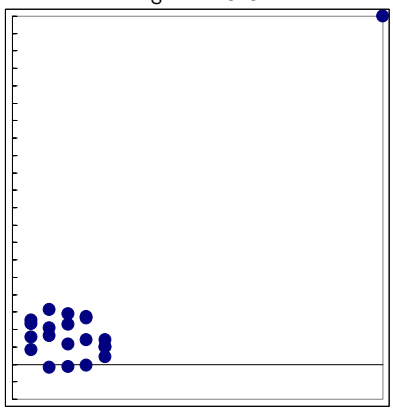
Problems:

The problem of "outliers" or inhomogeneity

r = - 0.01
r_S = - 0.01

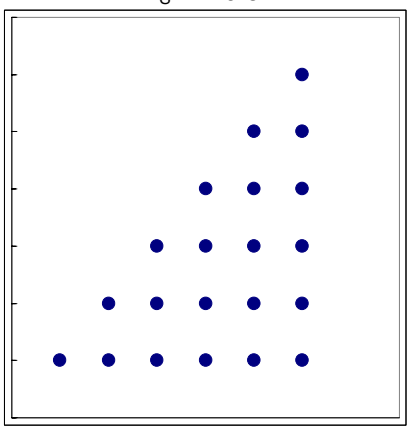


r = 0.9
r_S = - 0.01

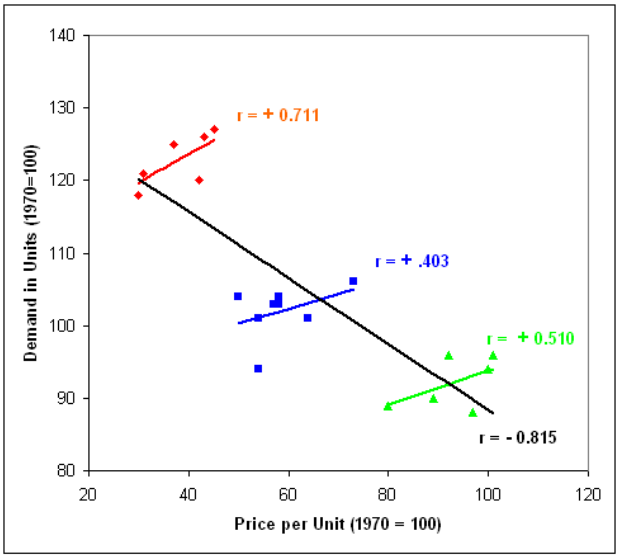


Heteroscedacity

r = 0.5
r_S = 0.5



Simpson's Paradox



Definitions (formal)

Pearson correlation r

$$\frac{1}{n-1}\sum_i\left(\frac{(x_i-\bar{x})}{s_x}\right)\left(\frac{(y_i-\bar{y})}{s_y}\right)$$

Spearman correlation r_S

$$1-\frac{6\sum_i d_i^2}{n(n^2-1)}$$

with d_i = rank difference of the ith pair

Caution

Association does not imply causation

(Moore 2006, p.160)

Quantitative Bivariate

Regression

Mere fitting data via regression is a subject of descriptive statistic.

One can always fit lines, functions, curves, splines,… to quantitative data without any assumptions or restriction.

It is just compression process in data description.

Example:
Instead of 1000 data pairs you describe your data by three characteristics: linear, slope=3 , intersection=100

You will decrease the compression factor if the fitting form has many parameters

Example
You can describe 3 data pairs with a polynomial function of 2nd degree which has 3 parameters ($y=ax^2+bx+c$):

Correlation is a symmetric property	**r(y;x)=r(x;y)**		
Regression forces you to define a direction	**y = f(x)**	**Left side variable = y**	**Right side variable = x**

Used but not always appropriate are:

"dependent",	"effect" ,	"response"	"endogene"	"predicant"	variable for the left side and
"independent",	"cause",	"explanatory",	"exogene"	"predictor"	variable for the right side.,

Regression techniques

OLS Ordinary least-squares (Carl Friedrich Gauss 1800) (not resistant)

WLS Weighted least-squares (e.g. points in areas of higher variance get less weight in order to decrease their influence on the parameter estimation)

Quantile The sum of absolute residuals against a chosen quantile is minimized.

