Correlation & Regression

Statistics Tutorial

Day 8

Prabesh Dhakal

2020 June 04

WHAT ARE WE DOING TODAY?

RECAP + Q&A

We briefly revisit the contents from last week.

Correlation and Regression

EXERCISE

We apply what we learned.

Q&A and Recap

Please ask if you have any questions now.

Otherwise, we can move on to the recap.

$t - Test$

Test Statistic:

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

1. **One Sample t-Test**
   - Check if the sample mean differs statistically from a hypothesized population mean
2. **Paired t-Test**
   - Compare means of two samples of same object/category/…
3. **Independent t-Test**
   - Compare means of two independent samples in order to determine whether the associated population means differ significantly

$F - Test$

F-Test is used to **compare variance** of two groups and check if they are different from each other.

$$H_0: ratio\ of\ variance = 1$$
$$H_1: ratio\ of\ variance \neq 1$$

Test Statistic:

$F = \frac{s_1^2}{s_2^2}$ Degrees of freedom: $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$

Main assumption: data is normally distributed

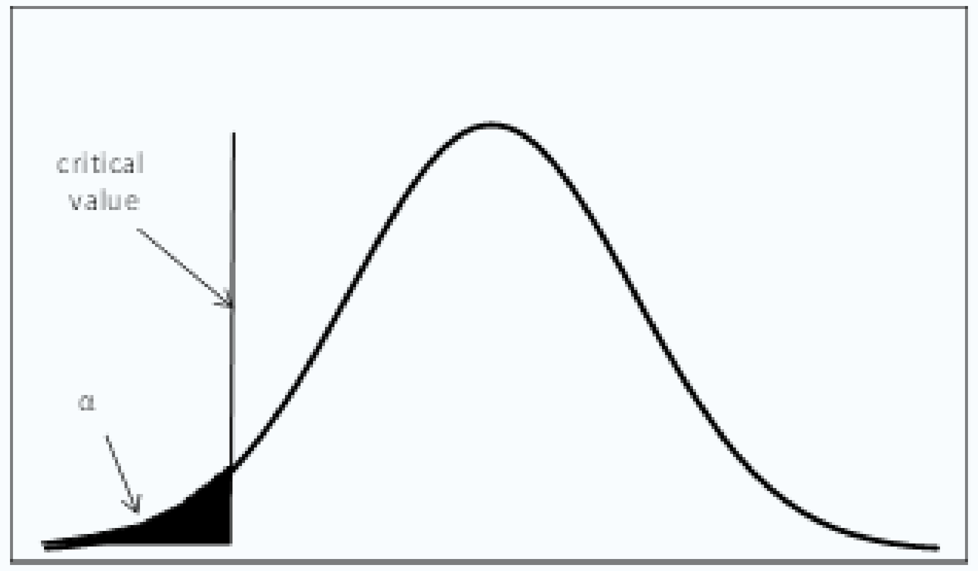
Problem: very weak if data deviates from a normal distribution

