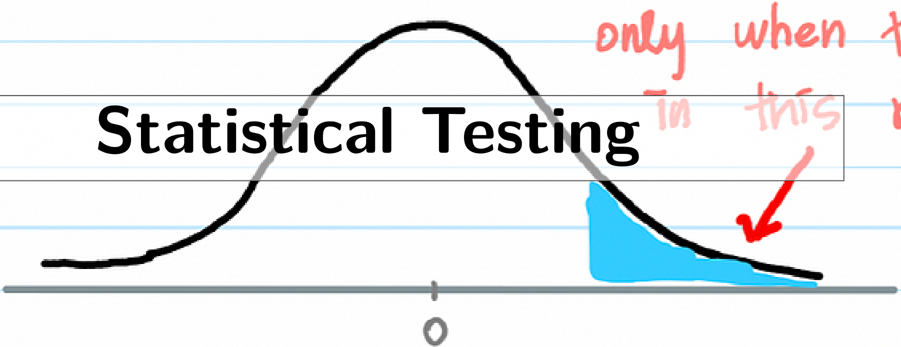
Statistical Testing



We reject H

only when the

in this regi

$\alpha = 20\%$

Learning outcomes:

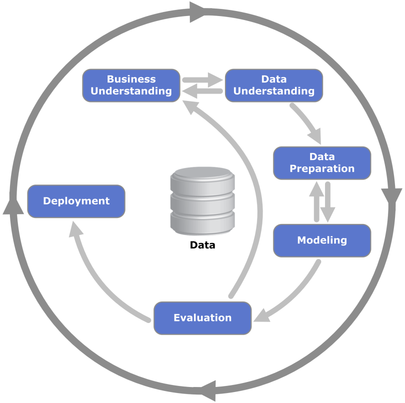


Figure: The CRISP-DM process.

- Review the logic of a hypothesis test;
- Explain how a hypothesis test uses probability;
- Choose appropriate hypothesis tests for comparing data of different types.

Recapitulation: the basic logic of a hypothesis test

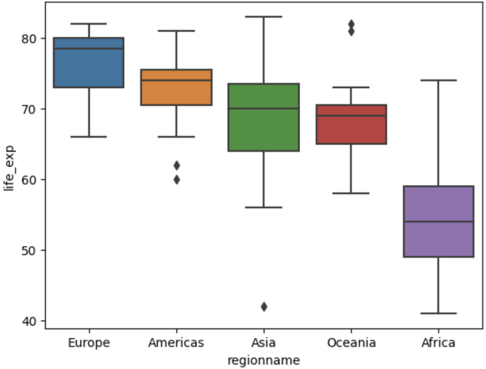
- Assume something to be true about the world (the null hypothesis);
- Collect data;
- Summarize the information in the data (the test statistic);
- If the data is...
    1. ...really unlikely (p-value < 0.05) given what you assumed to be true about the world, then conclude that what you assumed (the null hypothesis) is wrong;
    2. ...not that unlikely (p-value > 0.05) given what you assumed to be true about the world, then do not conclude it is wrong.

Numerical vs categorical w/ two categories

- **Research question**: does life expectancy depend on location?
- **Hypothesis**: life expectancy in Europe is greater than life expectancy in Asia.
- How to investigate?

Numerical vs categorical w/ two categories

- **Research question**: does life expectancy depend on location?
- **Hypothesis**: life expectancy in Europe is greater than life expectancy in Asia.
- How to investigate? Start with visualization. But how do we know a difference is 'real'?



Numerical vs categorical w/ two categories: Two sample t-test

- A two-sample t-test is a hypothesis test that:
  - Compares a categorical variable with two values vs a numerical variable;
  - Answers the question of whether the means of the two groups defined by the categorical variable differ;

- Parts of the two-sample t-test:
  1. H_0: the means of the two groups are equal;
  2. H_A: the means of the two groups are NOT equal;
  3. Test statistic/distribution:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\mu_1 =$ group 1 mean.
$\mu_2 =$ group 2 mean.
$s_1 =$ group 1 sample standard deviation.
$s_2 =$ group 2 sample standard deviation.
$n_1 =$ group 1 size.
$n_2 =$ group 2 size.
test statistic follows a student's t distribution.