Example:

Consider that you want to estimate the influence of prenatal care (pc) on the birth weight of babies (bw).

In an ordinary OLS regression you get an estimate which is best over the whole range of pc and bw (but sensible to extreme values and maybe not differentiated enough)

With a WLS regression you might want to correct that the spread for low bw is larger than for high bw.

In a quantile regression you could compare the estimate for the lower 5% birth weights, with the estimate for the upper 5% birth weights. Also you can estimate the prenatal influence on the middle 90% birth weights.

Quantitative Bivariate

Regression OLS

Least-squares

Least-squares methods search estimates so that the sum of squared deviations is minimized.

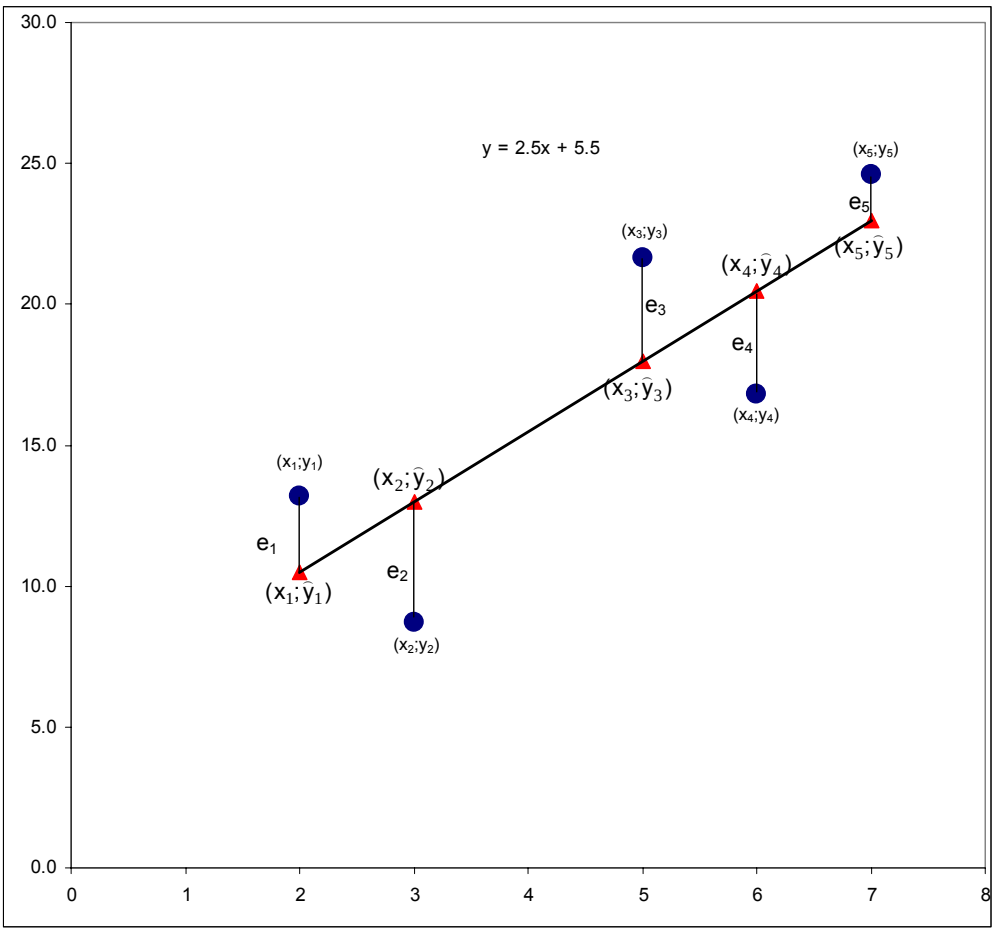
Example:
If $(x_i;y_i)$ are 5 data pairs and we want to fit a function $f(x) = y = ß_1x + ß_0$

