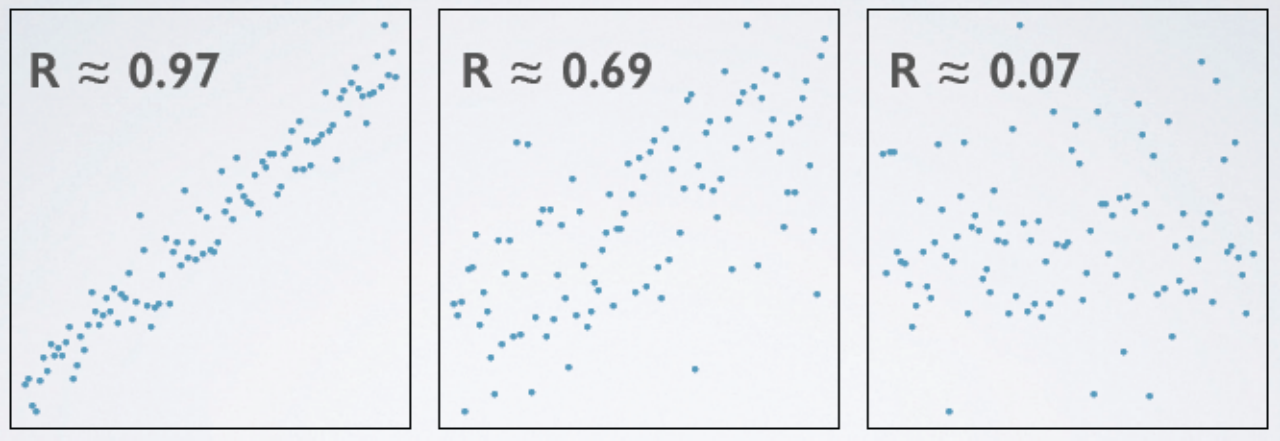
Data Analysis in Astronomy and Physics

Lecture 8: Correlation

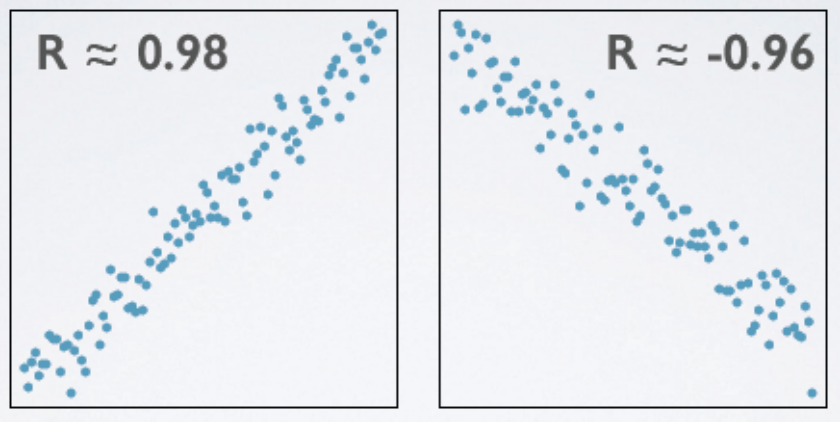
M. Röllig

Introduction

- Correlation describes the strength of the linear association between two variables.
  - Denoted as **R**, or **r**, or ρ. (population: ρ, sample: r or R)
  - The magnitude (absolute value) of the correlation coefficient measures the strength of the linear association between two numerical variables.

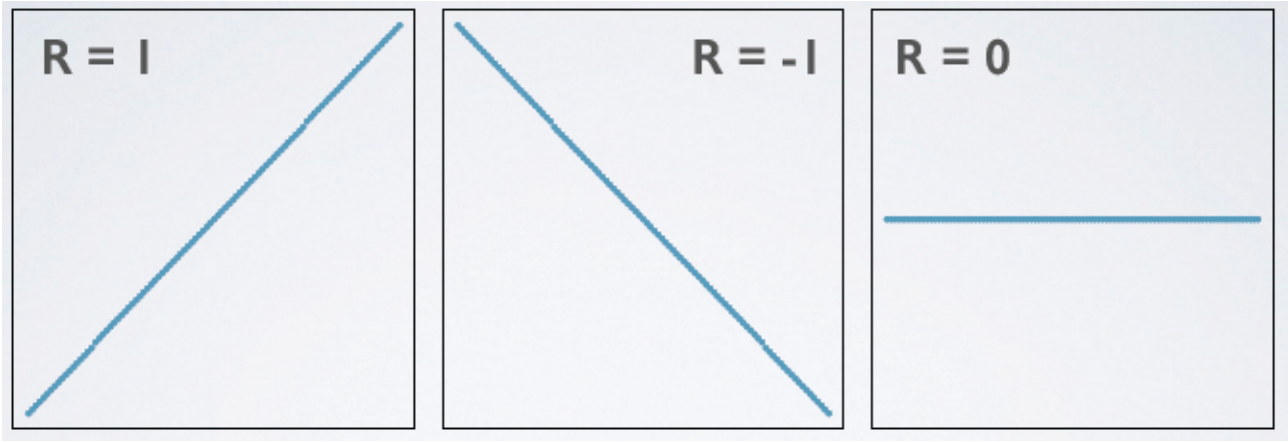


  - The sign of the correlation coefficient indicates the direction of association

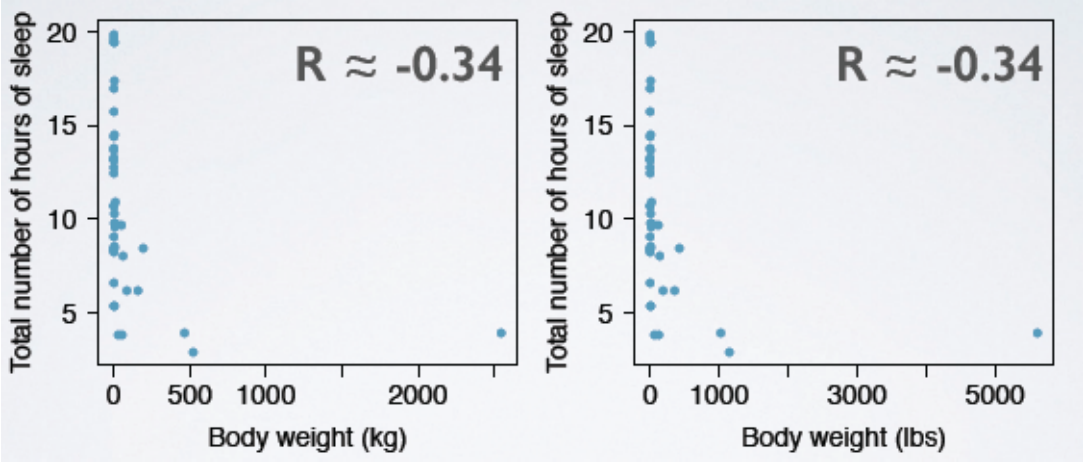


Introduction

- The correlation coefficient is always between -1 (perfect negative linear association) and 1 (perfect positive linear association).
    - R = 0 indicates no linear relationship.

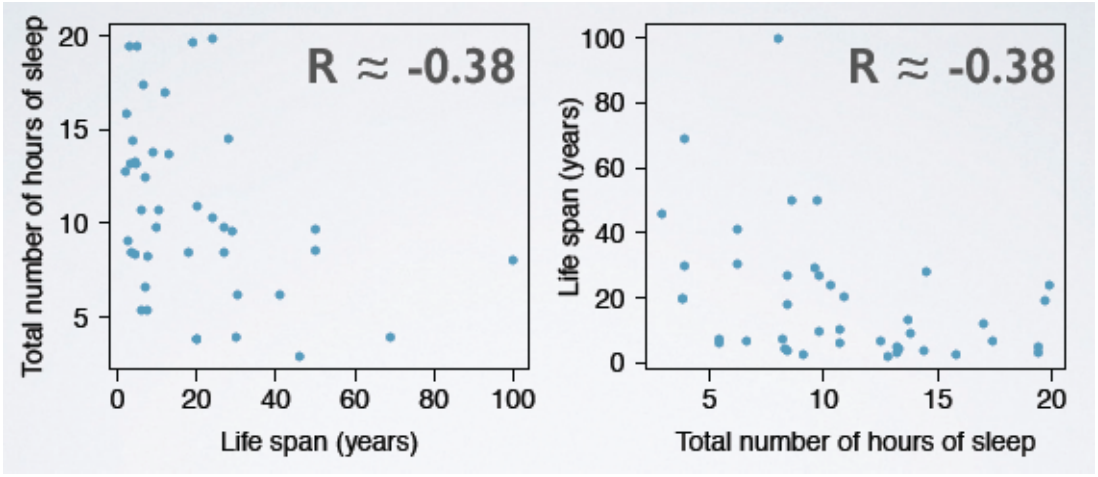


- The correlation coefficient is unit-less, and is not affected by changes in the center or scale of either variable (such as unit conversions)