  4. Rejection criterion: p-value < 0.05.

Numerical vs categorical w/ two categories: Two sample t-test

- A two-sample t-test is a hypothesis test that:
  - Compares a categorical variable with two values vs a numerical variable;
  - Answers the question of whether the means of the two groups defined by the categorical variable differ;

- Parts of the two-sample t-test:
  1. H_0: the means of the two groups are equal;
  2. H_A: the means of the two groups are NOT equal;
  3. Test statistic/distribution:

$$|t| = \left| \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \right|$$

$\mu_1 =$ group 1 mean.
$\mu_2 =$ group 2 mean.
$s_1 =$ group 1 sample standard deviation.
$s_2 =$ group 2 sample standard deviation.
$n_1 =$ group 1 size.
$n_2 =$ group 2 size.
test statistic follows a student's t distribution.

4. Rejection criterion: p-value < 0.05.

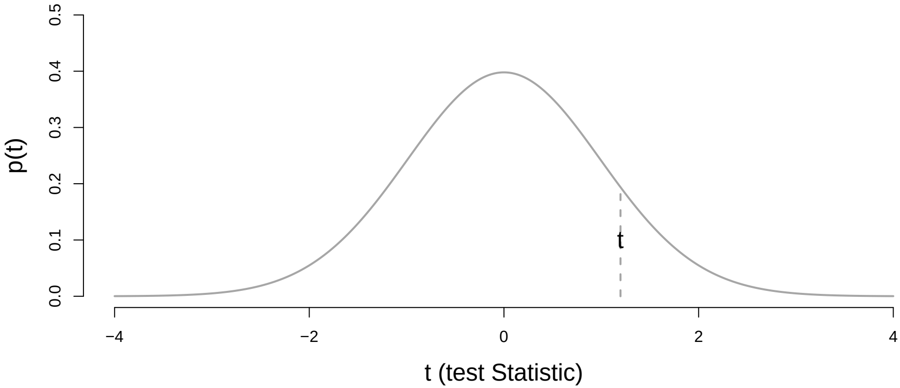
Quick aside – let's link hypothesis testing up with probability

A *p*-value is...

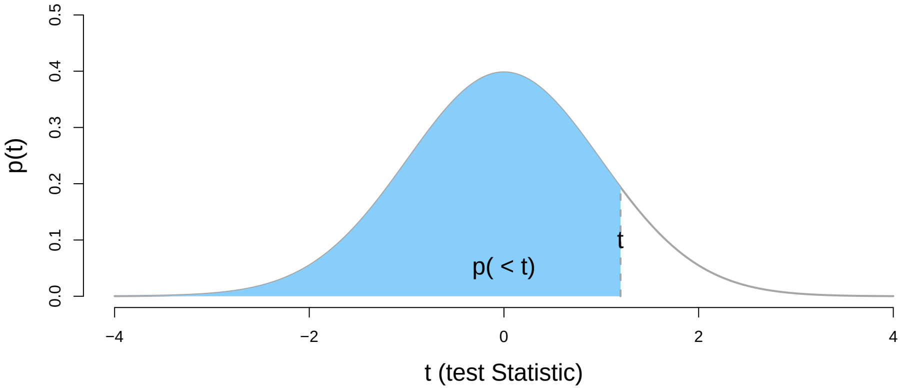
Quick aside – let's link hypothesis testing up with probability

A *p*-value is... the probability of observing something more extreme than the test statistic you got given that the null hypothesis is true.

