Correlation, Regression & ANOVA

Math & Stats Tutorial
Day 7

Prabesh Dhakal
27 May 2019

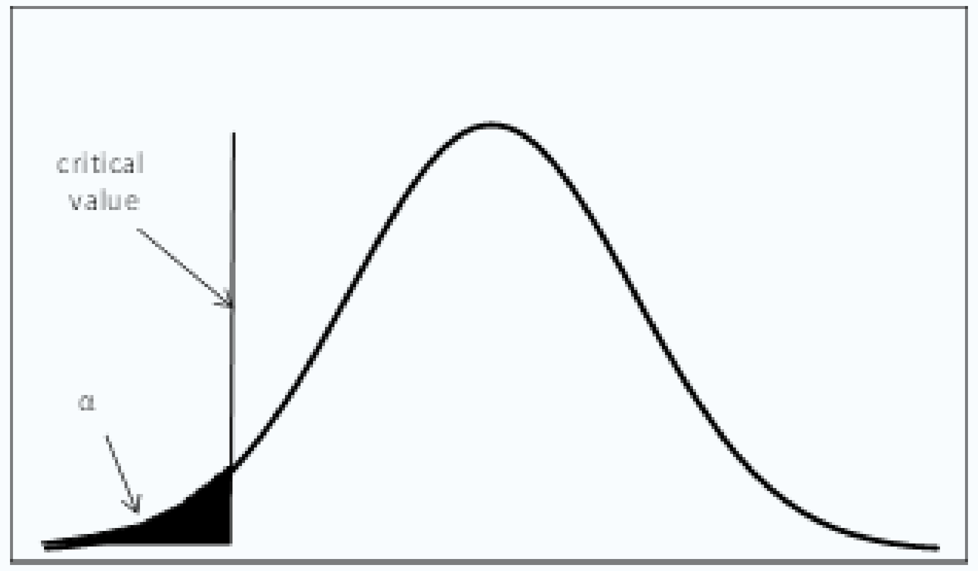
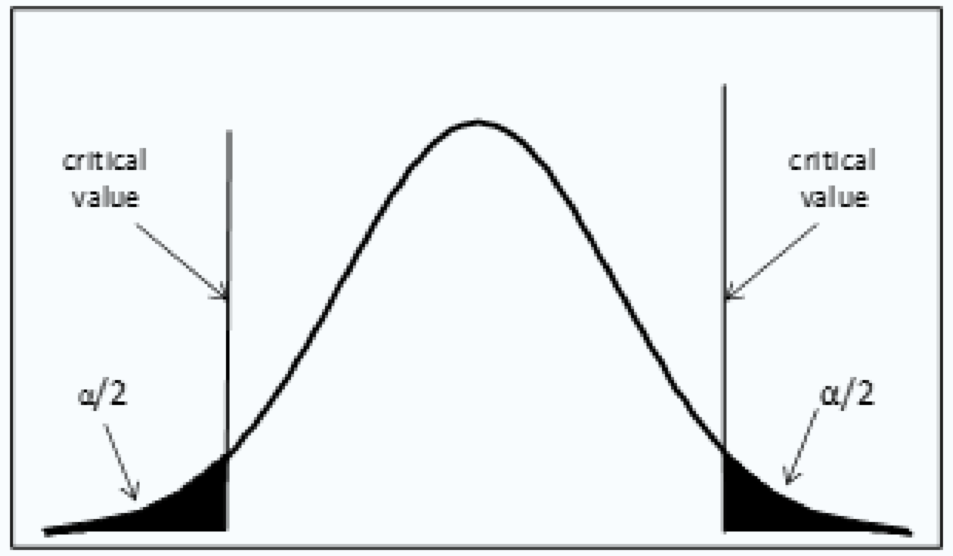
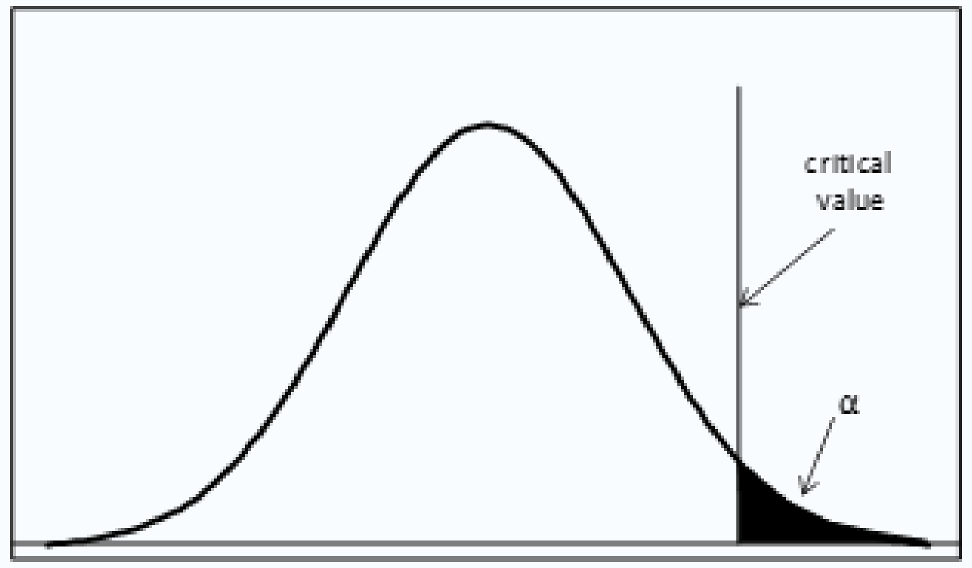
REVIEW