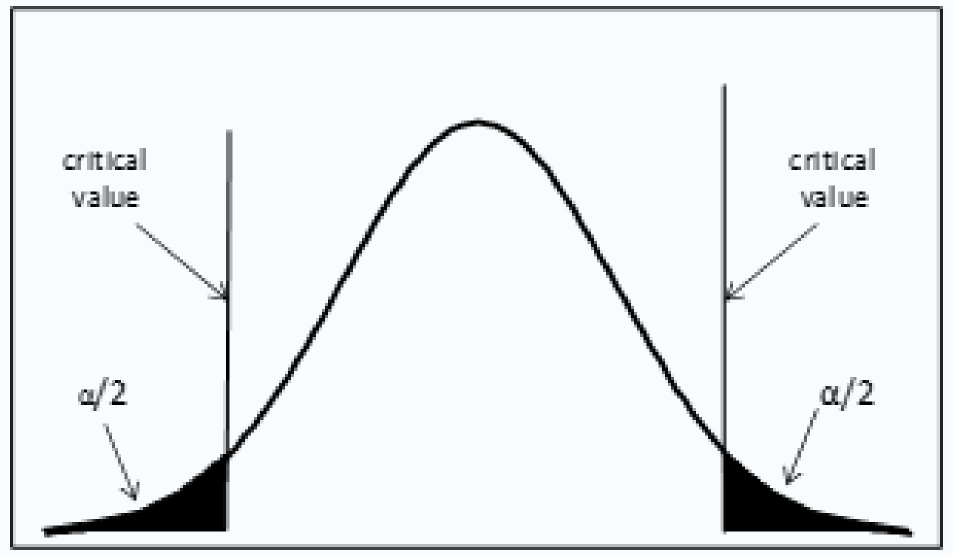
Hypothesis Testing

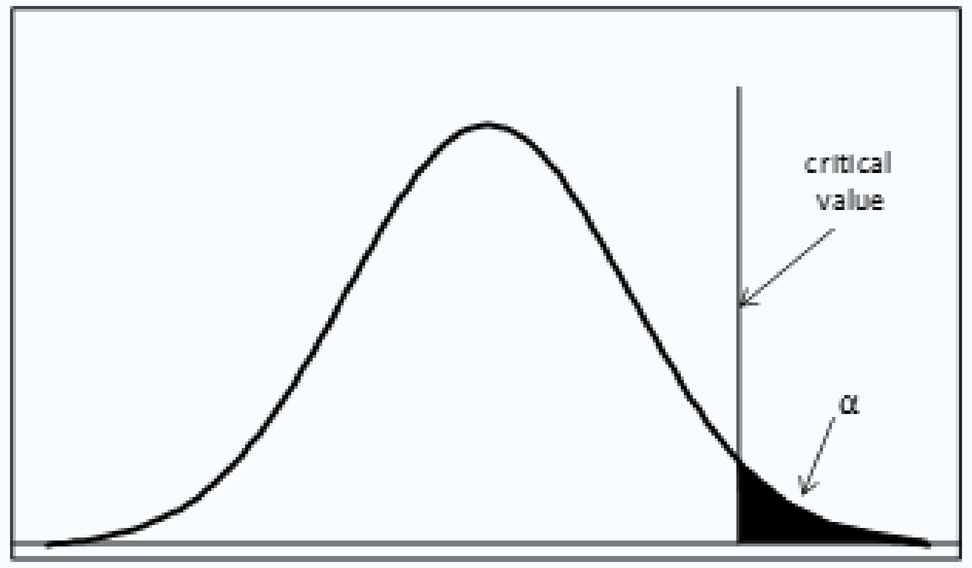
Additional Notes

1 tailed vs 2 tailed Tests

One Tailed Test (Left Tail)	Two-Tailed Test	One Tailed Test (Right Tail)
$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$
When population parameter is believed to be **lower** than the assumed one.	It determines whether the sample tested falls **within or outside** a certain range of values.	When population parameter is believed to be **higher** than the assumed one.
Reject H_0 if **test statistic < critical value**	Reject H_0 if **test statistic < critical value or** **test statistic > critical value**	Reject H_0 if **test statistic > critical value**