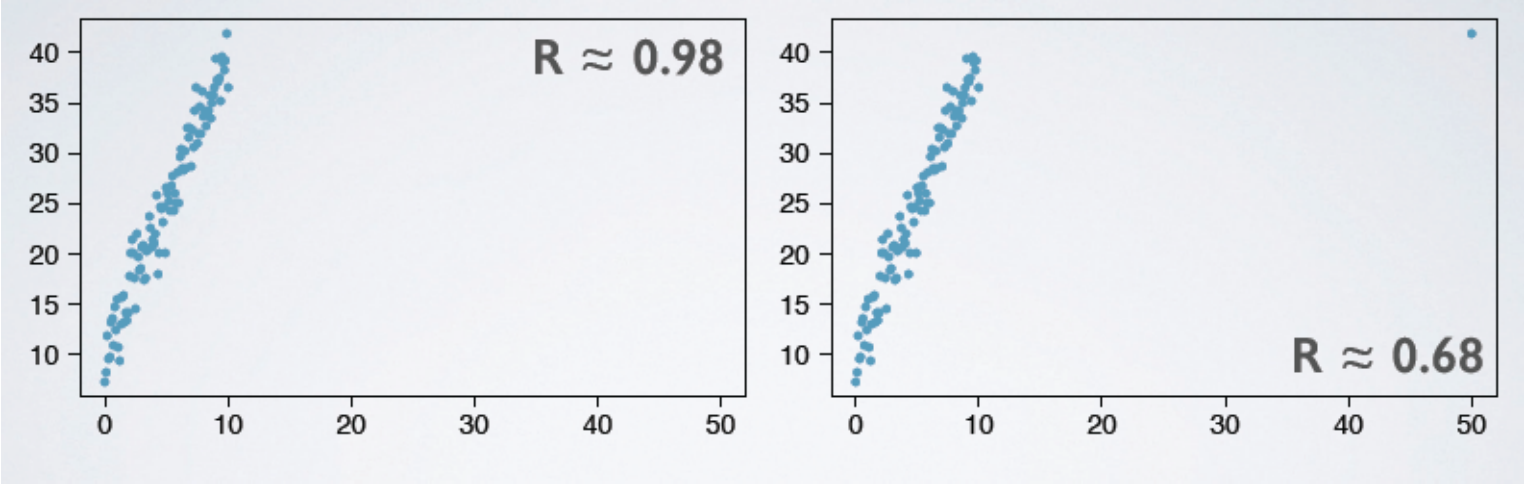


Introduction

- The correlation of X with Y is the same as of Y with X

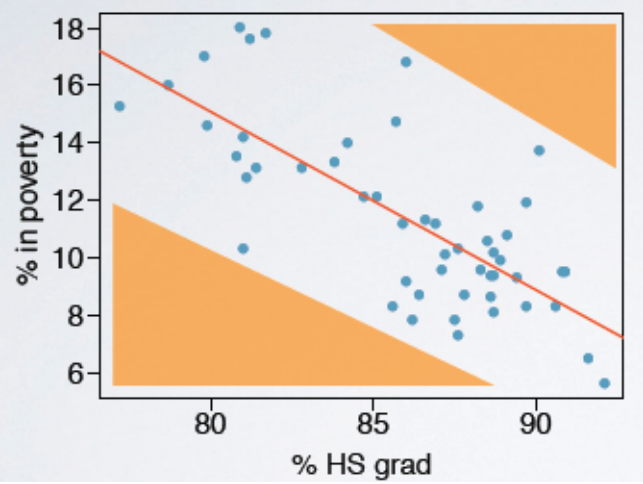


- The correlation coefficient is sensitive to outliers



Quiz

Given the following graph, what is the best guess for the correlation between % in poverty and % HS grad?



1. 0.6

2. -0.75

3. -0.1

4. 0.02

5. -1.5

Introduction

When we make a set of measurements, it is instinct to try to correlate the observations with other results. We might wish

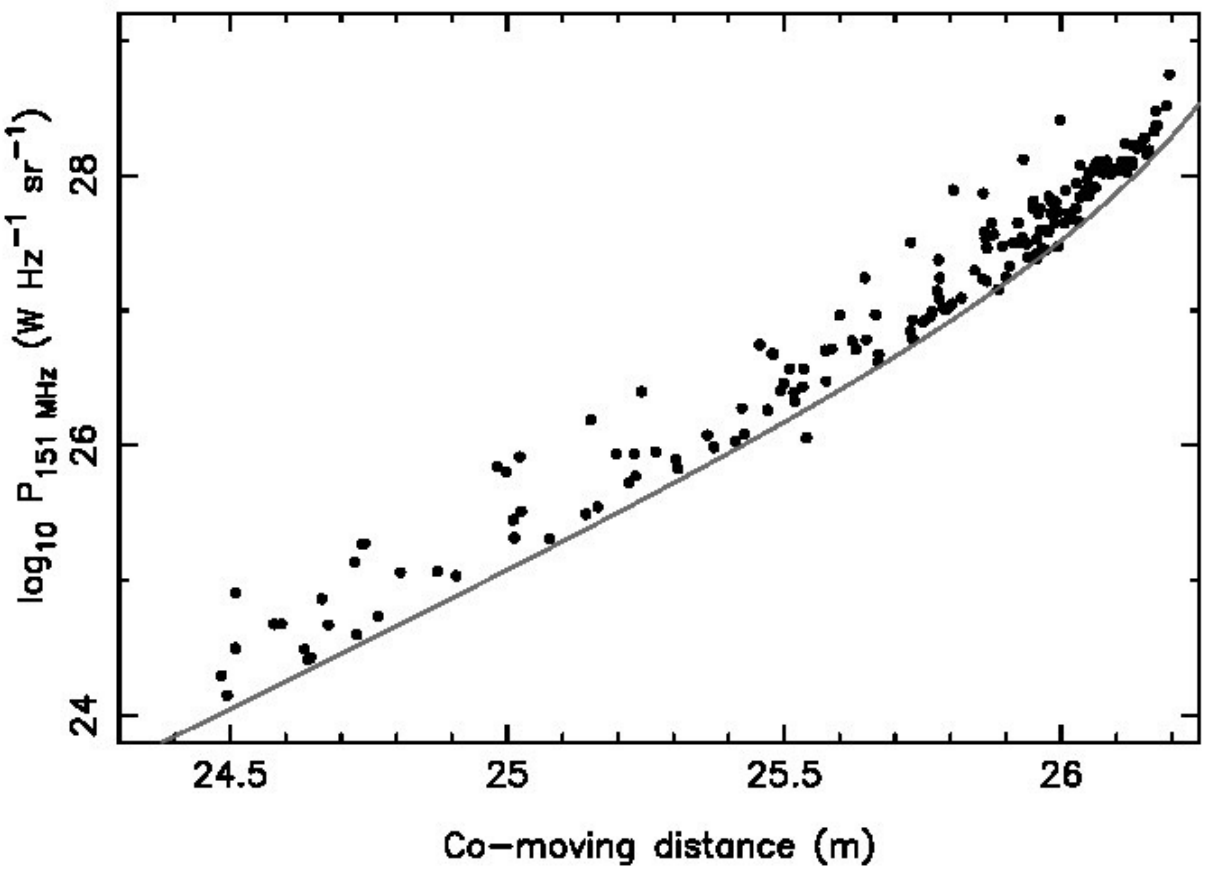
1. to check that other observers' measurements are reasonable,
2. to check that our measurements are reasonable,
3. to test a hypothesis, perhaps one for which the observations were explicitly made,
4. in the absence of any hypothesis, any knowledge, or anything better to do with the data, to find if they are correlated with other results in the hope of discovering some New and Universal Truth.

DANGER AHEAD!

Introduction

Suppose that we have plotted something against something, on a Fishing Expedition.

1. Does the eye see much correlation? If not, formal testing for correlation is probably a waste of time.
2. Could the apparent correlation be due to selection effects? Consider for instance the beautiful correlation obtained by Sandage (1972): 3CR radio luminosities vs distance.



Introduction

Is this a luminosity evolution for radio sources? Are the more distant objects (at earlier epochs) clearly not the more powerful?

No! The sample is flux-limited; the solid line shows the flux-density limit of the 3CR catalogue. The lower right-hand region can never be populated; such objects are too faint to show above the limit of the 3CR catalogue.

But the upper left? Provided that the luminosity function (the true space density in objects per Mpc^3) slopes downward with increasing luminosity, the objects are bound to crowd towards the line. The only conclusion from the diagram! The space density of powerful radio sources is less that the space density of the weaker sources.

In order to populate the upper left region of the plot, we would need to sample large spheres about us to have a chance to encounter a powerful radio source. There are only low-luminosity sources to be seen at low redshifts because there's not enough volume to pick up the high-fliers.

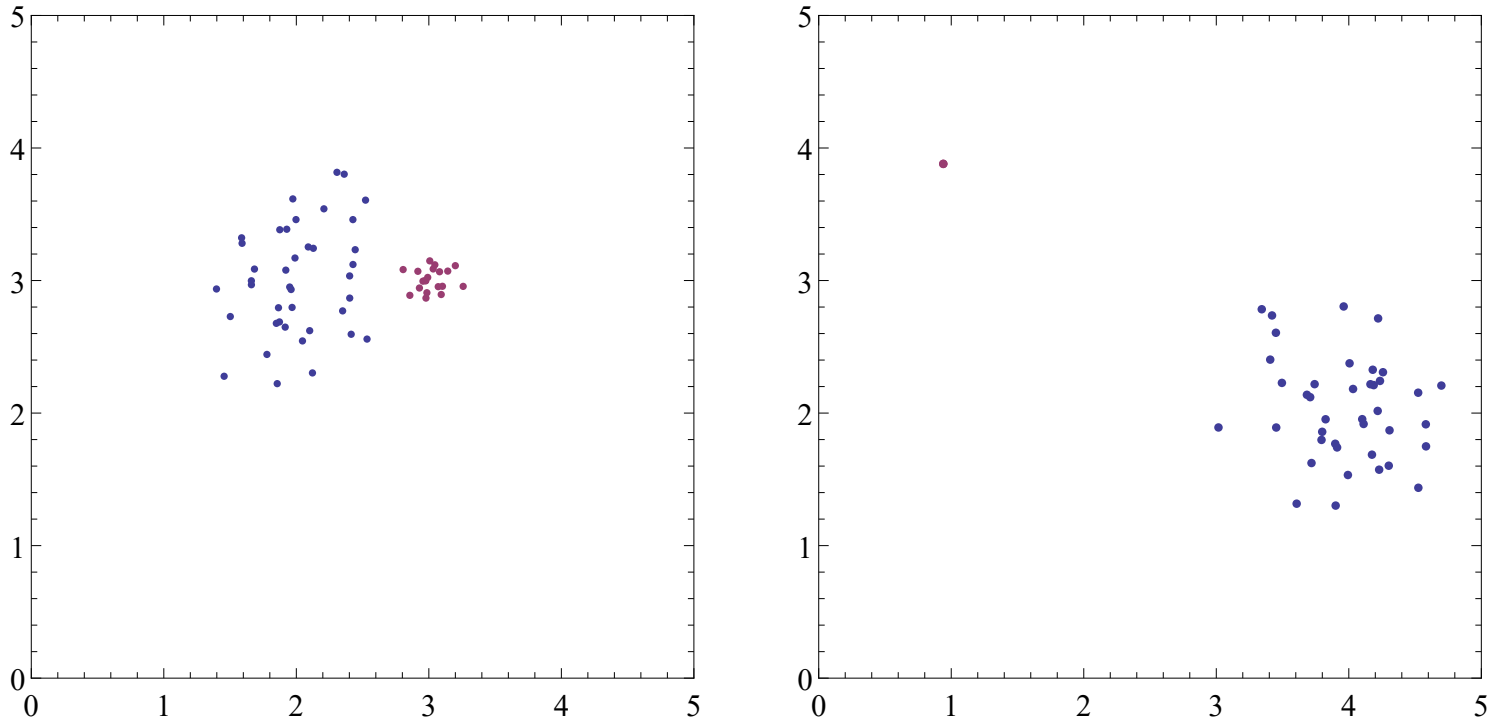
This applies to any proposed correlation for variables with steep probability distribution functions dependent upon one of the variables plotted.

