DSO 530 Week1 Technical Notes: Review of Basic Statistics II

Recall

- Discussed measures of <u>center</u> of a distribution:
  - e.g., Mean, Median

- Discussed measures of <u>spread</u> of distribution:
  - e.g., SD, IQR

Recall: Variance and SD

- Variance:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}$$

- SD $= s = \sqrt{s^2} = \sqrt{Variance}$

Why Taking *Squares*?

- Sum of deviations (not squared) is just 0.
- Squaring the deviations converts the negative deviations to positive numbers
- So Variance is strictly positive as long as the variable does not take a single value only

Why divide by *n-1* and not *n*?

- It's unimportant if n is large.
- Dividing by n-1 gives an unbiased estimate of variance. (More on this later…)

What happens when $n = 1$??

More practice

100, 100, 100, 100, 100, 100, 100

$s = ?$

90, 90, 90, 100, 110, 110, 110

$s = ?$

IQR and SD(s)

IQR

- Robust measure (same as the median)
- Has the same units as the observations

s

- NOT a robust measure (same as the mean)
- Has the same units as the observations
- s=0 if and only if all the observations are equal

Five Number Summary

The smallest observation, the first quartile, the median, the third quartile, and the largest observation.

Min Q1 M_d Q3 Max

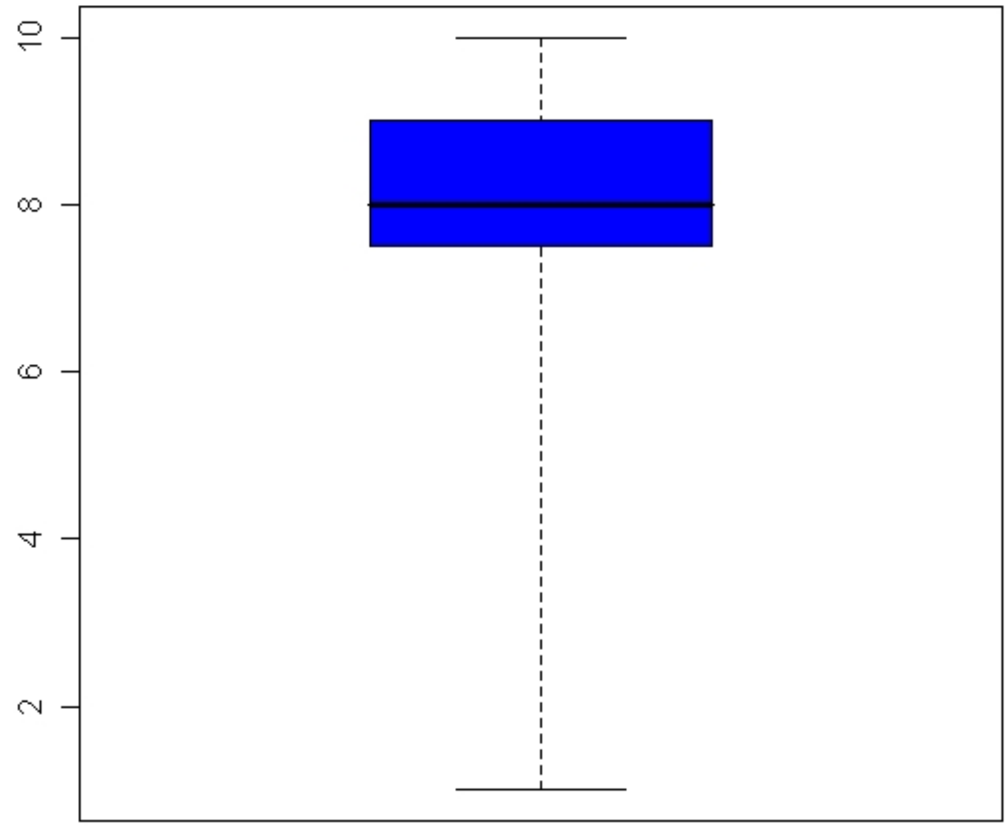
Customer satisfaction data

Minimum	Q1	Median	Q3	Maximum
1	7.5	8	9	10

Can graph this summary using a **boxplot**

Boxplot (simplest form)

