Measures of association

Petr Ocelík

MEBn5033 Introduction to Quantitative Data Analysis

15th December 2020

Outline

- Jaccard’s similarity index
- Kendall’s tau correlation coefficient
- Pearson’s r correlation coefficient
- R time!

Measures of association (MA)

- There are **many measures of association (MA)**
- Correlation coefficients represent just one of the subsets of the MA

- Do not reduce MA to correlation or even Pearson's *r*

- **MA** measure the size (and/or direction) of associations between the variables of interest
- MA typically range within <0,1> or <-1,1> intervals

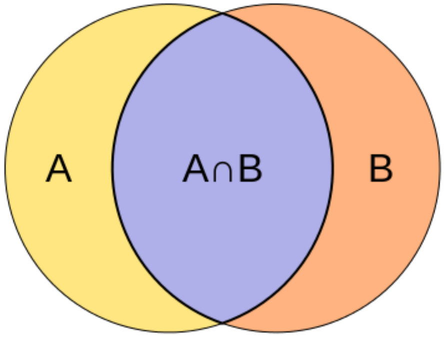
- Correlation is not a causation
- Causation can be based on different types of associations

Measures of association (MA)

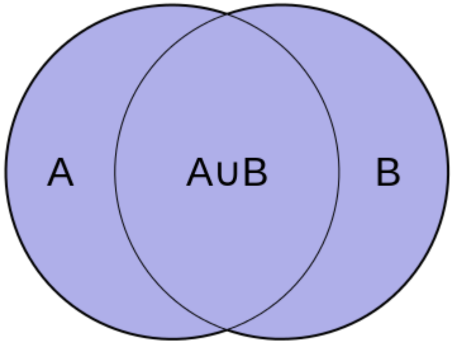
level of measurement	coefficient
nominal	**Jaccard's index**
ordinal	Kendall's tau
metric (interval & ratio)	Pearson's rho

Jaccard (similarity) index

- J used for **categorical binary data** (e.g., gender)
- Measures similarity between two samples



		sample B	
		present	absent
sample A	present	A ∩ B	A – B
	absent	B – A	∉ A ∪ B



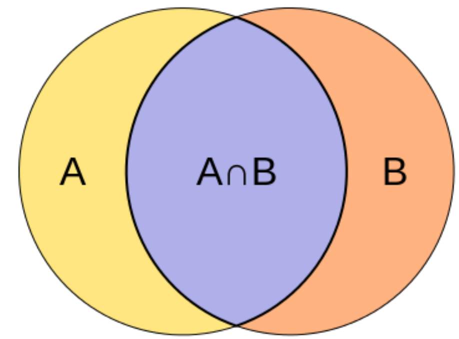
- J = the **size of the intersection** (A ∩ B) by the **size of the union** (A + B = A ∪ B) of the samples
- J = A ∩ B / (A ∪ B)
- Does not account for observations missing in both samples (∉ A ∪ B)

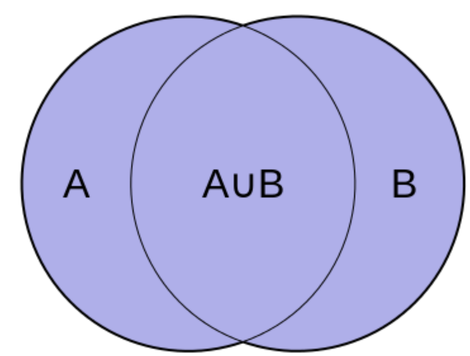
wikimedia commons

Jaccard (similarity) index: example

- Let's assume there are **1091 int. environ. NGOs** worldwide
- What is the similarity of the CR and Germany based on presence/absence of int. environ. NGOs?

IENGOs		Czech Republic	
		present	absent
Germany	present	21 ($A \cap B$)	56 ($A - B$)
	absent	13 ($B - A$)	1001 ($\notin A \cup B$)





- $J = A \cap B / (A \cup B)$
- $J = 21 / (21 + 56 + 13) = 21 / 90 = \mathbf{0.23} = 23\%$

wikimedia commons

Measures of association (MA)

level of measurement	**coefficient**
nominal	Jaccard's index
ordinal	**Kendall's tau**
metric (interval & ratio)	Pearson's rho

Kendall's tau correlation coefficient

- **Kendall's tau** (τ) used for ordinal data (e.g., attitude scales)

- **A non-parametric** measure of association between two ordinal variables
- Accommodates also small samples and many values with the same order/ranking

- **Ranges within <-1,1>**
  - Perfect agreement (variables are identically ordered) = 1
  - Perfect inversion (variables are ordered in exactly reversed way) = -1
  - No ordered relationship = 0

- KT represents the degree of concordance between two ordinal variables
  - $\boldsymbol{\tau_a}$ does not correct for tied values
  - $\boldsymbol{\tau_b}$ corrects for tied values

- **E.g.:** is there an ordered association between the income level and acceptance of climate change?

Hypothesis testing: Kendall's τ

- **H0:** There is no ordered association between variables X and Y; **H0:** $\tau <= 0$
- **HA:** There is positive ordered association between variables X and Y; **HA:** $\tau > 0$

- **H0:** There is no ordered association between level of income (X) and level of acceptance of climate change (Y); **H0:** $\tau <= 0$
- **HA:** There is positive ordered association between level of income (X) and level of acceptance of climate change (Y); **HA:** $\tau > 0$

- Critical value (CV) τ sets threshold (level of test) between **statistically in/significant values of the test statistic** at a pre-defined level of **α** (0.05)
- Observed test statistic τ **>** CV τ**?**

- **p-value:** indicates probability of observing such, or even more extreme, **value of the test statistic** (τ) if **H0 holds**

cases (N)	X: income	Y: acceptance
A	1 (low)	1 (disagree)
B	2 (middle)	1 (disagree)
C	2 (middle)	2 (neutral)
D	3 (high)	3 (agree)

- We have **n*(n – 1)/2 pair combinations**; i.e., 4*(4-1)/2 = **6**
- Specifically: (A,B), (A,C), (A,D), (B,C), (B,D), (C,D)

- **Concordance:** $X_i > X_j$ AND $Y_i > Y_j$; or: $X_i < X_j$ AND $Y_i < Y_j$
- **Discordance:** $X_i > X_j$ AND $Y_i < Y_j$; or: $X_i < X_j$ AND $Y_i > Y_j$
- **Neither (tied values):** $X_i = X_j$ OR $Y_i = Y_j$
  - Pair (A,B) = neither (**tied**); ; $X_A < X_B$ & $\mathbf{Y_A = Y_B}$
  - Pair (A,C) = concordant; $X_A < X_C$ & $Y_A < Y_C$
  - Pair (A,D) = concordant; $X_A < X_D$ & $Y_A < Y_D$
  - Pair (B,C) = neither (**tied**); $\mathbf{X_B = X_C}$ & $Y_B < Y_C$
  - Pair (B,D) = concordant; $X_B < X_D$ & $Y_B < Y_D$
  - Pair (C,D) = concordant; $X_C < X_D$ & $Y_C < Y_D$

cases (N)	X: income	Y: acceptance
A	1 (low)	1 (disagree)
B	2 (middle)	1 (disagree)
C	2 (middle)	2 (neutral)
D	3 (high)	3 (agree)

- We have **n*(n – 1)/2 pair combinations**; i.e., 4*(4-1)/2 = **6**
  - Pair (A,B) = neither (tied)
  - Pair (A,C) = concordant
  - Pair (A,D) = concordant
  - Pair (B,C) = neither (tied)
  - Pair (B,D) = concordant
  - Pair (C,D) = concordant

