Probability and Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

May 7, 2019

Notes

Topics

Notes

Objectives

Objectives:

- Produce and interpret a scatterplot of bivariate numerical data.
- Compute and interpret the sample correlation between two numerical variables.

Notes

Scatterplots (12.1)

- ***Bivariate data*** consist of observations of **two variables** on each individual.
- Bivariate data can be used to investigate the **relationship** between the variables.
- The two variables usually play different roles:

  One of them, the ***explanatory*** (or ***independent***) variable, ”explains” variation in the other, which is called the ***response*** (or ***dependent***) variable.

Notes

Notes

Example

Consider a study of the **relationship** between **time spent studying** for an exam and the **exam score**.

The **explanatory variable** is **time spent studying** and the **response** is the **exam score**.

Notes

Example

Here are **lengths** (cm) and **weights** (g) of $n = 9$ female snakes.

Snake	Length	Weight
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145

Notes

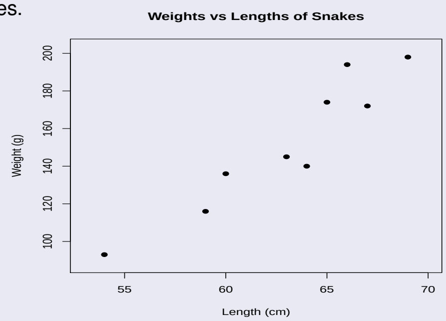
The **explanatory variable** is **length** and the **response** is **weight**.

Notes

- ***Scatterplot of Bivariate Numerical Data***: Shows the **relationship** between the two variables.
  - Plot each **bivariate observation** as a point, with the explanatory variable as the x-coordinate and the response as the y-coordinate.

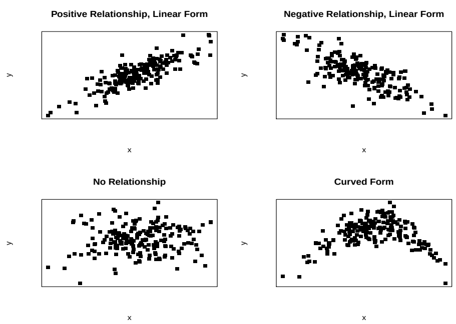
Example (Cont'd)

Here's the **scatterplot** of the **lengths** and **weights** of the snakes.

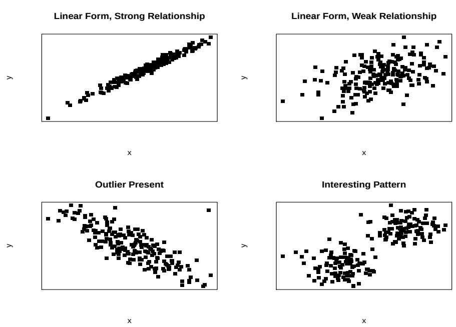


Notes

The figures below illustrate some common **scatterplot patterns**.



Notes



Notes

- **Terminology** used to describe **scatterplot patterns**:
  1. ***Form*** of the pattern (i.e. is it ***linear***, ***curved***, etc.?).
  2. The ***direction*** of the relationship between the two variables:
     - ***Positive***: Y tends to be *large* when X is *large* and *small* when X is *small* (the points in the plot slope upward to the right).
     - ***Negative***: Y tends to be *small* when X is *large* and *large* when X is *small* (the points in the plot slope downward to the right).
  3. The ***strength*** of the relationship (i.e. how distinct is the pattern?)
  4. ***Outliers*** or other ***interesting features***.

Notes

Correlation (12.5)

- **Notation** for a data set of n bivariate observations:

Observation	Explanatory Variable	Response Variable
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

Notes

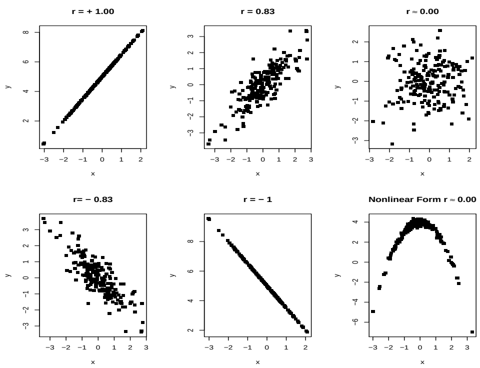
- When two variables exhibit (approximately) a **linear relationship**, we summarize that relationship by the ***sample correlation***, denoted r, defined as follows.

