Relations Between Two Variables

Ľubica and Ján Krauskovi, Dominik Heger

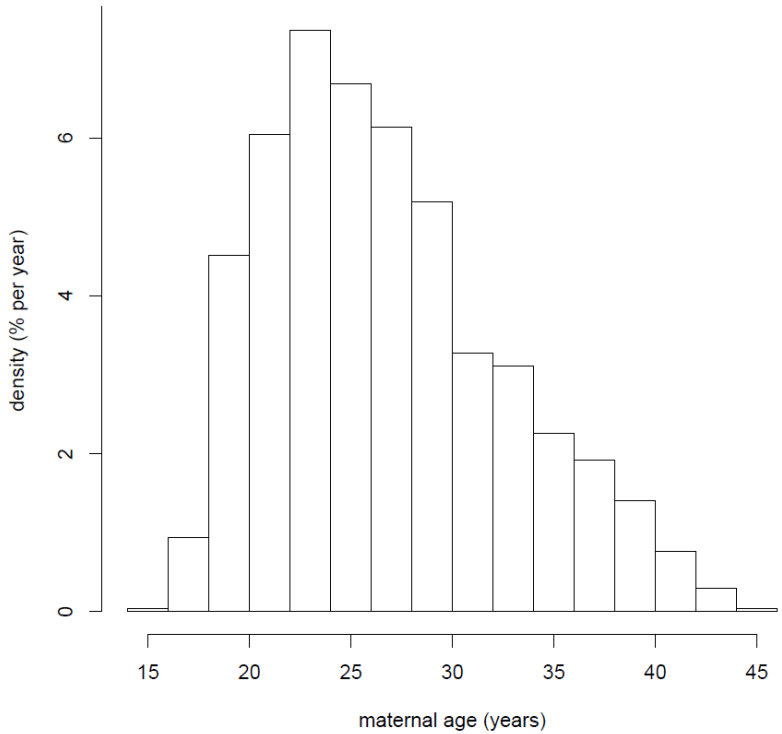
Masaryk University

hegerd@chemi.muni.cz

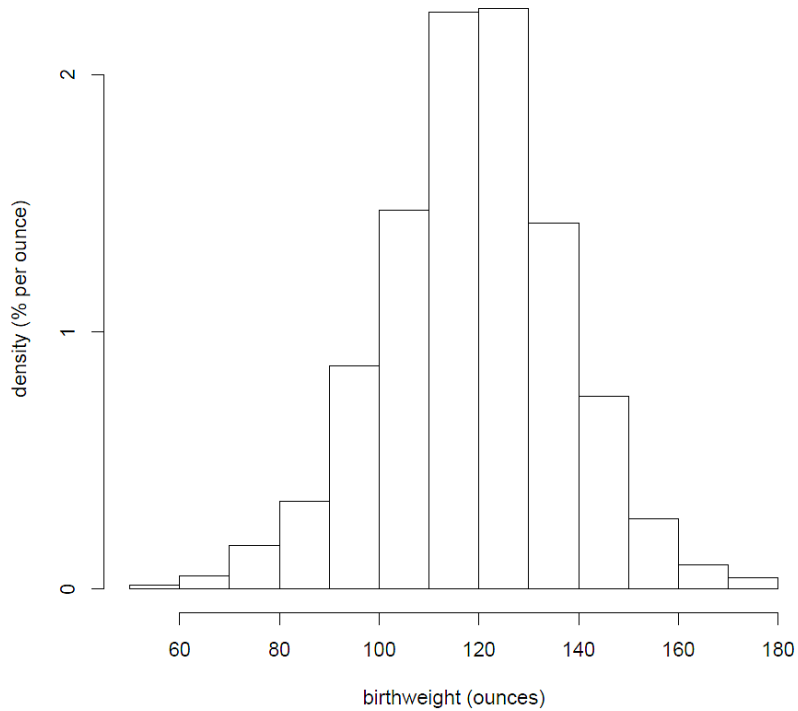
STDT06 Rel Betw 2 Variables

Multivariat Data

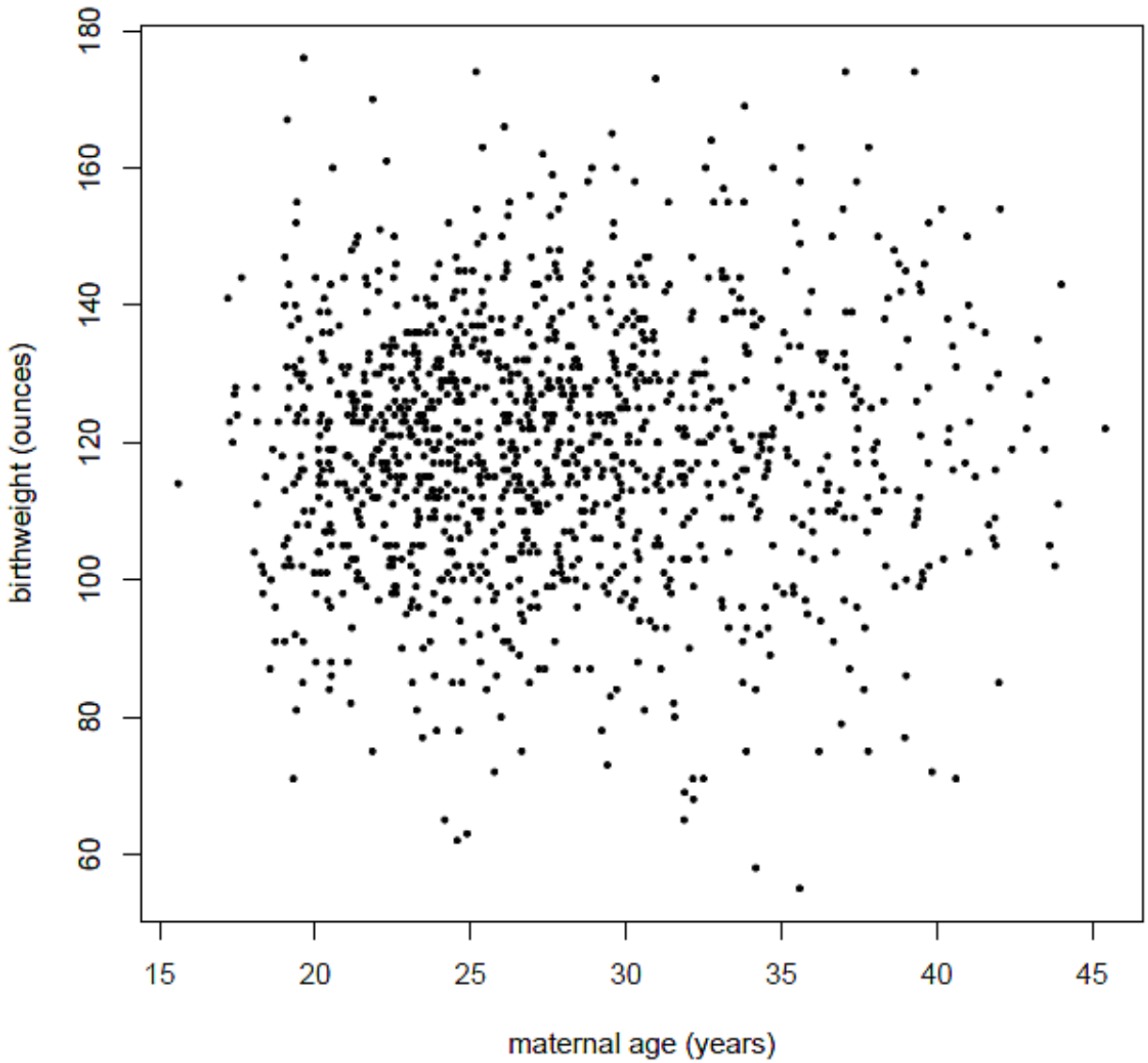
Maternal ages



Birthweights of their babies



Bivariate Data: Scatter Diagram



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.

Linear association:
roughly, the scatter diagram is clustered around a straight line.

Association

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

Linear association:
roughly, the scatter diagram is clustered around a straight line.

Positive association
Above average values of one variable tend to go with **above average** values of the other; the scatterplot **slopes up**.

Association

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

Linear association:
roughly, the scatter diagram is clustered around a straight line.

Positive association
Above average values of one variable tend to go with **above average** values of the other; the scatterplot **slopes up**.

Negative association
Above average values of one variable tend to go with **below average** values of the other; the scatterplot **slopes down**.