τ_a = (# of concordant pairs – # of discordant pairs) / # of all pairs

$\tau_a = n_c - n_d / ((n * (n - 1)) / 2)$

$\tau_a = 4 - 0 / ((4 * (4 - 1)) / 2) = 4 / 6 =$ **0.66**

- We have **n*(n – 1)/2 pair combinations**; i.e., 4*(4-1)/2 = **6**
  - Pair (A,B) = neither (tied)
  - Pair (A,C) = concordant
  - Pair (A,D) = concordant
  - Pair (B,C) = neither (tied)
  - Pair (B,D) = concordant
  - Pair (C,D) = concordant

$\boldsymbol{\tau_a}$ = (# of concordant pairs – # of discordant pairs) / # of all pairs

$\boldsymbol{\tau_b = (n_c - n_d) / \mathrm{sqrt}((N - n_1) * (N - n_2))}$

$N = (n * (n - 1))/2$; total # of pairs

$n_1 = t_1 * (t_1 - 1))/2$; t_1 = # of tied values in the first set/variable

$n_2 = t_2 * (t_2 - 1))/2$; t_2 = # of tied values in the second set/variable

$n_1 = 2 * (2 - 1)/2 = 1$ (income var: middle/middle)

$n_2 = 2 * (2 - 1)/2 = 1$ (attitude var: disagree/disagree)

$\boldsymbol{\tau_b}$ = (4 – 0) / sqrt((6 - 1)*(6 - 1)) = 4 / sqrt(25) = 4 / 5 = **0.8**

	Nominal α					
n	0.10	0.05	0.025	0.01	0.005	0.001
4	1.000	1.000	-	-	-	-
5	0.800	0.800	1.000	1.000	-	-
6	0.600	0.733	0.867	0.867	1.000	-
7	0.524	0.619	0.714	0.810	0.905	1.000
8	0.429	0.571	0.643	0.714	0.786	0.857
9	0.389	0.500	0.556	0.667	0.722	0.833
10	0.378	0.467	0.511	0.600	0.644	0.778
11	0.345	0.418	0.491	0.564	0.600	0.709
12	0.303	0.394	0.455	0.545	0.576	0.667
13	0.308	0.359	0.436	0.513	0.564	0.641
14	0.275	0.363	0.407	0.473	0.516	0.604
15	0.276	0.333	0.390	0.467	0.505	0.581
16	0.250	0.317	0.383	0.433	0.483	0.567
17	0.250	0.309	0.368	0.426	0.471	0.544
18	0.242	0.294	0.346	0.412	0.451	0.529
19	0.228	0.287	0.333	0.392	0.439	0.509
20	0.221	0.274	0.326	0.379	0.421	0.495
21	0.210	0.267	0.314	0.371	0.410	0.486
22	0.203	0.264	0.307	0.359	0.394	0.472
23	0.202	0.257	0.296	0.352	0.391	0.455
24	0.196	0.246	0.290	0.341	0.377	0.449
25	0.193	0.240	0.287	0.333	0.367	0.440
26	0.188	0.237	0.280	0.329	0.360	0.428
27	0.179	0.231	0.271	0.322	0.356	0.419
28	0.180	0.228	0.265	0.312	0.344	0.413
29	0.172	0.222	0.261	0.310	0.340	0.404

- n = # of observed pairs
- For **two-sided tests** = **α/2**

Decision on H0

- **Example:** Kendall's tau $\boldsymbol{\tau_b}$ = **0.8**
- Critical value (α = 0.05) = **0.73**

- Since $\boldsymbol{\tau_b}$ **value** = **0.8 > CV (α = 0.05)** = **0.73**, analogically

- **We reject H0:** Correlation coefficient τ is zero, *there is no ordered association between X and Y*; **H0:** $\tau <= 0$

- And support **HA:** Correlation coefficient τ is positive, *there is a positive ordered association between X and Y*; **HA:** $\tau > 0$

- Thus, there is ordered association between level of income and acceptance of climate change

Measures of association (MA)

level of measurement	coefficient
nominal	Jaccard's index
ordinal	Kendall's tau
metric (interval & ratio)	**Pearson's rho**

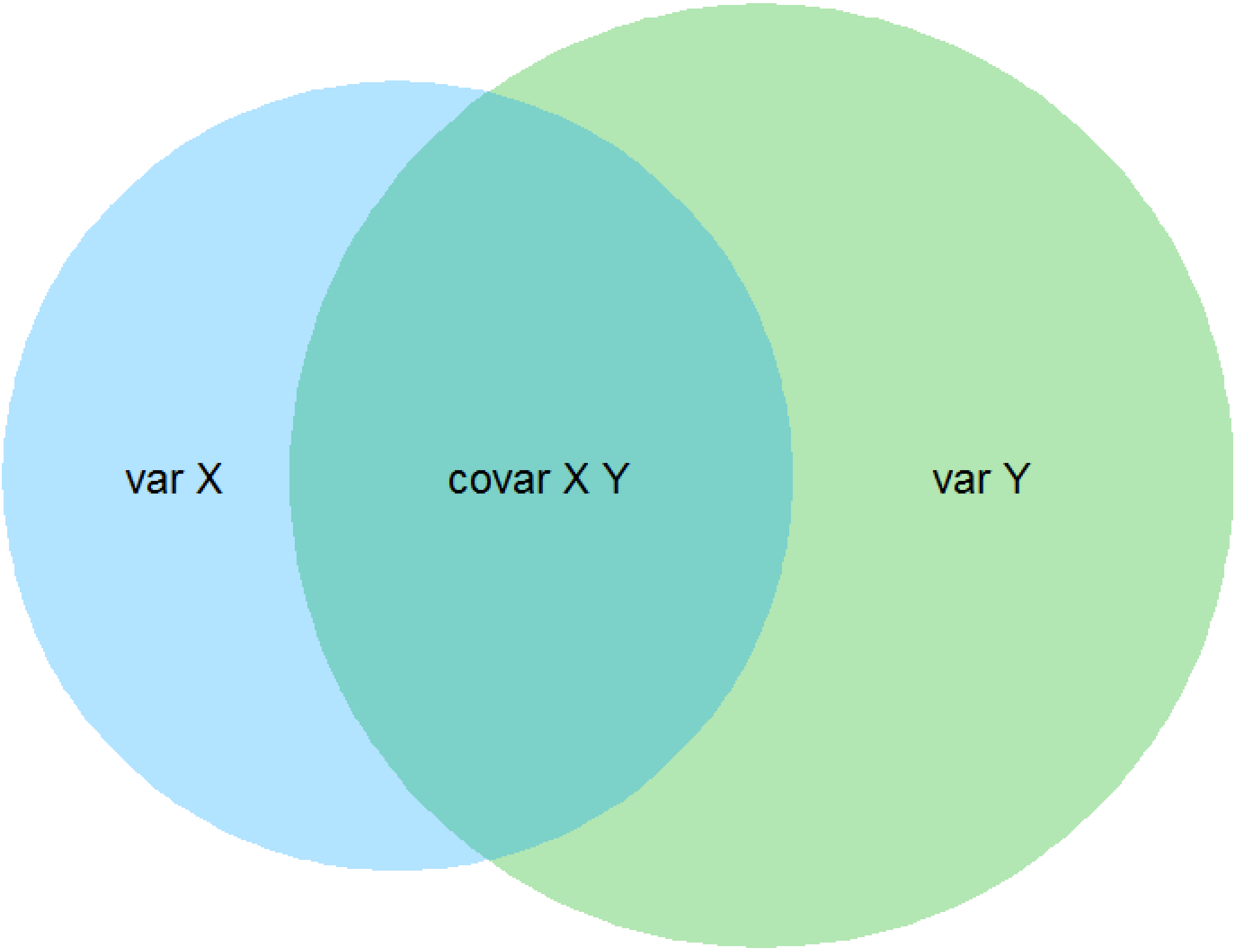
Pearson’s r correlation coefficient

- Pearson’s product-moment correlation coefficient (**r**)
- Pearson’s r measures the **strength and direction of the linear relationship between two variables**

- Ranges within <-1,1>
  - Perfect positive linear relationship = 1
  - Perfect negative linear relationship = -1
  - No linear relationship = 0