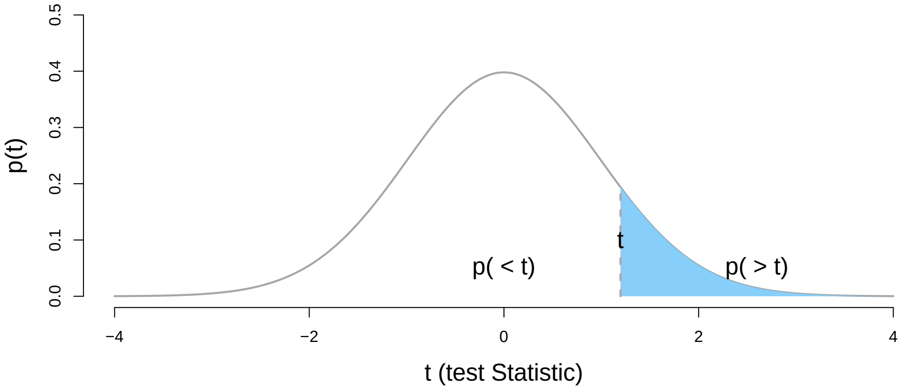
Suppose that **the null is true** and we sample from the population many times computing t each time...



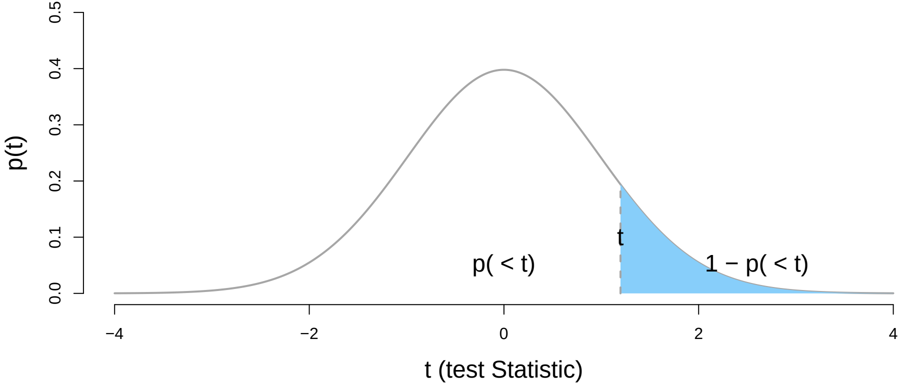
So if we observe t in a sample what is the probability we would observe something less than t in a new sample?



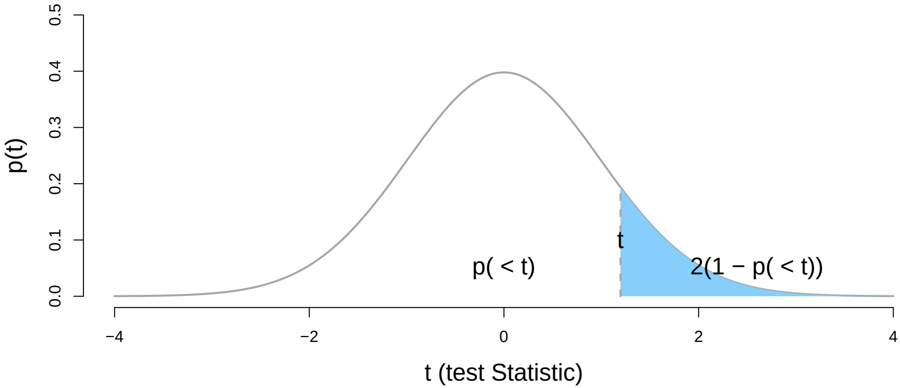
So if we observe t in a sample what is the probability we would observe something greater than t in a new sample?



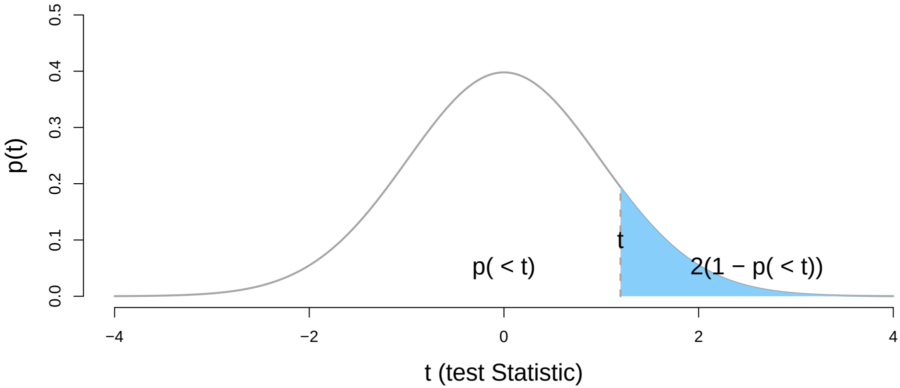
So if we observe t in a sample what is the probability we would observe something greater than t in a new sample?