Introduction

3. If we are happy about **2.**, we can try formal calculation of the significance of the correlation we're coming to this. But, if there is a correlation, does the regression line (the fit) make sense?
4. If we are still happy - is the formal result is realistic? Rule of Thumb - if 10 percent of the points are grouped by themselves so that covering them with the thumb destroys the correlation to the eye, then we should doubt it. Selection effects, data errors, or some other form of statistical conspiracy?

Introduction

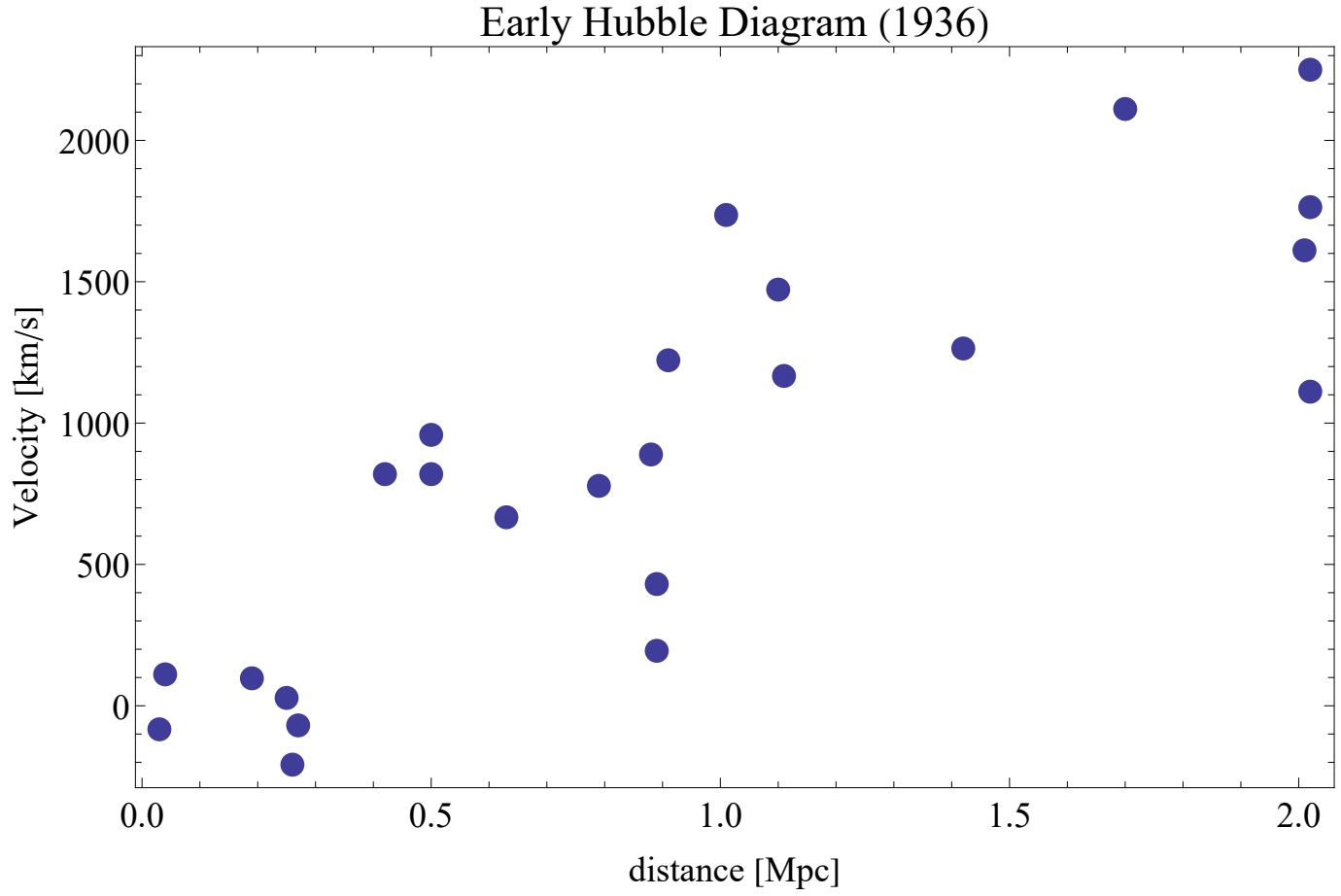
Out[ ]=



Suspect correlations: in each case formal calculation will indicate that a correlation exists to a high degree of significance. If still confident, remember that a **correlation does not prove causal connection**. Correlation may simply indicate the dependence on a third variable.

Introduction

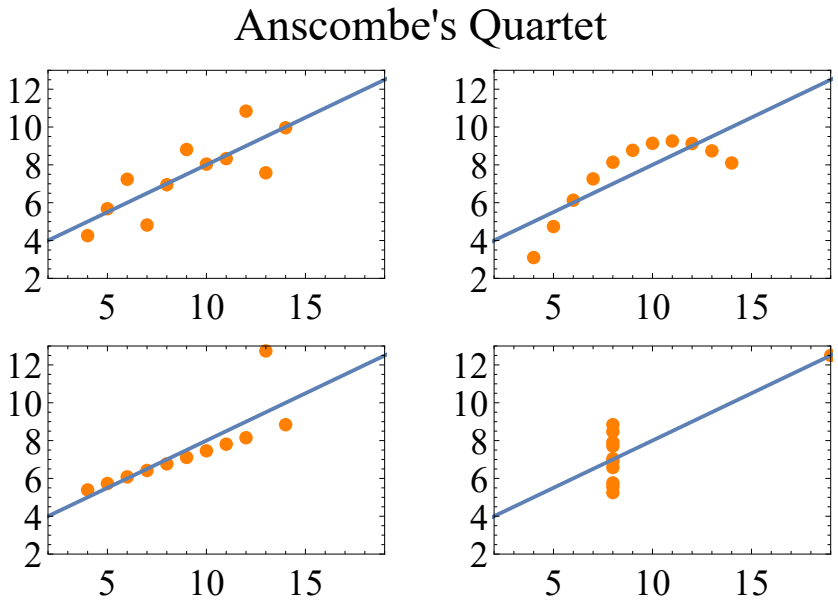
Still, there is data with meaningful correlation.



Anscombe's Quartet

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

Out[ ]=



Mean of each set	9
Variance in each set	11
Mean of y in each case	7.5
Variance in y in each case	4.122 or 4.127
Correlation between x and y in each set	0.816
Linear regression line in each case	y=3.00+0.500 x

Correlation - the standard definition

We have a set of measurements $\{X_i, Y_i\}$ and we ask (formally) if they are related to each other. What does 'related' mean? In general we model our data as a bivariate or joint Gaussian of **correlation coefficient** ρ:

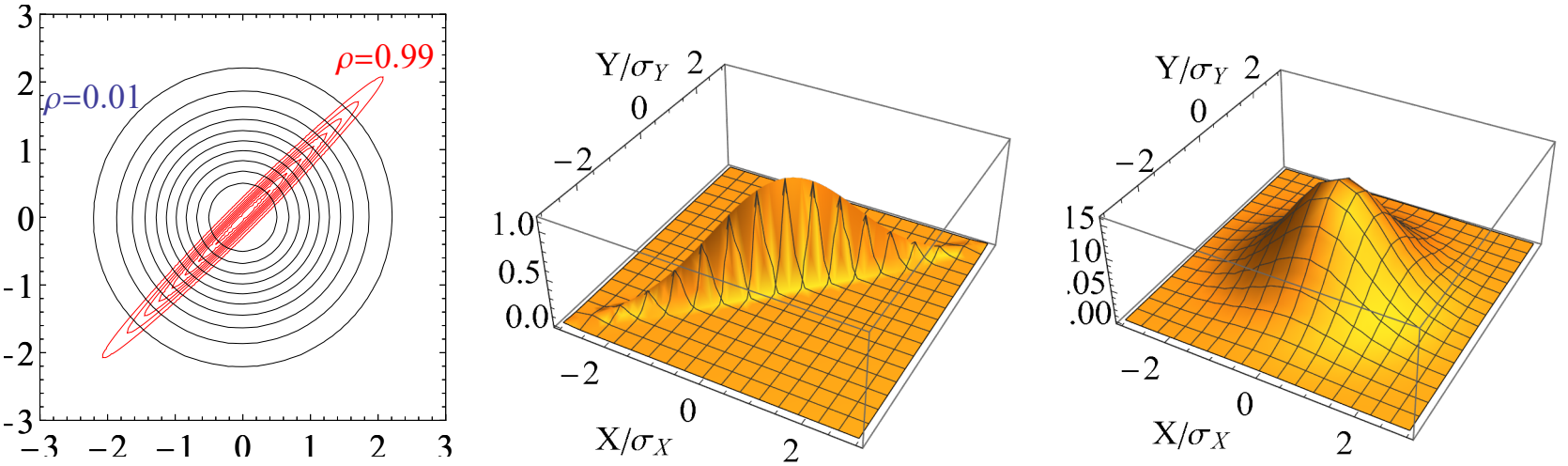
Out[ ]//TraditionalForm=

$$\mathcal{P}(x,\ y \mid \sigma_x,\ \sigma_y\ \rho) = \frac{1}{2\,\pi\,\sigma_x\,\sigma_y\,\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2\,(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\,\rho\,(x-\mu_x)\,(y-\mu_y)}{\sigma_x\,\sigma_y}\right)\right)$$

Correlation - the standard definition

if $\rho \to 0$ there is little correlation, while if $\rho \to 1$ the correlation is perfect. Negative values of $\rho \to$ 'anti-correlation'.

Out[•]=



The parameter **ρ is the correlation coefficient** (**of the population**) and is given by

Out[•]//TraditionalForm=

$$\rho = \frac{\text{Covariance}[X,\ Y]}{\sigma_X\, \sigma_Y}$$

Correlation - the standard definition

The correlation coefficient **for a sample** can be estimated by

$$r == \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2}}$$

r is known as the Pearson product-moment correlation coefficient (of the sample).

The contours will have dropped by $1/E$ from the value at the origin (assuming zero means) when

$$\frac{1}{1-\rho^2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\,\rho\,x\,y}{\sigma_x\,\sigma_y}\right) == 1$$

in matrix notation

Out[•]//TraditionalForm=

$$\begin{pmatrix} x & y \end{pmatrix} \frac{1}{1-\rho^2} \begin{pmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x\,\sigma_y} \\ -\frac{\rho}{\sigma_x\,\sigma_y} & \frac{1}{\sigma_y^2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 1$$

Correlation - the standard definition

The inverse of the central matrix is the ***covariance matrix*** (or *error matrix*).