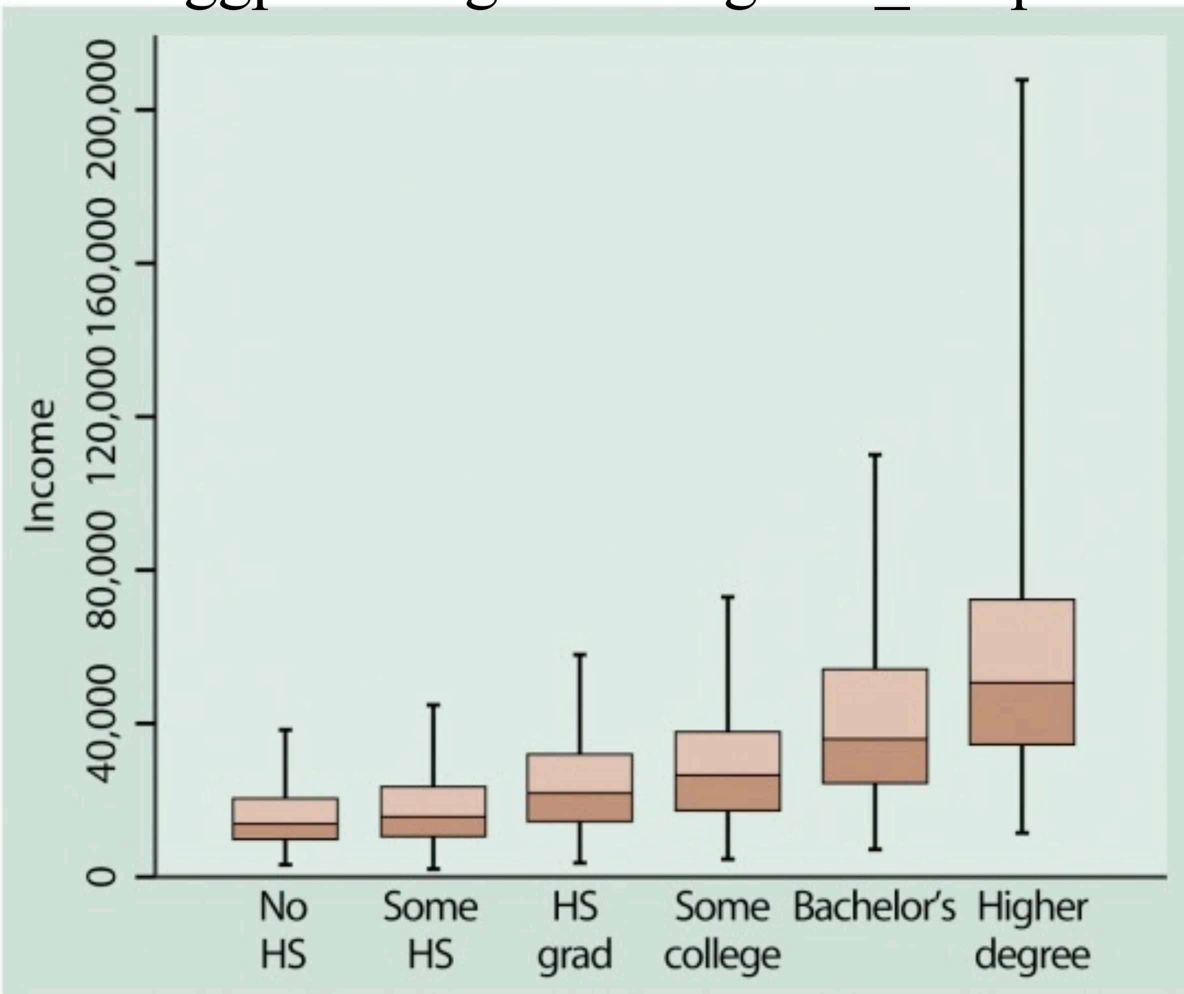
- The central box spans the quartiles Q1 and Q3.
- The line in the box marks the median M.
- The whiskers extend out to the smallest and largest observations



Customer satisfaction data

More on Boxplots

Boxplots are especially good for showing differences between distributions across groups.
(http://docs.ggplot2.org/0.9.3.1/geom_boxplot.html)



Choosing a Summary

- For a skewed distribution or a distribution with strong outliers the five number summary is usually better than mean and SD
- Use SD for the spread when you use the mean for the center

WARNING: Do not use only boxplots and numerical summaries to describe the shape of a distribution. Add a histogram. Why?