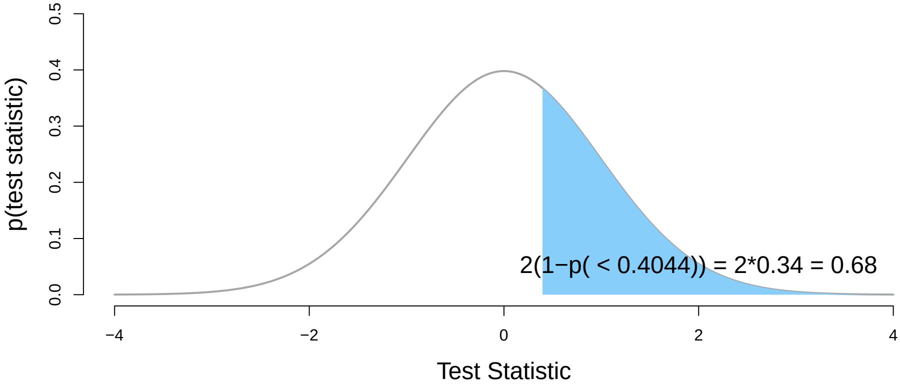
So if we observe t in a sample what is the probability we would observe something greater than $|t|$ in a new sample?



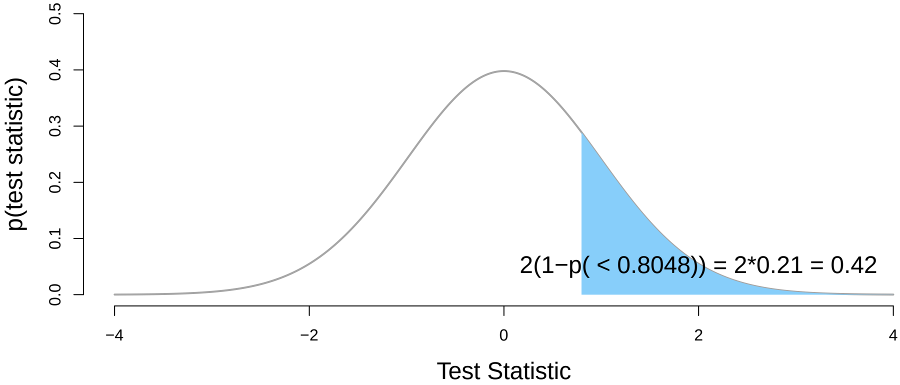
We just computed the probability of observing something more extreme than the test statistic you got given that the null hypothesis is true!



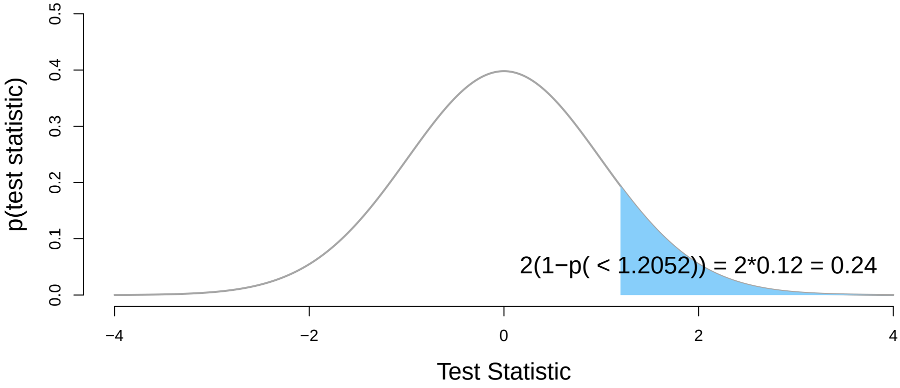
Suppose in our samples we compute $|t| = 0.4044$...