1. Critical Value/Rejection Region
2. Central Limit Theorem
3. Tails of tests
4. Tests: Shapiro-Wilk, t-Test, F-Test
5. Correlation
6. R-Studio Session

PLAN FOR TODAY

1. (One tailed t-Test)
2. Correlation
3. Regression
4. ANOVA
5. R-Studio Session

1 TAILED TEST

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**
		

WHEN IS A 1 TAILED TEST APPROPRIATE?

If you consider the consequences of missing an effect in the untested direction and conclude that they are negligible and in no way irresponsible or unethical, then you can proceed with a one-tailed test. For example, imagine again that you have developed a new drug. It is cheaper than the existing drug and, you believe, no less effective. In testing this drug, you are only interested in testing if it less effective than the existing drug. You do not care if it is significantly more effective. You only wish to show that it is not less effective. In this scenario, a one-tailed test would be appropriate.

Source: *UCLA: What are the differences between one-tailed and two-tailed tests?*

1 TAILED T-TEST (EXAMPLES)

How to do this by hand?

https://www.dau.mil/cop/ce/DAU%20Sponsored%20Documents/One%20tailed%20hypothesis%20test.pdf

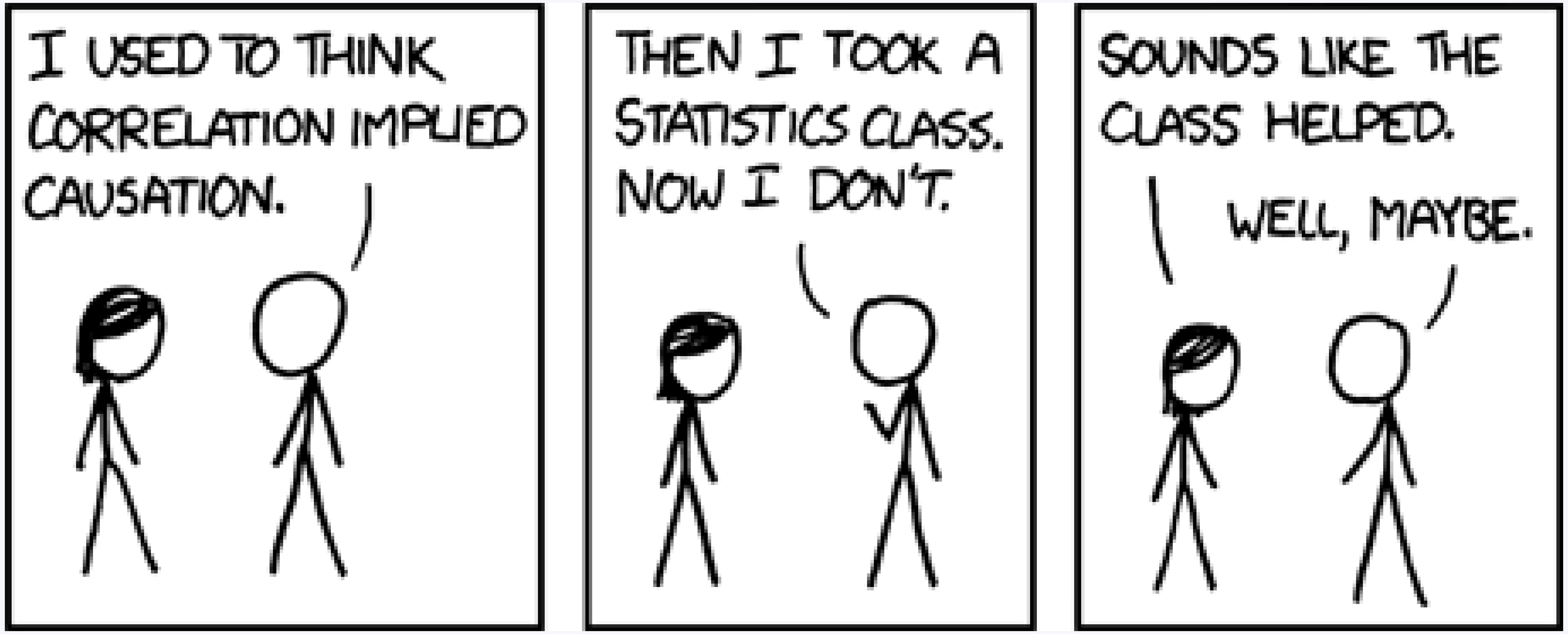
How to do this in R?

