Correlation

MSc Statistics

Correlation

- Measures the degree to which two (or more) variables change together
- Linear (straight line) relationship $\rightarrow$ *correlation*
- Nonlinear relationship $\rightarrow$ *association*
- Captured by a single number
- Correlation/association does *not* imply causation!

Pearson Correlation

- Measures the degree of *linear relationship* between two sets of *n* measurements,

  eg. weights W_i and heights H_i

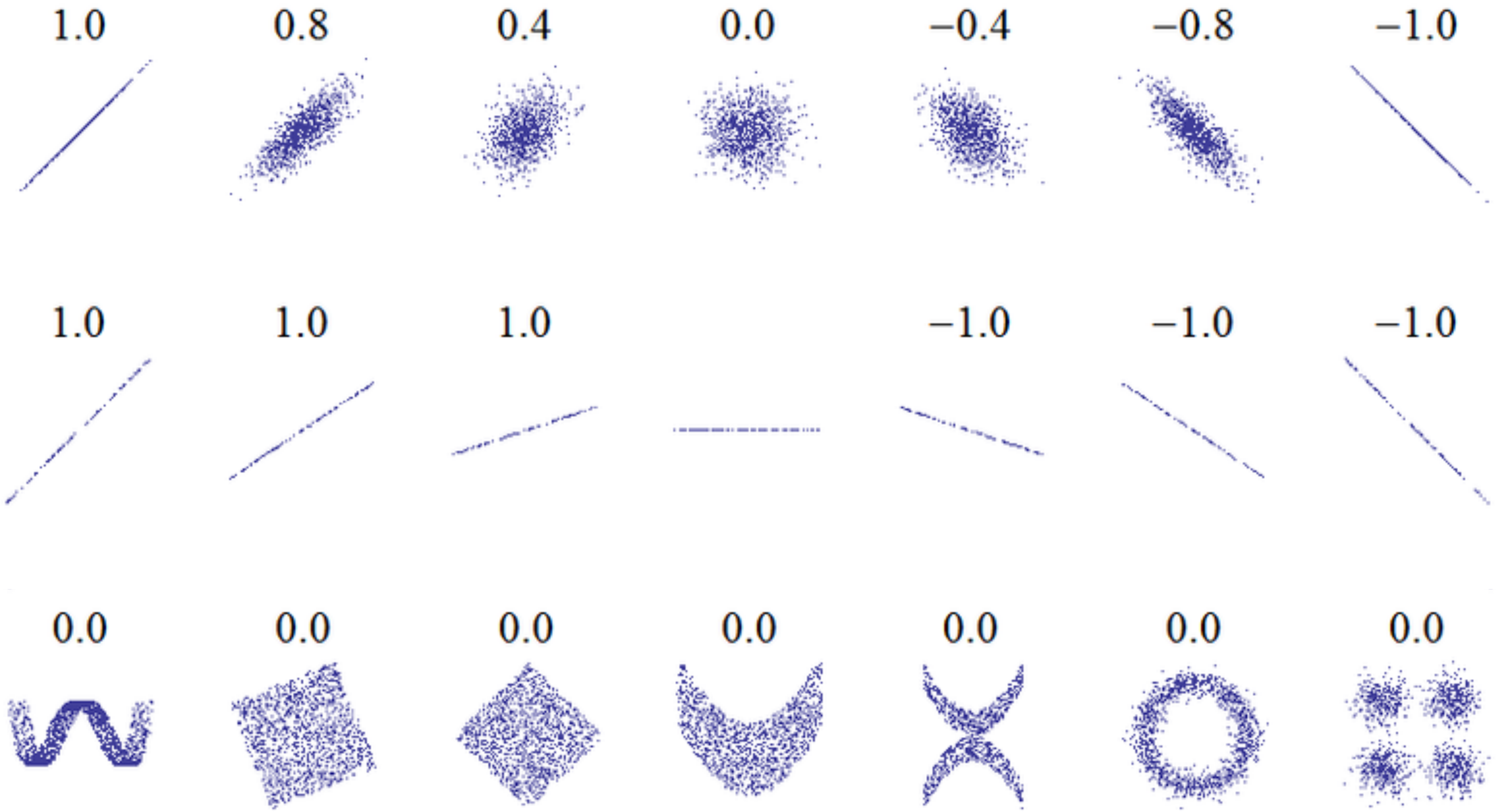
- Varies between -1 and 1

- The Pearson sample correlation coefficient, r is an **estimator** of the population coefficient, ρ *(rho)*

Interpretation of correlation

- $-1 \leq r \leq 1$
- $r = \mp 1$ indicates a deterministic linear relationship between X and Y.
- $r > 0$ indicates a positive linear relationship between X and Y.
- $r < 0$ indicates a negative linear relationship between X and Y.
- $r \approx 0$ indicates no linear association between X and Y.