- So far we have focused on individual variables

- Now we will study relationship between two variables
  - Two categorical variables
  - Two numerical variables

Two-way (i.e., contingency)tables

… are used to describe the relationship between two <u>categorical</u> variables. The tables contain counts or proportions (percentages)

Two-way tables

E.g. Cross-classification of a sample of 980 Americans by gender and party identification

rows: Party (D,I,R) *columns: Gender (F,M)*

# of female democrats in the sample

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

total # of democrats in the sample

total # of females in the sample

the total sample size

Notation

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

Party = **row variable**

Gender = **column variable**

Each combination of values of the two variables = **cell**

What is the total # of cells in the above table**?**

Joint distribution

A two-way table with proportions (or percentages) describes the ***joint distribution*** of the two variables.

Each cell gives the proportion of the total sample size

	F	M	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

	F	M	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

Marginal distribution

Distribution of a single variable in a two-way table
= ***marginal distribution***

	"Party"		"Gender"
D	45.3%	**F**	58.9%
I	12.2%	**M**	41.1%
R	42.5%		

rows: "**Party**" columns: "**Gender**"

	F	**M**	Total
D	28.5%	16.8%	45.3%
I	7.4%	4.8%	12.2%
R	23.0%	19.5%	42.5%
Total	58.9%	41.1%	100.0%

marginal distribution of "**Party**"

Conditional distribution

distribution of one variable after we condition on (i.e. restrict our attention to) the value of the other variable = ***conditional distribution***

E.g. What is the distribution (in our sample of 980) of party identification conditional on Gender = F **?**

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

Conditional distribution

…What is the distribution (in our sample of 980) of party identification conditional on Gender = F **?**

	F	**M**	Total
D	279	165	444
I	73	47	120
R	225	191	416
Total	577	403	980

D	279/577	**D**	48.4%
I	73/577	**I**	12.6%
R	225/577	**R**	39.0%

More definitions

- **Lurking Variable**: a variable that has an important effect but was overlooked

- **Simpson's Paradox**: a change in the direction of association between two variables when data are separated into groups defined by a third variable
- Berkeley Sex discrimination case: https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html

Another Example

Two-way table:

	Hospital A	Hospital B
Died	300	50
Survived	2700	950

Death status distributions are specified by the % died:

Hospital A: 300/3000 = 10%

Hospital B: 50/1000 = 5%

…continued

	Hospital A	Hospital B
Died	300	50
Survived	2700	950
Died:	10%	5%

Good condition

	Hospital A	Hospital B
Died	5	10
Survived	995	800
Died:	0.5%	1.2%

Bad condition

	Hospital A	Hospital B
Died	295	40
Survived	1705	150
Died:	14.8%	21.1%

Simpson's paradox:

association between two variables has a different direction from the association conditional on a third variable (lurking variable)

What is the lurking variable in our example?

	Hospital A	Hospital B
Died	300	50
Survived	2700	950
Died:	10%	5%

Good condition

	Hospital A	Hospital B
Died	5	10
Survived	995	800
Died:	0.5%	1.2%

Bad condition

	Hospital A	Hospital B
Died	295	40
Survived	1705	150
Died:	14.8%	21.1%

Response and explanatory variables

A **response** (dependent) variable is the variable of interest (measures the outcome of a study)

An **explanatory** (independent) variable explains or causes changes in the response variable

e.g. income and education (in years)

Note: definition works for both numerical and categorical variables

Scatterplot

- Shows the relationship between two numerical variables measured on the same individuals
- The values of one variable -> horizontal axis
- The values of the other variable -> vertical axis
- Each individual appears as a point in the plot
- Explanatory variable (if there is one) -> horizontal axis, Response -> vertical axis

