Correlation

Keegan Korthauer
Department of Statistics
UW Madison

Relationship Between Two Continuous Variables

- When we have measured two **continuous** random variables for each item in a sample, we can study the relationship between them

- For example, say we have a sample of houses that were sold recently and for each we know
  - the selling price
  - the square footage

- We suspect they might be related – how?

Linear Relationship

- If a plot of the ordered pairs shows a relatively straight line, the variables are said to have a **linear relationship**

- If we know the equation of the line that 'best fits' the data, then we can use it
  - to predict future observations
  - draw inferences about the relationship between the two variables

Bivariate Relationship

To study the relationship between two variables, we can start off as follows

- Graphical summary: scatterplot
- Numerical summary: **correlation coefficient**
  - Measures the strength of the linear relationship between two variables

Example – Housing Data

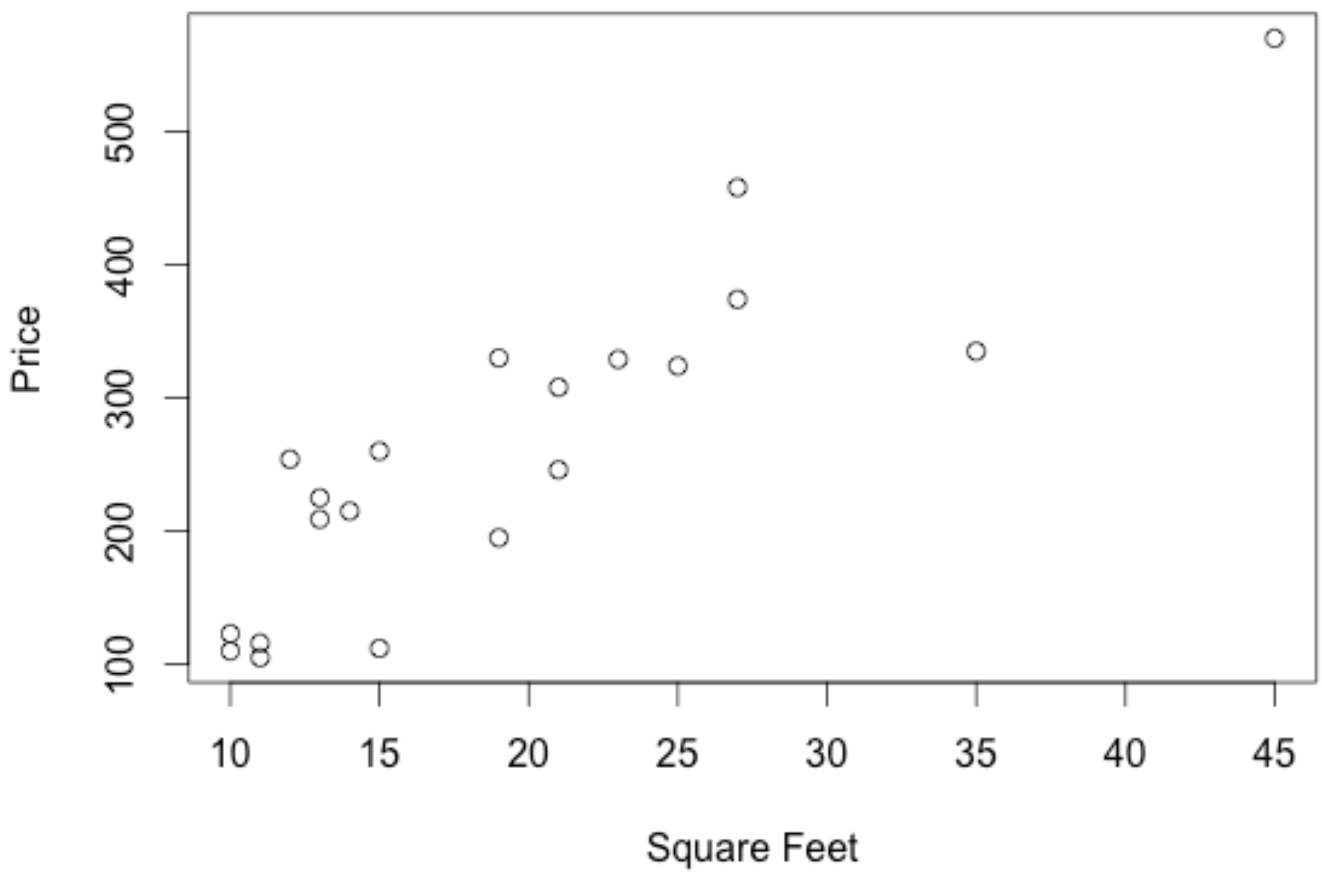
Here is a sample of 20 houses that were recently sold in Madison:

- Save as a text file with headers such as ‘Price’ and ‘Sqft’
- Read into R and make a scatterplot:

```
housing <- read.table("housing.txt",
    header=T)
attach(housing)
plot(Sqft, Price,xlab="Square Feet",
    ylab="Price")
```

Square Feet ($100ft^2$)	Price ($100K)
23	329
15	260
13	209
19	195
11	105
13	225
19	330
10	123
27	374
21	308
15	112
10	110
21	246
35	335
12	254
25	324
11	116
45	570
27	458
14	215

Example – Housing Data



How strong is the linear relationship?

Correlation Coefficient r

Let $(x_1,y_1),\ldots,(x_n,y_n)$ represent n points on a scatterplot

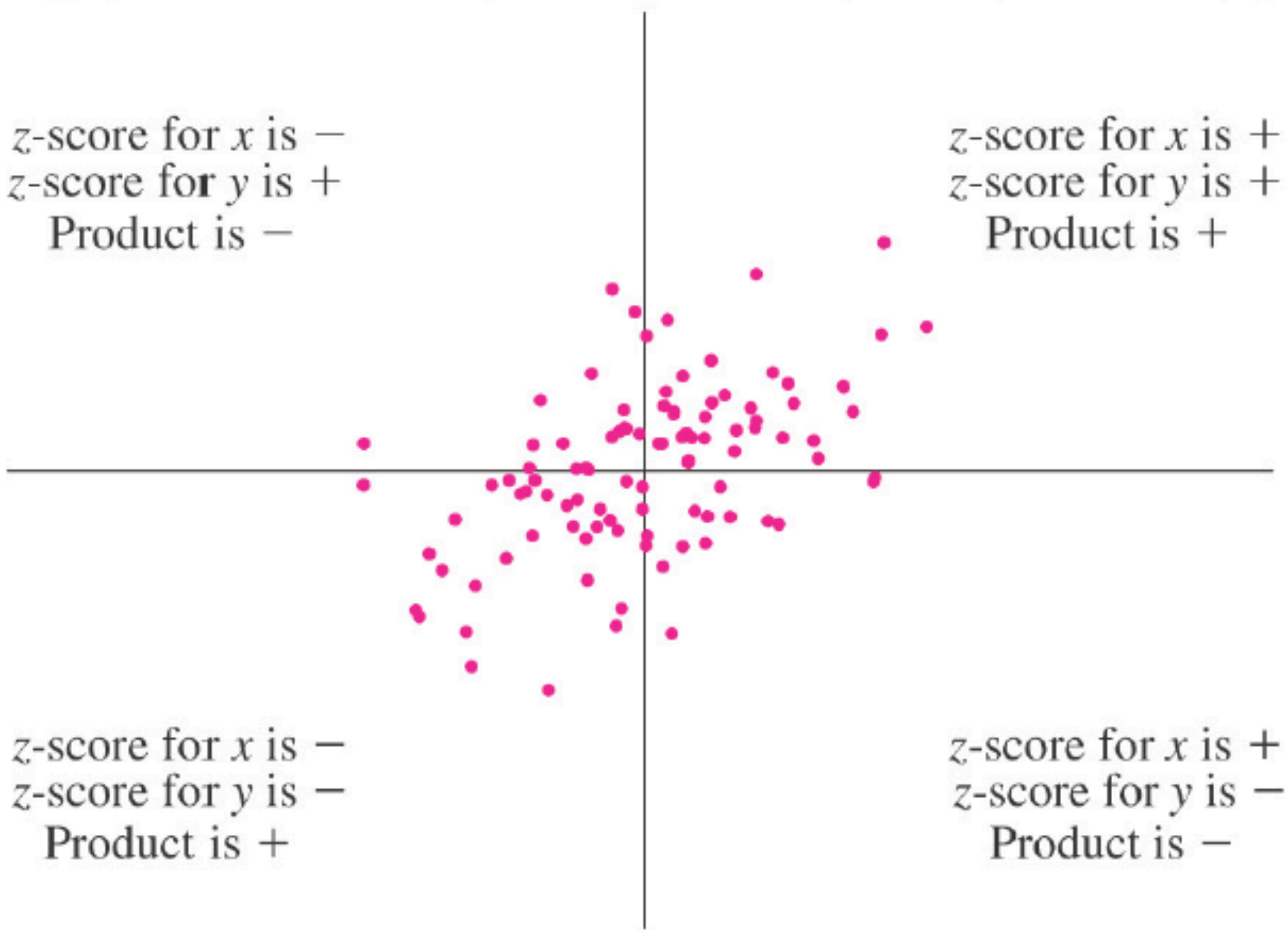
- Compute the means and standard deviations for the x's and y's
- Standardize the x's and y's (convert to z-scores):