We search $\hat{ß}_1$ and $\hat{ß}_0$ for and so that the sum of squares (SS) is minimal:

$$SS = \sum_{i=1}^{5}[y_i - f(x_i)]^2 = \sum_{i=1}^{5}[yi - (\hat{ß}_1 * x_i + \hat{ß}_0)]^2 = \sum_{i=1}^{5}[y_i - \hat{y}_i]^2 = \sum_{i=1}^{5} e_i^2$$

Residual: e_i = observed$_i$ - predicted$_i$

$\hat{ß}_1$=2.5 and $\hat{ß}_0$ = 5.5 are estimates that minimize the sum of squared deviations



Quantitative Bivariate

Correlation and regression

Connection between correlation r and linear regression

In case of OLS linear regression the connection between r and the estimate for the slope is
slope = correlation(x;y) * standard deviation(y) / standard deviation(x)

$$\hat{ß}_1 = r\frac{s_y}{s_x}$$

In the above example
$s_x = 2.1$; $s_y = 6.5$; $r = 0.8$

$$\hat{ß}_1 = 0.8 * \frac{6.5}{2.1} = 2.5$$

Therefore estimations from y on x is generally not the same as regression from x on y

Connection between correlation r and least-squares regression

$$r^2 = \frac{\text{variance of predicted values } \hat{y}}{\text{variance of original values } y}$$

If the $r^2 = 1$ then all original y values = predicted y values (fell exactly on the line)

The r squared always gives the percentage of explained variation of y by OLS

Example:
in the above example r = 0.8 and $r^2 = 0.64$

$$r^2 = \frac{\text{variance of predicted } \hat{y} = s_{\hat{y}}^2 = 5.2^2 = 27.04}{\text{variance of original values } y = s_y^2 = 6.5^2 = 42.25} = 0.64$$

64% of the variation of y is explained by the least-squares regression of y on x

Caution

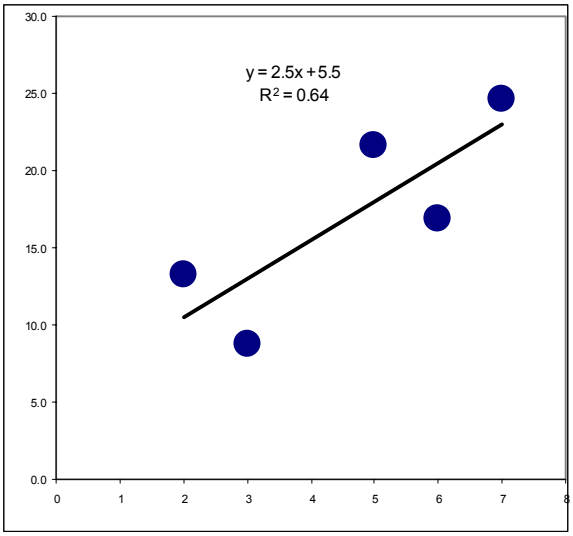
Association does not imply causation

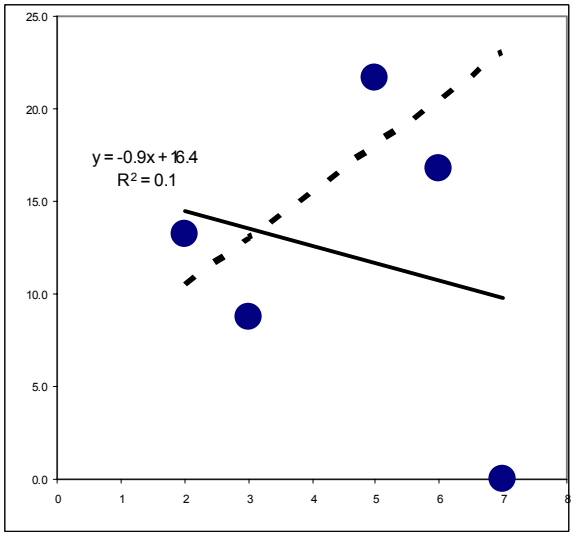
(Moore 2006, p.160)
"An association between an explanatory variable x and a response variable y
even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y."