PARAMETRIC VS NON-PARAMETRIC TESTS

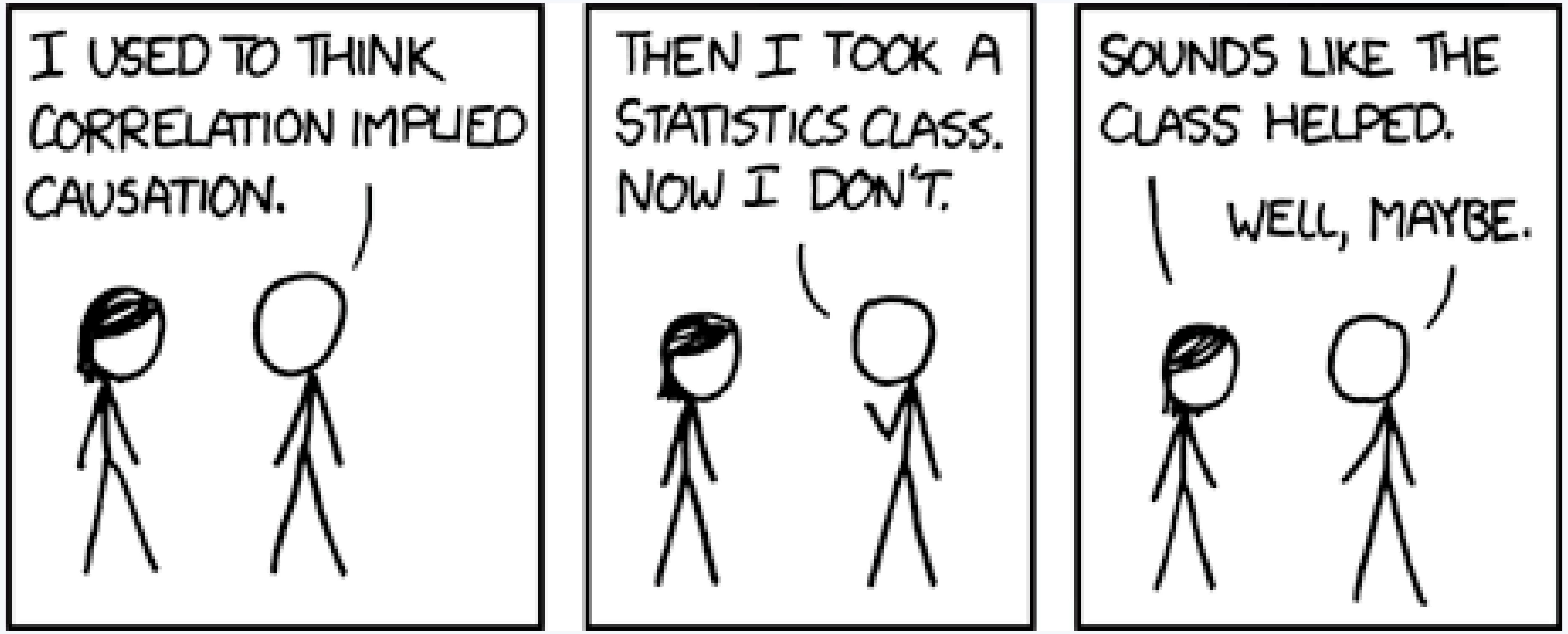
Parametric Tests:

Make assumptions about the parameters of the population distribution from which the sample is drawn. Mostly, normality is also assumed

Non-parametric Tests:

Also called “distribution-free” as they don’t make any assumptions about parameters of the population distribution.

Parametric Test	Non-parametric Test Equivalent
One Sample t-Test	Wilcoxon signed-rank Test
Two-sample t-Test	Wilcoxon 2-sample rank-sum Test
F-Test	Levene’s Test / Fligner-Kileen Test
Pearson Correlation	Spearman Correlation



CORRELATION

COVARIANCE

Covariance is a measure of how much two variables *vary* together.

$$cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x - \overline{x})(y - \overline{y})$$

Two important things to notice:

1. **Sign** of covariance:
   - <u>Positive covariance</u>: the two variables move together
   - <u>Negative covariance</u>: the two variables move inversely
2. **Magnitude** of covariance: not easy to interpret

CALCULATING COVARIANCE

Steps:

1. Form a table with these columns:

   $\mathbf{x}$, $\mathbf{y}$, $(\mathbf{x} - \bar{\mathbf{x}})$, $(\mathbf{y} - \bar{\mathbf{y}})$, and $(\mathbf{x} - \bar{\mathbf{x}}) * (\mathbf{y} - \bar{\mathbf{y}})$

2. Add all the values from $(\mathbf{x} - \bar{\mathbf{x}}) * (\mathbf{y} - \bar{\mathbf{y}})$ column
3. Divide the sum from step 2 by the number of observation $(\boldsymbol{n} - \mathbf{1})$

$$cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x - \bar{x})(y - \bar{y})$$

$$= \frac{1}{6-1} * (-114.33)$$

$$= -22.866$$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) * (y - \bar{y})$
2	86.00	-1.67	34.67	-57.78
5	77.00	1.33	25.67	34.22
4	43.00	0.33	-8.33	-2.78
6	23.00	2.33	-28.33	-66.11
1	56.00	-2.67	4.67	-12.44
4	23.00	0.33	-28.33	-9.44
3.67	**51.33**			**-114.33**

(PEARSON'S) CORRELATION

Pearson's correlation – correlation – is a normalized version of covariance.

$$cor(x, y) = \frac{cov(x, y)}{\sigma_x * \sigma_y}$$

<u>The correlation coefficient</u>

- measures the strength of the linear relationship between two quantitative variables
- value lies between [-1, +1] (whereas, covariance can have any value)