Examples of Correlation and Non-Linear Relationships

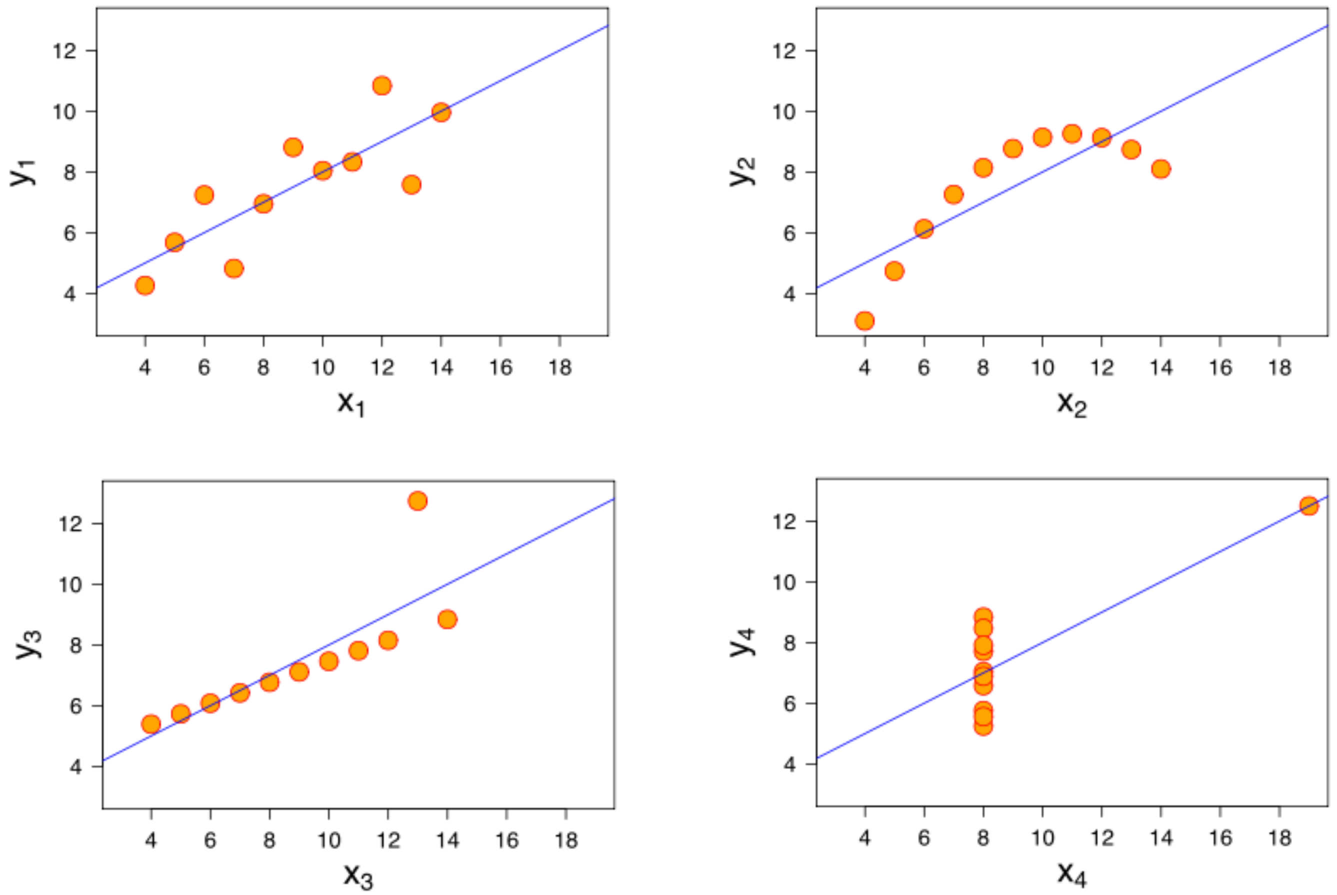


Data Quality

- The measure of correlation is highly sensitive to anomalies in the data:
  - Outliers
  - Clustered data points
  - Nonlinearity
  - Spurious correlations/associations

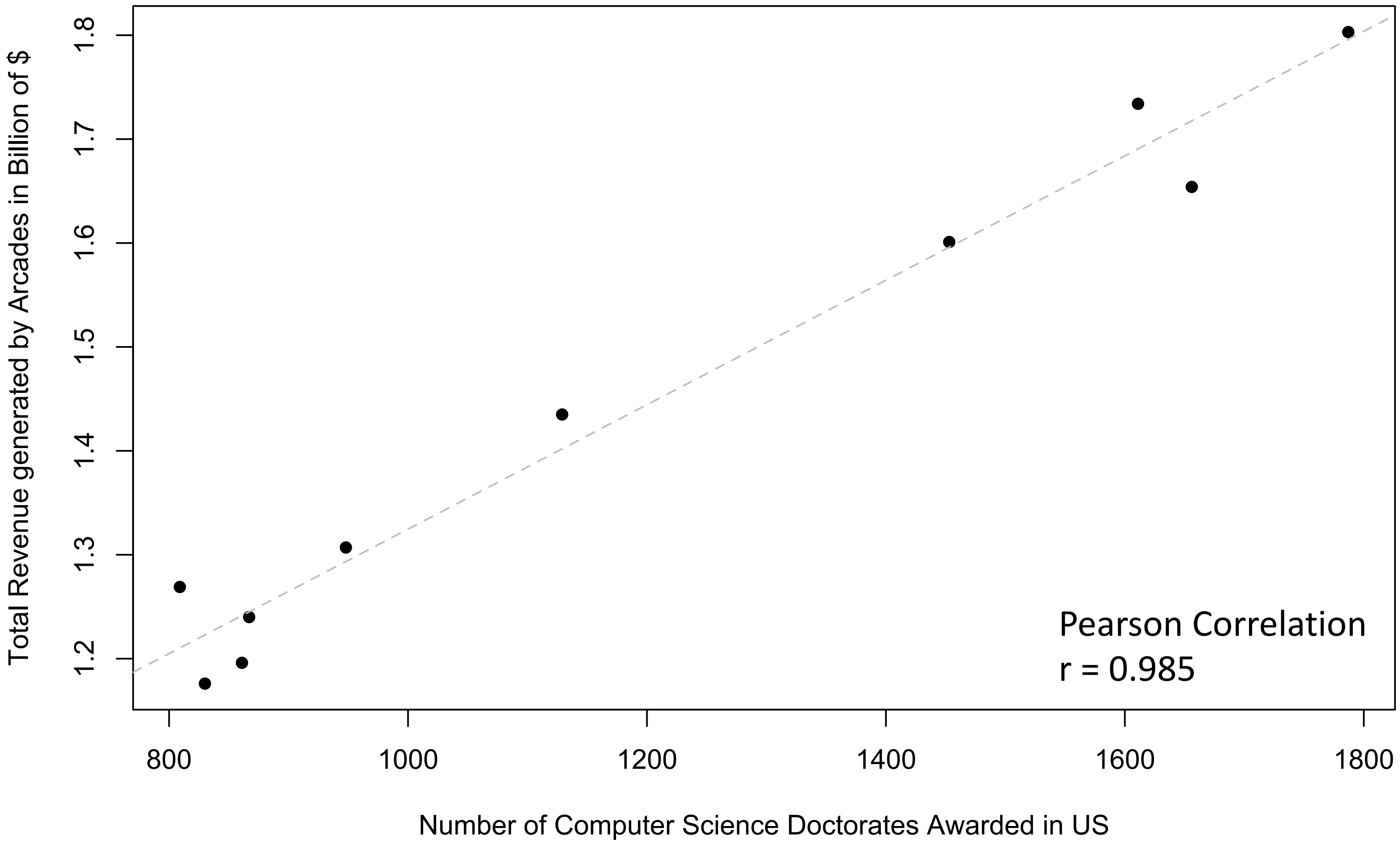
Look at the Data!

