Popisná statistika dvou proměnných

Dominik Heger

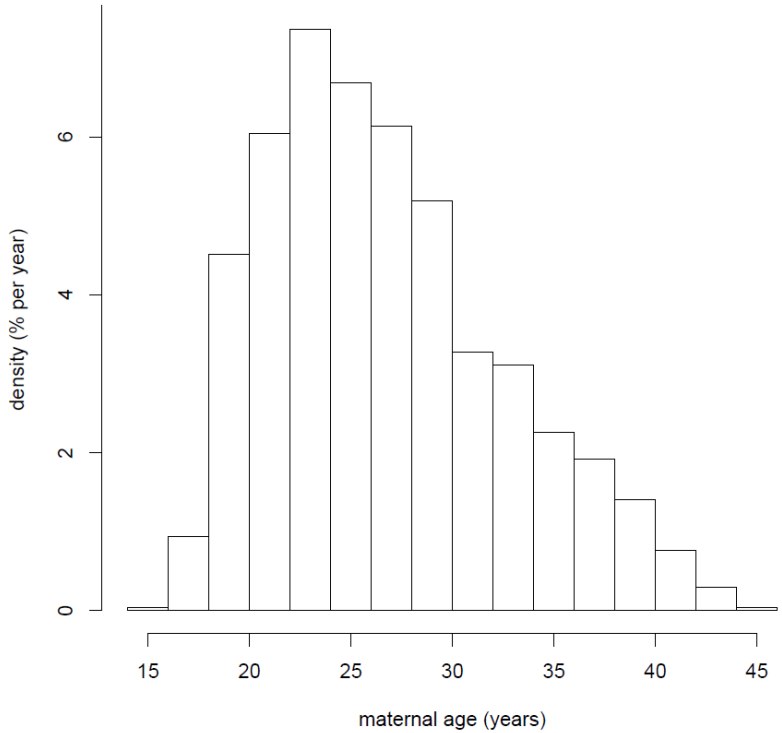
Masaryk University

hegerd@chemi.muni.cz

STDT06 Dvě proměnné

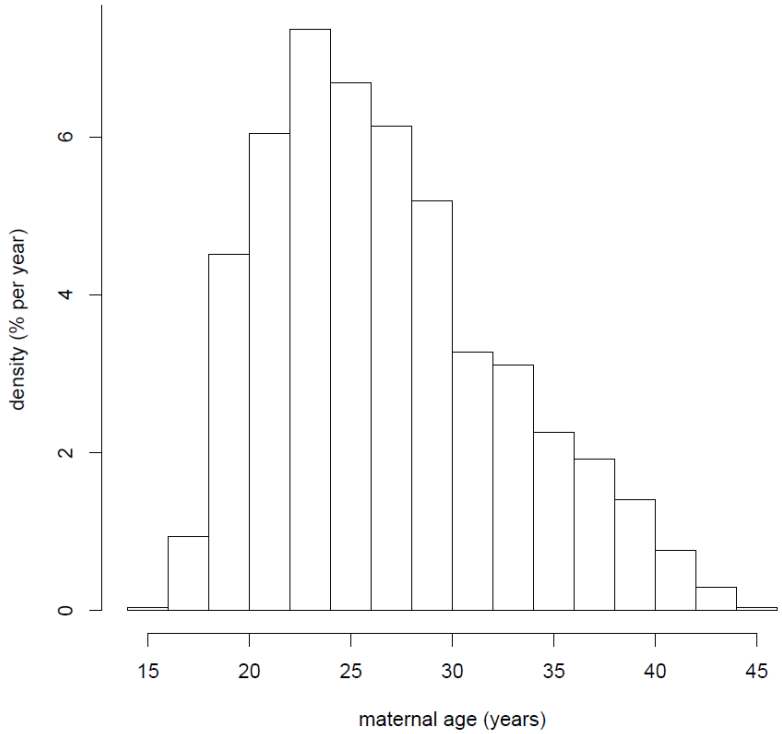
Dava více proměnných