Correlation: The correlation between two variables

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of the x_i's and y_i's, respectively, and s_x and s_y are their sample standard deviations.

Notes



Notes

- The following **properties of the correlation** r help us interpret its value:
  1. The value of r will always lie **between -1.0 and 1.0**.
  2. The **sign** of r tells us the **direction** of the relationship between X and Y:
     - Positive r values indicate a **positive** relationship.
     - Negative r values indicate a **negative** relationship.

Notes

3. The value of r also tells us how **strong** the relationship between X and Y is:
   - r values **near zero** imply a very **weak** relationship or none at all.
   - r values **close to -1.0** or **1.0** imply a very **strong** linear relationship.
   - The extreme values $r = -1.0$ and $r = 1.0$ occur only when there's a **perfect linear** relationship.
4. The value of r doesn't depend on which variable is labeled X and which is labeled Y.

Notes

5. r has no units of measure (e.g. it's not measured in inches or pounds or dollars, even if the data are measured such units).
6. The value of the r is **unaffected** by a (linear) **change of measurement scale** of either X or Y (e.g. converting from Celsius to Fahrenheit).
7. r only measures the strength of the **linear relationship** between X and Y. In particular, curved relationships often have r near zero.
8. r is **not resistant** to outliers.

Notes

Example (Cont'd)

Here are the summary statistics for the **lengths** and **weights** of the $n = 9$ snakes:

	Lengths	Weights
Mean	$\bar{x} = 63.00$	$\bar{y} = 152.00$
Standard Deviation	$s_x = 4.64$	$s_y = 35.34$

Compute the correlation between length and weight and interpret the result.

Notes

The **correlation** between **length** and **weight** is

$$
\begin{aligned}
r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \\
&= \frac{1}{9-1}\left[\left(\frac{60-63}{4.64}\right)\left(\frac{136-152}{35.34}\right)+\left(\frac{69-63}{4.64}\right)\left(\frac{198-152}{35.34}\right)\right. \\
&\quad \left. + \cdots + \left(\frac{63-63}{4.64}\right)\left(\frac{145-152}{35.34}\right)\right] \\
&= \mathbf{0.944},
\end{aligned}
$$

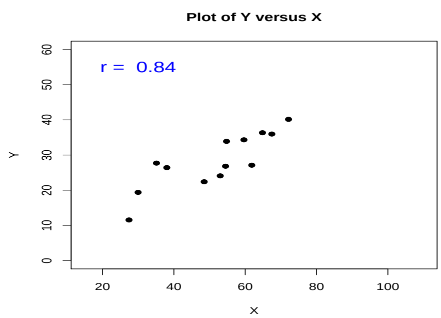
which is consistent with the **strong**, **positive** linear relationship seen in the scatterplot.

Notes

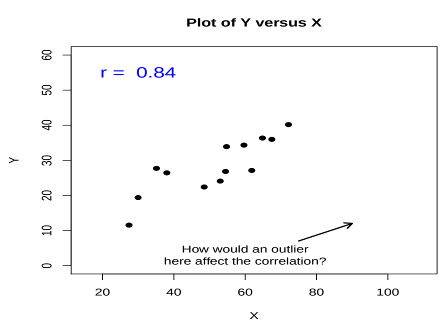
- The next plots show that the **correlation** r is **not resistant** to outliers.

  The location of the outlier in the scatterplot, relative to the rest of the data, determines the affect that the outlier has on the correlation.