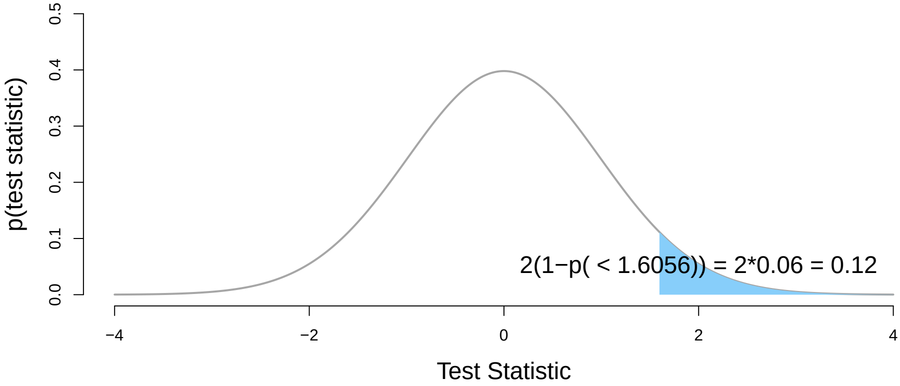
Suppose in our samples we compute $|t| = 0.8048$...



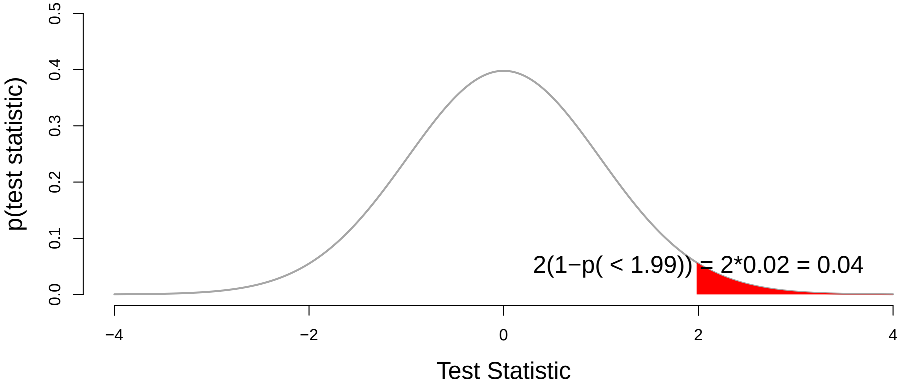
Suppose in our samples we compute $|t| = 1.2052$...



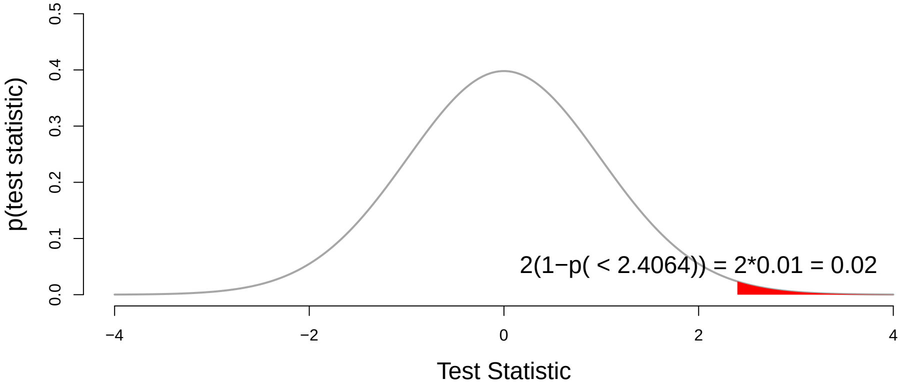
Suppose in our samples we compute $|t| = 1.6056$...



