

Notes

Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver
ngrevsta@msudenver.edu

November 10, 2019

Nels Grevstad

Notes

Topics

- 1 Graphing Bivariate Numerical Data: Scatterplots
- 2 Summarizing a Bivariate Relationship: The Sample Correlation

Nels Grevstad

Notes

Objectives

Objectives:

- Produce a scatterplot from bivariate data.
- Interpret the sample correlation r .
- **(Optional)** Compute the sample correlation from bivariate data.

Nels Grevstad

Notes

Graphing Bivariate Numerical Data: Scatterplots (14.1,

14.2)

Introduction

- **Bivariate data** are data for which **two variables** are measured on each individual.

Nels Grevstad

Example

The following **bivariate** data represent **lengths** (cm) and **weights** (g) of $n = 9$ female snakes.

Snake	Length (cm)	Weight (g)
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

Nels Grevstad

- Bivariate data are used to investigate the **relationship** between the two variables (e.g. to investigate how one changes with the other).
- The two variables usually play different roles.
The **explanatory** variable is the one that *explains* differences or *causes* changes in the **response** variable.

Nels Grevstad

Scatterplots

- The most useful way to display *bivariate* data is with a **scatterplot**. To construct one:
 1. Determine which variable (if any) is the **explanatory** variable and which is the **response**.
 2. Plot each individual as a **point** in an (x, y) coordinate system, with the **explanatory variable** on the ***x*-axis** and the **response** on the ***y*-axis**.
 3. Label the axes and add a title.

Nels Grevstad

Example

A **scatterplot** of the **lengths** and **weights** of snakes from is shown on the next slide.

The **explanatory** variable (**length**) is on the ***x*-axis** and the **response** (**weight**) is on the ***y*-axis**.

Snake number three from the data set is highlighted.

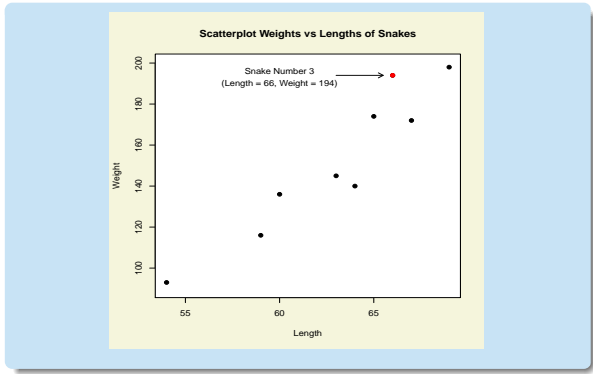
Nels Grevstad

Notes

Notes

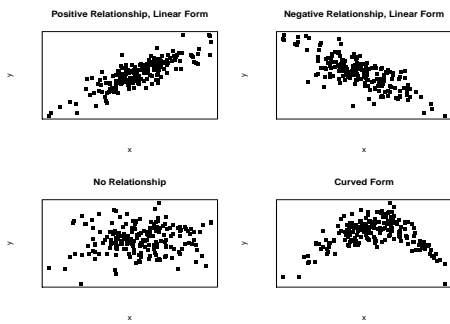
Notes

Notes

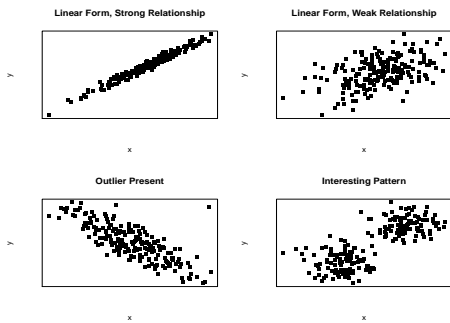


Nels Grevstad

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



Nels Grevstad



Nels Grevstad

• **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**.

Nels Grevstad

Notes

Notes

Notes

Notes

Summarizing a Bivariate Relationship: The Sample Correlation (14.4)

- The **sample correlation**, denoted r , is a statistic that **summarizes the strength** of a **linear relationship** between two quantitative (numerical) variables.