- All these datasets have the same mean, variance, correlation coefficient and regression line:

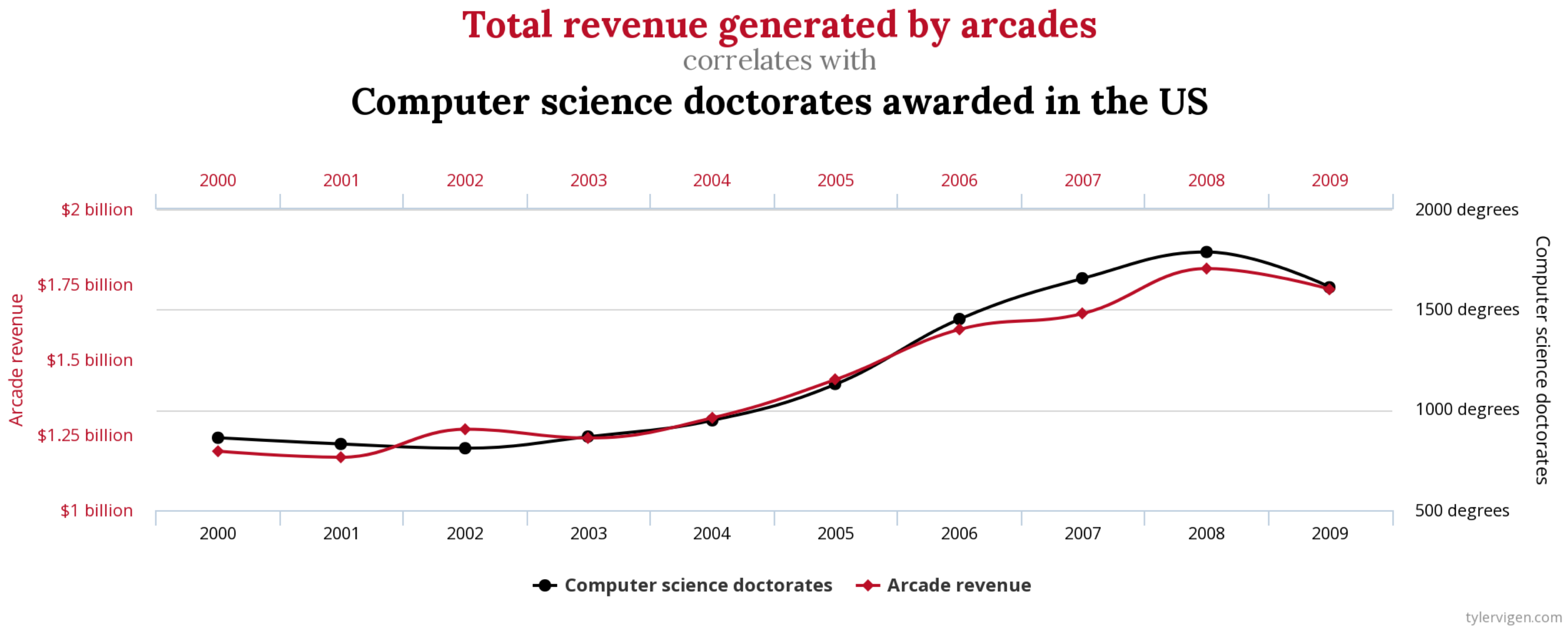


- Anscombe, Francis J. (1973) Graphs in statistical analysis. American Statistician, 27

Correlation is not causation



Correlation is not causation



Data sources: U.S. Census Bureau and National Science Foundation

Example of Pearson’s correlation

The height (in) and weight (lb) of four randomly selected women was recorded:

ID	Height (in)	Weight (lb)
1	67	120
2	62	172
3	64	167
4	65	145

The Correlation Coefficient:

- Pearson correlation written in terms of the original measurements

$$r = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=0}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=0}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=0}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=0}^{n} y_i^2 - n\bar{y}^2)}}$$

- In our case, these were the weights, W_i and the heights, H_i:

$$r = \frac{\sum_{i=0}^{n}(W_i - \bar{W})(H_i - \bar{H})}{\sqrt{\sum_{i=0}^{n}(W_i - \bar{W})^2}\sqrt{\sum_{i=0}^{n}(H_i - \bar{H})^2}}$$

Example of Pearson's correlation

Let X=Height and Y=Weight. Summary statistics are:

$\bar{x} = 64.5$ and $s_x = 2.081666$

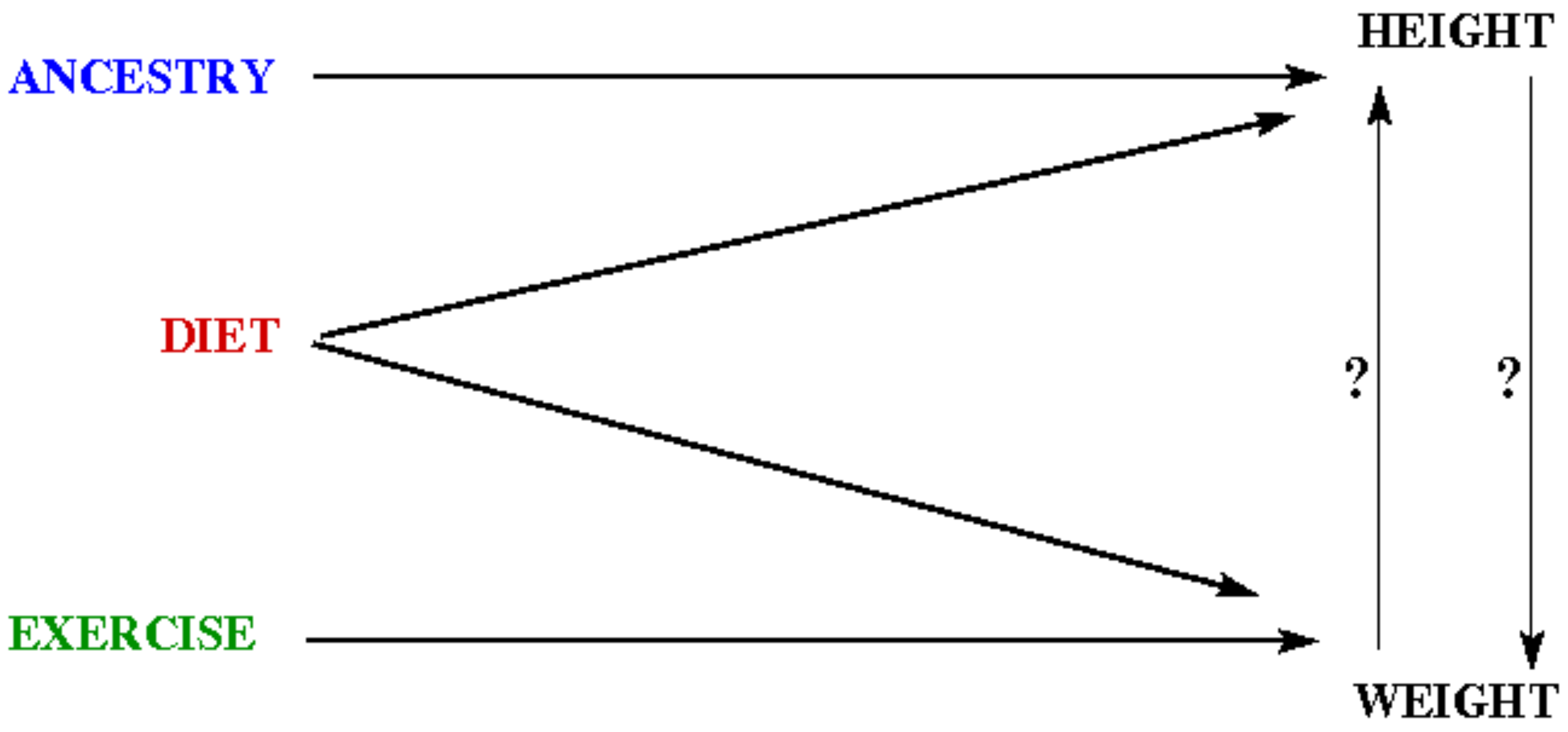
$\bar{y} = 151$ and $s_y = 23.76272$

$$r = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=0}^{n}(y_i - \bar{y})^2}}$$

The sample correlation is $r = -0.9501469$.

Dependency Structure

- Correlation is not causation...



Assessing Significance I

- The correlation coefficient r is an estimator of the population coefficient ρ

- Null Hypothesis: $\rho = 0$

- Assumption:
  - the variables X and Y are normally distributed
  - X and Y are independently and identically distributed

Assessing Significance II

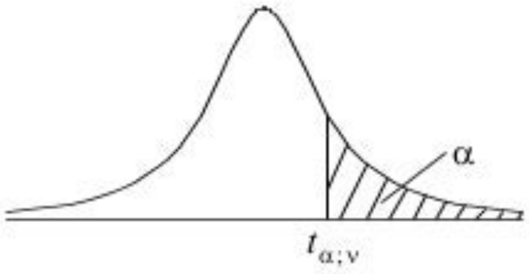
- Consider the distribution of the quantity:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

- If $\rho = 0$, this is distributed as Student's *t*, with n-2 degrees of freedom

$\rightarrow$ we can obtain the critical value of *t* and hence the critical value of *r*

Table of the Student's t-distribution



The table gives the values of $t_{\alpha;\nu}$ where
$\Pr(T_\nu > t_{\alpha;\nu}) = \alpha$, with ν degrees of freedom

ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922

Example of Pearson's correlation

```
In [624]: from scipy.stats import pearsonr

In [625]: height=[67,62,64,65]
     ...: weight=[120,172,167,145]
     ...:
     ...: corr= pearsonr(height, weight)
     ...: corr
Out[625]: (-0.9501468513565026, 0.049853148643497436)
```

- pearsonr gives the estimate for the correlation for the sample and the p-value (2 –sided) for testing the Null Hypothesis: $\rho = 0$

Non-parametric correlation

- If the distribution of (X, Y) deviates strongly from bivariate normal, it may be better to consider non-parametric alternatives to assessing Pearson's correlation coefficient.

- Or, if measures of association that are not necessarily linear are of interest, alternative measures of association can be considered.

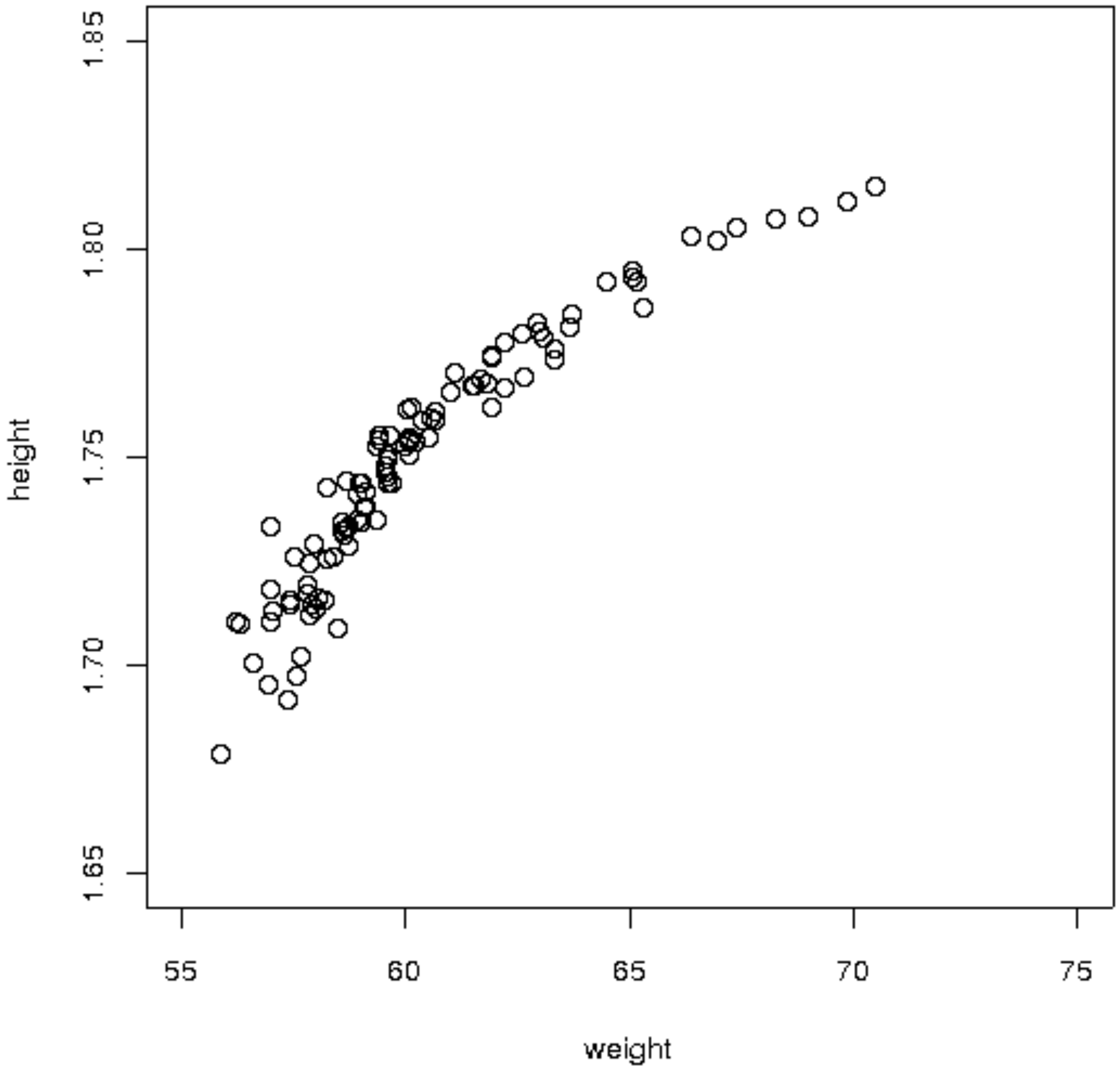
- Correlation: refers specifically to the linear relationship between X and Y.