- Value does not depend on variables’ units

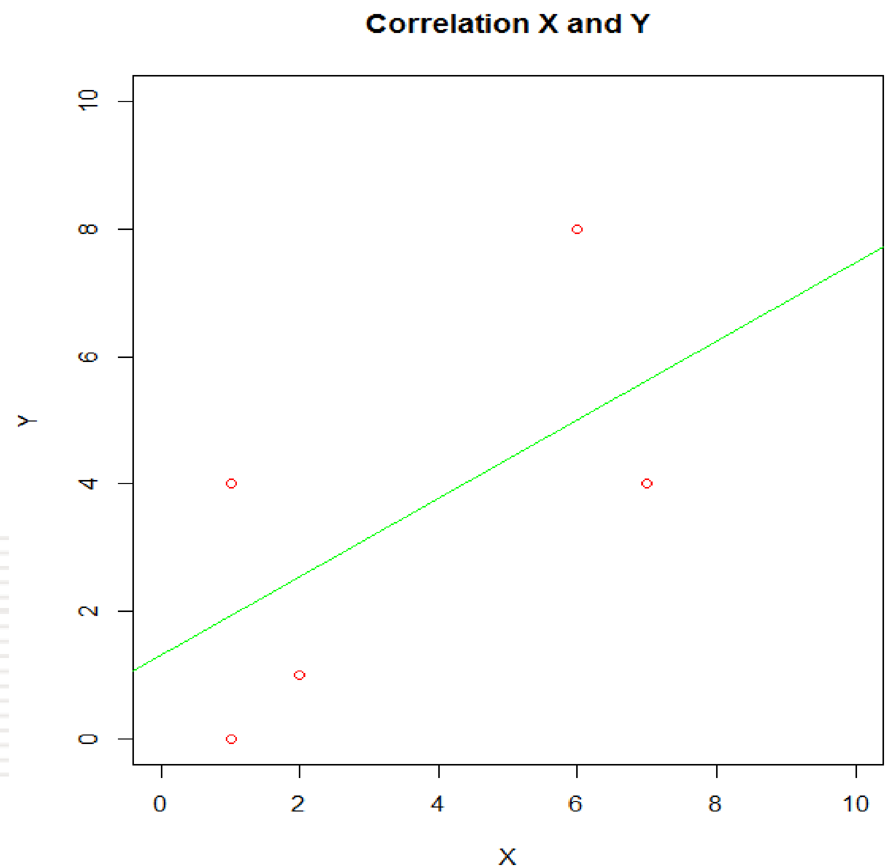
- r = covariance / combined total variance

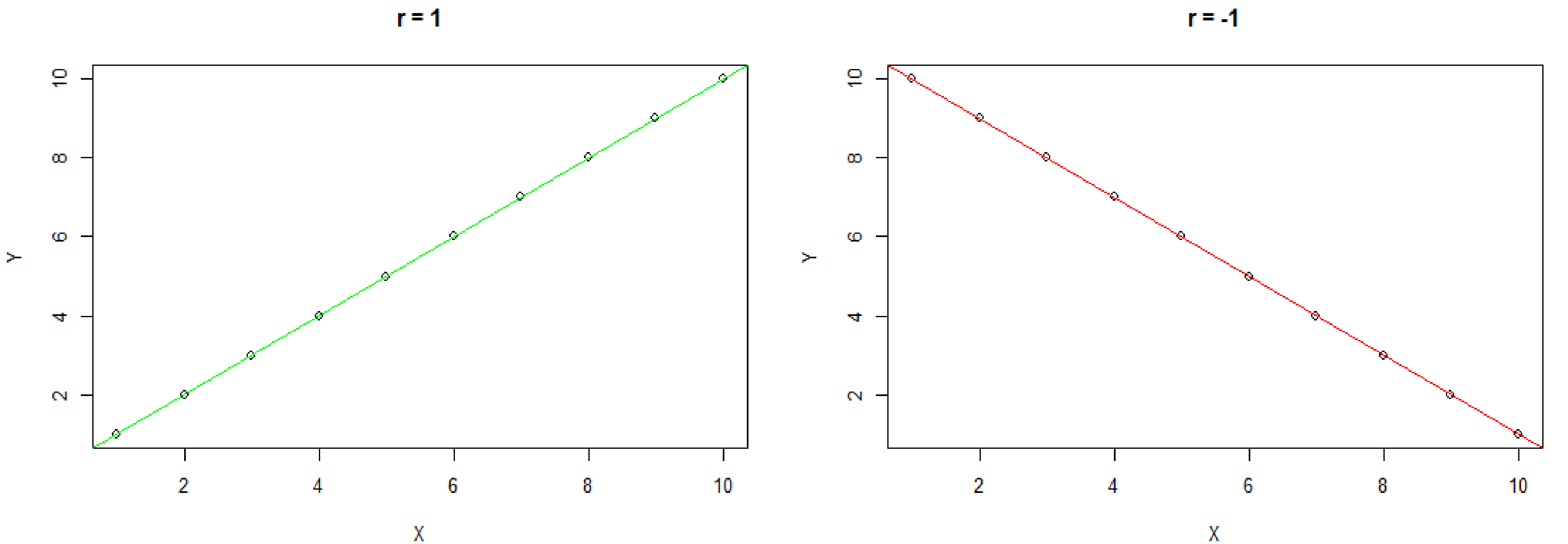


r = covariance(X, Y) / total variance(X, Y)

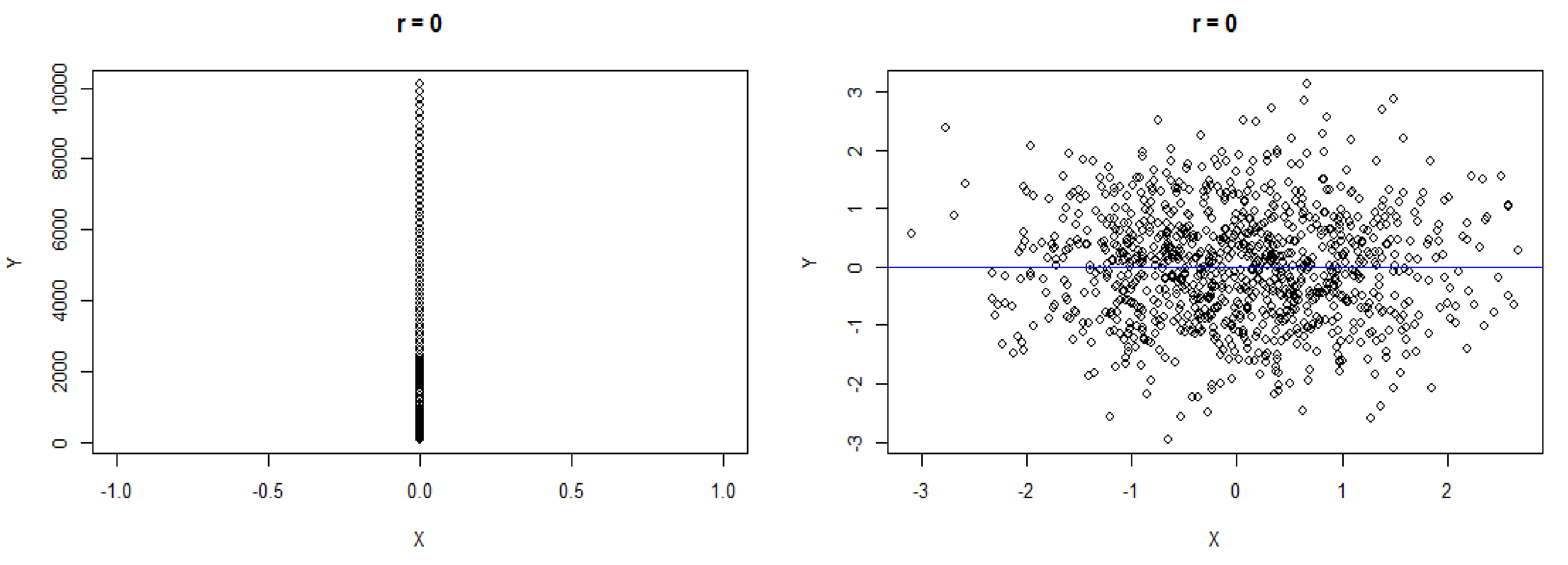
r = cov(X, Y) / sqrt(var(X) * var(Y))