Nels Grevstad

(Optional)

- Suppose we have bivariate numerical data:

Individual	Explanatory Variable X	Response Variable Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
\vdots	\vdots	\vdots
n	x_n	y_n

Nels Grevstad

(Optional)

Sample Correlation: Use the following to compute r :

$$r = \frac{\frac{1}{n-1} \{ (x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \dots + (x_n - \bar{x})(y_n - \bar{y}) \}}{s_x \times s_y}$$

$$= \frac{\frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})}{s_x \times s_y}$$

where \bar{x} and \bar{y} are the **sample means** and s_x and s_y the **sample standard deviations** of the x and y variables, respectively.

Nels Grevstad

Interpreting the Correlation

- The following **properties** of the **correlation** r help us **interpret** its value:
 - The value of r will always lie **between -1.0 and 1.0**.
 - 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.

Nels Grevstad

Notes

Notes

Notes

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 (i.e. the points in the scatterplot lie **exactly** on a **straight line**).

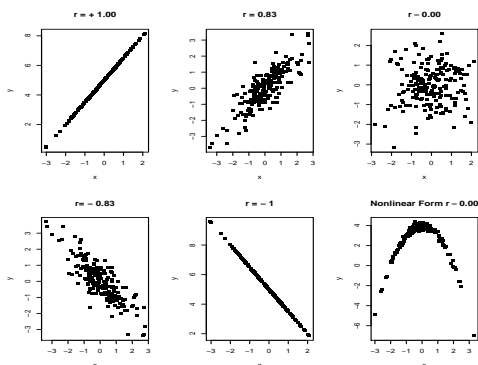
Nels Grevstad

4. The value of r doesn't depend on which of the two variables is labeled x and which is labeled y (i.e. r makes no distinction between explanatory and response variables).
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), it's just a number between -1.0 and 1.0.
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).

Nels Grevstad

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.
9. The correlation requires that both variables be **quantitative** (numerical), so that the calculation of r makes sense.
10. A **correlation** between two variables, even if it's strong, **doesn't** necessarily **imply** a **cause and effect** relationship. The relationship may be due to a third variable "lurking" in the background.

Nels Grevstad



Nels Grevstad

Notes

Notes

Notes

Notes

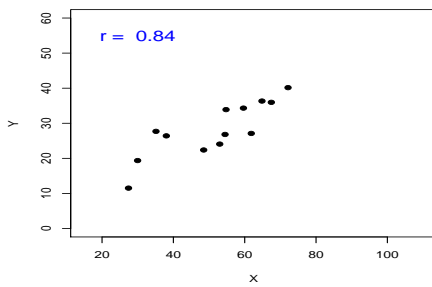
- The next plots show that the **correlation r is not resistant to outliers.**