Věk matek

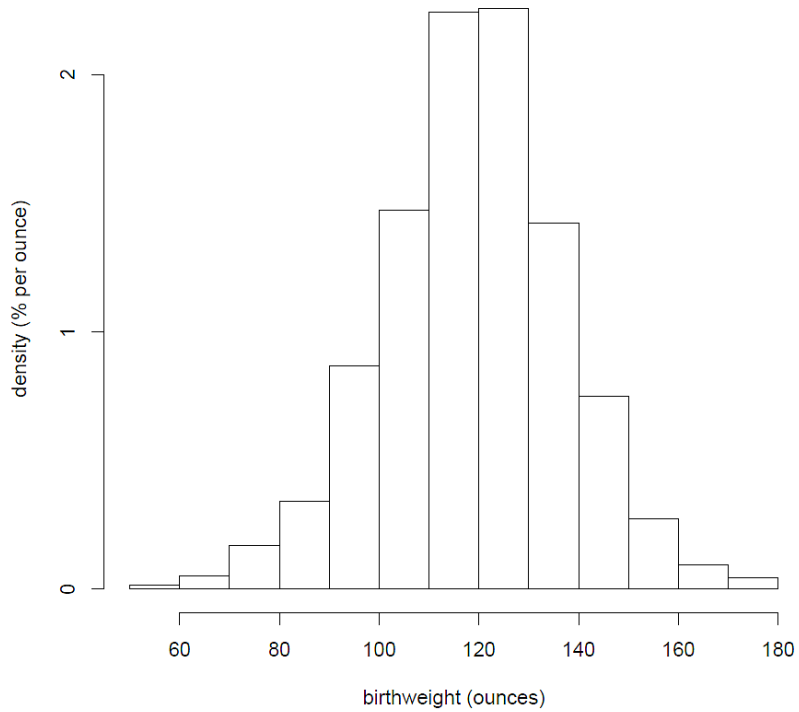


Dava více proměnných

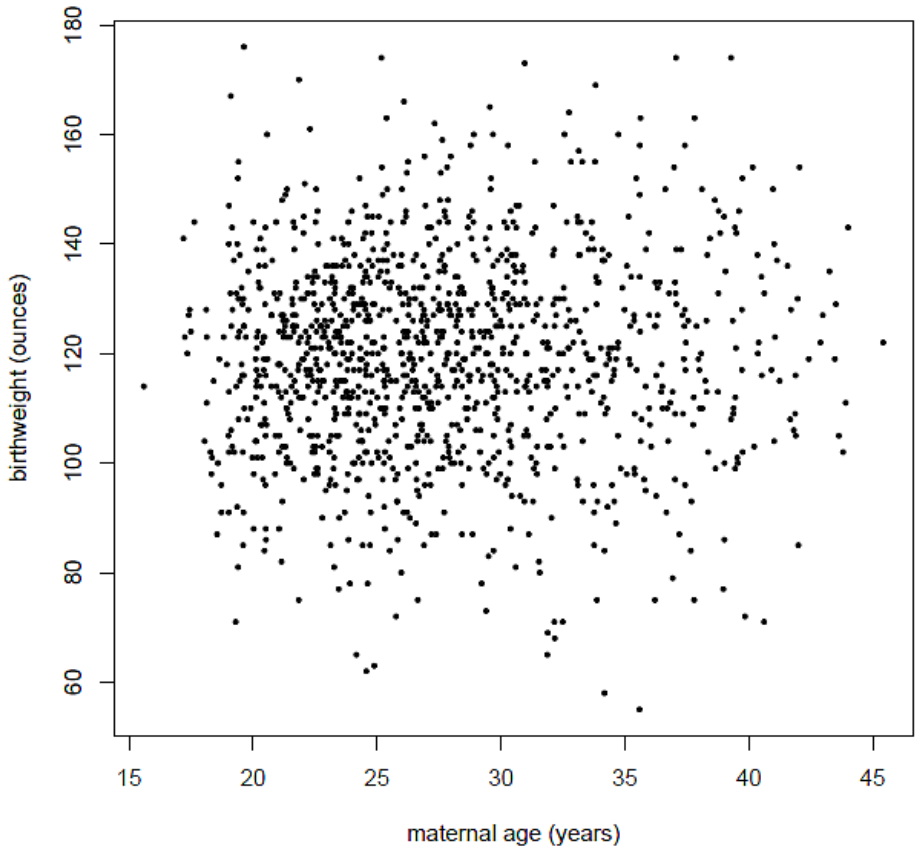
Věk matek



Váha novorozeňat



Data dvou proměnných: rozptylový graf (Scatter Diagram)