Describing Scatterplots

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

Describing Scatterplots

Point of averages 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.

Describing Scatterplots

Point of averages 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

Describing Scatterplots

Point of averages 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

Describing Scatterplots

Point of averages 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

Describing Scatterplots

Point of averages 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.

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.
- Ratio of a Na^+/K^+ is the same as a size of the moons of ...

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

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)

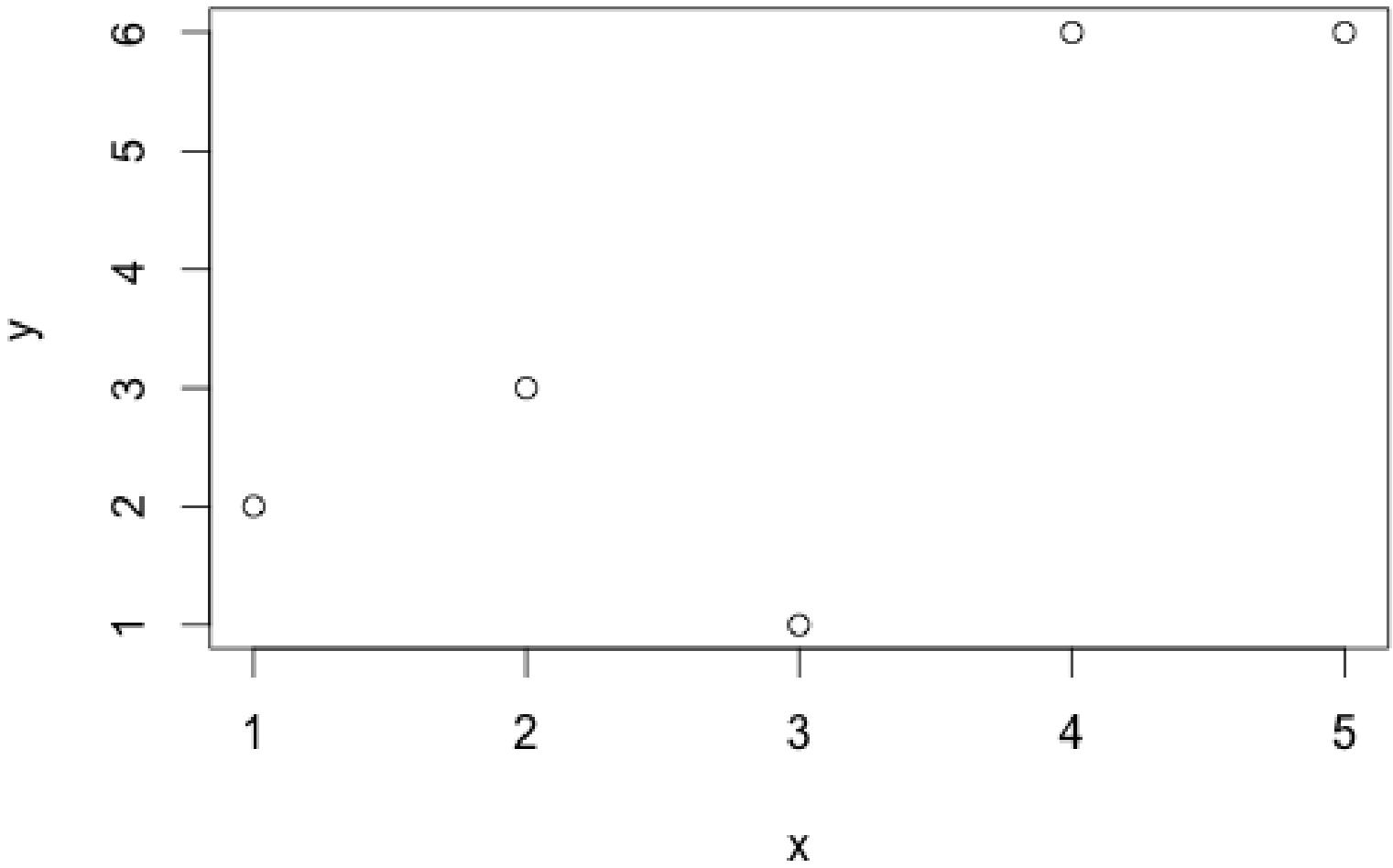
quantifies the linear association:

- 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)?



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/