Hypotheses:

$\boldsymbol{H_0}: \mu \leq 50$ and $\boldsymbol{H_1}: \mu > 50$

```
data = c(52.7, 53.9, 41.7, 71.5, 47.6, 55.1, 62.2, 56.5, 33.4, 61.8, 54.3,
         50.0, 45.3, 63.4, 53.9, 65.5, 66.6, 70.0, 52.4, 38.6, 46.1, 44.4,
         60.7, 56.4)
```

```
t.test(data, mu = 50, alternative = ‘greater’)
```



CORRELATION

COVARIANCE

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

- Sign of covariance:
  - Positive covariance: the two variables move together
  - Negative covariance: the two variables move inversely
- Magnitude of covariance: not easy to interpret

CALCULATING COVARIANCE

Given two continuous variables x and y, covariance between the two variables is given by:

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

Steps:

1. Form a table with these columns:

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

2. Add all the values from $(x - \bar{x}) * (y - \bar{y})$ column

3. Divide the sum from step 2 by the number of observation (x)

(PEARSON'S) CORRELATION

Correlation coefficient measures the strength of the linear relationship between two quantitative variables.

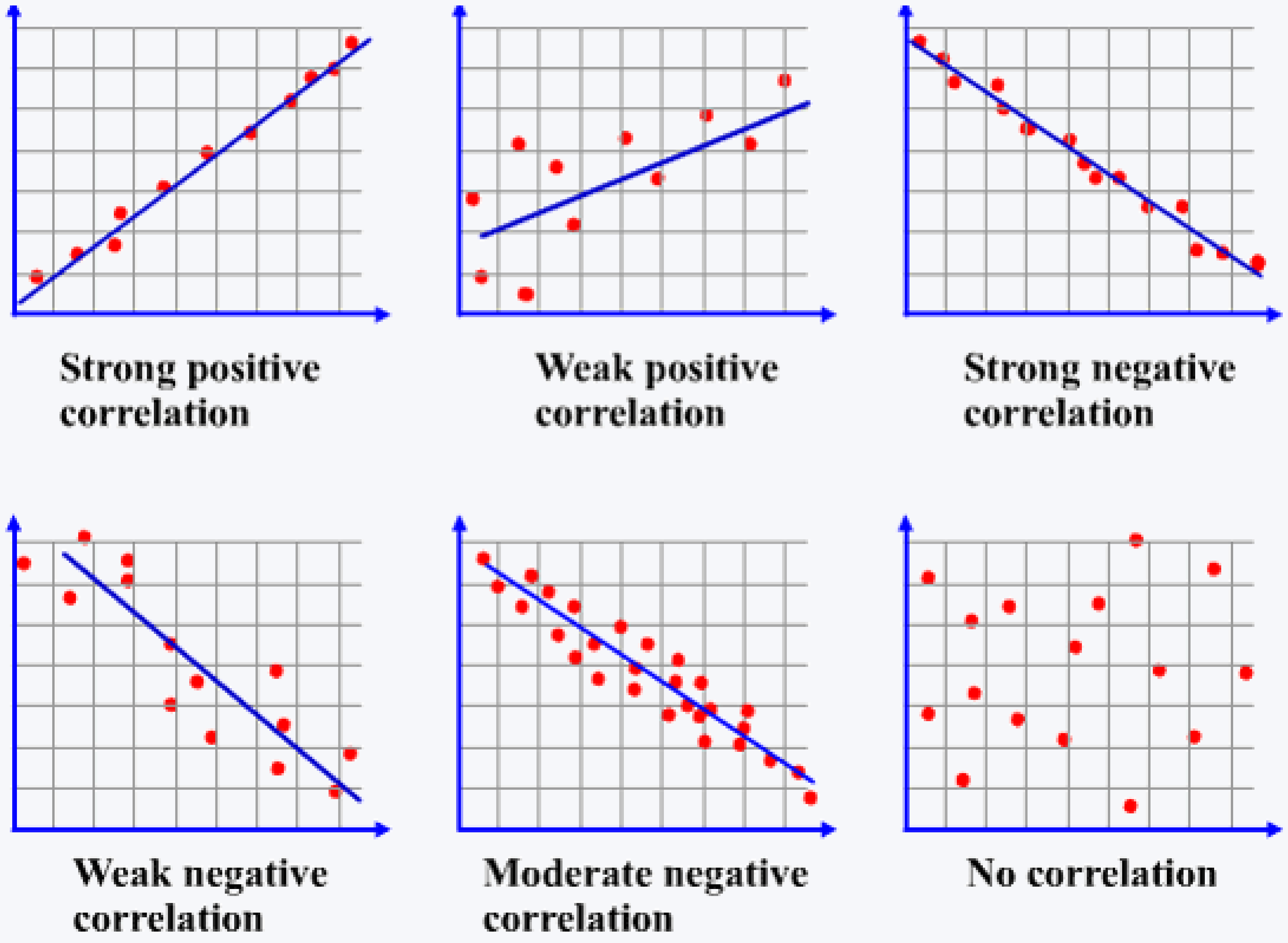
It is a normalized version of covariance.