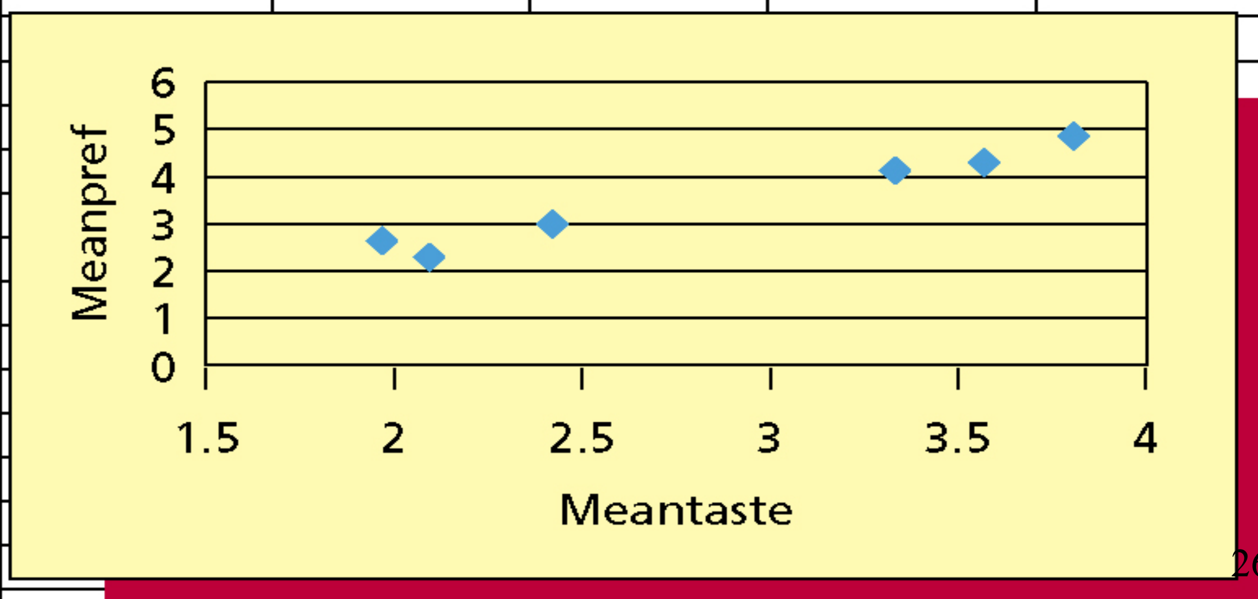
- To add a categorical variable to the scatterplot, you can use a different color or symbol for each category

http://docs.ggplot2.org/current/aes_group_order.html

Example

Restaurant Ratings: Mean Preference vs Mean Taste

	A	B	C	D	E	F
1	Restaurant	Meantaste	Meanconv	Meanfam	Meanprice	Meanpref
2	Borden Burger	3.5659	2.7005	2.5282	2.9372	4.2552
3	Hardee's	3.329	3.3483	2.7345	2.7513	4.0911
4	Burger King	2.4231	2.7377	2.3368	3.0761	3.0052
5	McDonald's	2.0895	1.938	1.4619	2.4884	2.2429
6	Wendy's	1.9661	2.892	2.3376	4.0814	2.5351
7	White Castle	3.8061	3.7242	2.6515	1.708	4.7812
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						



Scatterplot

- look for overall pattern and striking deviations from that pattern

- describe the overall pattern by form, direction, and strength of the relationship

- outlier is an individual value that falls outside the overall pattern of the relationship

Scatterplot

- form: clusters, linear association, etc.
- direction: positive association, negative association
- strength: how close the data points follow the form

- positive association: above-average values of one variable accompany above-average values of the other, and below-average values also tend to occur together

- negative association: above-average values of one variable accompany below-average values of the other and vice versa