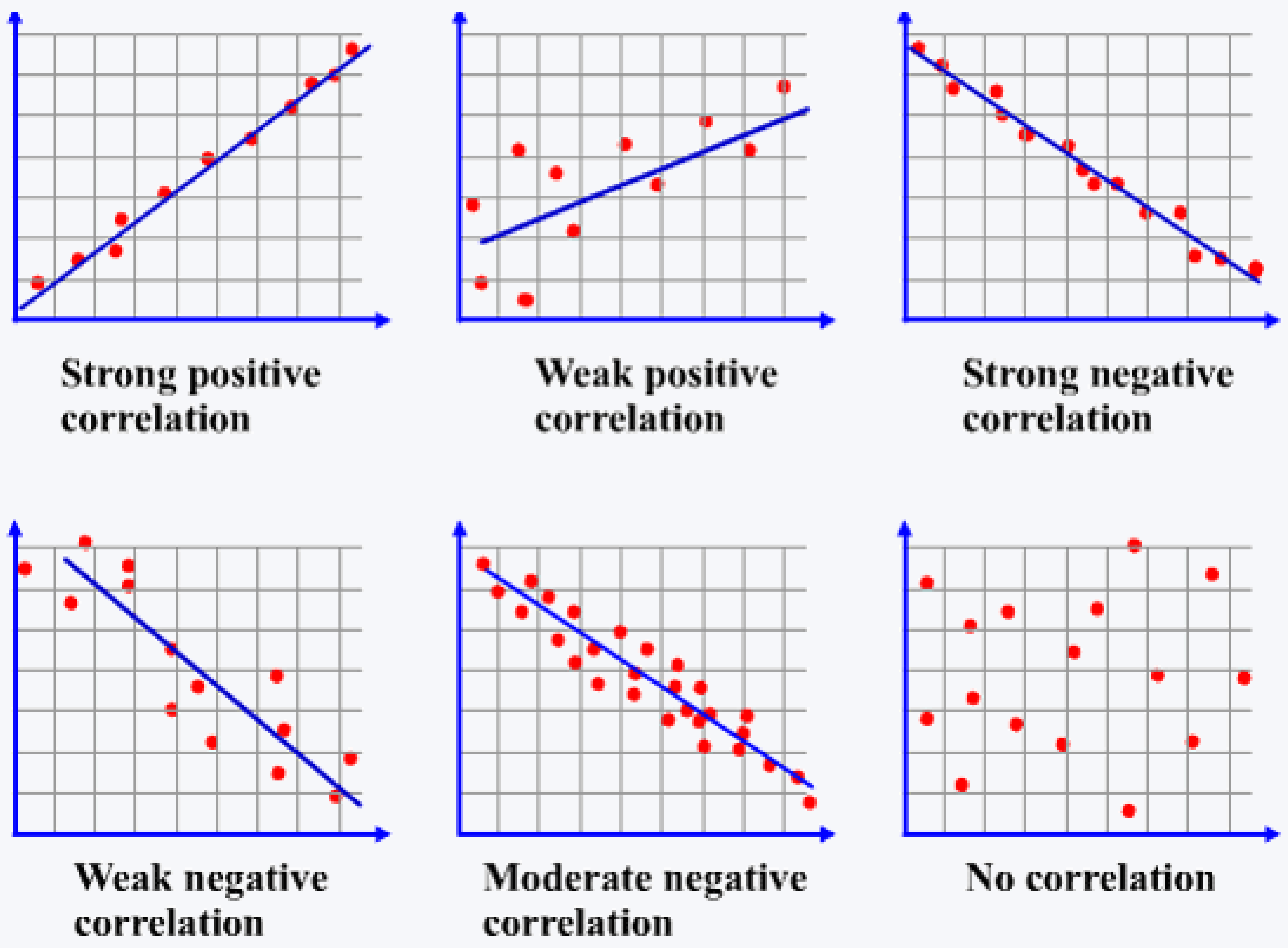
(Keep in mind that you will find different versions of formula for calculating correlation.)