→ The figure of correlation lies between [-1, +1], whereas covariance can take any value

CORRELATION ASSUMPTIONS

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

INTERPRETING CORRELATION FIGURES



Task:

Can you guess the correlation coefficients correctly based on the graphs?

Let‘s see who can do it better, students or tutor!

www.guessthecorrelation.com

CALCULATING CORRELATION

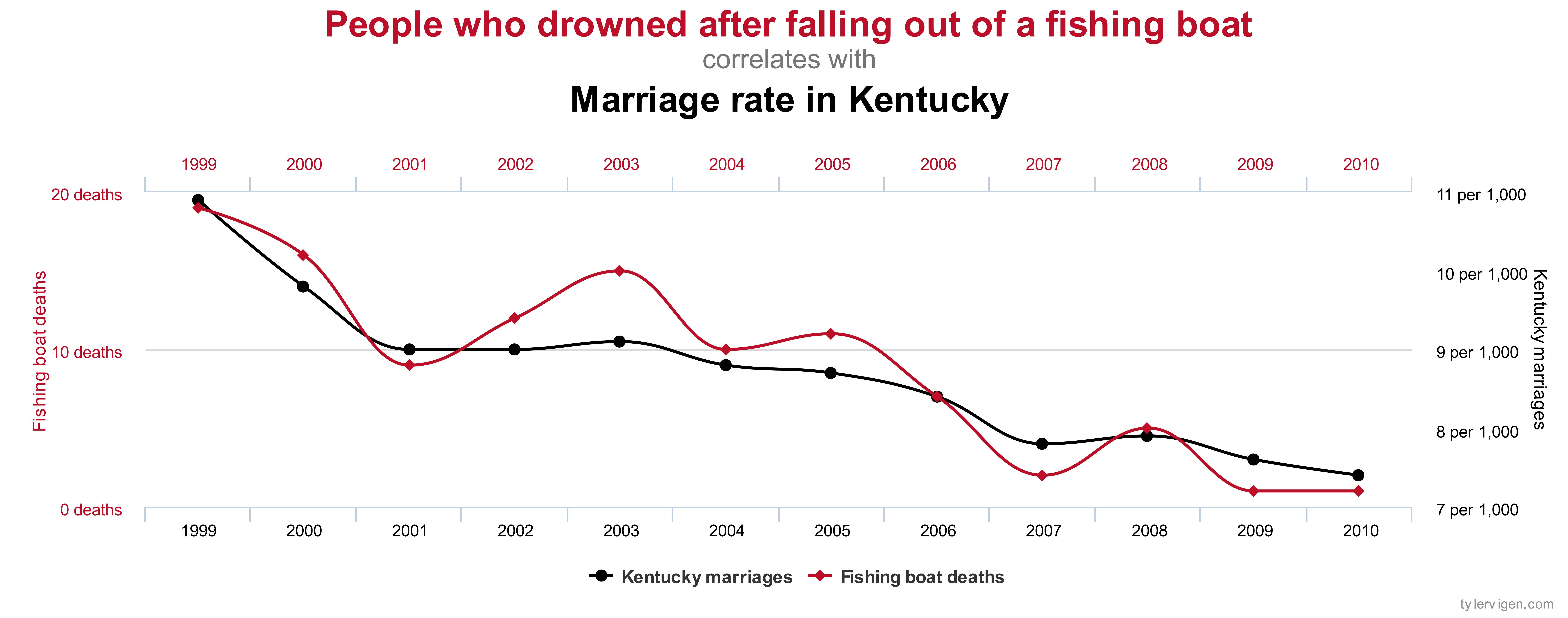
$$cor(x, y) = r = \frac{cov(x, y)}{s.d.(x) * s.d.(y)}$$