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}}$$





http://guessthecorrelation.com/



- Kellstedt & Whitten (2013). Example of correlation between incumbent party vote share (Y) and GDP change of the finished electoral period (X).
- Statistically significant correlation r = 0.574

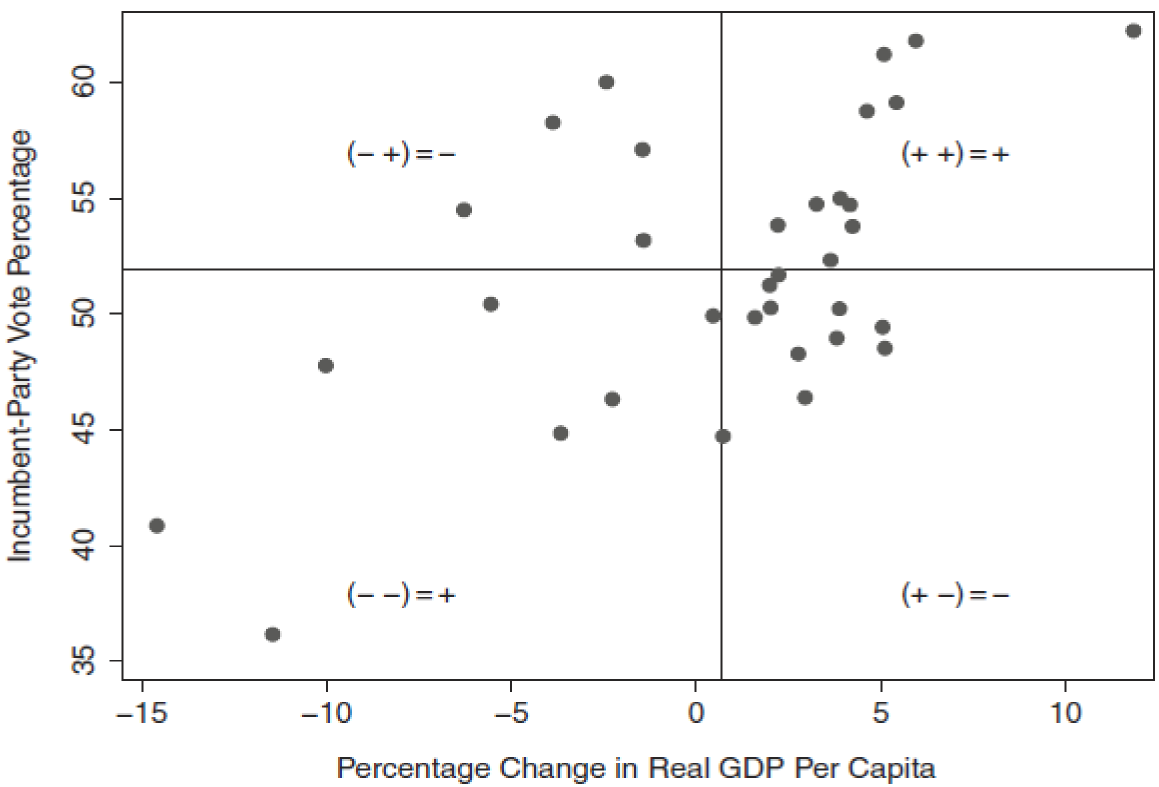


Figure 7.4. Scatter plot of change in GDP and incumbent-party vote share with mean-delimited quadrants.

Pearson’s r: description

Pearson’s r strength	Description
0.00–0.19	very weak
0.20–0.39	weak
0.40–0.59	moderate
0.60–0.79	strong
0.80–1.00	very strong

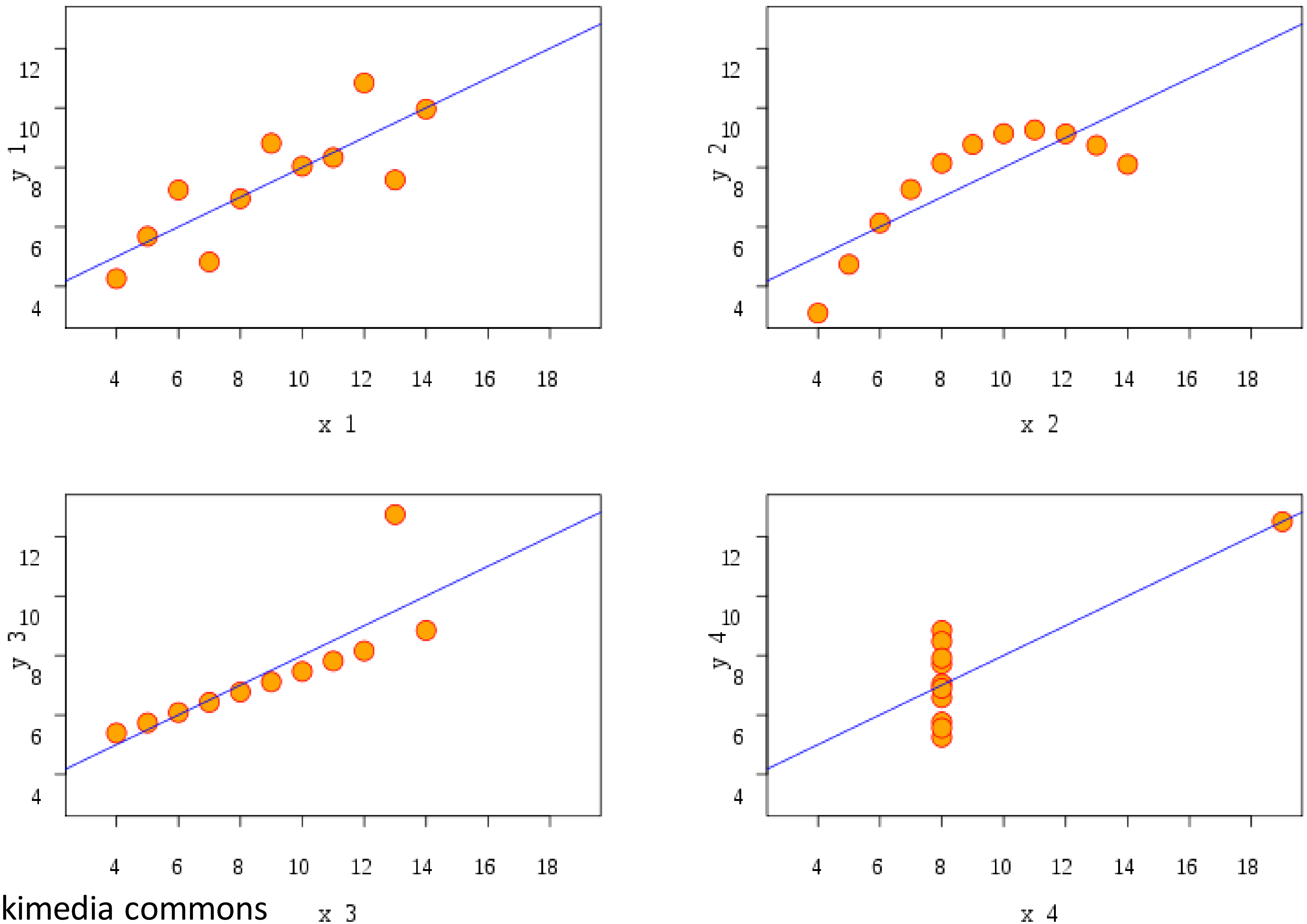
Always context dependent!