Out[•]//TraditionalForm=

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \text{cov}(x,\ y) \\ \text{cov}(x,\ y) & \sigma_y^2 \end{pmatrix}$$

The off-diagonal elements of the covariance matrix (the ***sample*** covariance of N observations) can be estimated by

Out[•]//TraditionalForm=

$$\text{cov}(X_i,\ Y_i) = \sigma_{X_i X_j} = \frac{1}{\text{N} - 1} \overline{(X_i - \overline{X})(X_j - \overline{X})} = \frac{1}{\text{N} - 1} \sum_{k=1}^{\text{N}} (X_{i,k} - \overline{X_i})(X_{j,k} - \overline{X_j})$$

Correlation - the standard definition

The reason the **sample** covariance matrix has $N - 1$ in the denominator rather than N is essentially that the **population** mean $E(X)$ is not known and is replaced by the **sample** mean $\overline{X}$. If the **population** mean $E(X)$ is known, the analogous unbiased estimate is given by

Out[ ]//TraditionalForm=

$$\operatorname{cov}(X_i,\ Y_i) = \sigma_{X_i\,X_j} = \frac{1}{\mathrm{N}}\overline{\left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)} = \frac{1}{\mathrm{N}}\sum_{k=1}^{\mathrm{N}}\left(X_{i,k} - \overline{X_i}\right)\left(X_{j,k} - \overline{X_j}\right)$$

Correlation - the standard definition

Covariance is a measure of how much two random variables change together. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the smaller values, the covariance is positive.

In the opposite case, when the greater values of one variable mainly correspond to the smaller values of the other, i.e., the variables tend to show opposite behavior, the covariance is negative.

The sign of the covariance therefore shows the tendency in the linear relationship between the variables.

The magnitude of the covariance is not easy to interpret. The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

Example:

Let's calculate the covariance of three data points $\{x_i, y_i\}$, i.e. the covariance of the corresponding x- and y-vectors X and Y.

Consider the data points: **{0,0},{2,3},{10,9}**

Out[•]= $\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{1}{3}(0+2+10) = 4 \qquad \bar{y} = \frac{1}{3}\sum_{i=1}^{3} y_i = \frac{1}{3}(0+3+9) = 4$

Out[•]//TraditionalForm=

$$\sigma_X = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(x_i-\bar{x})^2} = \sqrt{\frac{1}{3}\left((-4)^2+(-2)^2+6^2\right)} = \sqrt{\frac{1}{3}(16+4+36)} = 4.32$$

Out[•]//TraditionalForm=

$$\sigma_Y = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(y_i-\bar{y})^2} = \sqrt{\frac{1}{3}\left((-4)^2+(-1)^2+5^2\right)} = \sqrt{\frac{1}{3}(16+1+25)} = 3.74$$

Out[•]//TraditionalForm=

$$\mathrm{cov}(X,\,Y) = \frac{1}{3-1}\sum_{i=1}^{3}(x_i-\bar{x})\,(y_i-\bar{y}) = \frac{1}{2}\left((0-4)\,(0-4)+(2-4)\,(3-4)+(10-4)\,(9-4)\right) = \frac{1}{2}(16+2+30) = 24$$

Example 2:

Now consider the data points: **{0,0},{2,3},{10,0}**

Out[ ]= $\bar{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{1}{3}(0+2+10) = 4$ $\quad\quad$ $\bar{y} = \frac{1}{3}\sum_{i=1}^{3} y_i = \frac{1}{3}(0+3+0) = 1$

Out[ ]//TraditionalForm=

$$\sigma_X = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(x_i-\bar{x})^2} = \sqrt{\frac{1}{3}\left((-4)^2+(-2)^2+6^2\right)} = \sqrt{\frac{1}{3}(16+4+36)} = 4.32$$

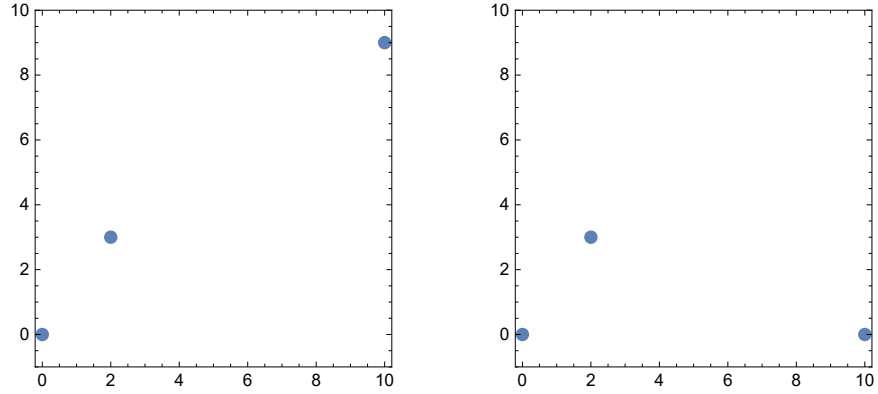
Out[ ]//TraditionalForm=

$$\sigma_Y = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(y_i-\bar{y})^2} = \sqrt{\frac{1}{3}\left((-1)^2+(-2)^2+(-1)^2\right)} = \sqrt{\frac{1}{3}(1+4+1)} = 1.414$$

Out[ ]//TraditionalForm=

$$\text{cov}(X,\,Y) = \frac{1}{3-1}\sum_{i=1}^{3}(x_i-\bar{x})\,(y_i-\bar{y}) = \frac{1}{2}\left((0-4)\,(0-1)+(2-4)\,(3-1)+(10-4)\,(0-1)\right) = \frac{1}{2}(4-4-6) = -3$$

Out[ ]=



Generate bivariate random numbers

To generate numbers obeying a bivariate (or even multivariate) Gaussian, with given σ_i and ρ_i:

1. Set up error matrix: $\mathrm{C} = \begin{pmatrix} \sigma_x^2 & \rho\, \sigma_x\, \sigma_y \\ \rho\, \sigma_x\, \sigma_y & \sigma_y^2 \end{pmatrix}$

2. Find eigenvalues λ_x and λ_y and eigenvectors $\vec{e}_x = \begin{pmatrix} e_{x,1} \\ e_{x,2} \end{pmatrix}$ and $\vec{e}_y = \begin{pmatrix} e_{y,1} \\ e_{y,2} \end{pmatrix}$ of C

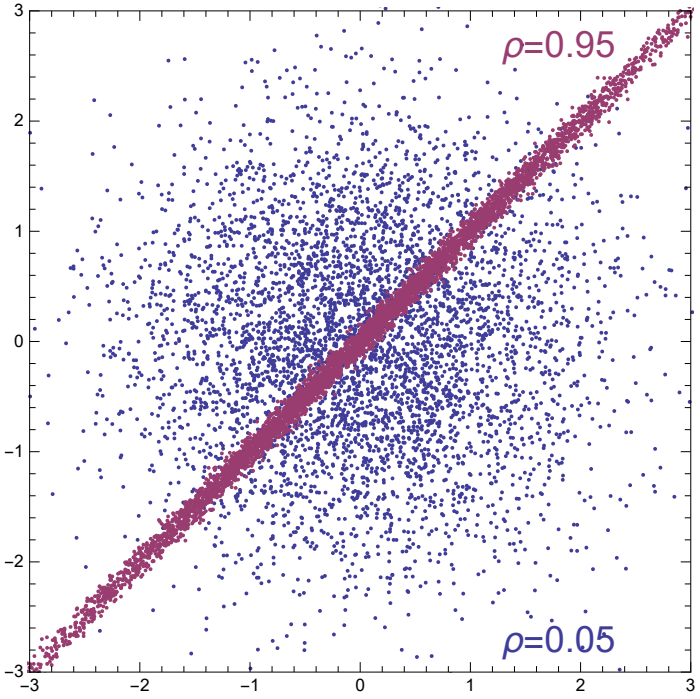
3. Create the transformation matrix T, the matrix that diagonalizes the covariance matrix, using the eigenvectors as column vectors

$T = \begin{pmatrix} e_{x,1} & e_{y,1} \\ e_{x,2} & e_{y,2} \end{pmatrix}$

Generate bivariate random numbers

1. Variate $\{X', Y'\}$ pairs (uncorrelated) from a Gaussian distribution with variances equal to the two eigenvalues: $\sigma_{x'} = \lambda_x$ and $\sigma_{y'} = \lambda_y$.

2. Compute the $\{X', Y'\}$ pairs (correlated with ρ) using: $\begin{pmatrix} X \\ Y \end{pmatrix} = [T] \begin{pmatrix} X' \\ Y' \end{pmatrix}$



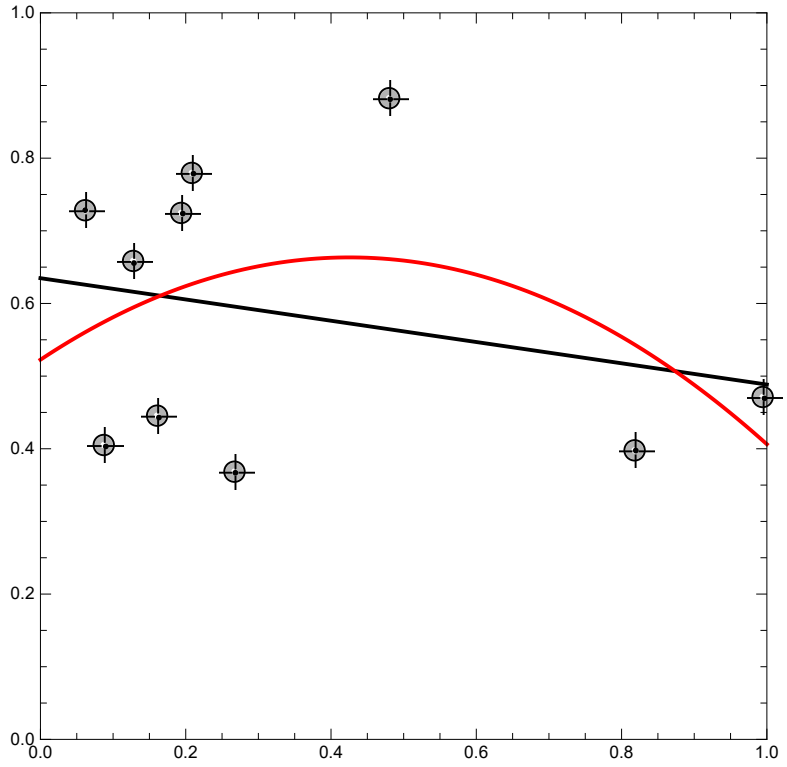
Correlation test shows different correlation index as required!!