$$\frac{(x_i - \bar{x})}{s_x} \text{ and } \frac{(y_i - \bar{y})}{s_y}$$

- Finally, calculate r as the average of the products of the z-scores (divided by n-1 instead of n)

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

Correlation Coefficient in a Diagram



Alternative Formulae for r

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}$$

Often the easiest to compute by hand

Example

Let x = square feet, y = cost. Given the following summary data, compute the correlation coefficient r for the housing example (n=20):

$$\bar{x} = 19.3,\ \bar{y} = 259.9$$

$$\sum_{i=1}^{n} x_i y_i = 119{,}156$$

If we didn't have those summary data, we could use the cor() function in R:

```
>cor(Sqft, Price)
[1] 0.8808653
```

$$\sum_{i=1}^{n} x_i^2 = 9036,\ \sum_{i=1}^{n} y_i^2 = 1{,}639{,}188$$

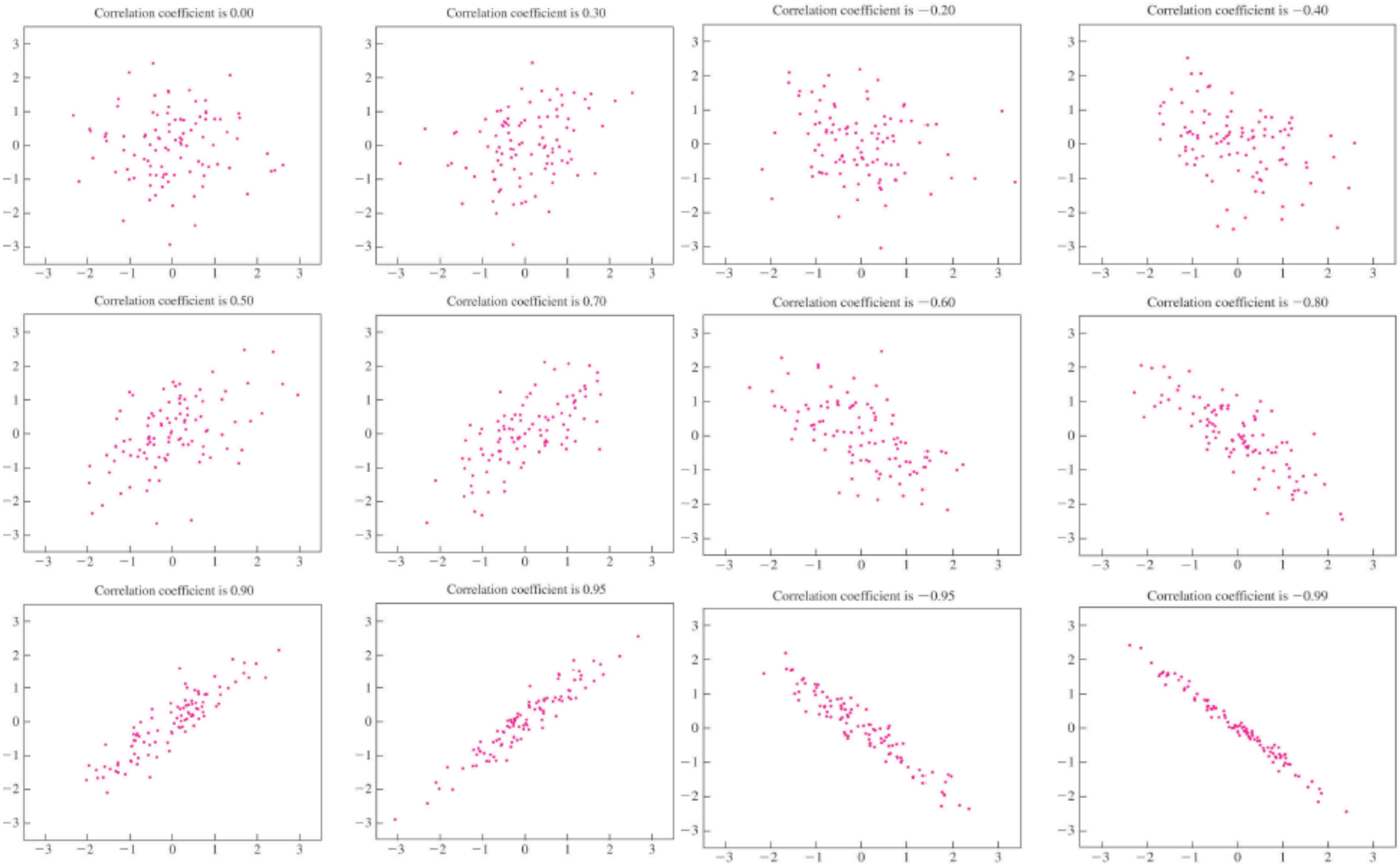
Properties of r

- r is unitless
- r is always between -1 and 1
- When r is exactly 1 or -1, all points fall exactly on a straight line
- r > 0: **positive** linear relationship/slope
  - greater values of one variable are associated with greater values of the other
- r < 0: **negative** linear relationship/slope
  - greater values of one variable are associated with smaller values of the other

More Properties of r

- Values of r close to 1 or -1 indicate a **strong** linear relationship
  - values of r close to 0 indicate a **weak** linear relationship
- When r is exactly 0, we say that the two variables are **uncorrelated**
  - Likewise, when $r \neq 0$ we say they are correlated
  - Note that **uncorrelated** is **not** the same as **independent**
- Correlation does not change if we
  - Multiply each value of a variable by a constant
  - Add a constant to each value of a variable
  - Interchange x and y