Evans 1996

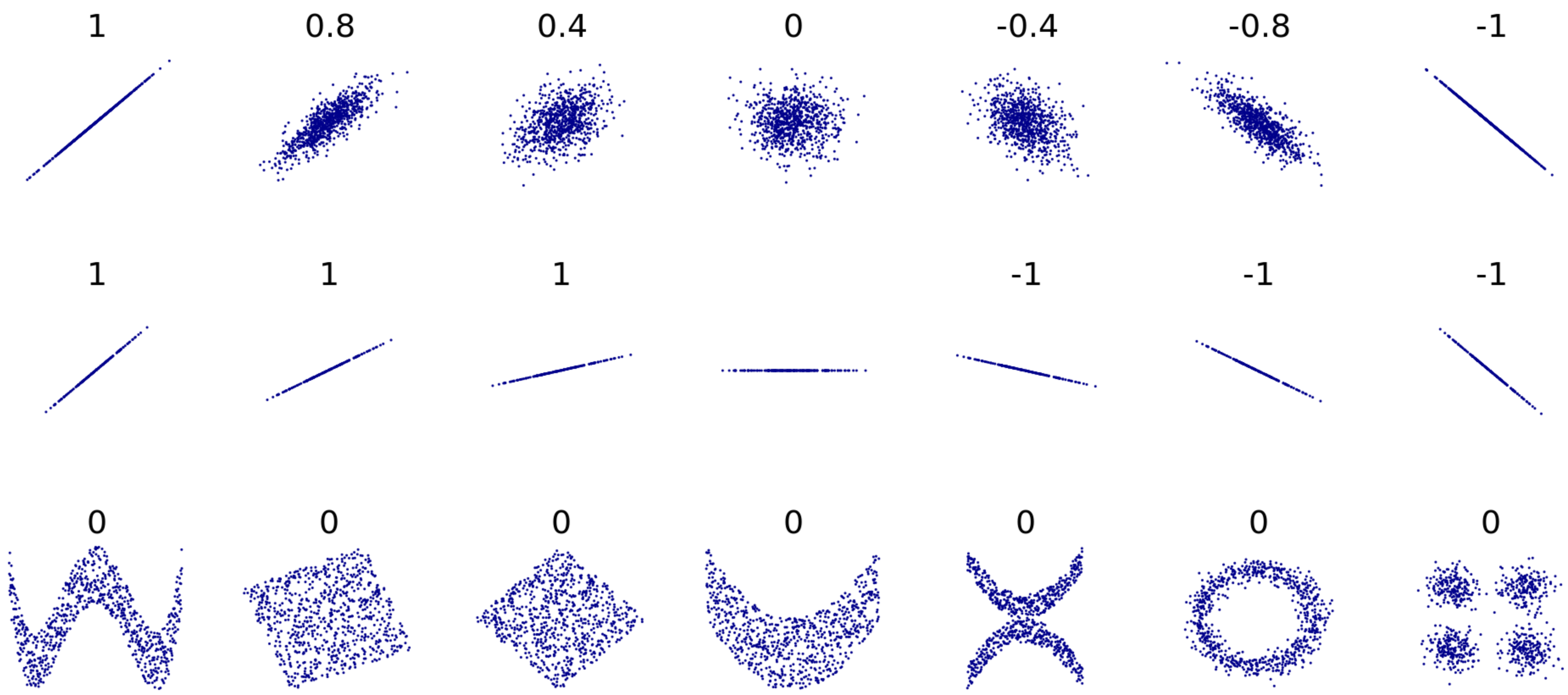
Pearson’s r: assumptions and limitations

- Metric level of measurement
- **Linear relationship between X and Y**
- Homoscedasticity (independence of variance)
- Sensitivity to outliers