Steps:

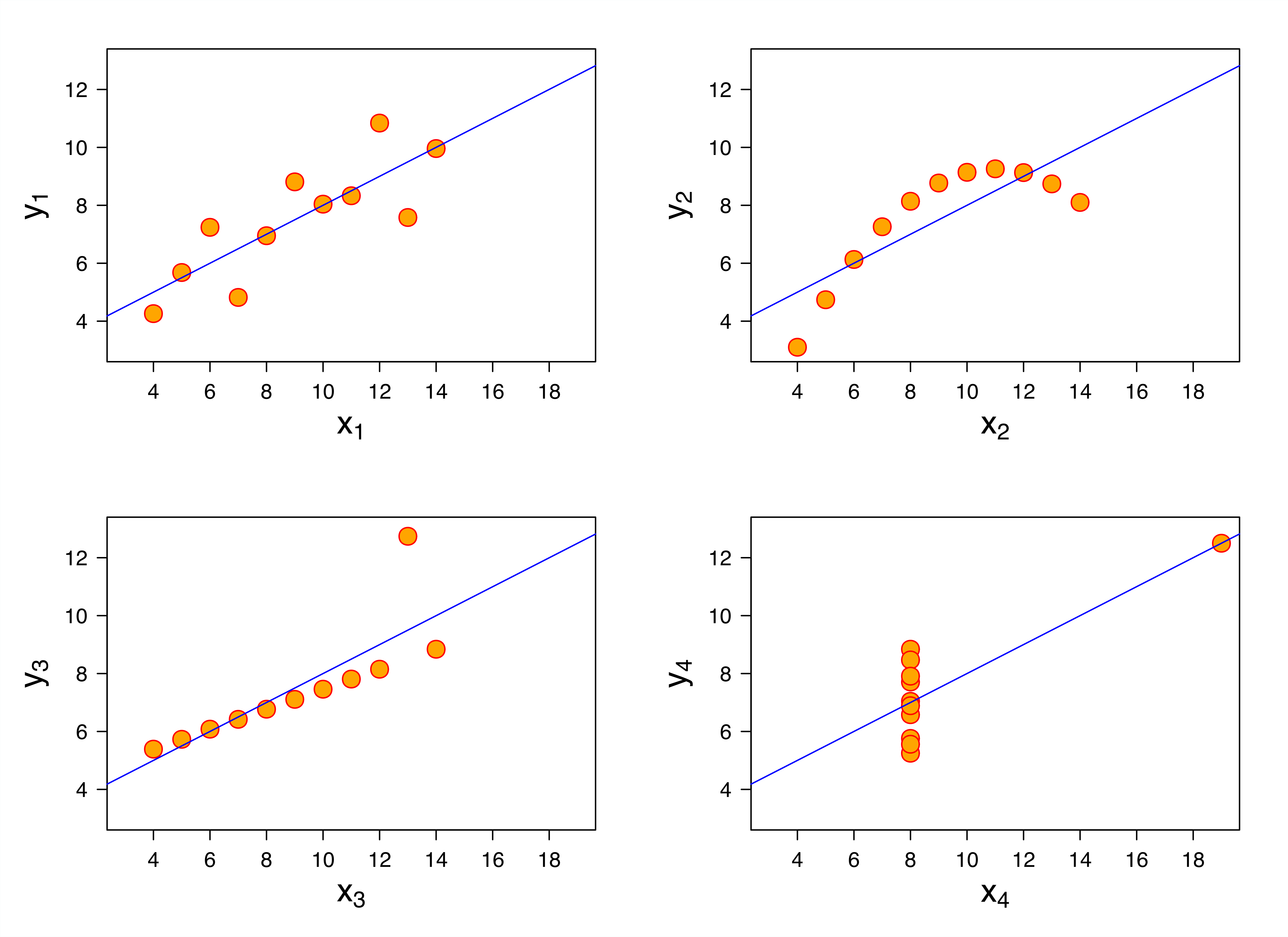
1. Calculate the covariance
2. Calculate standard deviation for both the variables
3. Divide the covariance figure by a multiple of the two standard deviations