- Association: refers generally to the relationship between X and Y.

The Spearman Coefficient

- The Spearman coefficient measures correlation between *rank ordered* data
- Can handle non–linear (monotonic) data
- The Spearman coefficient is *not* an estimator of any simple population parameter (non-parametric)

Example: Non Linear Data



Computing the Spearman Coefficient

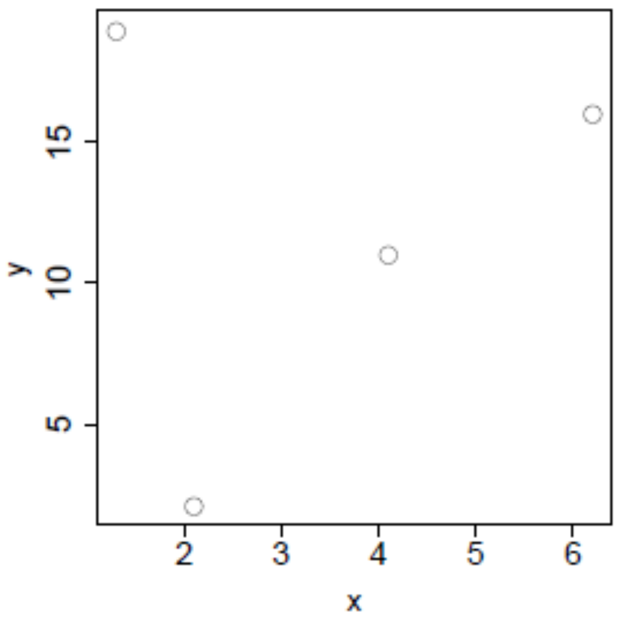
- The procedure is as follows:

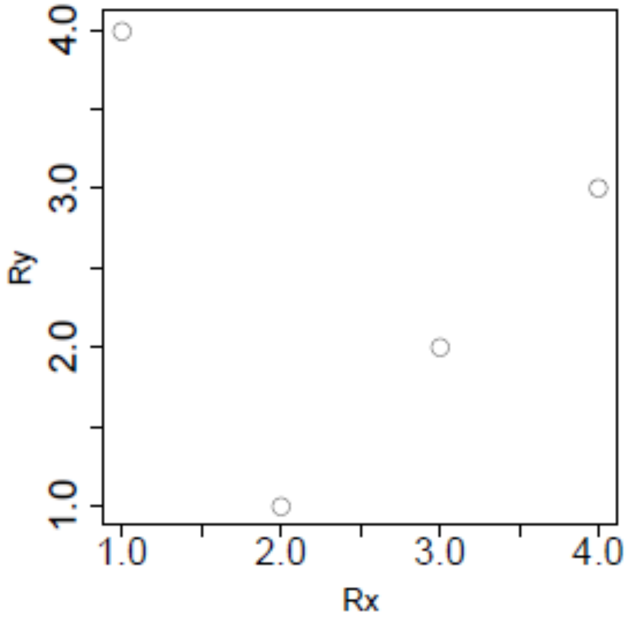
1. Rank the x-values in ascending order.
2. Rank the y-values in ascending order.
3. For each pair of rankings, calculate d^2, the square of the difference between the rankings.
4. Spearman’s Rank Correlation coefficient, r’, is found using the formula:

$$r' = 1 - \frac{6 \sum d^2}{n(n^2-1)}$$

Example for Spearman’s rank correlation

X	Y	Rank X	Rank Y
4.1	11	3	2
6.2	16	4	3
1.3	19	1	4
2.1	6	2	1





Spearman’s Rank Correlation in python

```
In [630]: from scipy.stats import spearmanr



In [631]:
     ...: x=[4.1,6.2,1.3,2.1]
     ...: y=[11,16,19,6]
     ...: corr_spear= spearmanr(x, y)
     ...: corr_spear
Out[631]: SpearmanrResult(correlation=-0.19999999999999998, pvalue=0.8)
```

- What is null hypothesis of the test here?
- What do we conclude?

Spearman’s Rank Correlation in python

- Here we are doing a two-sided test.
- H_0: There is no association between X and Y
- H_1: There is a monotonic association between X and Y.