Suppose in our samples we compute $|t| = 1.9900\ldots$ **Reject null!**



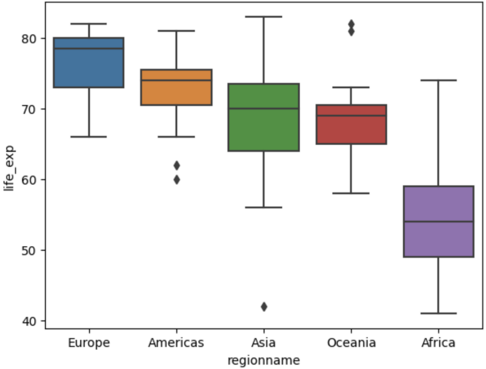
Suppose in our samples we compute $|t| = 2.4064$... **Reject null!**



Numerical vs categorical w/ many categories

- **Research question**: does life expectancy (in years) differ by continental region?

- **Hypothesis**: life expectancy is different in different continental regions.



Numerical vs categorical w/ many categories: Kruskal-Wallis test

- A Kruskal-Wallis test is a hypothesis test that:
  - Compares a categorical variable with multiple values vs a numerical variable;
  - Answers the question of whether at least one group defined by the categorical variable differs from at least one other;

- Parts of the Kruskal-Wallis test:
  1. H_0: medians of all of the groups are equal;
  2. H_A: the medians at least two groups are not equal;
  3. Test statistic/distribution:

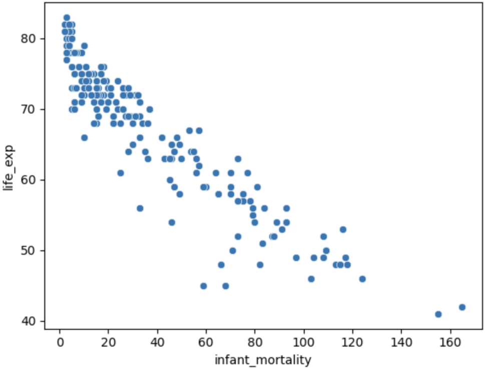
$$H = (N-1)\frac{\sum_{i=1}^{g} n_i \left(\frac{\sum_{j=1}^{n_i} r_{ij}}{n_i} - \frac{N+1}{2}\right)^2}{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (r_{ij} - \frac{N+1}{2})^2}$$

$g =$ number of groups.
$N =$ number of observations.
$n_i =$ number of observations in group i.
$r_{ij} =$ rank of observation j in group i.
test statistic follows a χ^2 distribution.

  4. Rejection criterion: p-value < 0.05.

Numerical vs numerical

- **Research question**: does life expectancy depend on infant mortality?
- **Hypothesis**: if infant mortality goes up then life expectancy goes down.
- How to investigate? Start with visualization.

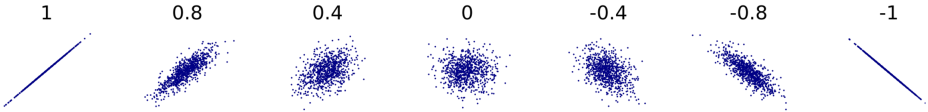


Numerical vs numerical: Pearson's correlation

- Pearson's correlation coefficient:
  - Measures strength/direction of the linear relationship btwn two numerical variables;
  - Answers the question of whether the two variables move together (positive correlation) or opposite (negative correlation);
  - Values lie between $+1$ and -1 – computed as:

$$R = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}}$$

$x_i =$ observation i for variable x.
$y_i =$ observation i for variable y.
$\mu_x =$ the sample mean of variable x.
$\mu_y =$ the sample mean of variable y.
$N =$ number of obs.



Numerical vs numerical: Pearson's correlation

- Pearson's correlation coefficient:
  - Measures strength/direction of the linear relationship btwn two numerical variables;
  - Answers the question of whether the two variables move together (positive correlation) or opposite (negative correlation);
  - Values lie between $+1$ and -1 – computed as:

$$R = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 \sum_{i=1}^{N}(y_i - \mu_y)^2}}$$

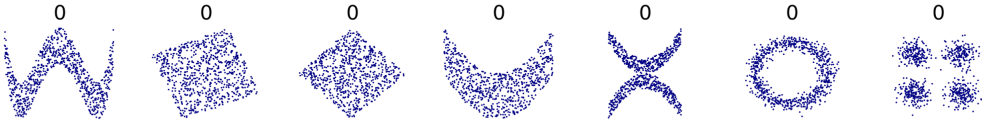
$$\begin{aligned} x_i &= \text{ observation } i \text{ for variable } x. \\ y_i &= \text{ observation } i \text{ for variable } y. \\ \mu_x &= \text{ the sample mean of variable } x. \\ \mu_y &= \text{ the sample mean of variable } y. \\ N &= \text{ number of obs.} \end{aligned}$$