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

SPURIOUS CORRELATIONS



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

CORRELATION vs CAUSATION

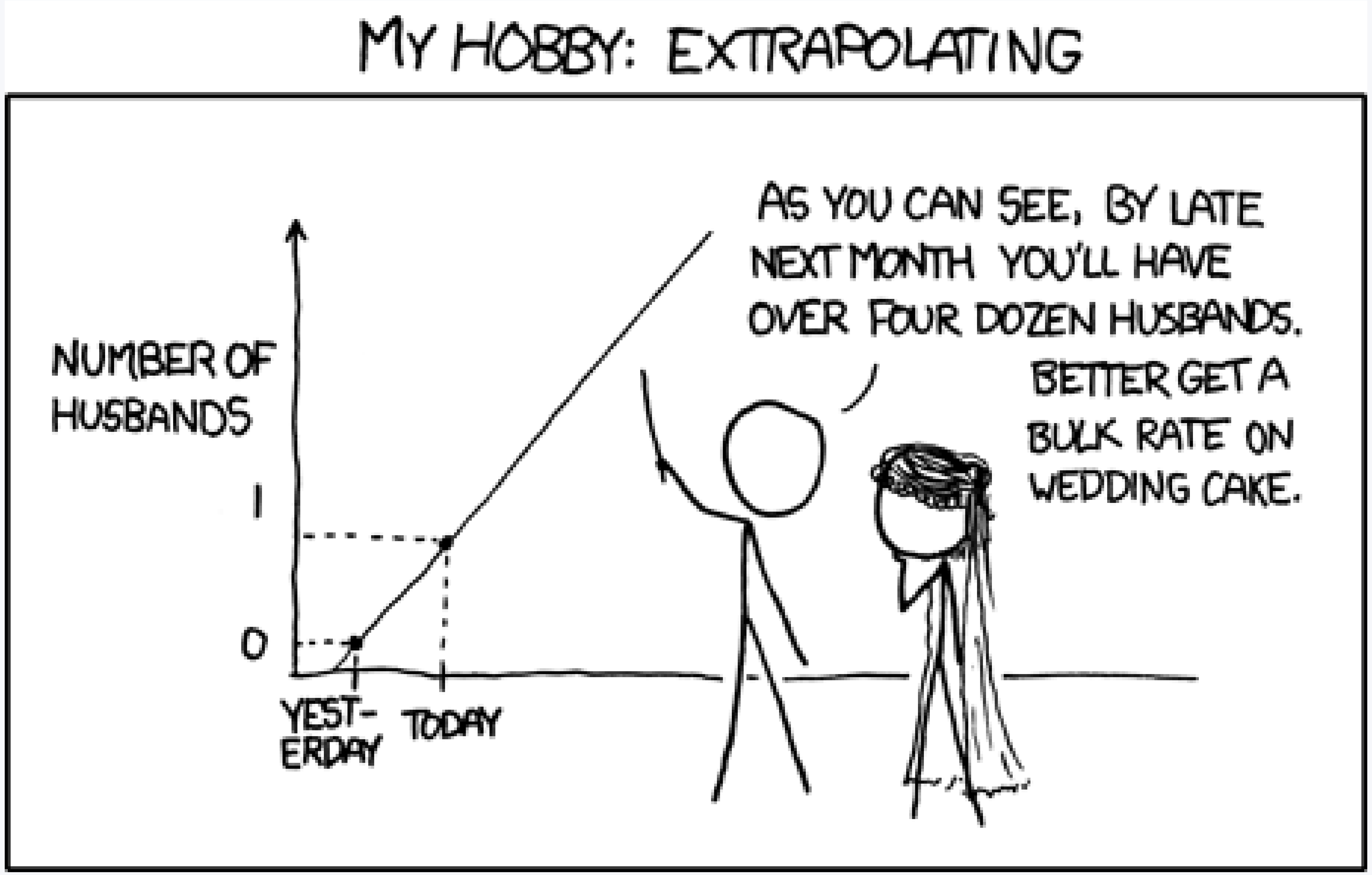
Correlation ≠ Causation

Correlation

"X and Y tend to be observed at the same time"

Causality:

"X causes Y"