- The p-value calculated using permutations
- Help from “spearmanr”:

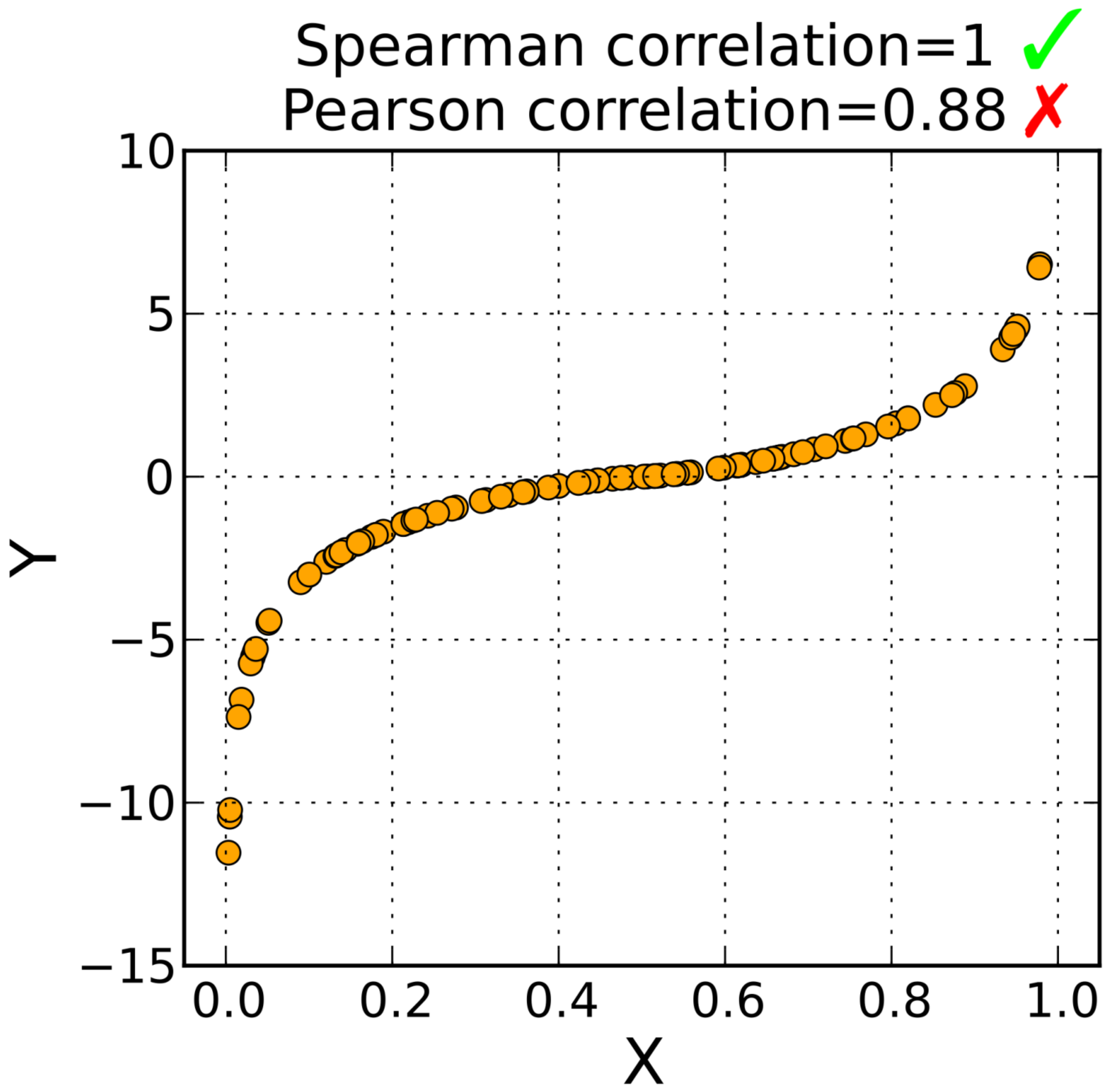
The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Spearman correlation at least as extreme as the one computed from these datasets. The p-values are not entirely reliable but are probably reasonable for datasets larger than 500 or so.

Calculate p-value using permutations

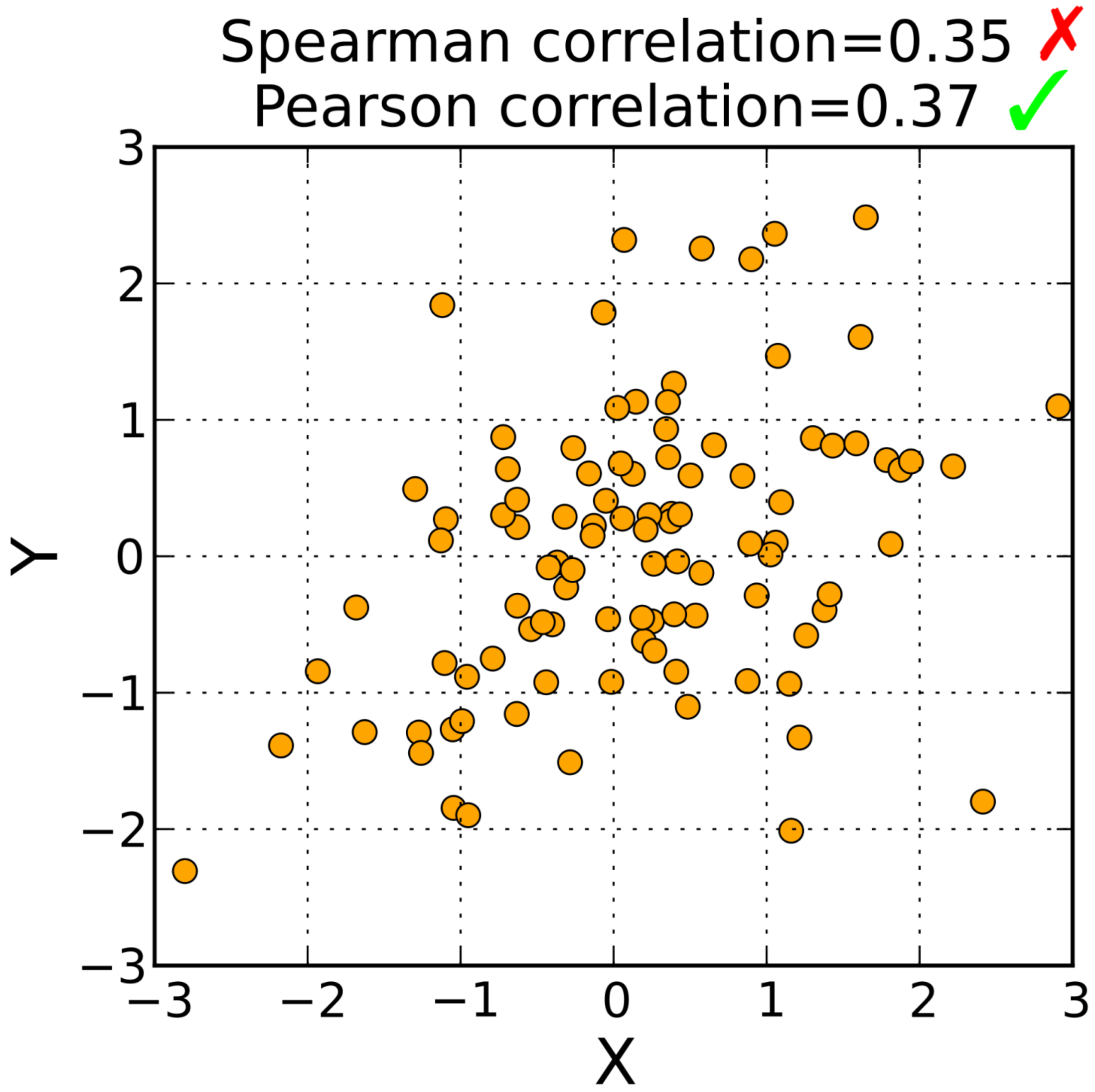
Steps:
1. Calculate Spearman’s correlation coefficient, rho, for the sample of data. (It is estimated that $r_s = -0.2$ in our data)
2. Permute the Y's among the X's in the n! possible ways.
3. For each permutation, calculate *rho*.
4. Find the P-value using the distribution of permuted *rho* values (as extreme or more extreme).