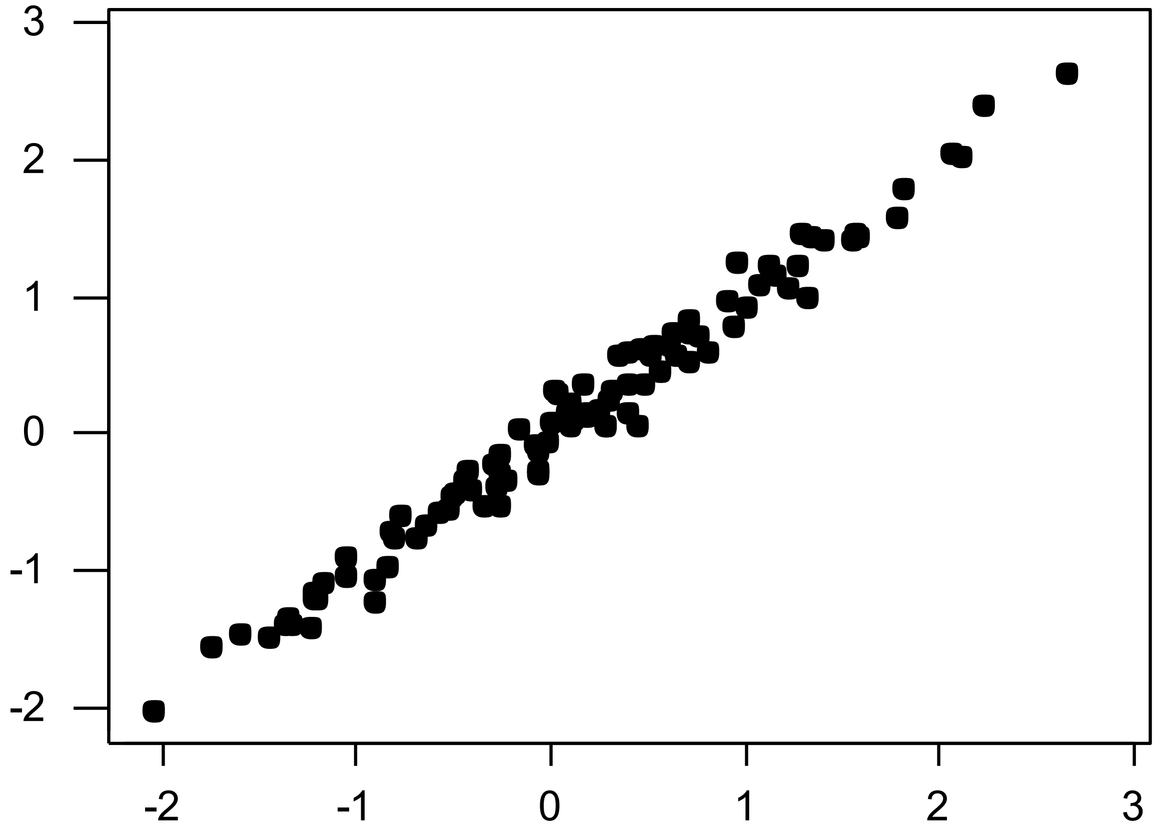
Correlation (*r*)

- measures the direction and strength of the linear relationship between two <u>numerical</u> variables

- is always between –1 and 1

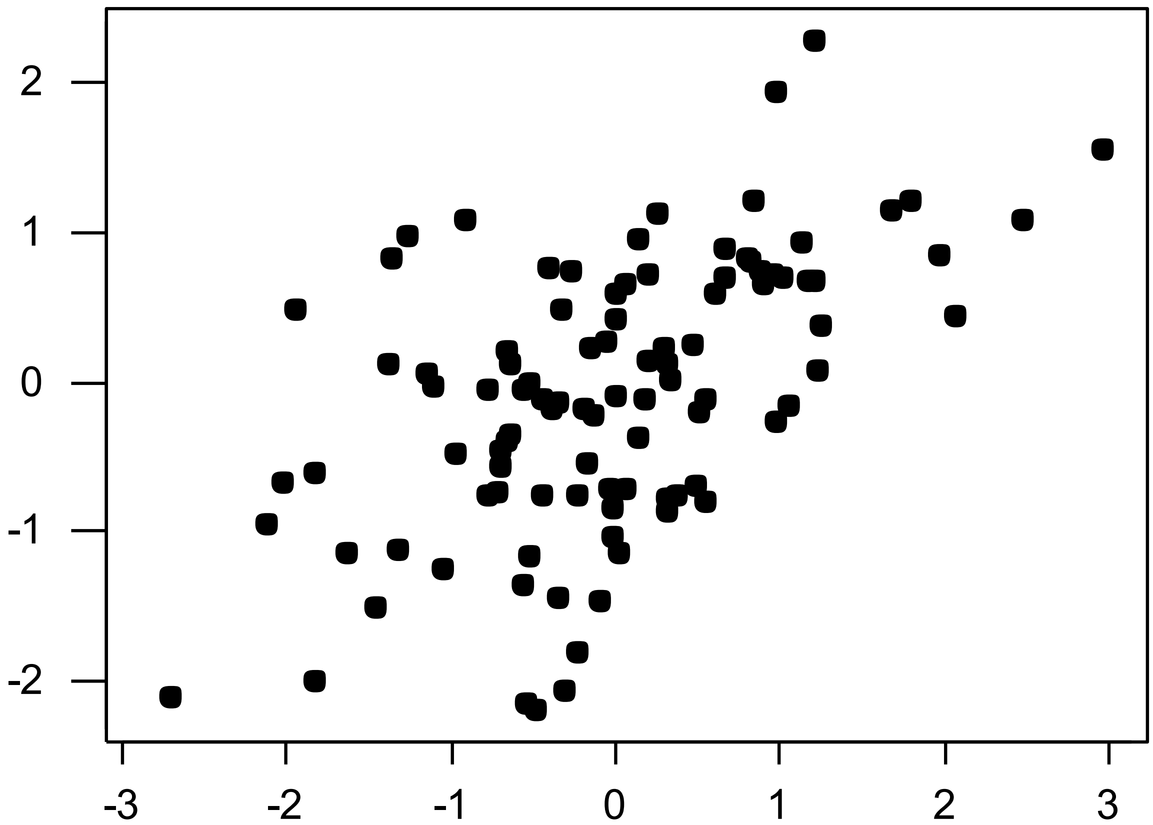
- the strength increases as you move away from 0 to either –1 or 1