- Parts of the Pearson's correlation test:
  1. H_0: variables are uncorrelated so that R equals 0;
  2. H_A: variables are correlated so that R does NOT equal 0;
  3. Test statistic: $t = R\sqrt{\frac{N-2}{1-R^2}}$, follows a student's t distribution;
  4. Rejection criterion: p-value < 0.05.

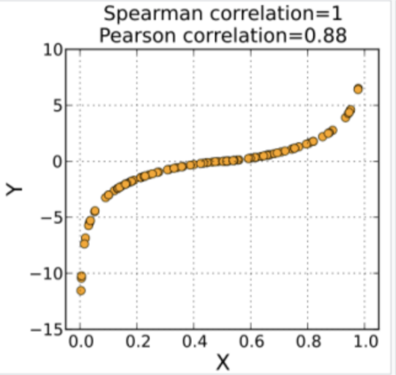
WARNING 1:

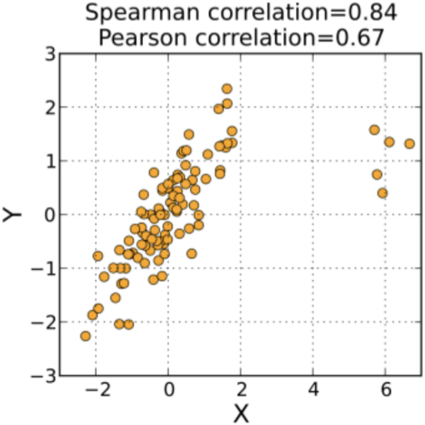
- Pearson's correlation measures strength/direction of the **linear** relationship btwn two numerical variables;

- If Pearson's correlation is near 0, might conclude that there is no relationship between the variables, but there could be a **non-linear** relationship;

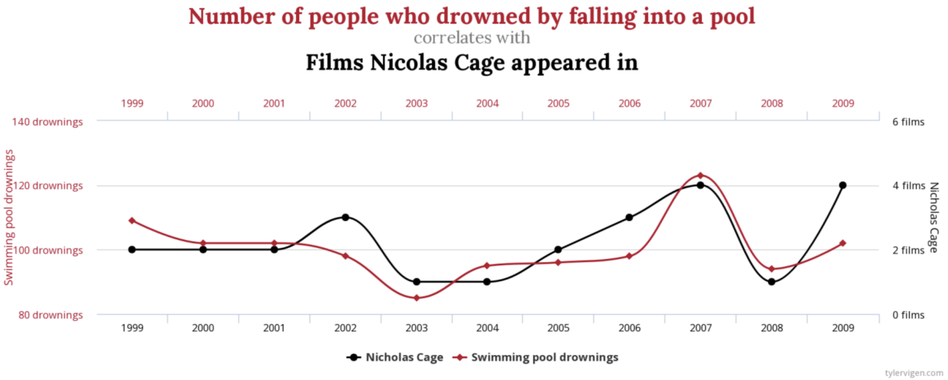


WARNING 1: if you suspect nonlinearity, use Spearman's correlation





WARNING 2: correlation $\neq$ causation



Categorical vs categorical: χ^2 test of independence

- A χ^2 test of independence is a hypothesis test that:
  - Compares two categorical variables;
  - Measures whether there is ‘dependence’ between then;
- Parts of the χ^2 test of independence:
  1. H_0 : There is no association between the two categorical variables/The two categorical variables are independent.
  2. H_A : There is an association between the two categorical variables. The two categorical variables are dependent.
  3. Test statistic/distribution: see next slide...
  4. Rejection criterion: p-value < 0.05.