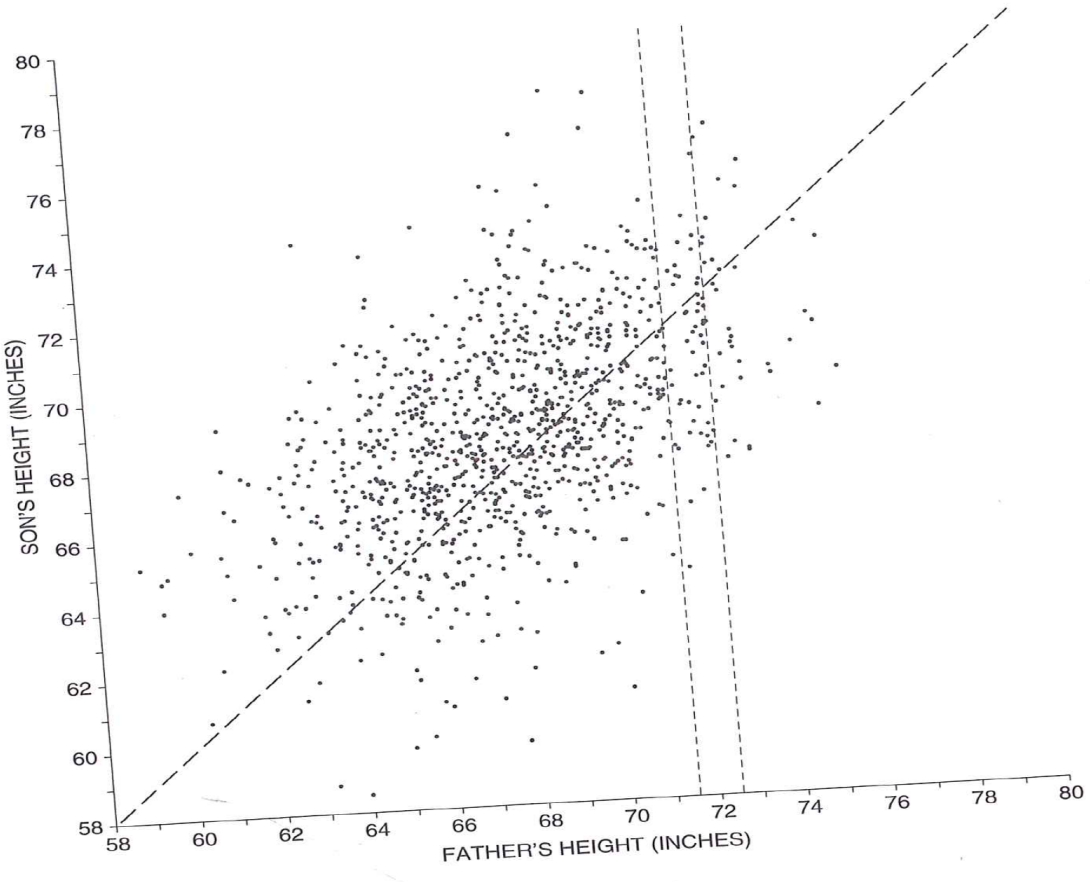


We call such a plot a scatterplot of Y versus X or a scatterplot of Y against X. Scatterplot is one of the best way to study association.

Data dvou proměnných: rozptylový graf

120 CORRELATION [CH. 6]

Figure 1. Scatter diagram for heights of 1,078 fathers and sons. Shows positive association between son's height and father's height. Families where the height of the son equals the height of the father are plotted along the 45-degree line $y = x$. Families where the father is 72 inches tall (to the nearest inch) are plotted in the vertical strip.



1

[1] Freedman, Pisani, Purves: Statistics

Asociace

Asociace

O **asociaci** můžeme hovořit tehdy, když úzkém výřezu na X-ové ose je rozptyl v Y menší než je SDy.

Asociace

O **asociaci** můžeme hovořit tehdy, když úzkém výřezu na X-ové ose je rozptyl v Y menší než je SDy.

Lineární asocice:
rozptylový graf je nakupen okolo přímky.

Asociace

O **asociaci** můžeme hovořit tehdy, když úzkém výřezu na X-ové ose je rozptyl v Y menší než je SDy.

Lineární asocice:
rozptylový graf je nakupen okolo přímky.

Kladná asociace
Nadprůměrné hodnoty jedné proměnné mají tendenci se acosiovat s **nadprůměrnými** hodnotami druhé proměnné; rozptylový graf **roste**.

Asociace

O **asociaci** můžeme hovořit tehdy, když úzkém výřezu na X-ové ose je rozptyl v Y menší než je SDy.

Lineární asocice:
rozptylový graf je nakupen okolo přímky.

Kladná asociace
Nadprůměrné hodnoty jedné proměnné mají tendenci se acosiovat s **nadprůměrnými** hodnotami druhé proměnné; rozptylový graf **roste**.

Záporná asociace
Nadprůměrné hodnoty jedné proměnné mají tendenci se acosiovat s **podprůměrnými** hodnotami druhé proměnné; rozptylový graf **klesá**.

Popis rozptylového grafu

Bod průměrů in the scatter plot is the point with coordinates [mean of X, mean of Y] $= [\bar{X}, \bar{Y}]$.

Popis rozptylového grafu

Bod průměrů in the scatter plot is the point with coordinates [mean of X, mean of Y] $= [\bar{X}, \bar{Y}]$.
The **point of averages** is a measure of the "center"of a scatterplot, quite analogous to the mean as a measure of the center of a list.

Popis rozptylového grafu

Bod průměrů in the scatter plot is the point with coordinates [mean of X, mean of Y] $= [\bar{X}, \bar{Y}]$.
The **point of averages** is a measure of the "center"of a scatterplot, quite analogous to the mean as a measure of the center of a list.

1. Linearity and Nonlinearity

Popis rozptylového grafu

Bod průměrů in the scatter plot is the point with coordinates [mean of X, mean of Y] $= [\bar{X}, \bar{Y}]$.
The **point of averages** is a measure of the "center"of a scatterplot, quite analogous to the mean as a measure of the center of a list.

1. Linearity and Nonlinearity
2. Homoscedasticity and Heteroscedasticity

Popis rozptylového grafu

Bod průměrů in the scatter plot is the point with coordinates [mean of X, mean of Y] $= [\bar{X}, \bar{Y}]$.
The **point of averages** is a measure of the "center"of a scatterplot, quite analogous to the mean as a measure of the center of a list.

1. Linearity and Nonlinearity
2. Homoscedasticity and Heteroscedasticity
3. Outlier

Popis rozptylového grafu

Bod průměrů in the scatter plot is the point with coordinates [mean of X, mean of Y] = $[\bar{X}, \bar{Y}]$.
The **point of averages** is a measure of the "center"of a scatterplot, quite analogous to the mean as a measure of the center of a list.