Examples of Various Levels of r



Correlation Coefficient Measures Linear Association ONLY

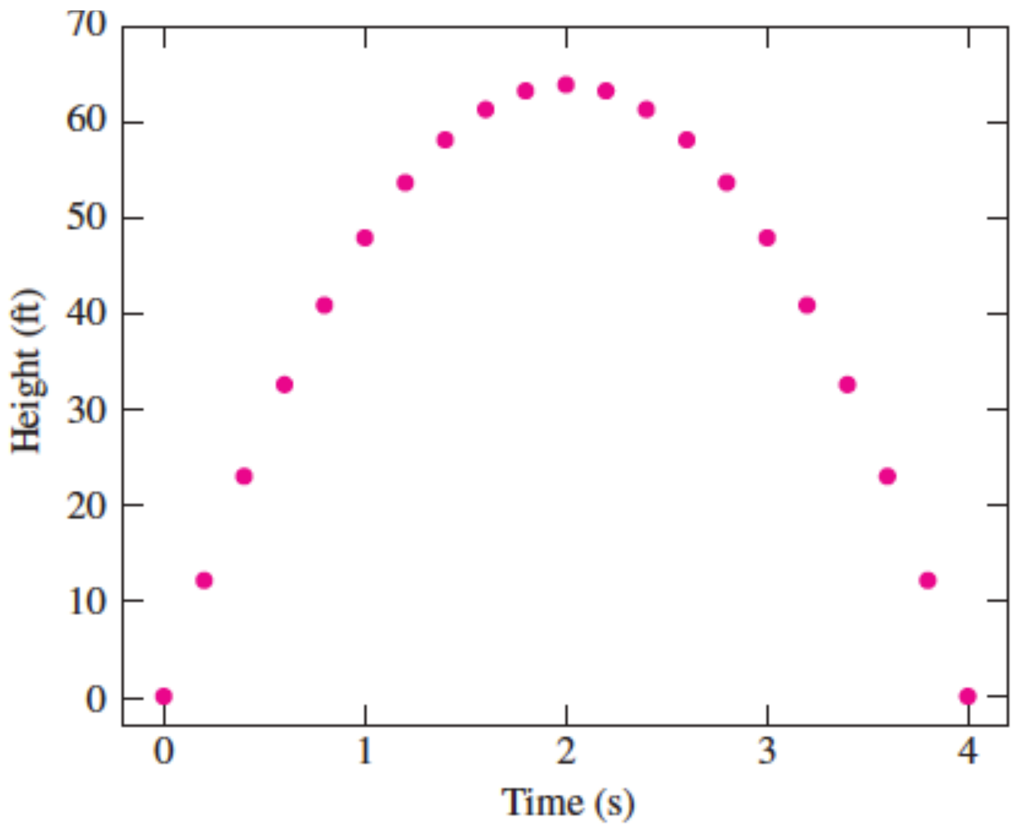
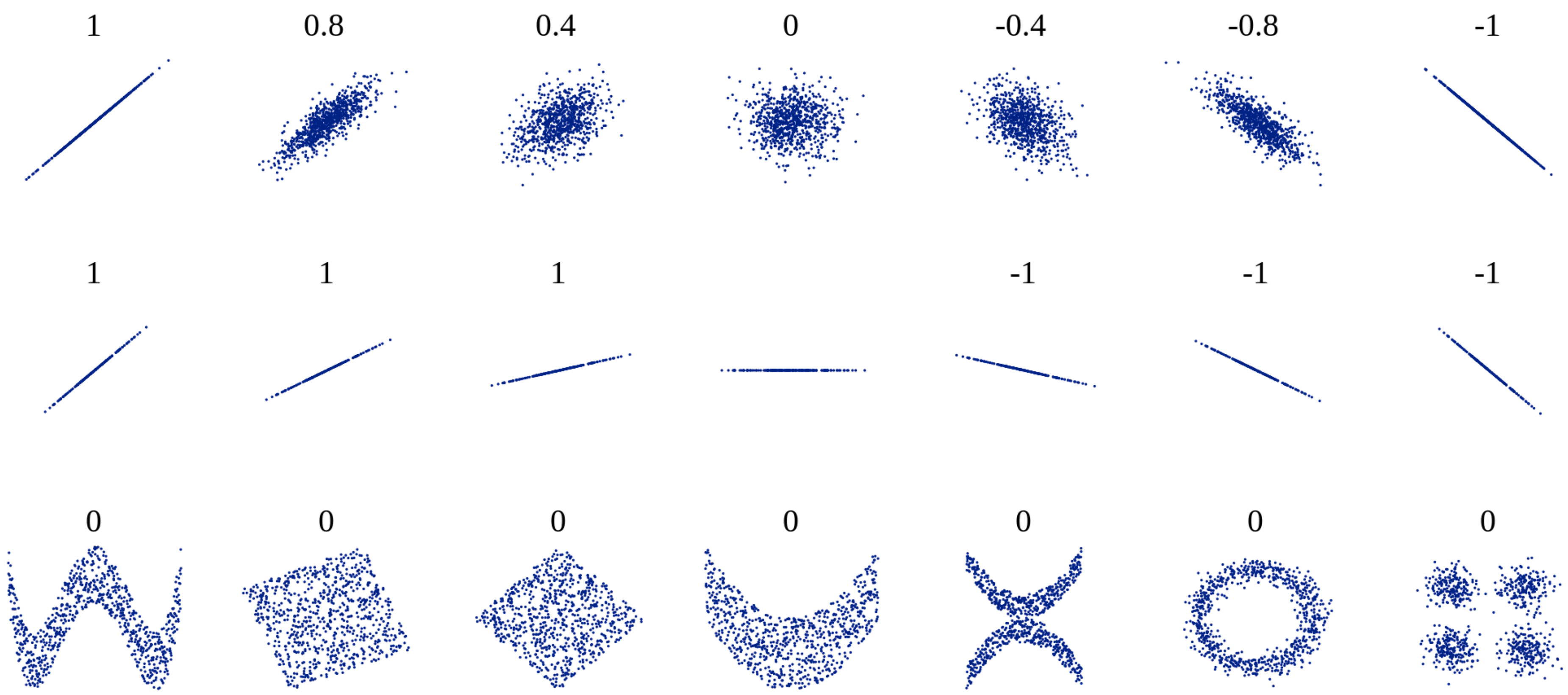