Anscombe's quartet



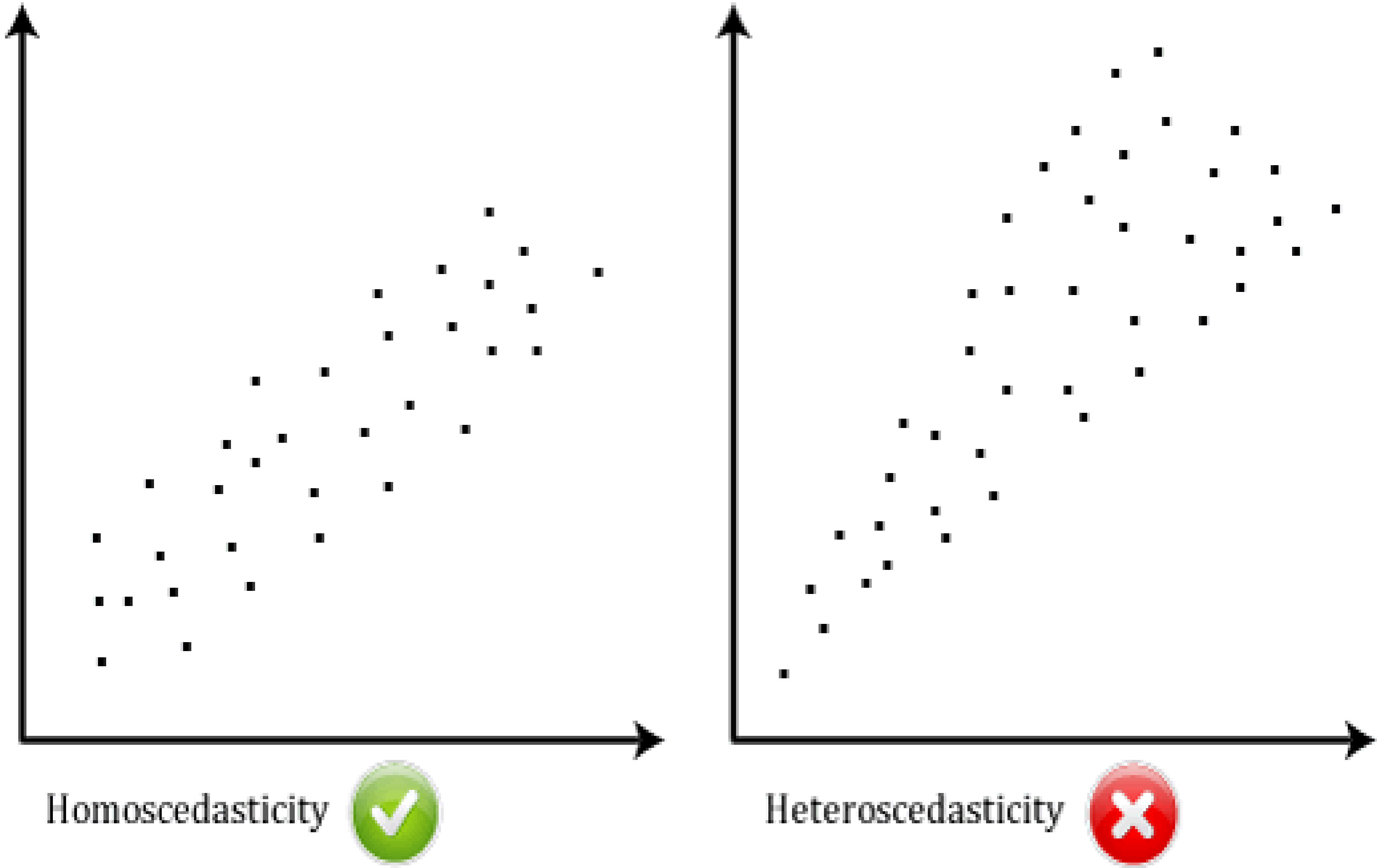
wikimedia commons



wikimedia commons

Pearson’s r: assumptions and limitations

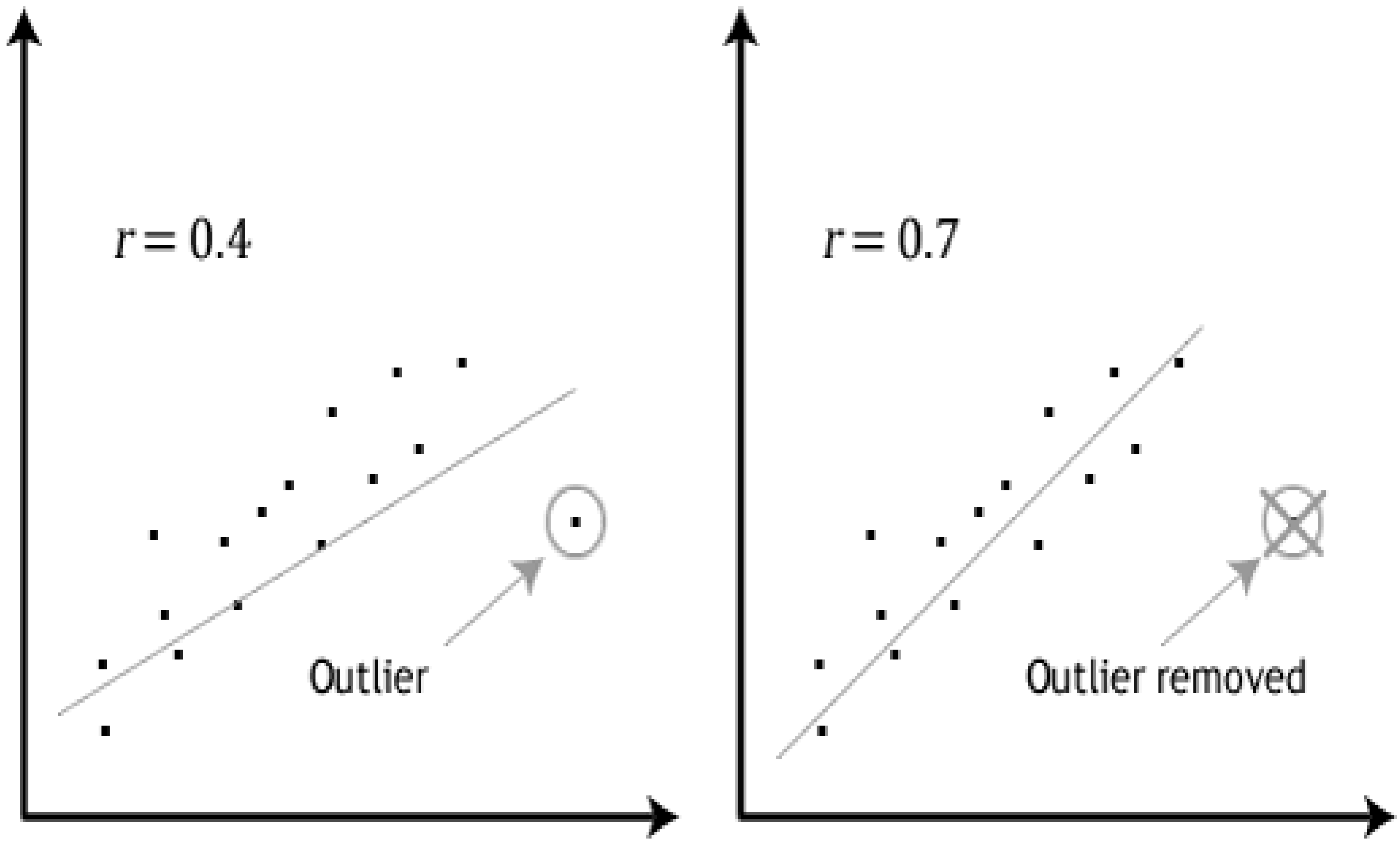
- Metric level of measurement
- Linear relationship between X and Y
- **Homoscedasticity (independence of variance)**
- Sensitivity to outliers



stats.stackexchange.com

Pearson's r: assumptions and limitations

- Metric level of measurement
- Linear relationship between X and Y
- Homoscedasticity (independence of variance)
- **Sensitivity to outliers**



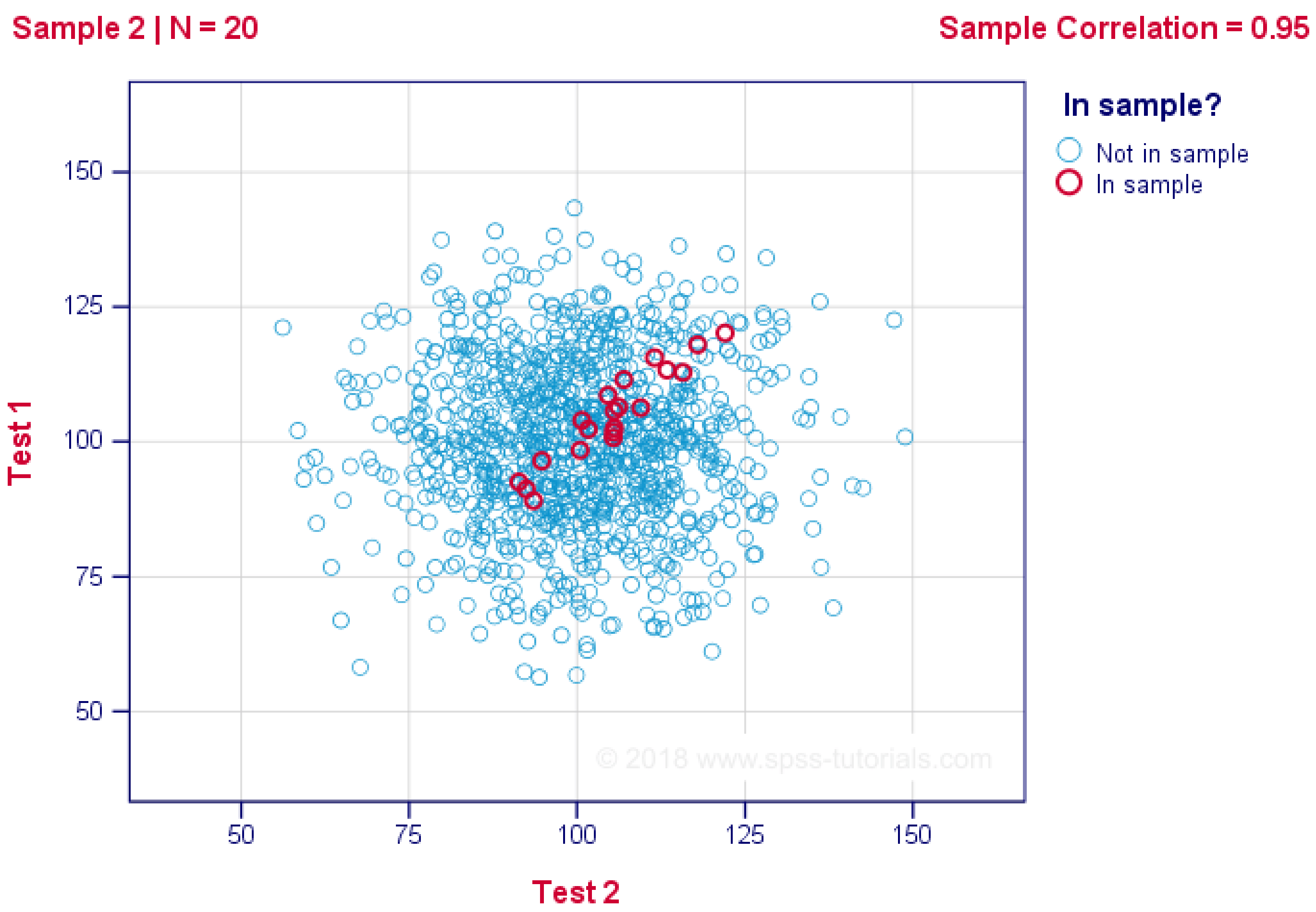
statistics.leard.com

Hypothesis testing: Pearson's r

- H0: There is no linear relationship (correlation *r*) between X and Y
- HA: There is a linear relationship (correlation *r*) between X and Y