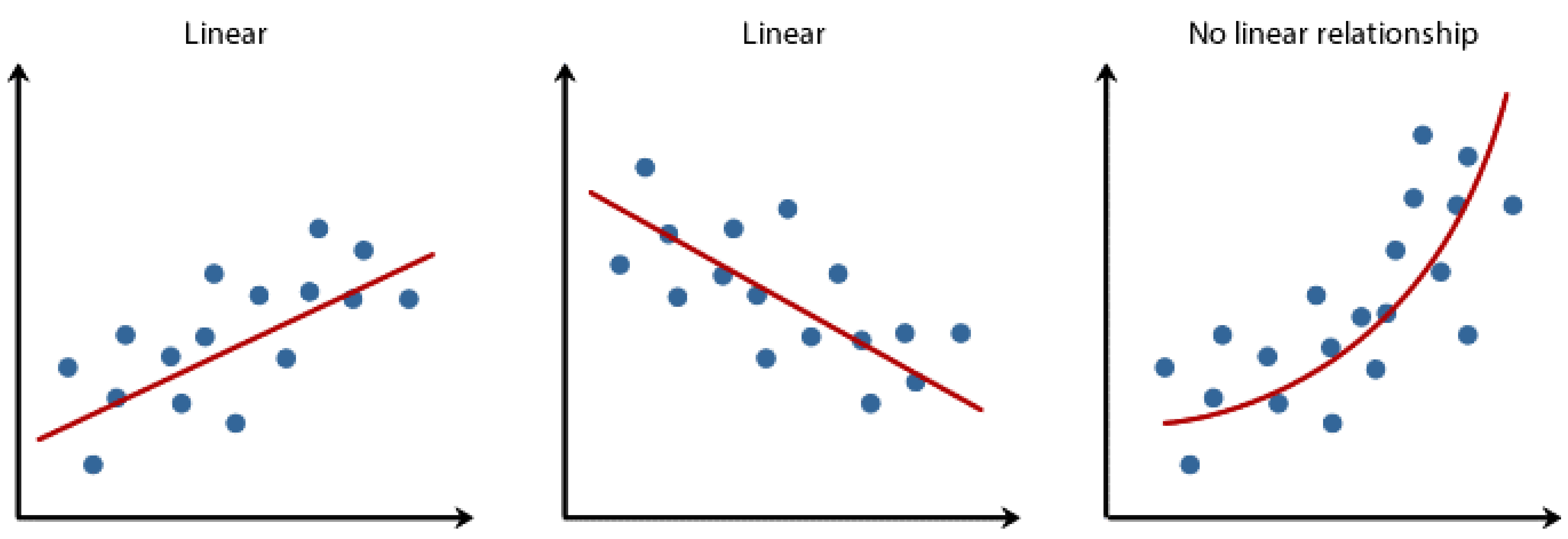

MY HOBBY: EXTRAPOLATING
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY

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:

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

Error *or* Residuals

Dependent variable

Regression coefficients

Independent variable

REGRESSION $\left(y = \beta_0 + \beta_1 x + \varepsilon \right)$

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 or q-q plot
3. Homoskedasticity of residuals → plot of residuals

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

GOODNESS OF FIT OF REGRESSION MODEL

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

R-Squared is the statistic that 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.

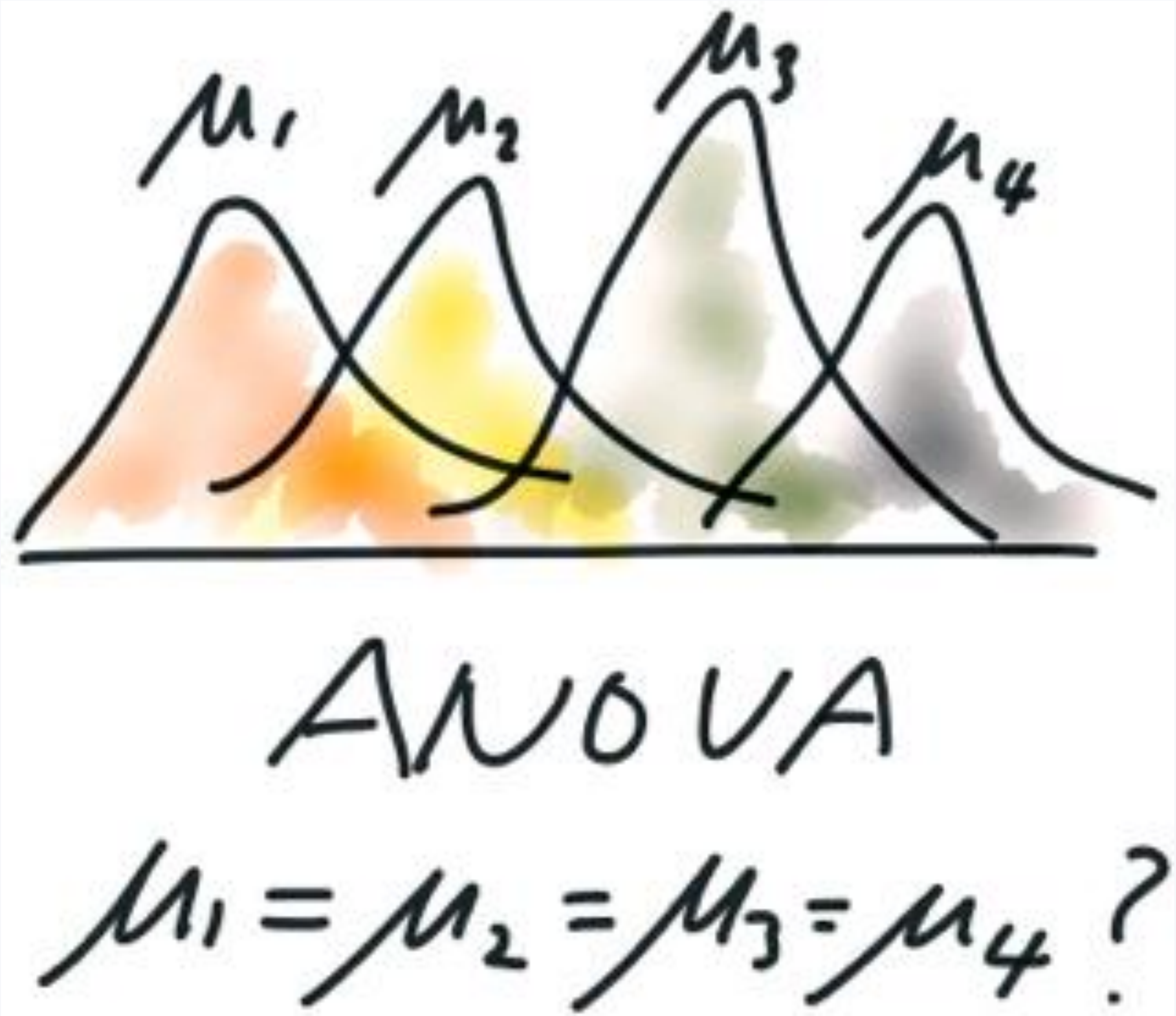
For simple regression model (only 2 variables), R-squared is also the squared value of correlation figure r.