(PEARSON'S) CORRELATION

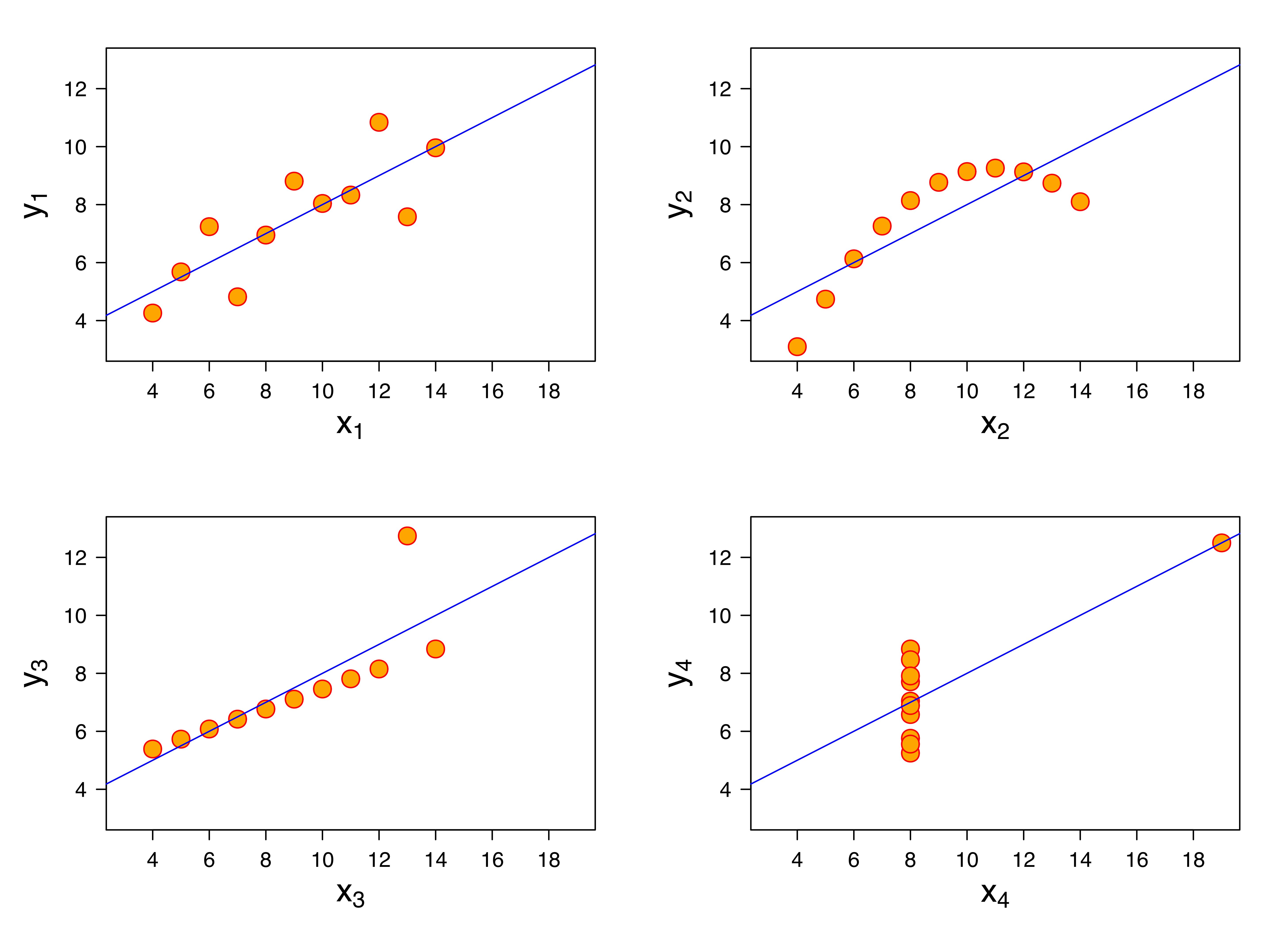
Assumptions:

1. Observations are continuous
2. Variables follow a normal distribution
3. Variables have a linear relationship

$$cor(x, y) = \frac{cov(x, y)}{\sigma_x * \sigma_y}$$



ANSCOMBE'S QUARTET

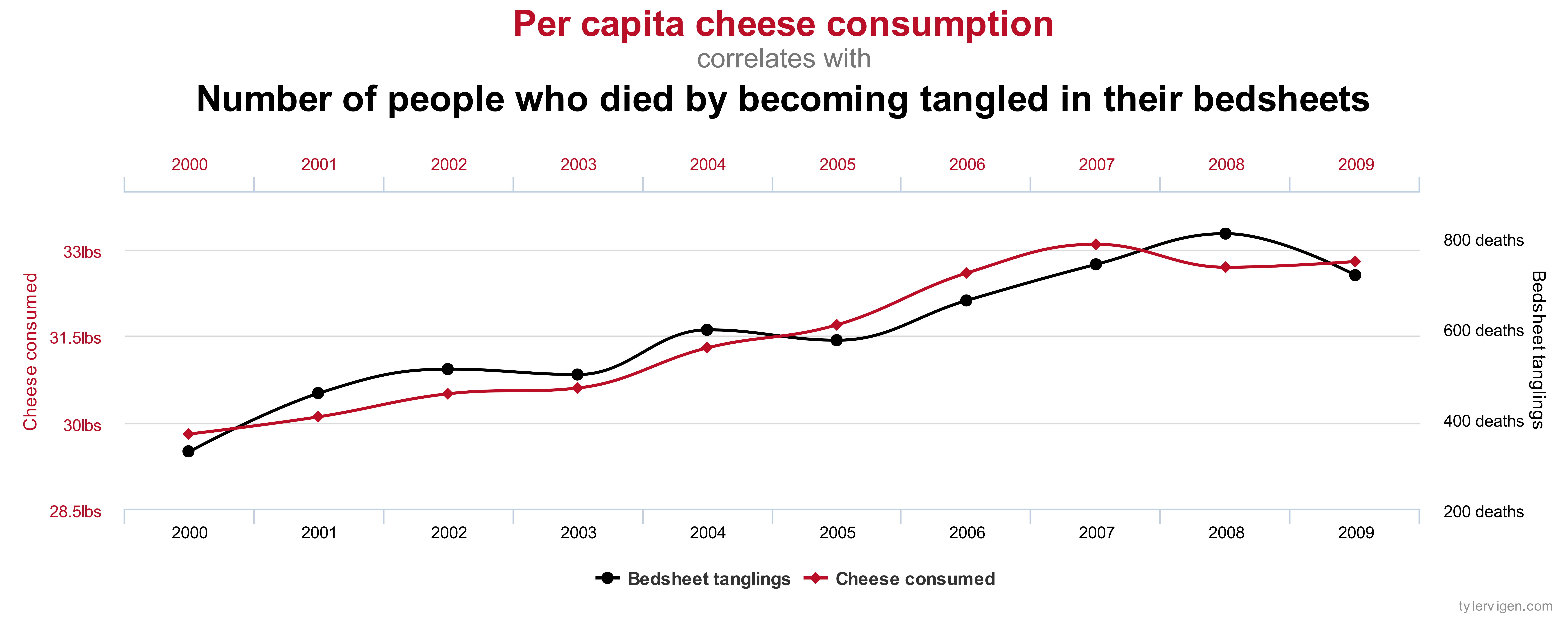


For all 4 datasets:

Mean of x $=$ 9
Var. of x $=$ 11
Mean of y $=$ 7.5
Var. of y $=$ 4.125

cor(x, y) $=$ 0.816

SPURIOUS CORRELATIONS



https://www.tylervigen.com/spurious-correlations

MAIN MESSAGE

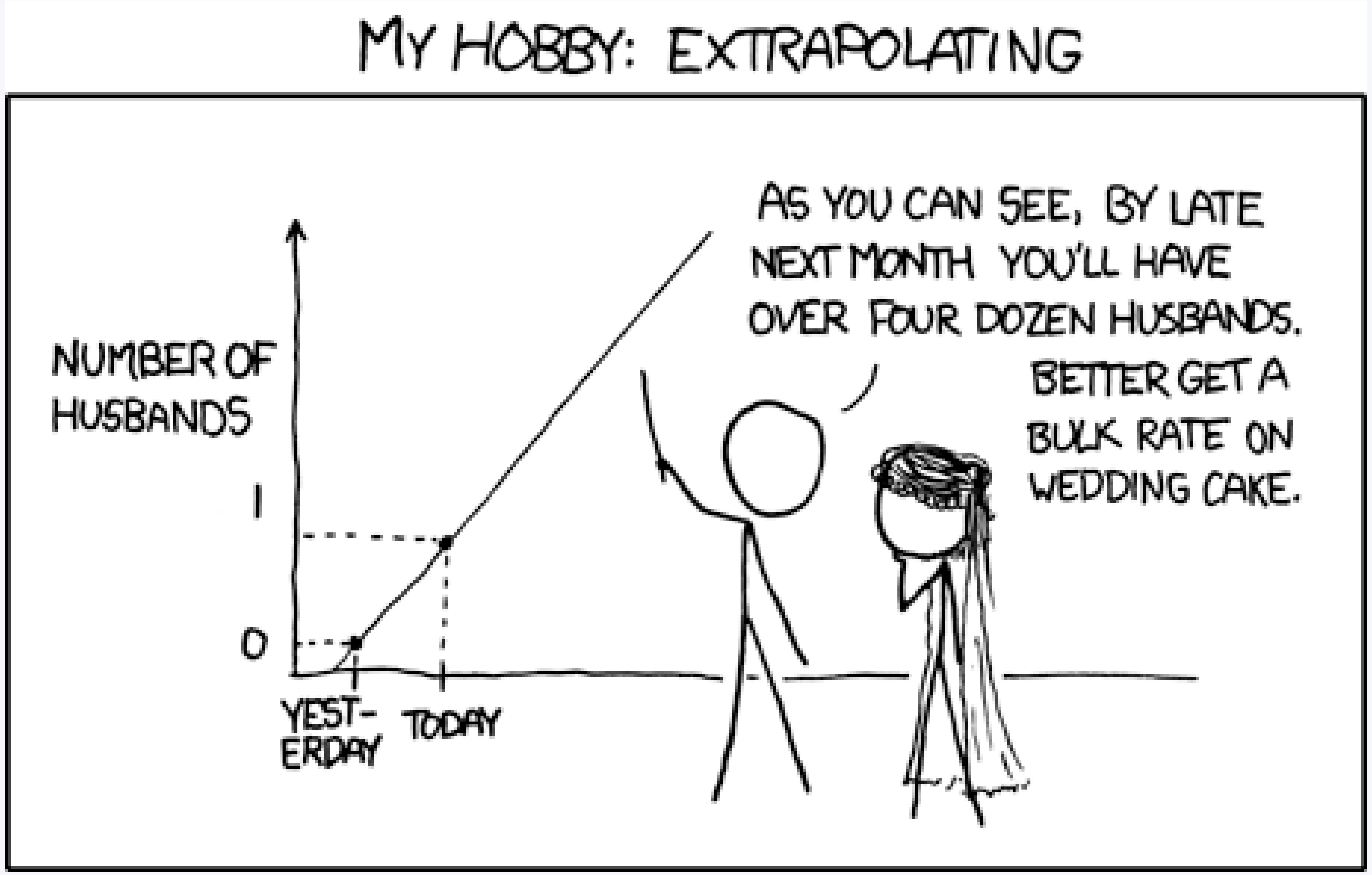
Correlation $\neq$ Causation

Correlation

“X and Y tend to be observed at the same time”

Causality

“X causes Y”