Example

Out[•]=

reset | random

Press [Alt]+[LeftMouse] to add/remove data points



X	Y	X − $\bar{X}$	Y − $\bar{Y}$
0.481	0.881	0.140	0.296
0.211	0.779	−0.131	0.194
0.268	0.367	−0.073	−0.218
0.196	0.724	−0.146	0.139
0.062	0.728	−0.280	0.144
0.129	0.656	−0.213	0.071
0.820	0.398	0.478	−0.187
0.163	0.443	−0.178	−0.142
0.997	0.469	0.655	−0.115
0.090	0.403	−0.252	−0.182

$\bar{X}$=0.341586

$\bar{Y}$=0.58477

σ[X] =0.323144

σ[Y] =0.18834

Cov[X,Y] =−0.0152919

ρ=−0.251259

Multivariate-Gaussian Distribution

The Gaussian distribution defined over a $\mathcal{D}$-dimensional vector x of continuous variables is given by

Out[ ]//TraditionalForm=

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\mathcal{D}/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^{\mathsf{T}}.\Sigma^{-1}.(x-\mu)\right)$$

where the $\mathcal{D}$-dimensional vector μ is called the mean, the $\mathcal{D} \times \mathcal{D}$ matrix Σ is called the covariance, and $|\Sigma|$ denotes the determinant of Σ.

Conditional Gaussian distributions

An important property of the multivariate Gaussian distribution is that if two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian. Similarly, the marginal distribution of either set is also Gaussian.

Suppose x is a $\mathcal{D}$-dimensional vector with Gaussian distribution $\mathcal{N}(x|\mu,\Sigma)$ and that we partition x into two disjoint subsets x_a and x_b. Without loss of generality, we can take x_a to form the first M components of x, with x_b comprising the remaining $\mathcal{D}$–M components, so that

Out[ ]//TraditionalForm= $$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \quad \text{and,} \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

and of the covariance matrix Σ given by

Out[ ]//TraditionalForm= $$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

Conditional Gaussian distributions

The inverse of the covariance matrix is known as the precision matrix. Because the inverse of a symmetric matrix is also symmetric, we see that Λ_{aa} and Λ_{bb} are symmetric, while
$\Lambda_{ab}{}^{T} = \Lambda_{ba}$. Attention: Λ_{aa} is not simply given by the inverse of Σ_{aa}.

Out[ ]//TraditionalForm=
$$\Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

It turns out, that the expression for the conditional distribution $p(x_a \mid x_b)$ will again be Gaussian with

Out[ ]//TraditionalForm=
$$p(x_a \mid x_b) = \mathcal{N}\left(x \mid \mu_{a|b}, (\Lambda_{aa})^{-1}\right)$$

Out[ ]//TraditionalForm=
$$\mu_{a \mid b} = \mu_a - \Lambda_{ab} (\Lambda_{aa})^{-1} (x_b - \mu_b) = \mu_a + \Sigma_{ab} (\Sigma_{bb})^{-1} (x_b - \mu_b)$$

Out[ ]//TraditionalForm=
$$\Sigma_{a \mid b} = \Sigma_{aa} - \Sigma_{ab} (\Sigma_{bb})^{-1} \Sigma_{ba}$$

Conditional Gaussian distributions

We have seen that if a joint distribution $p(x_a, x_b)$ is Gaussian, then the conditional distribution $p(x_a \mid x_b)$ will again be Gaussian. Now we turn to a discussion of the marginal distribution given by

Out[ ]//TraditionalForm= $$p(x_a) = \int p(x_a, x_b)\, d x_b$$

without derivation, the mean and covariance are given by

Out[ ]//TraditionalForm= $$\mathbb{E}(x_a) = \mu_a$$

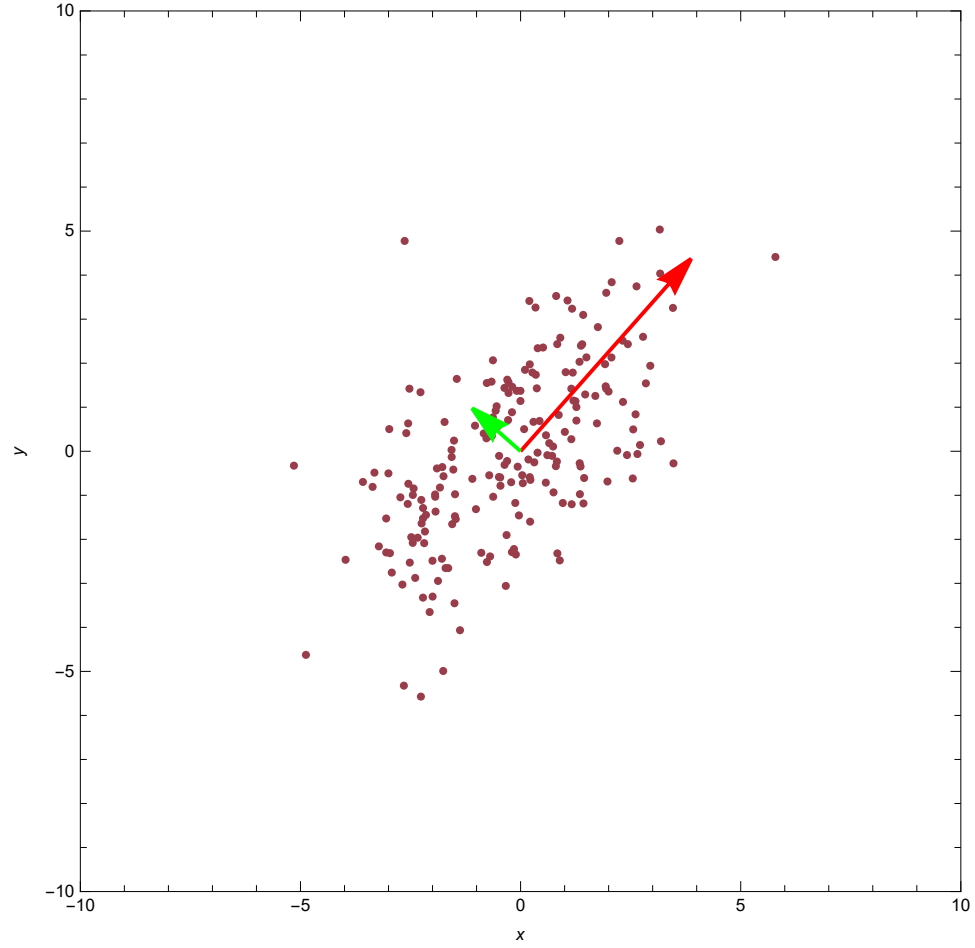
Out[ ]//TraditionalForm= $$\text{Covariance}[x_a] = \Sigma_{aa}$$

We see that for a marginal distribution, the mean and covariance are most simply expressed in terms of the partitioned covariance matrix.

Geometric interpretation of covariance matrix

Variance can only be used to explain the spread of the data in the directions parallel to the axes of the feature space. The diagonal spread is captured by the covariance

Out[•]=



$$\sigma(x,y) = E[(x - E[x])(y - E[y])]$$

$$\Sigma = \text{cov} = \begin{pmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{pmatrix} = \begin{pmatrix} 3.393 & 2.1696 \\ 2.1696 & 3.91818 \end{pmatrix}$$

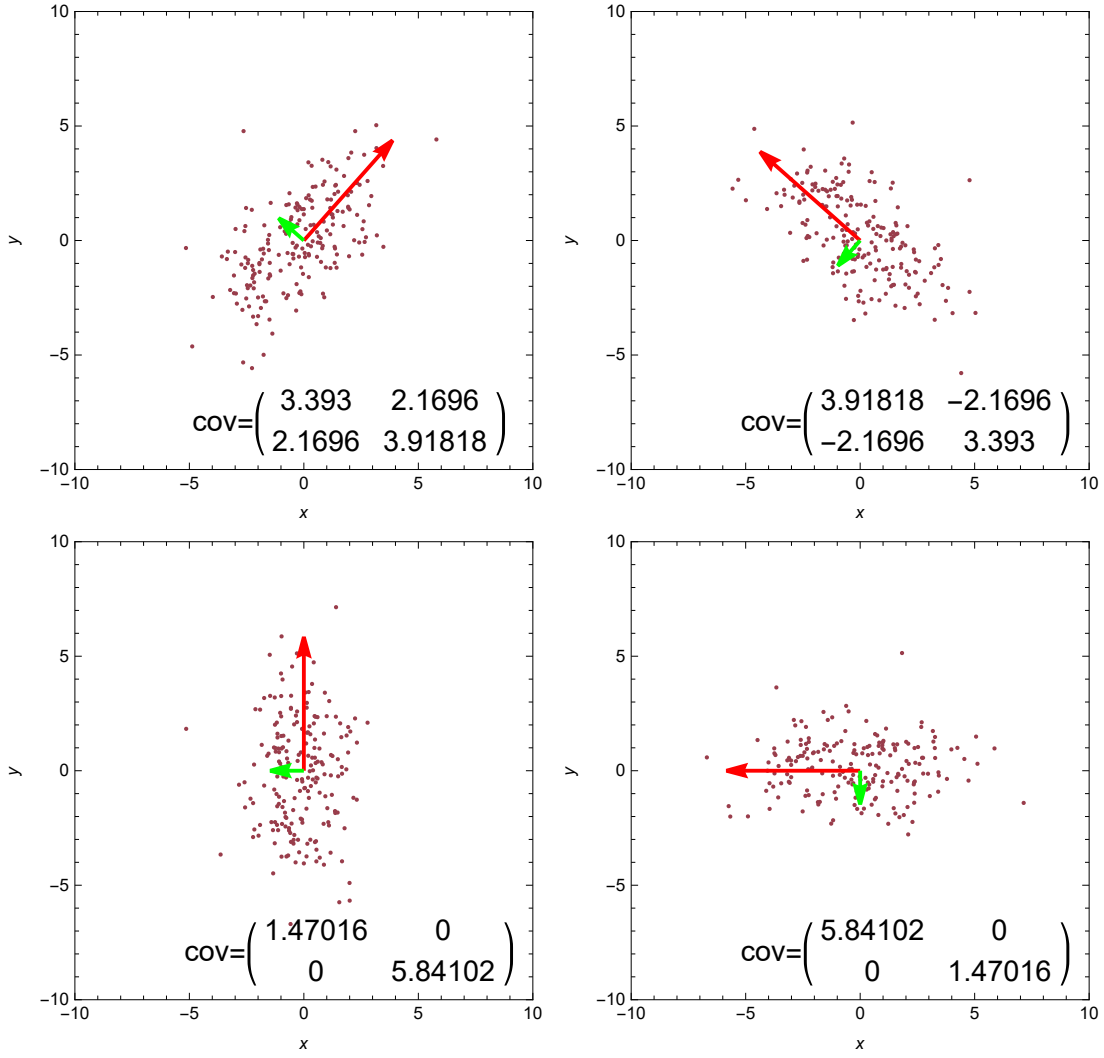
Geometric interpretation of covariance matrix

If x is positively correlated with y, y is also positively correlated with x. In other words, we can state that $\sigma(x, y) = \sigma(y, x)$. Therefore, the covariance matrix is always a symmetric matrix with the variances on its diagonal and the covariances off-diagonal. Two-dimensional normally distributed data is explained completely by its mean and its 2×2 covariance matrix. Similarly, a covariance matrix is used to capture the spread of three-dimensional data, and a 3×3 covariance matrix captures the spread of N-dimensional data.