REGRESSION CALCULATION

Performing regression by hand:

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

We will do more in R later.


μ_1
μ_2
μ_3
μ_4
ANOVA
$\mu_1 = \mu_2 = \mu_3 = \mu_4$?

ANOVA

Allows for comparison of means of more than two groups/categories.

- Remember: we used t-Test to compare means of 2 groups

$H_0: The\ means\ of\ all\ groups\ under\ consideration\ are\ equal$

$H_1: The\ means\ are\ not\ all\ equal$

ANOVA TERMINOLOGIES

- Factors
  = explanatory or independent variables
- Response
  = the dependent variable
- One-way ANOVA
  = one factor with two or more levels
- Two/Three-way Anova
  = two or more factors with two or more levels
- Factorial design
  = replication of each combination of levels in a multi-way ANOVA
  (enables study interaction of variables)

ANOVA ASSUMPTIONS

- Subjects are chosen via random sampling
- Response variable is normally distributed
- Population variance is the same across different groups (means can be different)

1 WAY ANOVA

- ‘yield.txt’ data set uploaded on MyStudy
- 2 Columns: Yield and Soil-Type

Response Variable: Yield

Factors: Soil

$k = 3$ soil types
$n = 10$

d.f. = $k*(n-1) = 3*(10 - 1) = 27$

Degrees of freedom (d.f.) is the sample size minus the number of parameters estimated from the data.

ANOVA RESULT FROM R

```
> model <- lm(yield ~ soil)
> anova(model)
Analysis of Variance Table
Response: yield
          Df   Sum Sq   Mean Sq  F value  Pr(>F)
soil      2    99.2     49.600   4.2447   0.02495 *
Residuals 27   315.5    11.685   ---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F - ratio = \frac{SS_{soil}/df_{soil}}{SS_{resid}/df_{resid}}$$

(in R:) Critical value at $\alpha = 0.05$ = **`qf(0.95, 2, 27) = 3.35`**

SUM OF SQUARES

- **SSY** = Total variation
  = Total Sum of Squares (also known as SST)
  = Total variation in the observation = SSA + SSE
- **SSA** = Explained variation
  = Sum of squares of differences between individual treatment means and the overall mean
- **SSE** = Unexplained variation
  = Error sum of squares
  = Sum of squares of the differences between data point and individual treatment means

SSA = SSY - SSE

SSA = SSY - SSE

- **SSY** = Total variation
- **SSA** = Explained variation
- **SSE** = Unexplained variation

- You can convert Sum of Squares into variances by dividing them by their **degrees of freedom**.

You want SSE to be smaller than SSY in your experiments.

R-Session

1. Download the R file for today
2. Download data files named `yield.txt`
3. (Remember where you downloaded the data files)
4. Open the downloaded R file (not the data file)

PLAN FOR NEXT WEEK

That's it for today! :-)

Next week, we are going to discuss:

1. ANOVA Cont.
2. Experiments and Research Papers

If you want to reach me, mail me at:

`prabesh.dhakal@stud.leuphana.de`