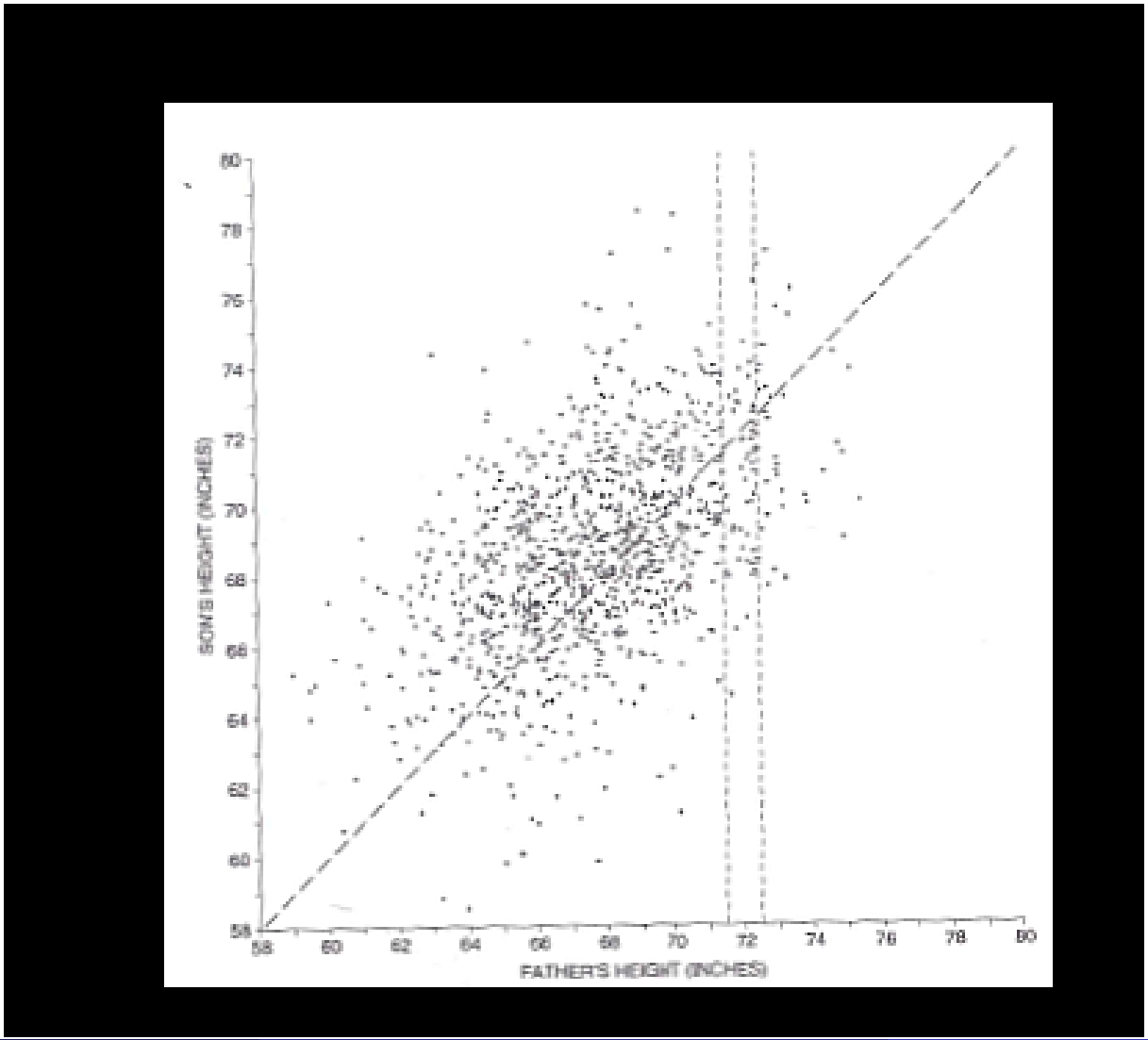
1. Linearity and Nonlinearity
2. Homoscedasticity and Heteroscedasticity
3. Outlier

If a scatterplot shows linear association (or no association), homoscedasticity, and no outliers, it is said to be football-shaped (**bivariant normal**).

Look on scatter diagram - see if there is a association, if it is linear and if there are outliers.

SD Line

SD Line goes through the point of averages and the points which are equal number of SDs away from average for both variables.

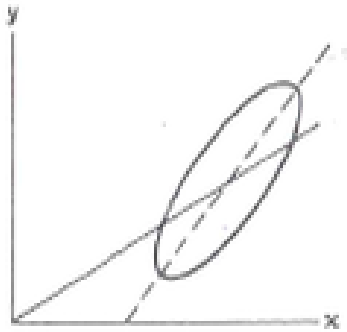


SD Line Exercise

Exercise Set C

1. True or false:
   (a) The SD line always goes through the point of averages.
   (b) The SD line always goes through the point (0, 0).

2. For the scatter diagram shown below, say whether it is the solid line or the dashed line which is the SD line.



3. One study on male college students found their average height to be 69 inches, with an SD of 3 inches. Their average weight was 140 pounds, with an SD of 20 pounds. And the correlation was 0.60. If one of these people is 72 inches tall, how heavy would he have to be to fall on the SD line?

Association

Association

There is a **association** if in the slice of X the scatter of Y is smaller than SDy.

Association

There is a **association** if in the slice of X the scatter of Y is smaller than SDy.