Permutation	Y1	Y2	Y3	Y4	Correlation
1	1	2	3	4	-0.6
2	1	2	4	3	-0.8
3	1	3	2	4	0
4	1	3	4	2	-0.4
5	1	4	2	3	0.4
6	1	4	3	2	0.2
7	2	1	3	4	-0.8
8	2	1	4	3	-1
9	2	3	1	4	0.4
10	2	3	4	1	-0.2
11	2	4	1	3	0.8
12	2	4	3	1	0.4
13	3	1	2	4	-0.4
14	3	1	4	2	-0.8
15	3	2	1	4	0.2
16	3	2	4	1	-0.4
17	3	4	1	2	1
18	3	4	2	1	0.8
19	4	1	2	3	-0.2
20	4	1	3	2	-0.4
21	4	2	1	3	0.4
22	4	2	3	1	0
23	4	3	1	2	0.8
24	4	3	2	1	0.6

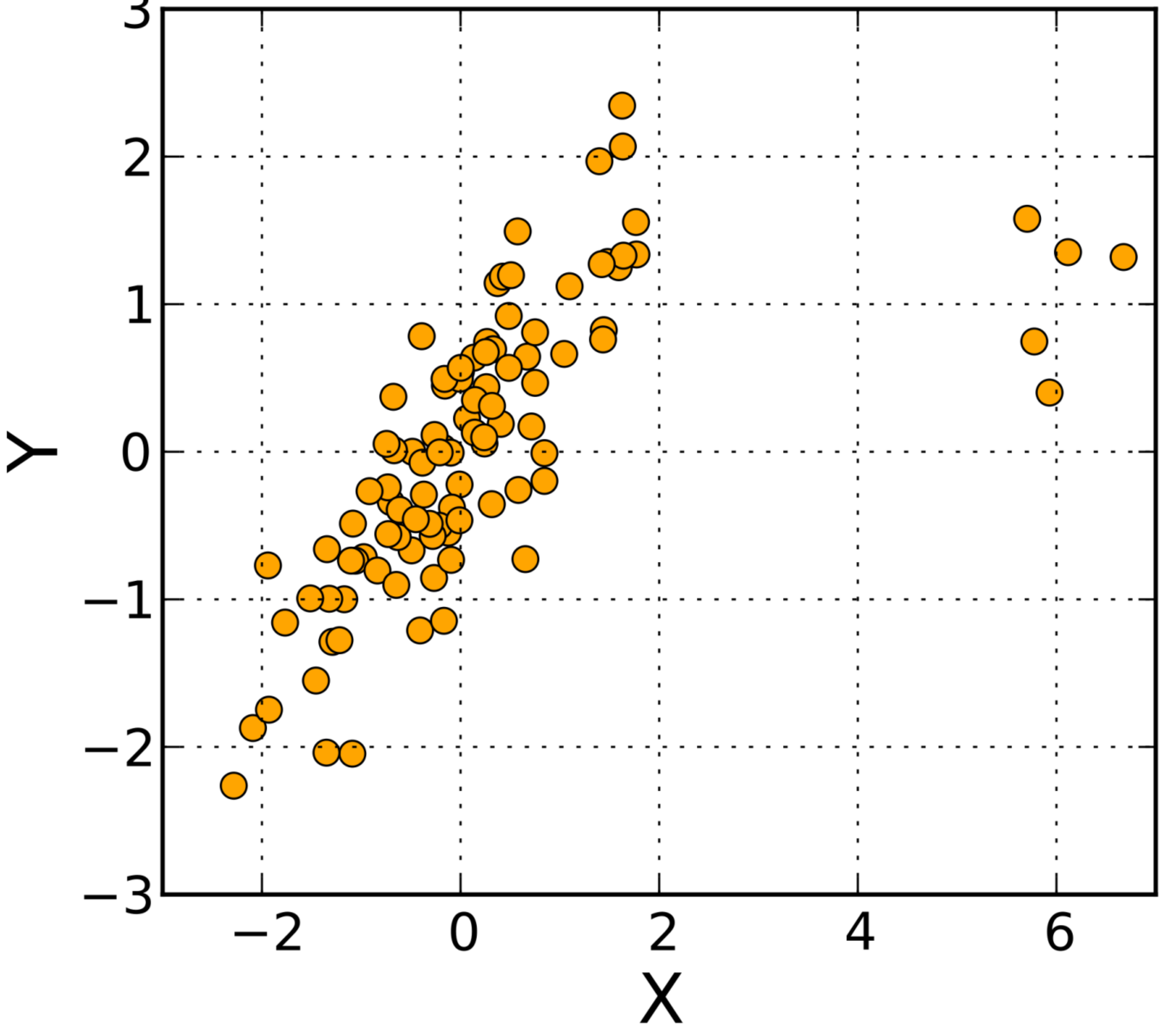
- Here we are doing a two-sided test.
- H_0: There is no association between X and Y
- H_1: There is a monotonic association between X and Y.
- To get the p-value, we first count the number of permuted correlations that are as extreme, or more extreme, than what we observed.
- We observed $r_s = -0.2$.
- For a two-sided test, we need to consider values both equal to or less than -0.2 and values greater than or equal to 0.2.
- 22 of the 24 permuted correlations are as extreme or more extreme than our observed correlation.
- Thus, our P-value $= 22/24 = 0.9167$.
- We fail to reject the null hypothesis, and conclude that we have no evidence that X and Y have a monotonic association.