Categorical vs categorical: χ^2 test of independence test statistic

	1	2	3	4	5	Total
Africa	34	11	4	4	0	53
Americas	0	3	7	16	9	35
Europe	2	16	9	11	9	47
Asia	0	1	5	11	25	42
Oceania	0	6	6	2	2	16
Total	36	37	31	44	45	193

These are observed counts – O_{ij}.

	1	2	3	4	5
Africa	$53 \times \frac{36}{193}$				
Americas	$35 \times \frac{36}{193}$				
Europe	$47 \times \frac{36}{193}$				
Asia	$42 \times \frac{36}{193}$				
Oceania	$16 \times \frac{36}{193}$				

What should this table look like if the null hypothesis is true (independence)?

Categorical vs categorical: χ^2 test of independence test statistic

	1	2	3	4	5	Total
Africa	34	11	4	4	0	53
Americas	0	3	7	16	9	35
Europe	2	16	9	11	9	47
Asia	0	1	5	11	25	42
Oceania	0	6	6	2	2	16
Total	36	37	31	44	45	193

These are observed counts – O_{ij}.

	1	2	3	4	5
Africa	9.88				
Americas	6.52				
Europe	8.76				
Asia	7.83				
Oceania	2.98				

What should this table look like if the null hypothesis is true (independence)?

Categorical vs categorical: χ^2 test of independence test statistic

	1	2	3	4	5	Total
Africa	34	11	4	4	0	53
Americas	0	3	7	16	9	35
Europe	2	16	9	11	9	47
Asia	0	1	5	11	25	42
Oceania	0	6	6	2	2	16
Total	36	37	31	44	45	193

These are observed counts – O_{ij}.

	1	2	3	4	5
Africa	9.88	$53 \times \frac{37}{193}$			
Americas	6.52	$35 \times \frac{37}{193}$			
Europe	8.76	$47 \times \frac{37}{193}$			
Asia	7.83	$42 \times \frac{37}{193}$			
Oceania	2.98	$16 \times \frac{37}{193}$			

What should this table look like if the null hypothesis is true (independence)?

Categorical vs categorical: χ^2 test of independence test statistic

	1	2	3	4	5	Total
Africa	34	11	4	4	0	53
Americas	0	3	7	16	9	35
Europe	2	16	9	11	9	47
Asia	0	1	5	11	25	42
Oceania	0	6	6	2	2	16
Total	36	37	31	44	45	193

These are observed counts – O_{ij}.

	1	2	3	4	5
Africa	9.88	10.16			
Americas	6.52	6.71			
Europe	8.76	9.01			
Asia	7.83	8.05			
Oceania	2.98	3.07			

What should this table look like if the null hypothesis is true (independence)?

Categorical vs categorical: χ^2 test of independence test statistic

	1	2	3	4	5	Total
Africa	34	11	4	4	0	53
Americas	0	3	7	16	9	35
Europe	2	16	9	11	9	47
Asia	0	1	5	11	25	42
Oceania	0	6	6	2	2	16
Total	36	37	31	44	45	193

These are observed counts – O_{ij}.

	1	2	3	4	5
Africa	9.88	10.16	8.51	12.08	12.36
Americas	6.52	6.71	5.62	7.98	8.16
Europe	8.76	9.01	7.55	10.72	10.96
Asia	7.83	8.05	6.74	9.58	9.79
Oceania	2.98	3.07	2.57	3.65	3.73

These are the expected counts – E_{ij}.

Categorical vs categorical: χ^2 test of independence test statistic

	1	2	3	4	5	Total
Africa	34	11	4	4	0	53
Americas	0	3	7	16	9	35
Europe	2	16	9	11	9	47
Asia	0	1	5	11	25	42
Oceania	0	6	6	2	2	16
Total	36	37	31	44	45	193

These are observed counts – O_{ij}.

	1	2	3	4	5
Africa	9.88	10.16	8.51	12.08	12.36
Americas	6.52	6.71	5.62	7.98	8.16
Europe	8.76	9.01	7.55	10.72	10.96
Asia	7.83	8.05	6.74	9.58	9.79
Oceania	2.98	3.07	2.57	3.65	3.73

These are the expected counts – E_{ij}.

The test statistic is:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$