Highly correlated variables



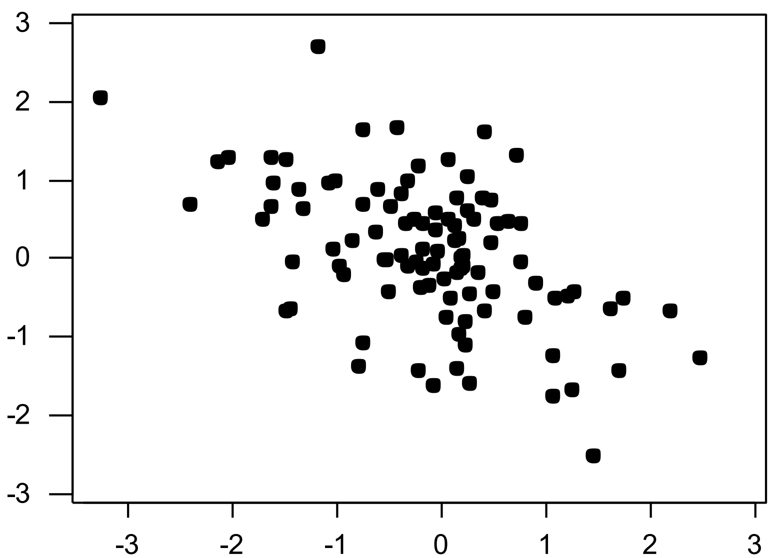
$r = 0.99$

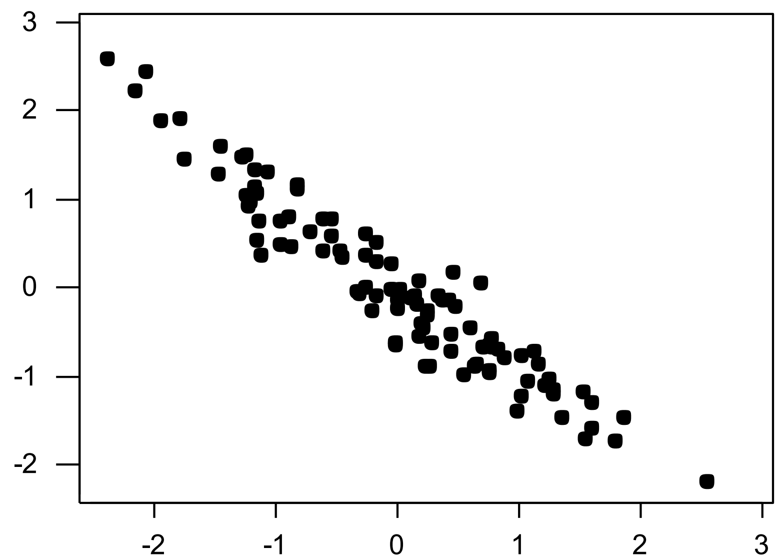
Moderate correlation



$r = 0.55$

Negative correlation

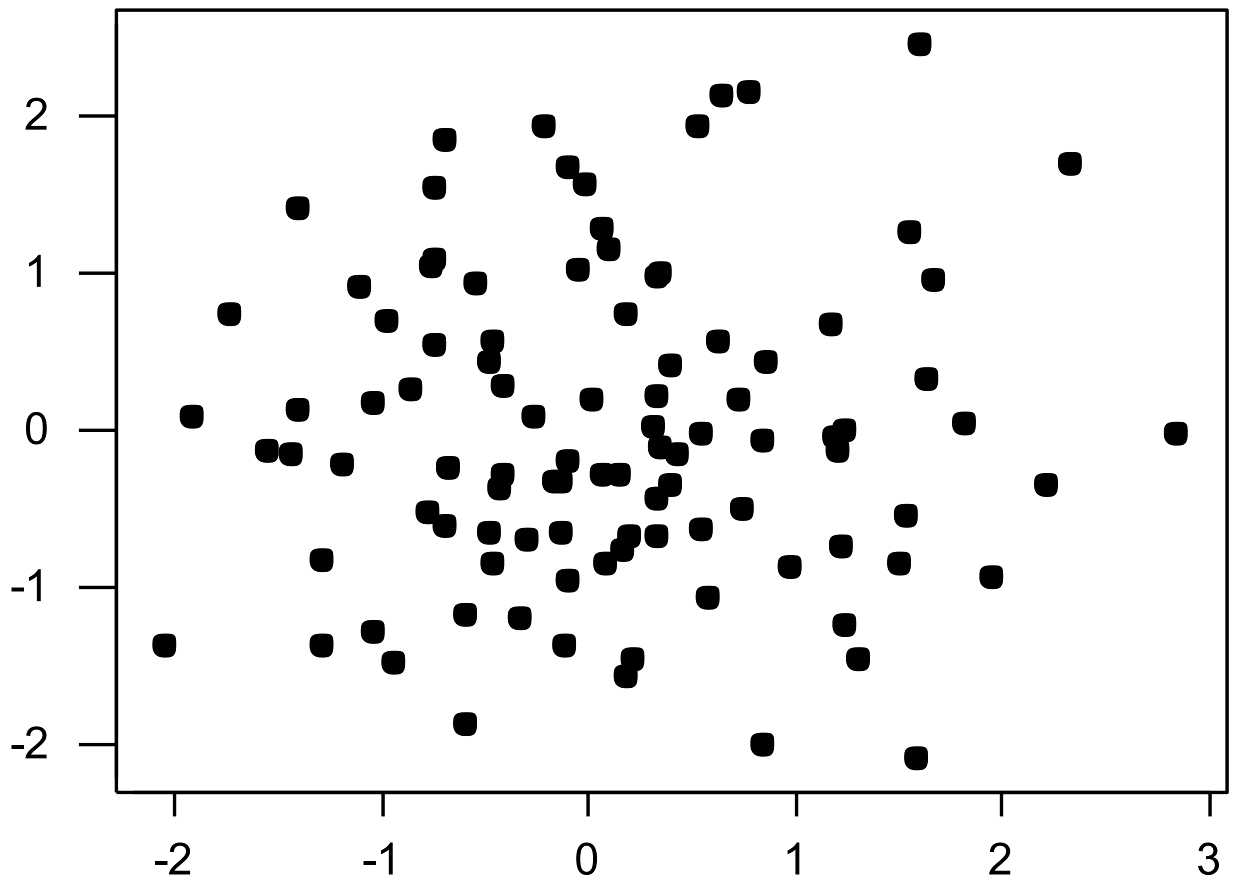




r = -0.52

r = -0.96

Zero correlation



Notes about correlation

- r makes no distinction between explanatory and response variables, it does not matter which variable you call X and which you call Y
- both variables have to be numerical
- r has no units of measurement
- r does not change if you change the units of measurement of the data (e.g. from *lbs* to *kg*)

Notes about correlation

- $r>0$ indicates positive association between the variables, $r<0$ indicates negative association

- extremes $r=-1$ and $r=1$ occur if and only if the points on a scatterplot lie exactly along a straight line

- r measures the strength of only the linear relationship, it does not describe curved relationships

- r is not robust

Example

1971 study: people who drink coffee a lot have higher incidence of bladder cancer

Correlation noticed. Causation?

Example continued

1993: A larger study concluded that after adjusting for the effects of smoking, no evidence was founded for increased risk from coffee.

Correlation does not imply causation