Quantitative Bivariate

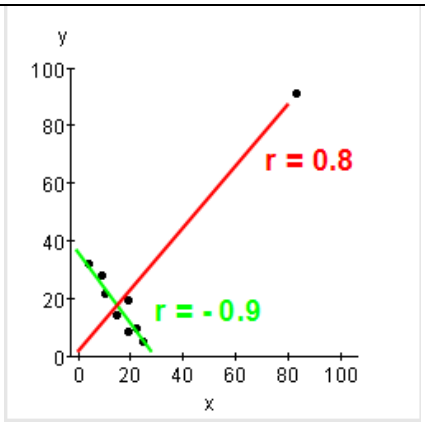
Regression residuals

OLS regression is not resistant

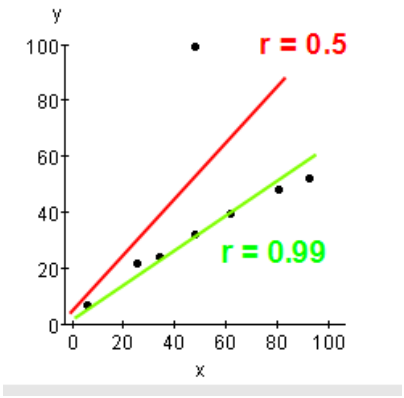




Detect influential points



An example for an influential point



An example for an "outlier"

The outlier in the left graph has an extreme influence in contrast to the outlier in the right graph. Left, the outlier would not produce a large residual while in the right example it would.

A point that is an outlier in x is often influential. Moore (2006, pp. 154-158)

Do not use correlation or regression to assess the agreement between two methods when scale and location matter

van Belle (2008, p. 76-80), Aspden (2005), Bland and Altman(1995,1986), Altman and Bland (1983)

If we have two methods X and Y of measuring the same quantity from which the true value remains unknown, we should not use the correlation between the two methods or a regression to asses the degree of agreement if location and scale matter.

For three reasons:

1. Association is not agreement: If one method is always higher than the other method (e.g. Y=a*X+b) then association would be strong (r=1) but the agreement could be poor.

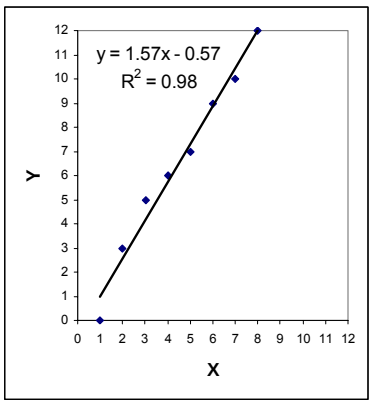
2. Sign of the correlation: The sign of the correlation would only give nontrivial information if it is negative because the two methods are measuring the same thing, so we might expect that one method produce high values when the other does and reverse.

3. Strength of the association: r will depend on the choice of objects.
   a. If the variation between objects (the spread of the axis) is higher than the variation in the measurements the correlation will be high.
   b. If we had many observations from similar objects (e.g. the same) the correlation would be small no matter how close the agreement is. See Altman and Bland (1983, p.308)

Westgard and Hunt (1973): "The correlation coefficient … is of no practical use in the statistical analysis of comparison data" (cited in Altman and Bland, 1983, p. 310)

Example:

X	Y
1	0
2	3
3	5
4	6
5	7
6	9
7	10
8	12



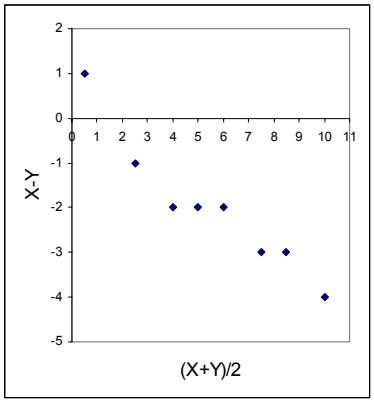
For these data correlation r=0.99 and regression are quite impressing although the agreement is poor.