FIGURE 7.7 The relationship between the height of a free-falling object with a positive initial velocity and the time in free fall is quadratic. The correlation is equal to 0.

Bottom row: r = 0



Outliers Have Strong Influence on r

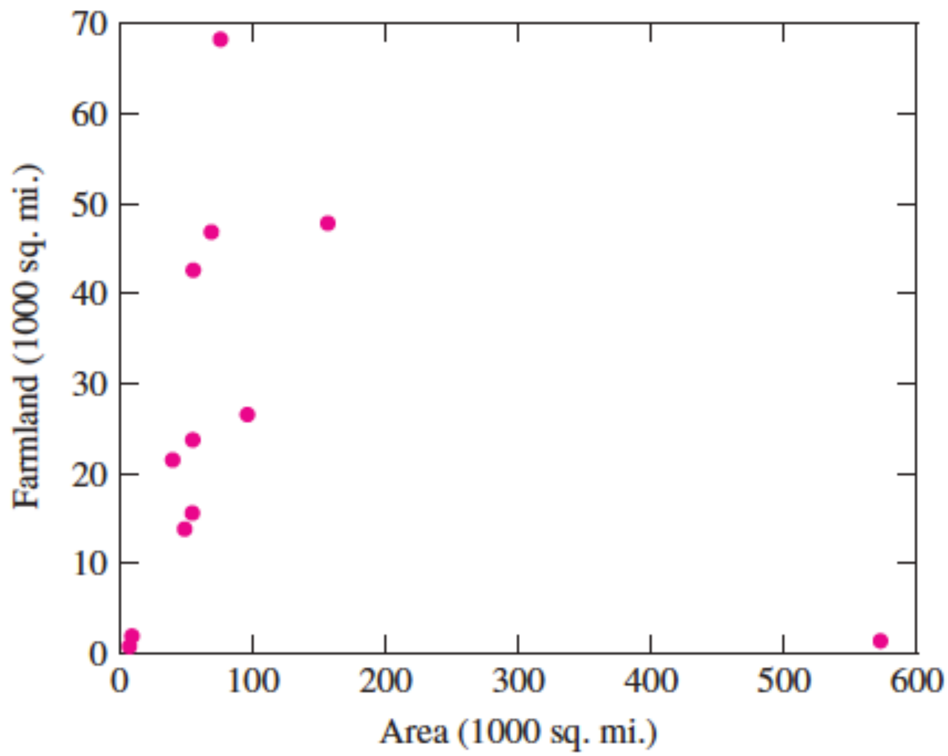
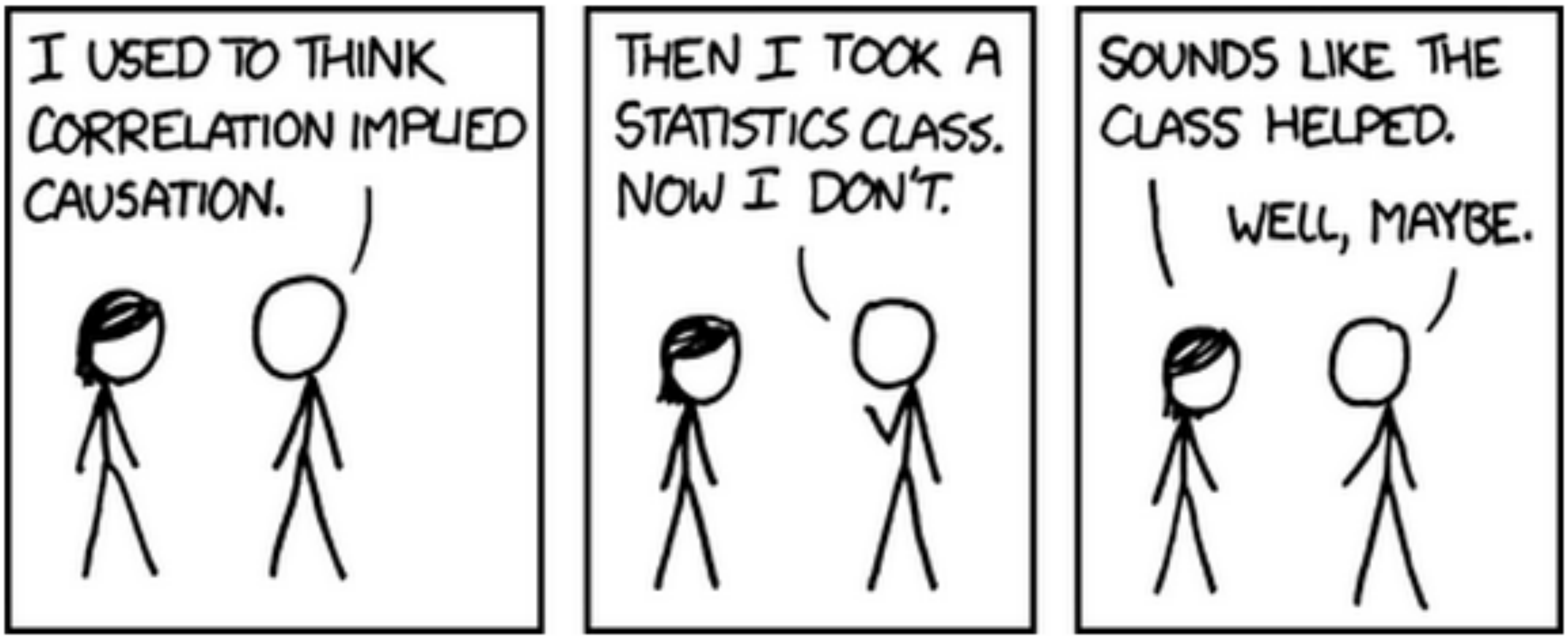


FIGURE 7.8 The correlation is −0.12. Because of the outlier, the correlation coefficient is misleading.