Positive association:
The individuals with larger than average values of one variable tend to have larger than average value of the other and individuals with smaller then average values of X tend to have smaller then average values of Y.

Optically examine if there is an association. If yes - is it linear? If yes - talk about correlation.

Korelace (je podmnožinou) $\subseteq$ asociace.
Correlation (is subset of, is included in) $\subseteq$ association.
Association (is superset of or includes) $\supseteq$ correlation.

Post Hoc Ergo Propter Hoc fallacy

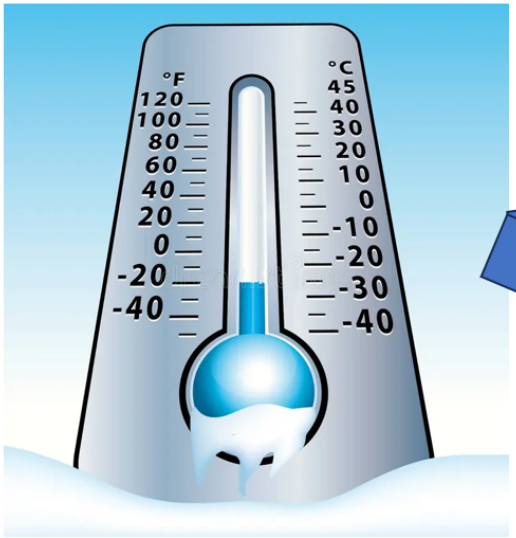
After this, therefore because of this.
Association between two variables is often used as evidence that there is a causal relationship between variables - **erroneously**.
NOT Truth = Fallacy:
If two things are associated, there is some causal relationship between them. One causes the other.

- Jeníček, Pepíček, Mařenka.
- Readibility and shoe size have positive association.
- Money spend on healthcare and life expectancy have negative association.
- Waxing of the car and its maximum speed have positive association.

The variables are related in some way, but that does not mean that one causes the other.
Association is not causation! What are the confounding factors?

Temperature is a Confounding Variable

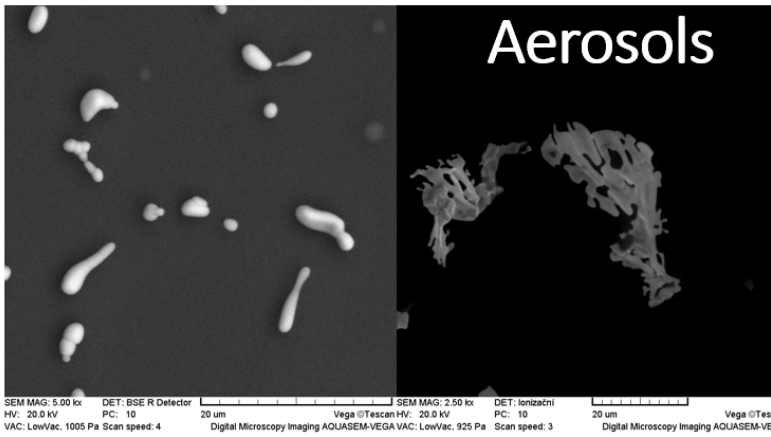
Low Temperature



Causation

Causation

Correlation



Correlation does not Imply Causality!

Post hoc, ergo propter hoc. (*lat.* „After this, therefore because of this.“)

https://www.japan.travel/en/sports/snow/snow-travel/lake-akan-frost-flowers/

How to calculate SOMETHING that would tell us how much linearly related are the data?

What can we use to get such a SOMETHING?

Correlation coefficient (r)

quantifies the linear association:

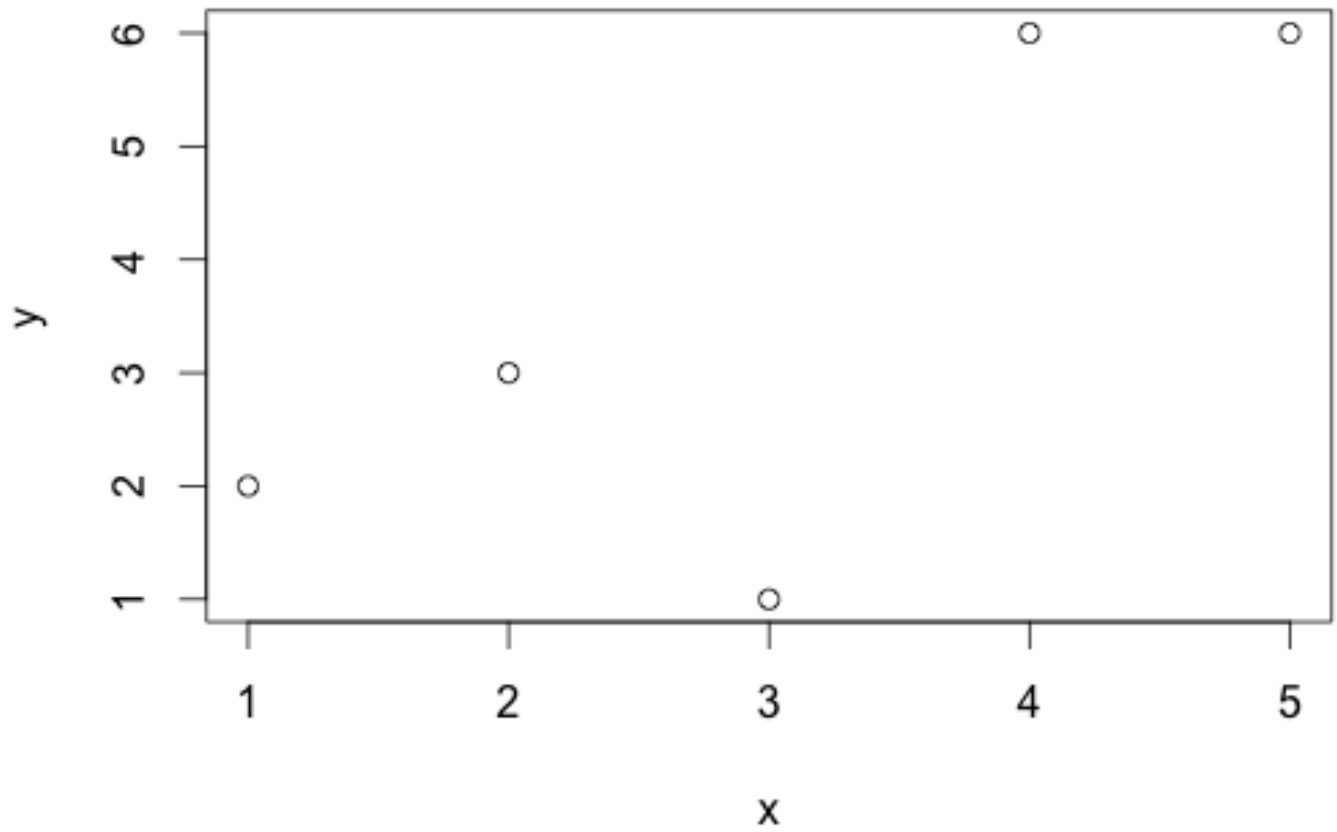
- Its sign tells us whether the scatterplot tilts up or down.
- Its magnitude tells us how tightly the data clusters around a straight line.
- If the points in a scatterplot of Y versus X fall on a horizontal line, r_{xy} is not defined.
- Correlation coefficient of X and Y is the average of the product of X and Y in standard units.

relation $\supseteq$ correlation

r has sense when association is:
linear, homoscedastic, without outliers.

Football-shaped scatterplot can be summarized with 5 numbers: **mean of X, mean of Y, SD of X, SD of Y and R (correlation coefficient)**

How to calculate correlation coefficient (r)?



How to calculate correlation coefficient (r)?

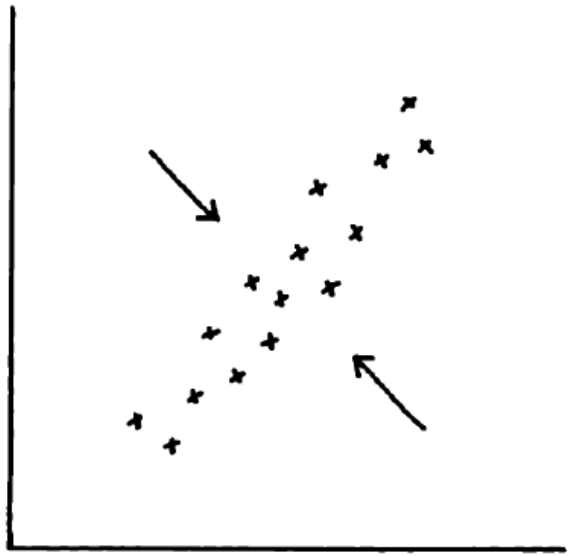
$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mu_y}{\sigma_y}\right) \times \left(\frac{x_i - \mu_x}{\sigma_x}\right)$$

How to calculate correlation coefficient (r)?

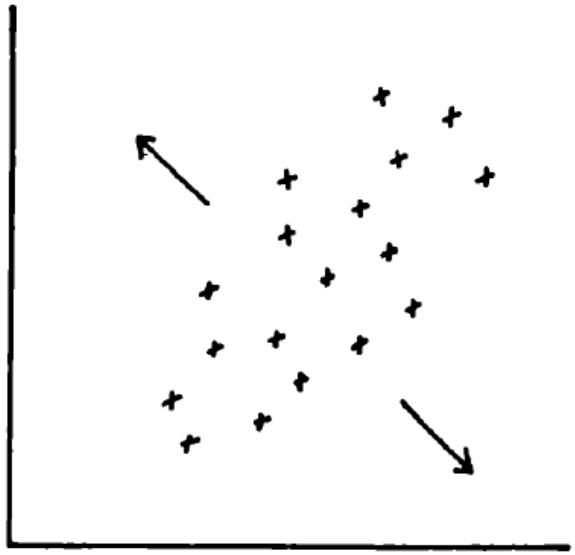
$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \mu_y}{\sigma_y}\right) \times \left(\frac{x_i - \mu_x}{\sigma_x}\right)$$

Figure 5. Summarizing a scatter diagram. The correlation coefficient measures clustering around a line.

(a) Correlation near 1 means tight clustering.