We should plot the data itself, but we should not fit a regression.

What could de done, is to insert the identity function.

How to do it:

X	Y	(X + Y) / 2	X - Y
1	0	0.5	1
2	3	2.5	-1
3	5	4	-2
4	6	5	-2
5	7	6	-2
6	9	7.5	-3
7	10	8.5	-3
8	12	10	-4



Plot X-Y against (X+Y)/2

(X+Y)/2 is as close as we can get for the "true" value.

Now you can see, that the second method is almost always larger then the first.

And, that it gets worse as "true" values become larger.

Cave:
But don't plot the difference against X or Y, because that could produce another artifact.
Bland and Altman(1995)

Remark:
If you are not aiming for the agreement but for the forecasting (calibration) Y from X then regression would be the appropriate method.

Do not correlate rates or ratios indiscriminately

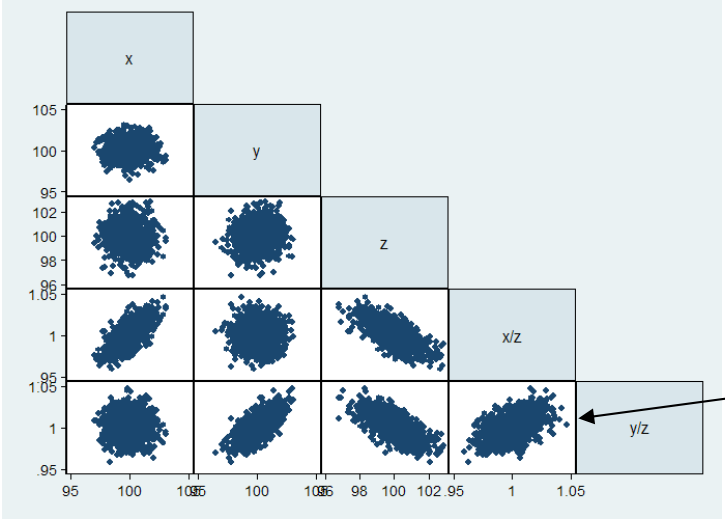
(van Belle, 2008, pp. 70-71)

Theoretical background see Kronmal (1993) and Nevill (1995).

Example (both sides are divided)

If X, Y and Z are mutually independent so that r(X,Y)=0 and r(X,Z)= 0 and r(Y,Z)=0 but r(X/Z; Y/Z) ≠ 0

Simulation: n=1,000; X, Y, Z are mutually independent Normal(100;1)



Pearson correlations

	x	y	z	x/z	y/z
x	1.00				
y	0.04	1.00			
z	-0.01	0.04	1.00		
x/z	0.71	0.00	-0.71	1.00	
y/z	0.03	0.70	-0.70	0.51	1.00

See STATA program in Appendix

Example (one side is divided)

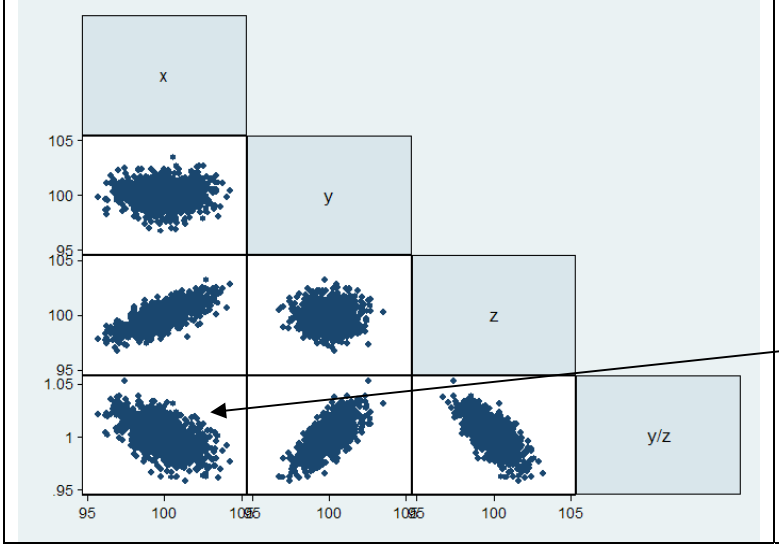
r(X;Y)=0 and r(X;Z) ≠0 then r(Y/Z; X) ≠ 0

This is a quite tricky situation. You would "normalize" **only one side** (Y) by a variable (Z) that is correlated to the second variable (X).