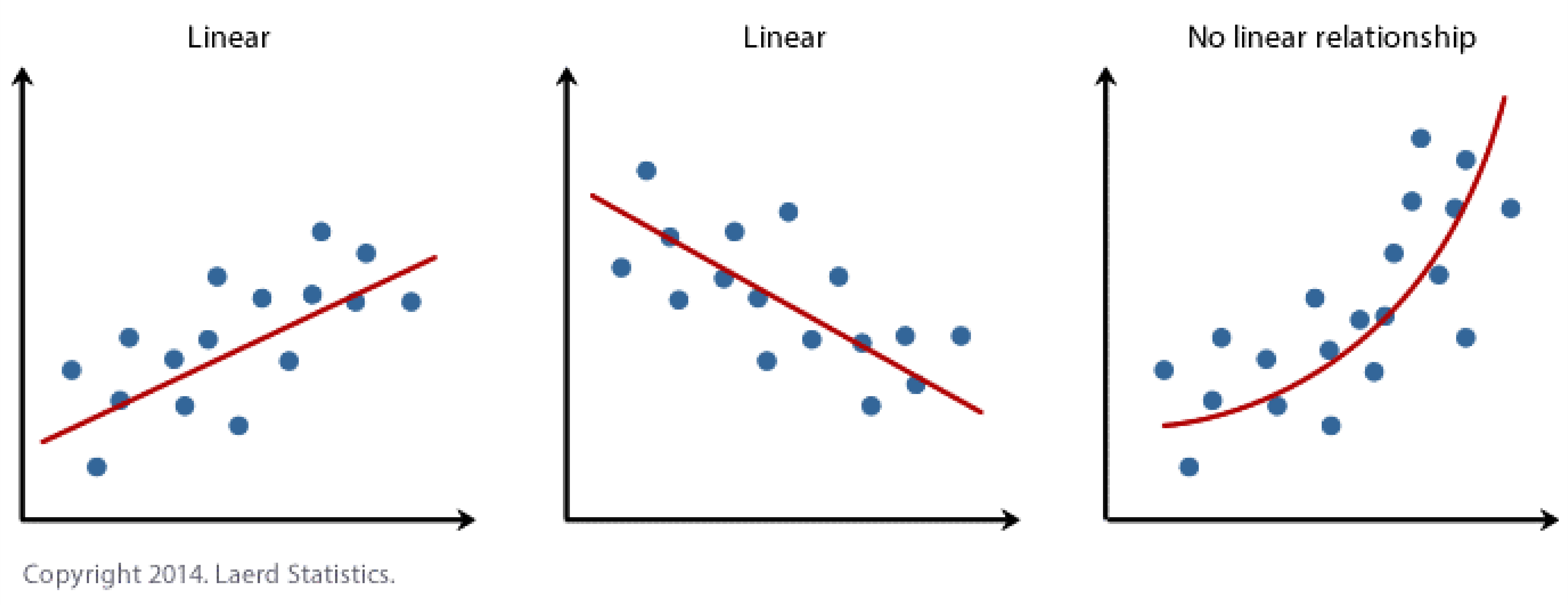

MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

REGRESSION

REGRESSION

Regression analysis = a statistical method for analyzing a relationship between two or more variables in such a manner that one variable can be predicted or explained by using information on others.

LINEAR RELATIONSHIP



REGRESSION

Variables: x and y are continuous, and follow a normal distribution

Objective: we want to predict y based on x $\leftrightarrow$ $(y \sim x)$

"Simple" Regression Model:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Dependent variable

Regression coefficients

Independent variable

Error *or* **Residuals**

REGRESSION $(\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x} + \boldsymbol{\varepsilon})$

Dependent Variable: depends on some other variable(s);

aka: response variable

Independent Variable(s): determine the value of dependent variable;

aka: predictor or explanatory variable

Objective:

estimating the **"right"** regression coefficients

What does *RIGHT* mean in this context?

The model with smallest **error** is the best model = the st. line that best fits the data.

REGRESSION DIAGNOSIS

1. Is there a linear relationship between the variables? → scatter plot
2. Are the residuals normally distributed? → histogram/Q-Q plot
3. Homoskedasticity of residuals → plot of residuals

   = residuals need to look uniformly scattered (no cones or obvious trends)
4. Coefficient of determination (next slide)

Coefficient of Determination (R-squared)

Intuition behind Goodness of Fit: how well does the model fit the data

R-Squared reflects goodness of fit for linear regression models.

It is also called the *coefficient of determination*.

$$R - squared = \frac{Explained\ Variation}{Total\ variation}$$

The R-squared value reflects the percentage of dependent variable variation that is explained by a linear model.

INTERPRETING R-SQUARED

R-squared value is always between 0%-100%

In general, higher R-squared = better the model fits the data.

For simple regression model (only 2 variables), R-squared is also the squared value of correlation figure r.

REGRESSION CALCULATION

Performing regression by hand:

https://www.youtube.com/watch?v=GhrxgbQnEEU

We will do more in R later.

Exercise

🛠 EXERCISE IN R 🛠

Download the R file for Day 8 and open it on RStudio. ☺

PLAN FOR NEXT WEEK

That's it for today! :-)

Next week, we are going to discuss:

- ANOVA

If you want to reach me, mail me at:

`prabesh.dhakal@stud.leuphana.de`