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

Nels Grevstad

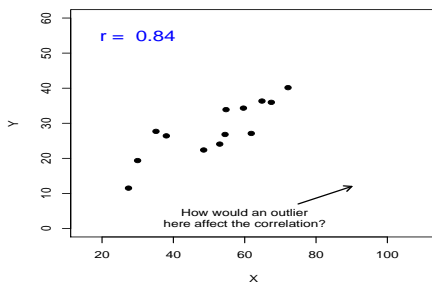
Plot of Y versus X



Notes

Nels Grevstad

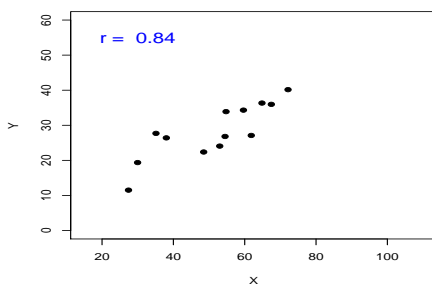
Plot of Y versus X



Notes

Nels Grevstad

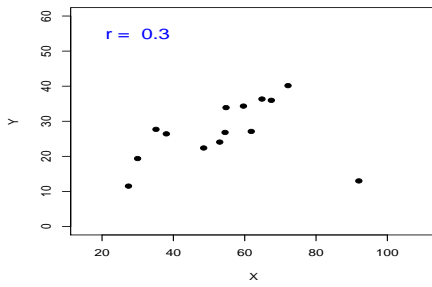
Plot of Y versus X



Notes

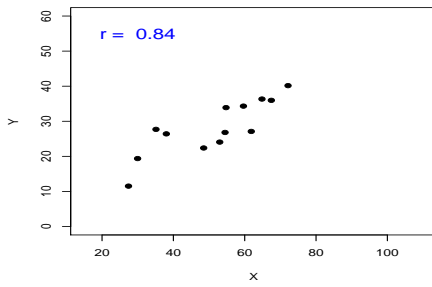
Nels Grevstad

Plot of Y versus X



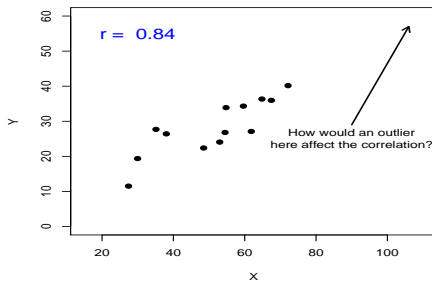
Nels Grevstad

Plot of Y versus X



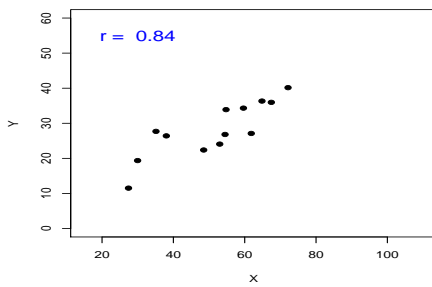
Nels Grevstad

Plot of Y versus X



Nels Grevstad

Plot of Y versus X



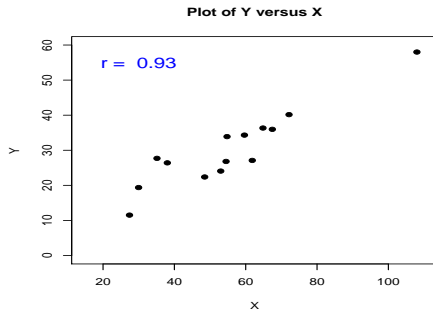
Nels Grevstad

Notes

Notes

Notes

Notes



Nels Grevstad

Notes

(Optional)

- Here's an example illustrating the computation and interpretation of r .

Nels Grevstad

Notes

Exercise

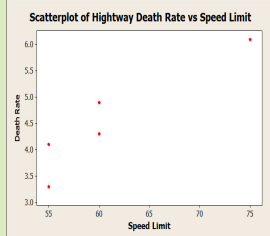
(Optional) The data in the table below come from the time when the U.S. still had a maximum speed limit of 55 mph. A scatterplot of the data is shown too.

An issue of concern was whether lower speed limits reduce the highway death rate.

Nels Grevstad

Notes

Country	Speed Limit (mph)	Death Rate (per 100 million veh. miles)
U.S.A.	55	3.3
Denmark	55	4.1
Canada	60	4.3
Australia	60	4.9
Italy	75	6.1



Nels Grevstad

Notes

- a) Calculate the correlation r .
- b) Explain why the value of r matches what's seen in the scatterplot.

Nels Grevstad

- Here's an example showing that correlation doesn't necessarily imply a cause and effect relationship x and y .

Nels Grevstad

Exercise

Suppose a study finds the correlation between beer sales and ice cream sales to be $r = 0.97$.

Does eating ice cream **cause** a thirst for beer? If not, suggest a possible "lurking" variable in the background that explains the relationship between beer and ice cream sales.

Nels Grevstad

Notes

Notes

Notes

Notes