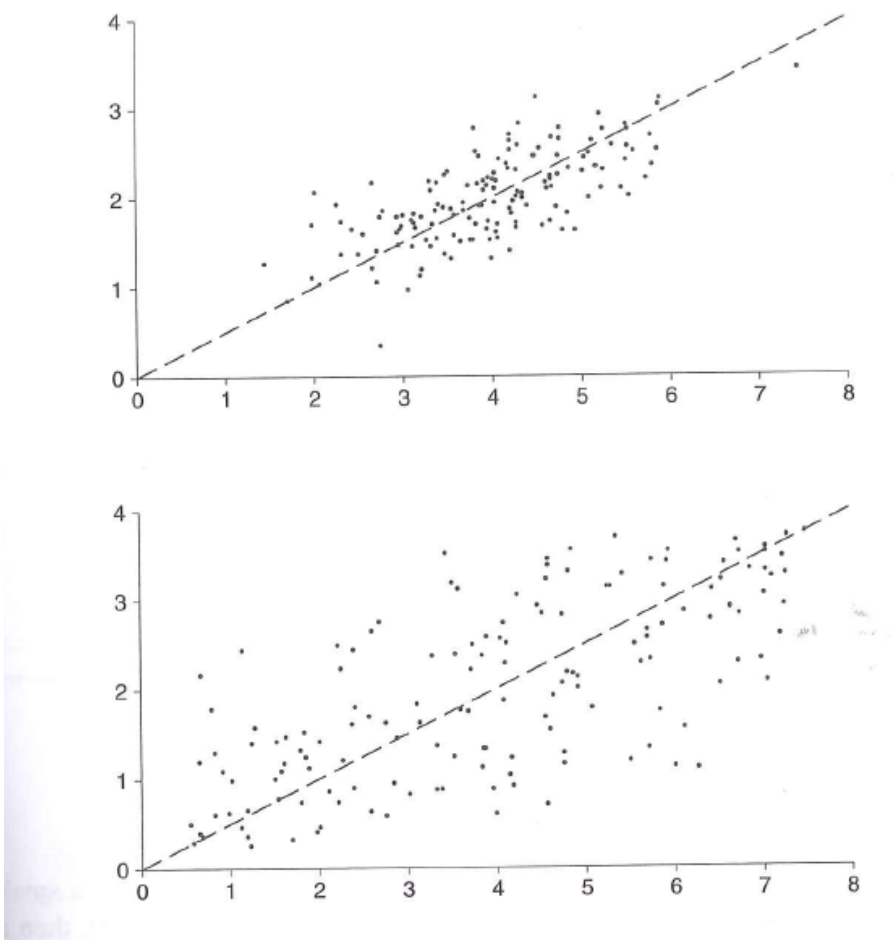


(b) Correlation near 0 means loose clustering.



Example of $r = 0.70$

Figure 3. The effect of changing SDs. The two scatter diagrams have the same correlation coefficient of 0.70. The top diagram looks more tightly clustered around the SD line because its SDs are smaller.



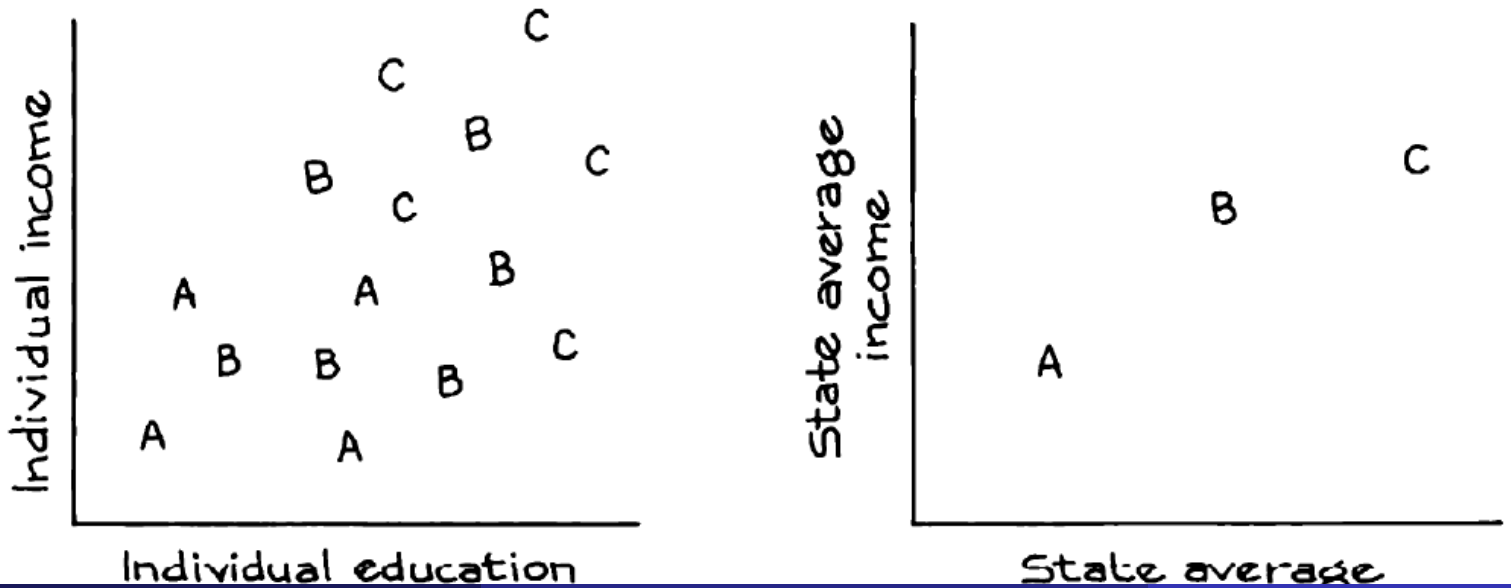
Ecological correlations

are correlation coefficients of averages across groups of individuals, rather than correlation coefficients for individuals.