Geometric interpretation of covariance matrix

The covariance matrix defines the shape of the data. Diagonal spread is captured by the covariance, while axis-aligned spread is captured by the variance.

Out[ ]=



Testing for correlation - Bayesian way

The bivariate Gaussian is a rather special case and does not help in everyday problems. It does not apply, for example, to data where the x-values are well-defined and there are 'errors' only in y, perhaps different at different x. In such cases we would use model-fitting, perhaps of a straight line. This is a different issue This is model-fitting, or parameter-estimation.

We will use Bayes' Theorem to extract the probability distribution for ρ from the likelihood of the data and suitable priors.

We want to know about ρ independently of any inference about the means and variances, we have to integrate these 'nuisance variables' out of the full posterior probability $\mathcal{P}(\rho, \sigma_x, \sigma_y, \mu_x, \mu_y \mid \text{data})$.

Testing for correlation - Bayesian way

For the bivariate Gaussian model, the result is given by Jeffreys (1961) as

Out[ ]//TraditionalForm=

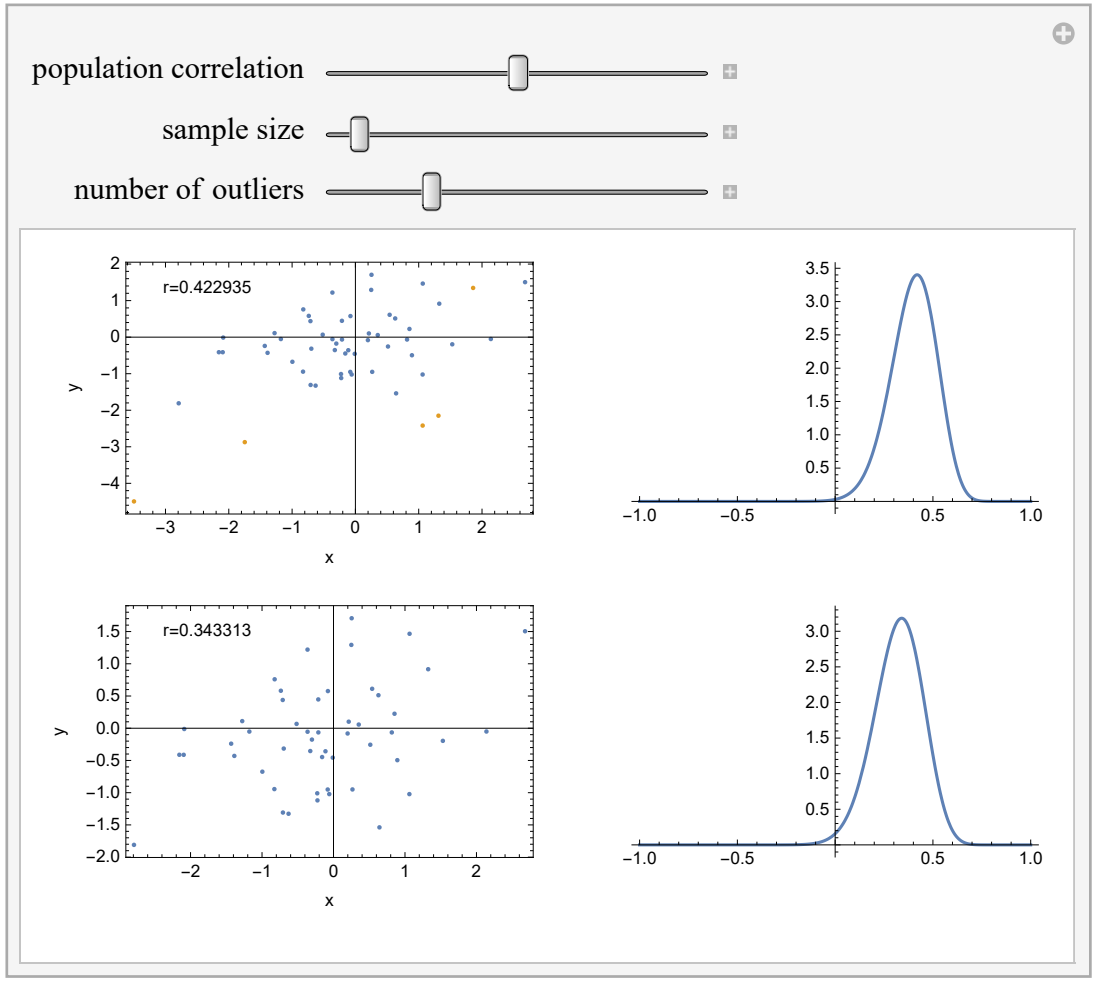
$$\mathcal{P}(\rho \mid \text{data}) \propto \frac{\left(1-\rho^2\right)^{\frac{n-1}{2}}}{(1-\rho\, r)^{n-\frac{3}{2}}}\left(1+\frac{1}{n-\frac{1}{2}}\,\frac{1+r\,\rho}{8}+\ldots\right)$$

The Bayesian test for correlation is simple: compute r from the $\{X_i, Y_i\}$, and calculate $\mathcal{P}(\rho)$ for the range of interest.

Example

Generate 50 samples from a bivariate Gaussian using true correlation coefficient of 0.5 and add some outliers, not accounted for by assuming a Gaussian.

Out[ ]=



Top panels: data with outliers, bottom panels: data with outliers removed → much better result!

The left panels show the data points, the corresponding Pearson correlation index *r* is given in the top left. The right panels show the probability distribution $\mathcal{P}(\rho \mid \text{data})$.

Example

We note, that the peak of $\mathcal{P}(\rho \mid \text{data})$ about corresponds to the r estimator computed from the data. So what is the advantage?

Given this probability distribution for ρ, we can answer questions like

what is the probability that $\rho > 0.5$?

what is the probability that ρ from data set A is bigger than ρ from data set B?

Jeffreys used a uniform prior for ρ, not obviously justifiable, and certainly not correct if ρ is close to 1 or -1. But in these cases a statistical test is a waste of time anyway.

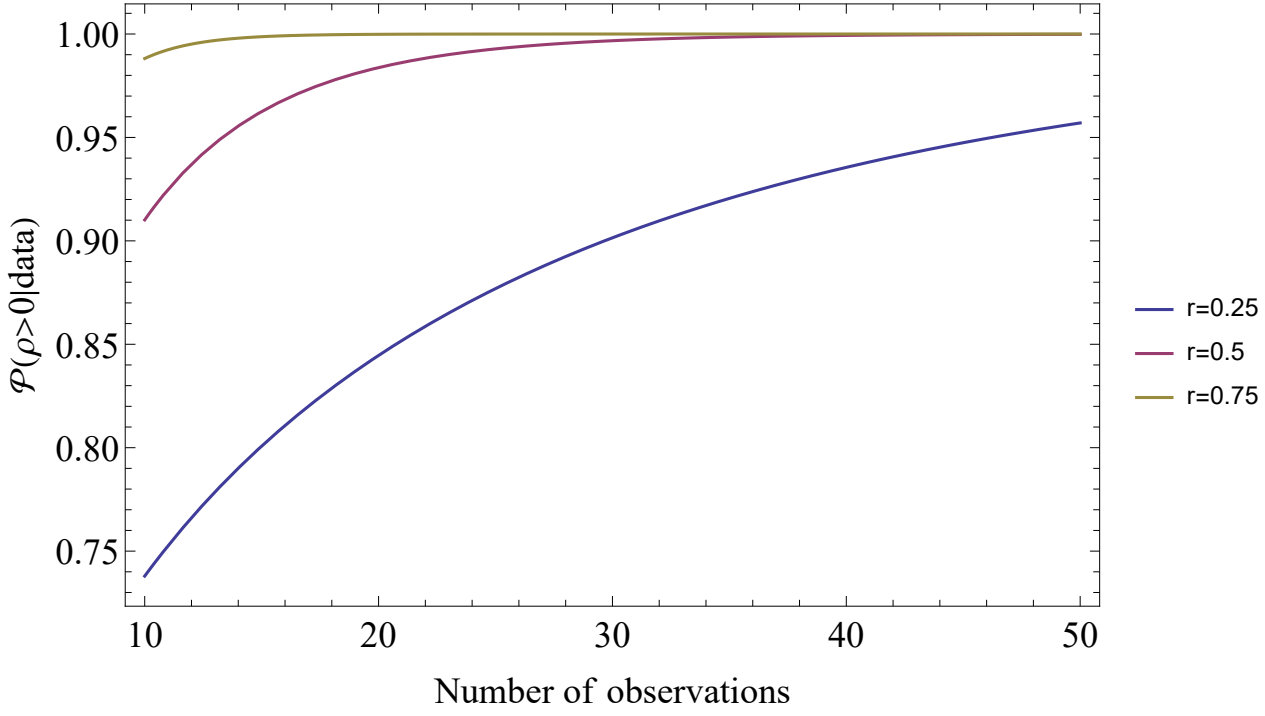
Example: Correlation sign as function of sample size

Out[ ]//TraditionalForm=

$$\mathcal{P}(\rho \mid \text{data}) \propto \frac{\left(1-\rho^2\right)^{\frac{n-1}{2}}}{(1-\rho\, r)^{n-\frac{3}{2}}}\left(1+\frac{1}{n-\frac{1}{2}}\left(\frac{1}{8}(1+r\,\rho)\right)+\ldots\right)$$