Warning: Correlation ≠ Causation

- **Confounder** – a third factor that is correlated with both x and y and results in spurious association between x and y
- Examples
  - Shoe size vs vocabulary: strong positive correlation between a child's shoe size and his/her vocabulary, but **age is the confounder** -> cannot conclude that a larger shoe size results in a large vocabulary
  - Ice cream sales vs drowning deaths: positive correlation, but **weather** is a confounder that influences both ice cream sales and drowning deaths -> cannot conclude that selling ice cream increases deaths due to drowning
- Controlled experiments reduce the risk of confounding



source: xkcd.com

Population Correlation Coefficient

- r is an estimate of the population correlation coefficient ρ (or $\rho_{X,Y}$)
- If X and Y are both normal variables (**not** necessarily independent) we can construct a test statistic that has a known distribution
- This allows us to find
  - Confidence intervals for ρ
  - Test a null hypothesis of the form H_0: $\rho = \rho_0$, H_0: $\rho \le \rho_0$, or H_0: $\rho \ge \rho_0$

Inference on the Population Correlation

- Let X and Y be bivariate normal and let ρ be the population correlation coefficient between X and Y.
- Let $(x_1,y_1),\ldots,(x_n,y_n)$ be a random sample from the joint distribution of X and Y
- Let r be the sample correlation of the n points; then

$$W = 0.5\ln\left(\frac{1+r}{1-r}\right) \sim N(\mu_W, \sigma_W^2)$$

$$\text{where } \mu_W = 0.5\ln\left(\frac{1+\rho}{1-\rho}\right) \text{ and } \sigma_W^2 = \frac{1}{n-3}$$

Inference on the Population Correlation

- To construct CIs for ρ, first find a CI for μ_W and then solve for ρ in the equation for μ_W :

$$\rho = \frac{e^{2\mu_W} - 1}{e^{2\mu_W} + 1}$$

- To perform a HT for ρ, use W as the test statistic and find the p-value using the normal distribution with mean/variance given on the previous slide

HT for population correlation

Let $(x_1,y_1),\ldots,(x_n,y_n)$ be a random sample from the joint distribution of X and Y where X and Y are bivariate normal. Let r be the sample correlation

1. Set up the null and alternative hypotheses (see table below)
2. State the significance level
3. Calculate the test statistic $W = 0.5\ln\left(\frac{1+r}{1-r}\right)$
4. Calculate the p-value, where the distribution of W is

$$W \sim N(\mu_W, \sigma_W^2) \text{ and } \mu_W = 0.5\ln\left(\frac{1+\rho}{1-\rho}\right) \text{ and } \sigma_W^2 = \frac{1}{n-3}$$

H_0	H_1	P-value
$\rho \le \rho_0$	$\rho > \rho_0$	Area to the right of *W*
$\rho \ge \rho_0$	$\rho < \rho_0$	Area to the left of *W*
$\rho = \rho_0$	$\rho \ne \rho_0$	Area to the left of -W plus area to the right of W

5. Make a conclusion

Examples 7.3 and 4

In a study of reaction times, the time to respond to a visual stimulus (x) and the time to respond to an auditory stimulus (y) were recorded for each of 10 subjects. Times were measured in ms. The results are presented in the following table.

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

Find a 95% confidence interval for the correlation between the two reaction times.

Find the P-value for testing $H_0: \rho \leq 0.3$ versus $H_1: \rho > 0.3$.

Next

- Review for Exam 2 Monday
  - Come with questions
  - Practice exam posted
- Exam 2 on Wednesday; remember to bring:
  - One 8.5x11” sheet (front/back) handwritten notes
  - Calculator
- No Homework due next Friday