Always use original data for correlation, NOT averages.

Most of the variability was taken away by averaging.

Beware arguments about association that rely on ecological correlations.



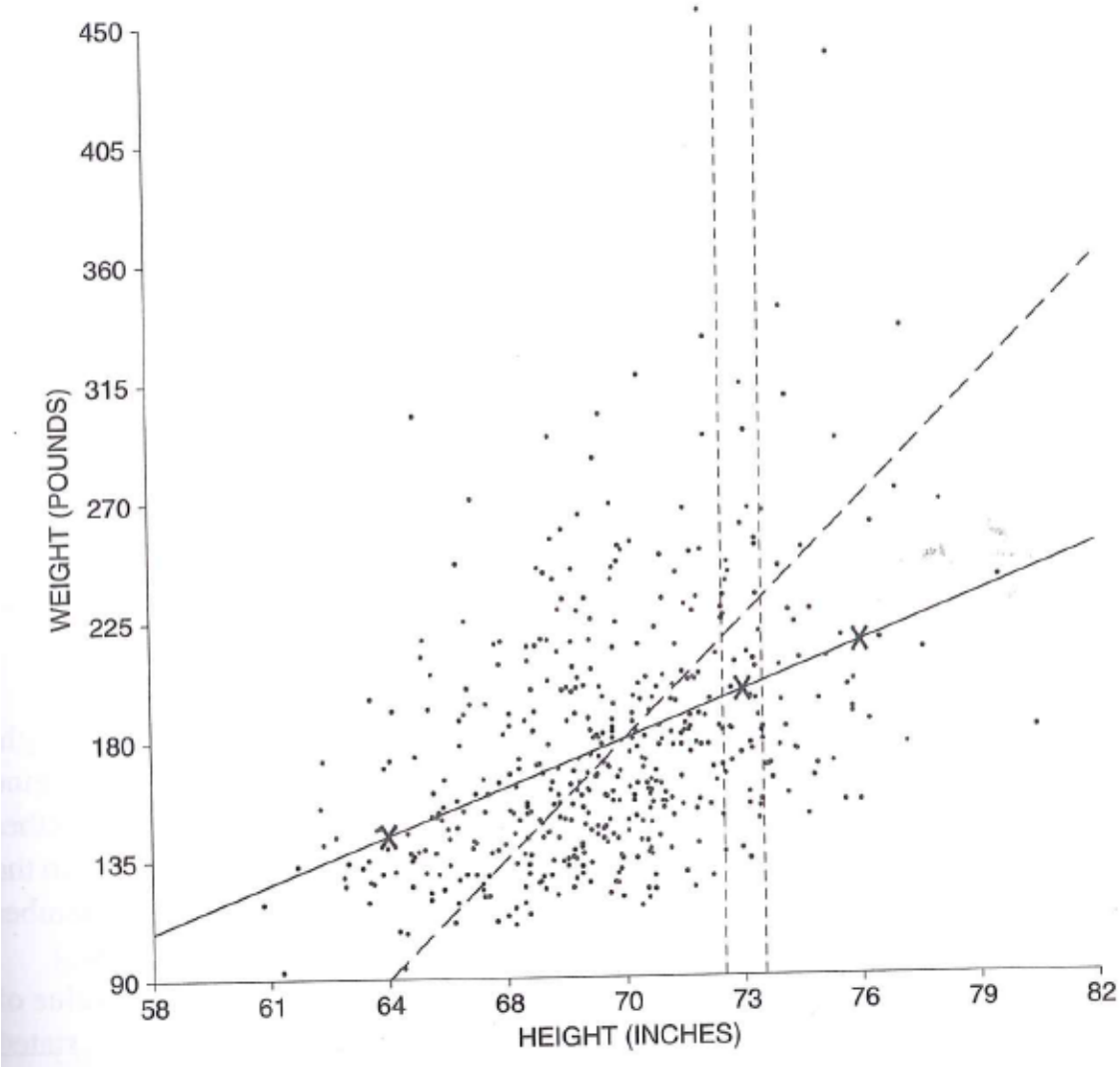
Correlation

Correlation is a tool of descriptive statistics. It is often confused with prediction and even with causal inference as such.

Infromation sources

https://courses.edx.org/courses/BerkeleyX/Stat_2.1x/
http://www.stat.berkeley.edu/ stark/SticiGui/
David Freedman, Robert Pisani, Foger Purves:Statistics

Regression


450
405
360
315
270
225
180
135
90
WEIGHT (POUNDS)
58
61
64
68
70
73
76
79
82
HEIGHT (INCHES)