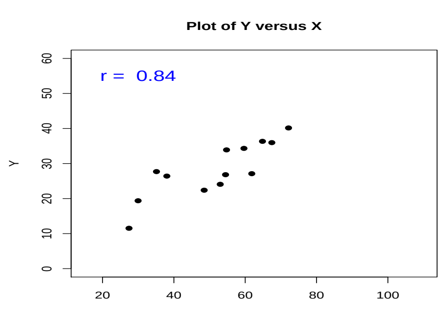
Notes



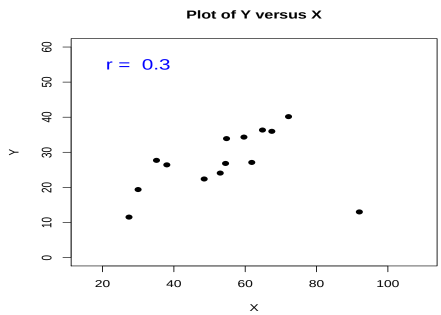
Notes



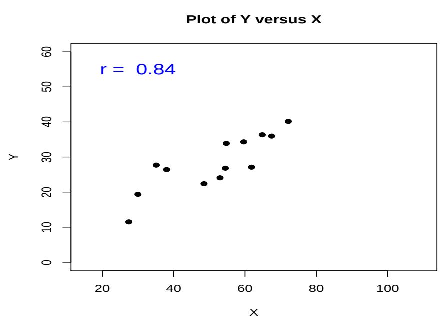
Notes



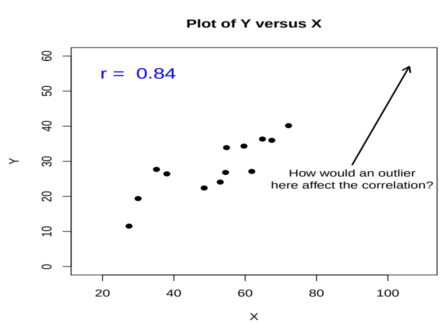
Notes



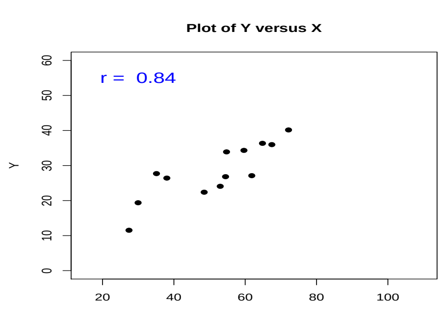
Notes



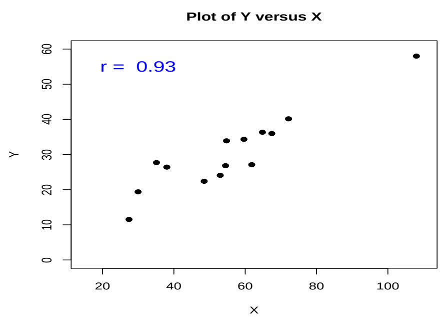
Notes



Notes



Notes



Notes

- A **correlation** between two variables, even if it's very strong, **doesn't imply a cause-and-effect relationship**.

  The relationship might instead be the result of one or more **confounding variables** "lurking" in the background (i.e. not measured).

  A ***confounding variable*** is a variable that's related to both X and Y. As the confounding variable changes, X and Y simultaneously change.

- The next two examples illustrate the notion of **confounding variables**.

Notes

Example

Data on the **number of TV sets per capita** and the **average life expectancy** for each of the world's nations shows a **strong positive correlation** between these two variables – nations with more TV sets have longer life expectancies.

Can we conclude that owning more TVs **causes** people to live longer? If not, what's the main **confounding variable**?

Here **wealth** is a **confounding variable** – nations with more TVs are wealthier, and wealth influences life expectancies (via more nutritious diets, better hospitals, etc.).

Notes

Example

A 1998 NIH study found that people aged 65 or older who **attend church** more often have lower incidences of high **blood pressure** than those who attend church less often.

An article about the study in *USA Today* (Aug. 11, 1998) stated:

> *"Attending religious services lowers blood pressure."*

This implies a **cause-and-effect** relationship between **church attendance** and **blood pressure**.

Is it reasonable to draw such a conclusion from the study?

Notes

It's known that **smoking** cigarettes and **drinking** alcohol can increase **blood pressure**, and people who **attend church** regularly may be **less likely** than others to **smoke** or **drink**.

Therefore **smoking** and **drinking** are possible **confounding variables** that may explain the observed relationship between **church attendance** and **blood pressure**.

Notes

Notes

Notes

Notes