Spearman correlation=1 ✓
Pearson correlation=0.88 ✗
10
5
0
−5
−10
−15
Y
0.0
0.2
0.4
0.6
0.8
1.0
X


Spearman correlation=0.35 ✗
Pearson correlation=0.37 ✓
Y
X
3
2
1
0
−1
−2
−3

Spearman correlation=0.84 ✗
Pearson correlation=0.67 ✗



Kendall's tau (τ)

Like Spearman's, Kendall's τ determines how close the association between two variables is to being monotonic.

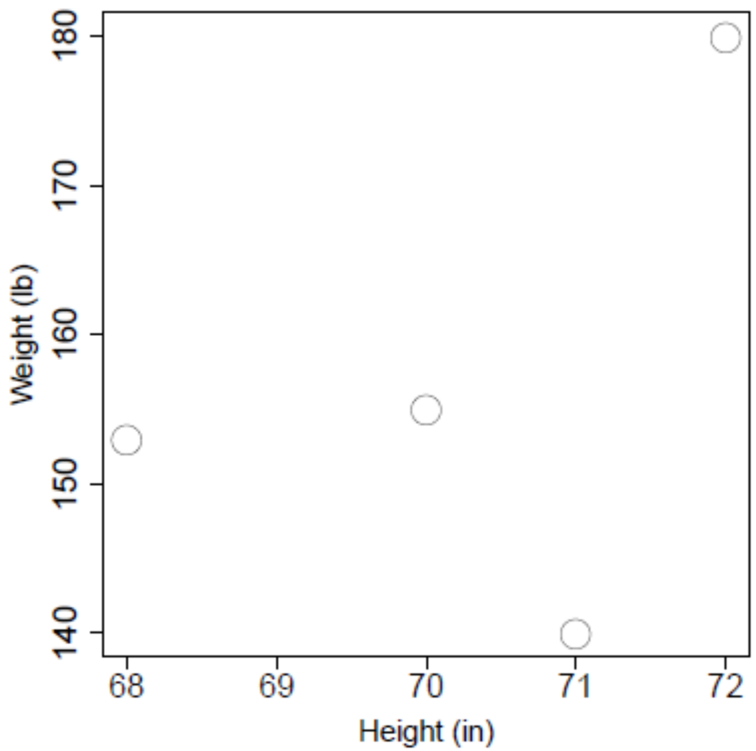
Steps:

1. For each pair of observations, record whether the slope between them is positive or negative.
2. Use this information to compute the estimate of Kendall's τ.
3. Find the p-value using the permutation method already mentioned

Example

The height (in) and weight (lb) of four randomly selected men was recorded.

ID	Height (in)	Weight (lb)
1	68	153
2	70	155
3	71	140
4	72	180



To find Kendall's tau:

Find the sign of the slope between each pair of observations.

1 vs 2 +

1 vs 3 -

1 vs 4 +

2 vs 3 -

2 vs 4 +

3 vs 4 +

Compute $\hat{\tau}_k = \frac{\sum_{i<j} sign[(X_i - X_j)(Y_i - Y_j)]}{\binom{n}{2}}$

$$= \frac{4-2}{\binom{4}{2}} = 0.3333$$

Kendall’s tau in python

```
from scipy.stats import kendalltau



men_h=[68,70,71,72]
men_w=[153,155,140,180]
corr_kend= kendalltau(men_h, men_w)
corr_kend
Out[633]: KendalltauResult(correlation=0.3333333333333334, pvalue=0.75)
```

Kendall's tau P-value

Here we are doing a two-sided test.

$H_0: \tau = 0$ versus $H_1: \tau \neq 0$

We observed $\hat{\tau}_k = 0.3333$.

The p-value is 0.75.

We fail to reject the null hypothesis, and conclude that we have no evidence that $\tau \neq 0$. i.e. there is no evidence here that there is an association between height and weight.

$R(X_1)$	$R(X_2)$	$R(X_3)$	$R(X_4)$	
1	2	3	4	
$R(Y_1)$	$R(Y_2)$	$R(Y_3)$	$R(Y_4)$	$\hat{\tau}_k$
1	2	3	4	1
1	2	4	3	0.67
1	3	2	4	0.67
1	3	4	2	0.33
1	4	2	3	0.33
1	4	3	2	0
2	1	3	4	0.67
2	1	4	3	0.33
2	3	1	4	0.33
2	3	4	1	0
2	4	1	3	0
2	4	3	1	-0.33
3	1	2	4	0.33
3	1	4	2	0
3	2	1	4	0
3	2	4	1	-0.33
3	4	1	2	-0.33
3	4	2	1	-0.67
4	1	2	3	0
4	1	3	2	-0.33
4	2	1	3	-0.33
4	2	3	1	-0.67
4	3	1	2	-0.67
4	3	2	1	-1

To get the P-value:

- Count the number of permuted correlations that are as extreme, or more extreme, than what we observed. We observed $\hat{\tau}_k = 0.3333$.

- For this two-tailed test, we need to consider values greater than or equal to 0.3333 or less than or equal to -0.3333. There are 18 of 24 permuted tau estimates as extreme or more extreme than our observed value.

- Thus, our P-value $= \frac{18}{24} = 0.75$.

Spearman vs Kendall

- Both non-parametric association tests
- Spearman's rho:
  - Usually have larger values than Kendall's Tau.
  - Calculations based on deviations.
  - Much more sensitive to error and discrepancies in data.
- Kendall's Tau:
  - Usually smaller values than Spearman's rho correlation.
  - Calculations based on concordant and discordant pairs.
  - Insensitive to error.
  - P values are more accurate with smaller sample sizes.

Summary I

- Correlation measures how two variables change together
- It does *not* imply causation
- It is sensitive to anomalies in the data
- Data should always be examined visually before doing a correlation analysis

Summary II

- The Pearson coefficient r measures linear relationships and varies between -1 and +1
- If both variables are normally distributed we can determine the statistical significance
- The Spearman coefficient and Kendall’s tau measures non-linear monotonic relationships and varies between -1 and +1