Example: Y=Number of churches Z=Population X=Amount of forest. Let's assume that the number of churches and the amount of forest are not correlated but that population and amount of forest are correlated.

Then Churches per capita and the amount of forest would be correlated while the number churches and the amount of forest is not.

Simulation: n=1,000; Y, Z are independent Normal(100;1) and X = Z + N(0;1)



Pearson correlations

	x	y	z	y/z
x	1.00			
y	0.00	1.00		
z	0.72	-0.01	1.00	
y/z	-0.50	0.70	-0.72	1.00

See STATA program in Appendix

How to handle such problems see van Belle (2008, p71).

References

Suggested textbooks and papers

Moore DS, McCabe GP (2006). *Introduction to the practice of statistics 5. ed.*. Freeman, New York.

Good PI, Hardin JW (2006). *Common Errors in Statistics (and How to Avoid Them)*, Wiley, New York.

van Belle G (2008). *Statistical Rules of Thumb*. Wiley, New York.

Cited in this overview

Aspden RM (2005). Agreement between two experimental measures or between experiment and theory. Journal of Biomechanics, 38, pp. 2136-2137.

Bland JM, Altman DG (1995). Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet, 346, pp. 1085-1087.

Bland JM Altman DG (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet, Volume 327, pp. 307-310.

Altman DG, Bland JM. (1983). Measurement in medicine: the analysis of method comparison studies. Statistician. 32, pp. 307-317.

Westgard JO, Hunt MR (1973). Use and interpretation of common statistical tests in method-comparison studies. Clinical Chemistry, 19, 49-57.

Nevill AM, Holder RL, McShane P and Kronmal RA (1995). Letter to the editors. Journal of the Royal Statistical Society. Series A, 158, 3, pp. 619-625.

Kronmal RA (1993). Spurious correlation and the fallacy of the ratio standard revisited (with discussion). Journal Royal Statistical Society, Series A, 156, 379-392.

Armitage P, Colton T (2005). *Encyclopedia of Biostatistics*.Vol.1-8, Wiley, New York.

Sahai H, Khurshid A (2002). *Dictionary of Statistics*. McGraw-Hill, Boston.

Other Internet Resources

Ender P (2000). Linear Statistical Models: Regression. < http://www.gseis.ucla.edu/courses/ed230bc1/ >.

Wikipedia Statistics Portal < http://en.wikipedia.org/wiki/Statistics >.

Wolfram Math World < http://mathworld.wolfram.com/ >.

Historical references:

Tukey JW (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading.

Bowley AL (1901). *Elements of Statistics*, (4th edition in 1920)[5]. London.

[5] "The *Elements of Statistics* is generally regarded as the first English-language statistics text-book" Wikipedia

Appendix

Program: (X; Y; Z) independent but (X/Z;Y/Z) dependent

```
clear
set obs 1000

gen x=rnormal(100,1)
gen y=rnormal(100,1)
gen z=rnormal(100,1)

gen xz=x/z
label var xz "x/z"

gen yz=y/z
label var yz "y/z"

graph matrix x y z xz yz, half
pwcorr x y z xz yz
```

Program: (X;Y) independent and (X;Z) dependent then (Y/Z;X) dependent

```
clear
set obs 1000

gen y=rnormal(100,1)
gen z=rnormal(100,1)

gen yz=y/z
label var yz "y/z"

gen x=z+rnormal(0,1)

graph matrix x y z yz, half
pwcorr x y z yz
```