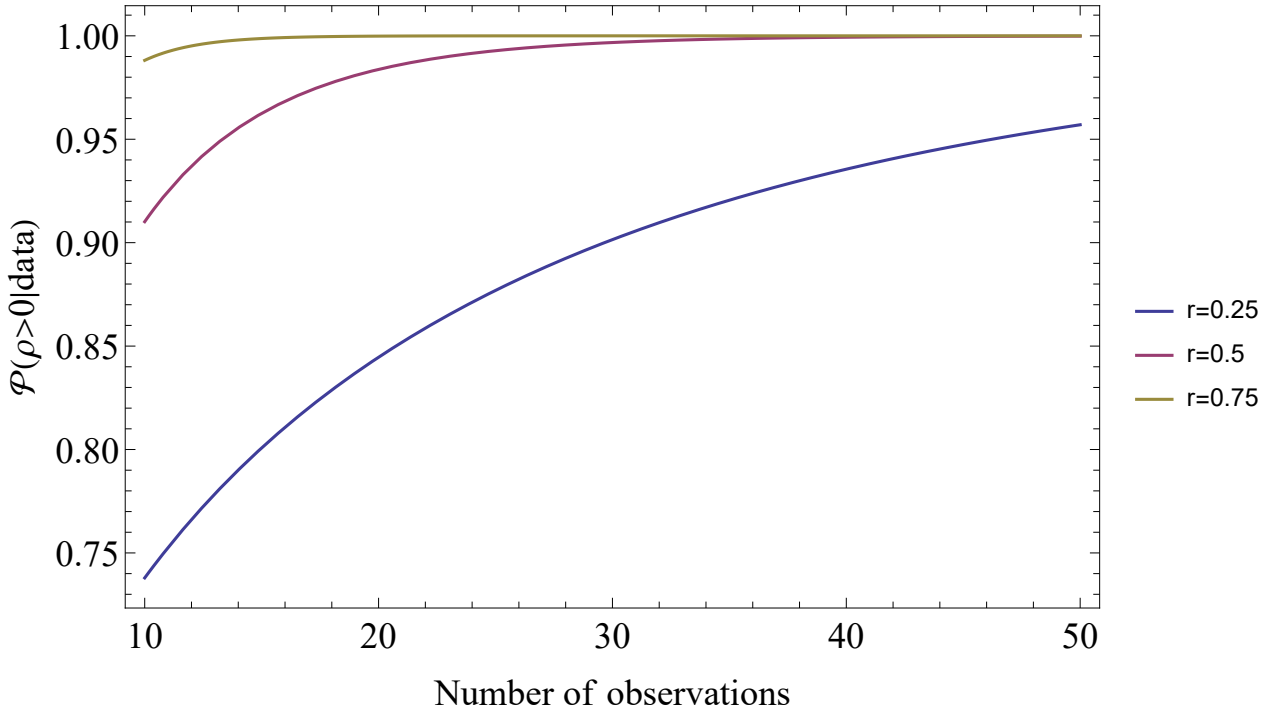
We note, that the sample size is a parameter of the above function. For a given, population correlation ρ we can calculate the probability that ρ is positive as a function of sample size. This tells us how much data we need to be confident of detecting correlations.

Out[ ]=



Example: Correlation sign as function of sample size

Out[ ]=



Here we plot $\int_0^1 \frac{\mathcal{P}(\rho,r,n)}{\int_{-1}^{1}\mathcal{P}(\rho,r,n)\,d\rho}\,d\rho$ as a function of n for values r = 0.25, 0.5, 0.75 (bottom to top line) assuming that the sample r reflects the true population ρ.

Correlation testing - Classical approach

The standard (parametric) test is to attempt to reject the null hypotheses, namely that $\rho = 0$ (i.e. no correlation).

1. first calculate *r*. The standard deviation in *r* is

Out[ ]//TraditionalForm=

$$\sigma_r = \frac{1 - r^2}{\sqrt{N - 1}}$$

2. To test the significance of a non-zero value for r, compute

Out[ ]//TraditionalForm=

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

which obeys the probability distribution of the Student's *t* statistics with $N - 2$ degrees of freedom.

Correlation testing - Classical approach

In the general case, i.e. the null hypothesis H_0 : $\rho = \rho_0$ we use the Fisher z transformation of the sample correlation r

1. Compute the Fisher's z-transformations:

Out[•]=
$$z = F(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right) = \tanh^{-1}(r)$$

with mean value

$$\overline{z} = \tanh^{-1}(\rho_0) + \frac{\rho_0}{2\,(N-1)} = \frac{1}{2}\left(\log\left(\frac{1+\rho_0}{1-\rho_0}\right) + \frac{\rho_0}{N-1}\right)$$

and standard error

$$SE = \frac{1}{\sqrt{N-3}}$$

Correlation testing - Classical approach

1. Thus, a Z-score is

Out[•]//TraditionalForm=

$$Z = \frac{\text{obs} - \text{mean}}{\text{SE}} = \left(F[r] - F[\rho_0] - \frac{\rho_0}{2\,(\text{N} - 1)}\right)\sqrt{\text{N} - 3}$$

under the null hypothesis of that $\rho = \rho_0$, given the assumption that the sample pairs are independent and identically distributed and follow a bivariate normal distribution. Thus an approximate p-value can be obtained from a normal probability table.

Correlation testing - Classical approach

1. The significance level at which a measured value of r differs from some hypothesized value ρ is given by:

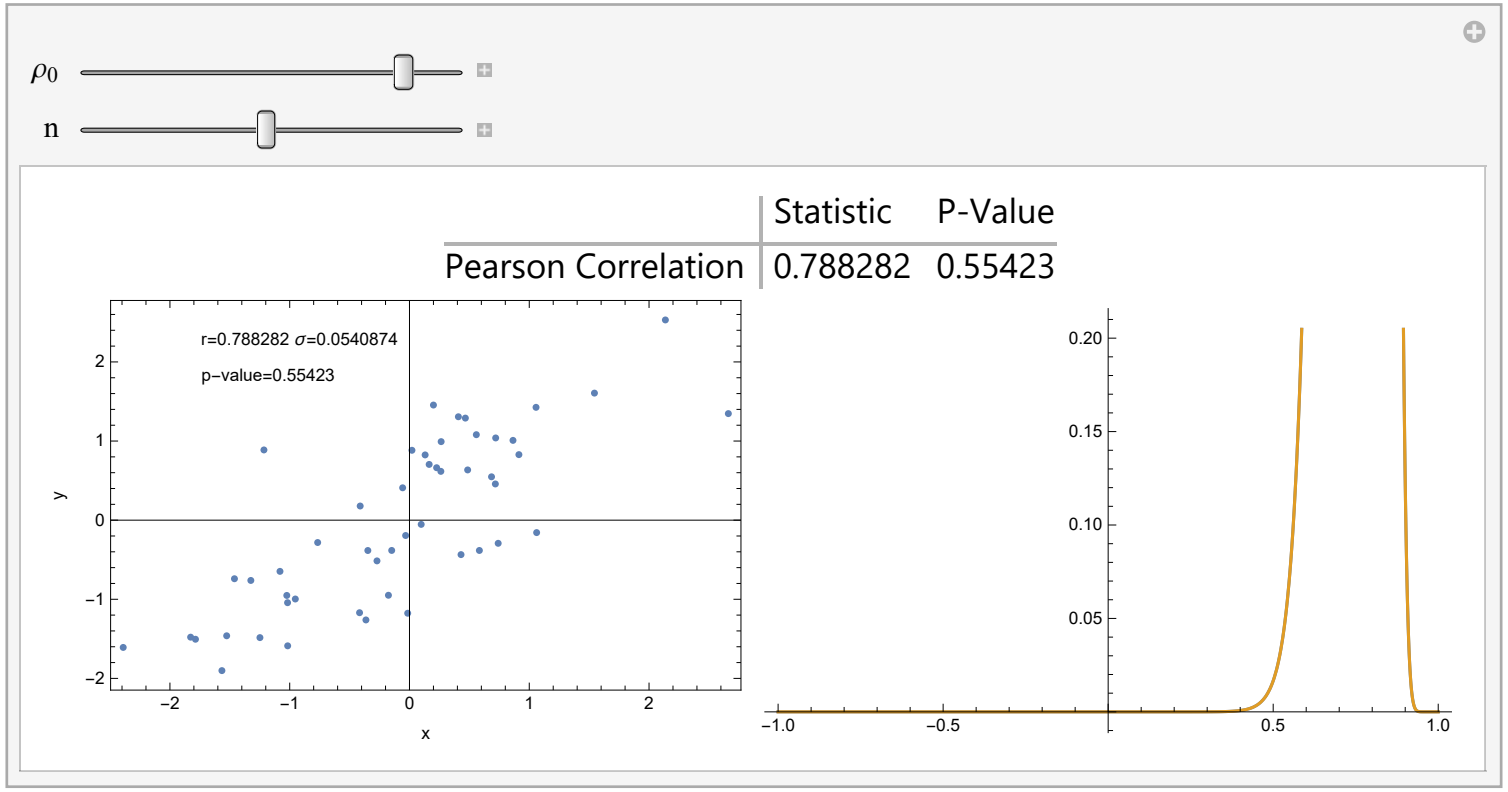
Out[•]= $\mathcal{P}_r(X > z) = \int_z^\infty \mathcal{N}(0, 1)\, \mathbb{d}x = \text{erfc}\left(\frac{|z|}{\sqrt{2}}\right)$ thus $\boxed{\mathcal{P}_r(X > z) = \text{erfc}\left(\frac{|z - \overline{z}|\ \sqrt{N - 3}}{\sqrt{2}}\right)}$

This is the probability, that in case of the null hypothesis, i.e. no correlation, $|r|$ should be larger that the observed value. This value is also called p-value. If the p-value is small, the null hypothesis can most likely be rejected. (Remains the question of what is a small enough p-value?)

Example

Given a bivariate data sample with the sample's correlation coefficient r we perform a hypothesis test on the bivariate sample with null hypothesis H_0 that the population correlation coefficient $\rho = \rho_0$, and alternative hypothesis H_a that ρ != ρ_0.

Out[ ]=



Correlation testing - constraints

- The inclusion in the test of values of unobserved values of *r* in the example before is problematic.
- The test is widely used, and is formally powerful. But:
  - the data must be on continuous scales
  - the relation between them must be linear. (How would we know this?)
  - the data must be drawn from Normally-distributed populations. (How ..?)
  - they must be free from restrictions in variability or groupings.
- There are parametric tests that help: the F-test for non-linearity and the Correlation Ratio test which gets around non-linearity.
- to circumvent the problems it is far better to go to **non-parametric tests**.