- H0: $r = 0$; $r <= 0$; $r >= 0$
- HA: $r \neq 0$ (two-sided hypothesis)
- HA: $r > 0$ (one-sided hypothesis, positive correlation)
- HA: $r < 0$ (one-sided hypothesis, negative correlation)

Hypothesis testing: Pearson's r

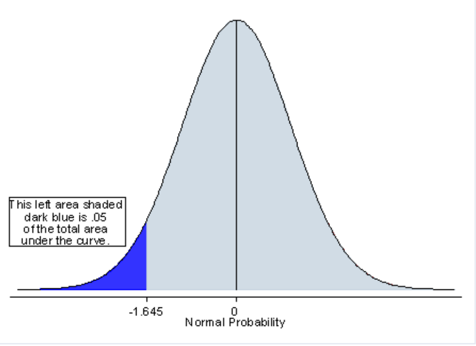


SPSS tutorial 2018

Hypothesis testing: Pearson's r

- **Theory:** voting behavior is influenced by socio-cultural cleavages (Norris & Inglehart 2019; Lipset & Rokkan 1967)

- **H0:** There is *no* correlation between Obama's election results in 2012 (X) and share of protestants (Y); $r(x, y) >= 0$
- **HA:** There *is a negative* correlation between Obama's election results in 2012 (X) and share of protestants (Y); $r(x, y) < 0$



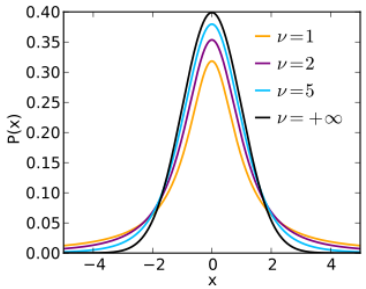
Hypothesis testing: Pearson's r

- **H0:** There is *no* correlation between Obama's election results in 2012 (X) and share of protestants (Y); $r(x, y) >= 0$
- **HA:** There *is a negative* correlation between Obama's election results in 2012 (X) and share of protestants (Y); $r(x, y) < 0$

- **Data:** 50 observations (U.S. states)

- How many **degrees of freedom?** For *r*: **n - 2**, i.e.: 50 - 2 = **48**

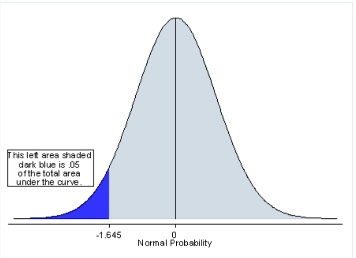
- **Testing level α** = 0.05 (5%) with corresponding critical **t-value** for **one-sided** negative hypothesis (α **=** 0.05, df = 48) = -**1.677**



Appendix: Critical Values Tables

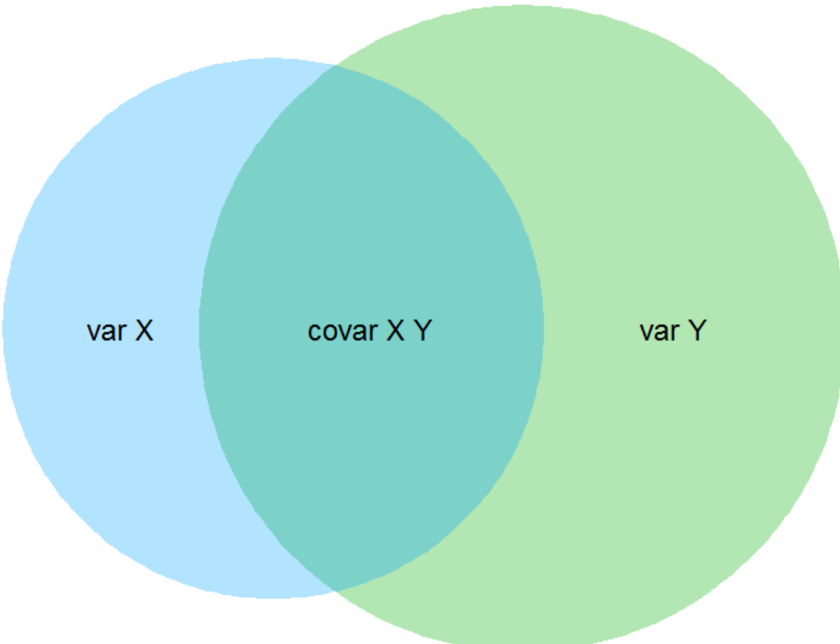
Degrees of Freedom (*df*)	80%	90%	95%	98%	99%
41	1.303	1.683	2.020	2.421	2.701
42	1.302	1.682	2.018	2.418	2.698
43	1.302	1.681	2.017	2.416	2.695
44	1.301	1.680	2.015	2.414	2.692
45	1.301	1.679	2.014	2.412	2.690
46	1.300	1.679	2.013	2.410	2.687
47	1.300	1.678	2.012	2.408	2.685
48	1.299	1.677	2.011	2.407	2.682
49	1.299	1.677	2.010	2.405	2.680
50	1.299	1.676	2.009	2.403	2.678
51	1.298	1.675	2.008	2.402	2.676
52	1.298	1.675	2.007	2.400	2.674
53	1.298	1.674	2.006	2.399	2.672
54	1.297	1.674	2.005	2.397	2.670
55	1.297	1.673	2.004	2.396	2.668
56	1.297	1.673	2.003	2.395	2.667
57	1.297	1.672	2.002	2.394	2.665
58	1.296	1.672	2.002	2.392	2.663
59	1.296	1.671	2.001	2.391	2.662
60	1.296	1.671	2.000	2.390	2.660
61	1.296	1.670	2.000	2.389	2.659
62	1.295	1.670	1.999	2.388	2.657
63	1.295	1.669	1.998	2.387	2.656
64	1.295	1.669	1.998	2.386	2.655
65	1.295	1.669	1.997	2.385	2.654
66	1.295	1.668	1.997	2.384	2.652
67	1.294	1.668	1.996	2.383	2.651
68	1.294	1.668	1.995	2.382	2.650
74	1.293	1.666	1.993	2.378	2.644
75	1.293	1.665	1.992	2.377	2.643
76	1.293	1.665	1.992	2.376	2.642
77	1.293	1.665	1.991	2.376	2.641
78	1.292	1.665	1.991	2.375	2.640
79	1.292	1.664	1.990	2.374	2.640
80	1.292	1.664	1.990	2.374	2.639
81	1.292	1.664	1.990	2.373	2.638
82	1.292	1.664	1.989	2.373	2.637
83	1.292	1.663	1.989	2.372	2.636

90% value for two-sided test, i.e. 95% for one-sided test



The table displays only positive t values. The Student's *t* distribution is symmetrical. It is thus unnecessary to list the same values for negative and positive *t* statistic. For HA: r < 0 we just add minus before.

- r = covariance X, Y / total variability X, Y



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}}$$

- Pearson’s r of Obama2012 and relig_prot = **-0.413**

Test statistic *t*

- **Does the *r*** (-0.413) **significantly differ from 0?**
- **Recall:** H0: r >= 0 ; 0 is assumed population average
- We use **t-test** for correlation coefficient *r* to find out.

- $t = \frac{signal}{noise}$; $t = \frac{r * \sqrt{n-2}}{\sqrt{1-r^2}}$; n = sample size

- $t = \frac{-0.413 * \sqrt{50-2}}{\sqrt{1-(-0.413)^2}} = \mathbf{-3.14}$

- **t-value** of Pearson's *r* is a **test statistic**

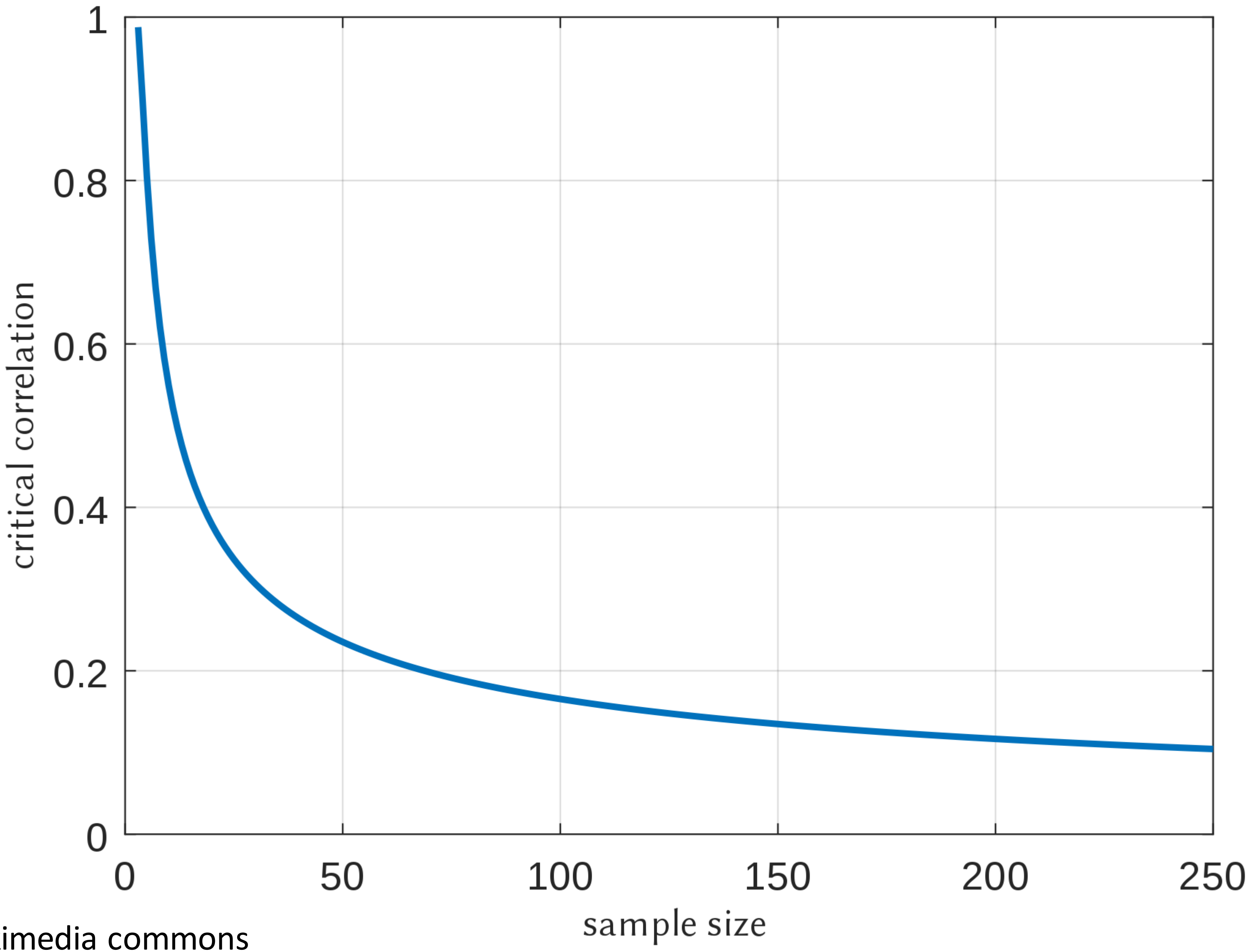
Decision on H0

- **H0:** There is *no* correlation between Obama's election results in 2012 (X) and share of protestants (Y); $r(x, y) >= 0$
- **HA:** There *is a negative* correlation between Obama's election results in 2012 (X) and share of protestants (Y); $r(x, y) < 0$

- Pearson's ***r*** = **-0.413**
- Test statistic ***t* = -3.14;** critical *t* value (α **=** 0.05) = **-1.677**
- Since ***t*** = **-3.14** **<** CV *t* (α = 0.05, df = 48) = **-1.677**, **we reject H0:** r >= 0 and support **HA:** *r* < 0.

- **p-value** = 0.003 (i.e.: 0.3%) indicates probability of observing such, or even more extreme, **value of the test statistic** (***t*** = **-3.14**) if **H0 holds**.

- **Thus:** There is a negative correlation (negative linear relationship) between Obama's election results in 2012 (X) and share of protestants (Y) at the 5% level of statistical significance



wikimedia commons

recode {car}

- `Recode` transforms into a numeric, character, or factor vectors according to recode specifications

`var`	numeric, character, or factor vector to be recoded
`recodes`	character: definition of recode specifications

```
recodes(vector, recodes="'Freq'='2'; 'Some'='1'; 'None'='0'")
```

plot {graphics}

- `plot` produces a two-dimensional graph

`x`	numeric vector: X coordinates of points in the plot
`y`	numeric vector: Y coordinates of points in the plot
`type`	type of plot to be drawn: "p" – point; "l" – lines; "n" – no plotting ("n" useful when plotting labels)
`xlim`	numeric: specifies start and end point of the X axis; e.g. xlim=c(0,100)
`ylim`	numeric: specifies start and end point of the Y axis; e.g. ylim=c(0,100)
`main`	character: a title of the plot
`xlab`	character: a title of the X axis
`ylab`	character: a title of the Y axis

text {graphics}

- `text` draws the strings given in the vector `labels` at the coordinates given by `x` and `y`

`x`	numeric vector: X coordinates of points in the plot
`y`	numeric vector: Y coordinates of points in the plot
`labels`	character vector: specifies the text to be displayed on the plot
`cex`	numeric: **c**haracter **ex**pansion factor (Default = 1)

- In the (common) case of labels overlaps – function `pointLabel` {maptools}.
- `pointLabel` uses same basic arguments `(x, y, labels)` as the `text` function.

table {base}

- `table` use the cross-classification to build a contingency table of the counts for each combination of factor levels (categories)

`...`	one or more objects that can be interpreted as factors
`exclude`	levels remove to all factors
`stringsAsFactors`	logical: should the classifying factors be returned as factors or strings (default = T)

cor {stats}

- `cor` computes correlation of x and y. For matrix: correlations of all pairs of rows/cols and diagonal. For matrices: col-wise pairs.

`x`	numeric: vector, matrix or data.frame
`y`	numeric: vector, matrix or data.frame (compatible dimensions to x)
`method`	character: "pearson", "kendall", "spearman"
`na.rm`	logical: should NA values be removed? (default = F)

cor.test {stats}

- `cor.test` tests for association between paired samples of x and y

`x`	numeric: vector
`y`	numeric: vector (compatible length to x)
`alternative`	character: "two.sided", "less", "greater"
`method`	character: "pearson", "kendall", "spearman"
`conf.level`	numeric: sets the significance threshold (default = 0.95)
`na.rm`	logical: should NA values be removed? (default = F)

corrplot {corrplot}

- `corr.plot` produces a graphical display of a correlation matrix including large number of additional arguments

`corr`	numeric: the correlation matrix to visualize
`method`	character: visualization methods = "circle" (default), "square", "ellipse", "number", "pie", "shade", "color"
`type`	character: type of plot = "full" (default), "lower", "upper"
`na.rm`	logical: should NA values be removed? (default = F)

cluster_similarity {clusteval}

- `cluster_similarity` calculates the specified similarity statistic based on co-memberships of the observations.

`labels1`	a vector of n clustering labels
`labels1`	a vector of n clustering labels
`similarity`	character: "jaccard", "rand"
`na.rm`	logical: should NA values be removed? (default = F)

Exercise

- Download the “MEBn5033_11_MA_practice_empty.R” from In-Class Exercises folder and follow the instructions