Correlation testing - Non-parametric

The best known **non-parametric** test consists of computing the **Spearman rank correlation coefficient**. (non-parametric covers techniques that do not rely on data belonging to any particular distribution OR techniques that do not assume that the structure of a model is fixed.)

Spearman rank correlation coefficient

1. For the N data pairs of $\{X_i, Y_i\}$, make rank tables of X_i and Y_i such that $\{\mathrm{XR}_i, \mathrm{YR}_i\}$, pairs represent the ranks for the i –th pair, $1 < \mathrm{XR}_i < N,\ 1 < \mathrm{YR}_i < N$.

2. Compute the Spearman rank correlation coefficient:

Out[•]=
$$r_s = \left(\frac{\mathrm{Cov}[\mathrm{XR}, \mathrm{YR}]}{s_{\mathrm{XR}}\, s_{\mathrm{YR}}}\right) == \left(\frac{\mathrm{N}^3 - \mathrm{N} - \frac{T_X}{2} - \frac{T_Y}{2} - 6\sum_{i=1}^{\mathrm{N}} (\mathrm{XR}_i - \mathrm{YR}_i)^2}{\sqrt{\left(\mathrm{N}^3 - \mathrm{N} - T_X\right)\left(\mathrm{N}^3 - \mathrm{N} - T_Y\right)}}\right) \quad \text{with} \quad T_\bullet = \sum_k \left(t_{\bullet,k}^3 - t_{\bullet,k}\right)$$

Here, $t_{\bullet,k}$ is the number of observations with the same rank (● stands for X and Y). If all ranks occur only once, the equation simplifies to

Out[•]//TraditionalForm=
$$r_s = 1 - 6\,\frac{\sum_{i=1}^{\mathrm{N}} (\mathrm{XR}_i - \mathrm{YR}_i)^2}{\mathrm{N}^3 - \mathrm{N}}$$

3. The range is $0 < r_s < 1$; a high value indicates significant correlation. To find how significant, refer the computed r_s to the table of critical values of r_s applicable for $4 <= N <= 30$. If r_s exceeds an appropriate critical value, the hypothesis that the variables are unrelated is rejected at that level of significance.

4. If N exceeds 30, compute

Out[•]//TraditionalForm=
$$t_r = r_s \sqrt{\frac{\mathrm{N} - 2}{1 - r_s^2}}$$

a statistic whose distribution for large N asymptotically approaches that of the t statistic with $N - 2$ degrees of freedom. The significance of t_r may be found from the t-distribution table, and this represents the associated probability under the hypothesis that the variables are unrelated.

5. In comparison with r, r_s has an efficiency of 91%. That means, we need 100 $\{X_i, Y_i\}$ pairs for r_s to reveal an existing correlation at the same level of significance which r attains for 91 $\{X_i, Y_i\}$ pairs

6. Moral: if in doubt, go non-parametric.

Table: critical values of r_s

Table A2.5. Critical values of r_s, the Spearman rank correlation coefficient

	Level of significance for one-tailed test								
	0.250	0.100	0.050	0.025	0.010	0.005	0.0025	0.0010	0.0005
	Level of significance for two-tailed test								
	0.500	0.200	0.100	0.050	0.020	0.010	0.005	0.002	0.001
$N = 4$	0.600	1.000	1.000	–	–	–	–	–	–
5	0.500	0.800	0.900	1.000	1.000	–	–	–	–
6	0.371	0.657	0.829	0.886	0.943	1.000	1.000	–	–
7	0.321	0.571	0.714	0.786	0.893	0.929	0.964	1.000	1.000
8	0.310	0.524	0.643	0.738	0.833	0.881	0.905	0.952	0.976
9	0.267	0.483	0.600	0.700	0.783	0.833	0.867	0.917	0.933
10	0.248	0.455	0.564	0.648	0.745	0.794	0.830	0.879	0.903
11	0.236	0.427	0.536	0.618	0.709	0.755	0.800	0.845	0.873
12	0.224	0.406	0.503	0.587	0.671	0.727	0.776	0.825	0.860
13	0.209	0.385	0.484	0.560	0.648	0.703	0.747	0.802	0.835
14	0.200	0.367	0.464	0.538	0.622	0.675	0.723	0.776	0.811
15	0.189	0.354	0.443	0.521	0.604	0.654	0.700	0.754	0.786
16	0.182	0.341	0.429	0.503	0.582	0.635	0.679	0.732	0.765
17	0.176	0.328	0.414	0.485	0.566	0.615	0.662	0.713	0.748
18	0.170	0.317	0.401	0.472	0.550	0.600	0.643	0.695	0.728
19	0.165	0.309	0.391	0.460	0.535	0.584	0.628	0.677	0.712
20	0.161	0.299	0.380	0.447	0.520	0.570	0.612	0.662	0.696
21	0.156	0.292	0.370	0.435	0.508	0.556	0.599	0.648	0.681
22	0.152	0.284	0.361	0.425	0.496	0.544	0.586	0.634	0.667
23	0.148	0.278	0.353	0.415	0.486	0.532	0.573	0.622	0.654
24	0.144	0.271	0.344	0.406	0.476	0.521	0.562	0.610	0.642
25	0.142	0.265	0.337	0.398	0.466	0.511	0.551	0.598	0.630
26	0.138	0.259	0.331	0.390	0.457	0.501	0.541	0.587	0.619
27	0.136	0.255	0.324	0.382	0.448	0.491	0.531	0.577	0.608
28	0.133	0.250	0.317	0.375	0.440	0.483	0.522	0.567	0.598
29	0.130	0.245	0.312	0.368	0.433	0.475	0.513	0.558	0.589
30	0.128	0.240	0.306	0.362	0.425	0.467	0.504	0.549	0.580
31	0.126	0.236	0.301	0.356	0.418	0.459	0.496	0.541	0.571
32	0.124	0.232	0.296	0.350	0.412	0.452	0.489	0.533	0.563
33	0.121	0.229	0.291	0.345	0.405	0.446	0.482	0.525	0.554
34	0.120	0.225	0.287	0.340	0.399	0.439	0.475	0.517	0.547
35	0.118	0.222	0.283	0.335	0.394	0.433	0.468	0.510	0.539
36	0.116	0.219	0.279	0.330	0.388	0.427	0.462	0.504	0.533
37	0.114	0.216	0.275	0.325	0.383	0.421	0.456	0.497	0.526
38	0.113	0.212	0.271	0.321	0.378	0.415	0.450	0.491	0.519
39	0.111	0.210	0.267	0.317	0.373	0.410	0.444	0.485	0.513
40	0.110	0.207	0.264	0.313	0.368	0.405	0.439	0.479	0.507

41	0.108	0.204	0.261	0.309	0.364	0.400	0.433	0.473	0.501
42	0.107	0.202	0.257	0.305	0.359	0.395	0.428	0.468	0.495
43	0.105	0.199	0.254	0.301	0.355	0.391	0.423	0.463	0.490
44	0.104	0.197	0.251	0.298	0.351	0.386	0.419	0.458	0.484
45	0.103	0.194	0.248	0.294	0.347	0.382	0.414	0.453	0.479
46	0.102	0.192	0.246	0.291	0.343	0.378	0.410	0.448	0.474
47	0.101	0.190	0.243	0.288	0.340	0.374	0.405	0.443	0.469
48	0.100	0.188	0.240	0.285	0.336	0.370	0.401	0.439	0.465
49	0.098	0.186	0.238	0.282	0.333	0.366	0.397	0.434	0.460
50	0.097	0.184	0.235	0.279	0.329	0.363	0.393	0.430	0.456

Example:

Take the following data:

Out[•]=

1	2	3	4	5	6	7	8	9	10

X	XR	Y	YR	d=XR−YR	d^2	$t_{X,k}$	$t_{Y,k}$	$t_{X,k}^3-t_{X,k}$	$t_{Y,k}^3-t_{Y,k}$
50	10	1.8	2	8	64	1	1	0	0
175	9	1.2	3.5	5.5	30.25	1	2	0	6
270	8	2.	1	7	49	1	1	0	0
375	7	1.	6	1	1	1	3	0	24
425	6	1.	6	0	0	1	–	0	–
580	5	1.2	3.5	1.5	2.25	1	–	0	–
710	4	0.8	9	−5	25	1	1	0	0
790	3	0.6	10	−7	49	1	1	0	0
890	2	1.	6	−4	16	1	–	0	–
980	1	0.85	8	−7	49	1	1	0	0

Out[•]//TraditionalForm=

$$r_s = 1 - 6\,\frac{\sum_{i=1}^{N}(XR_i - YR_i)^2}{N^3 - N} = 1 - 6\,\frac{285.5}{1000 - 10} = -0.73$$

Example:

The r_s value of -0.73 suggests a fairly strong negative relationship.

Since there are bindings present, i.e. we have entries with the same rank, the above relation returns a wrong result. The full computation is

Out[•]//TraditionalForm=

$$r_s = \frac{\text{Cov}(\text{XR}, \text{YR})}{s_{\text{XR}}\, s_{\text{YR}}} = \frac{\text{N}^3 - \text{N} - \frac{T_X}{2} - \frac{T_Y}{2} - 6\sum_{i=1}^{\text{N}} (\text{XR}_i - \text{YR}_i)^2}{\sqrt{\left(\text{N}^3 - \text{N} - T_X\right)\left(\text{N}^3 - \text{N} - T_Y\right)}}$$

$T_X = 0$ and $T_Y = 30$

Out[•]//TraditionalForm=

$$r_s = \frac{\text{N}^3 - \text{N} - \frac{T_X}{2} - \frac{T_Y}{2} - 6\sum_{i=1}^{\text{N}} (\text{XR}_i - \text{YR}_i)^2}{\sqrt{\left(\text{N}^3 - \text{N} - T_X\right)\left(\text{N}^3 - \text{N} - T_Y\right)}} = \frac{1000 - 10 - \frac{0}{2} - \frac{30}{2} - 6 \times 285.5}{\sqrt{1000 - 10 + 0}\ \sqrt{1000 - 10 - 30}} = -0.757013$$

Example:

As a comparison we can use the original definition of

Out[•]//TraditionalForm=

$$r_s = \frac{\text{Cov}(\text{XR}, \text{YR})}{s_{\text{XR}}\, s_{\text{YR}}}$$

Out[•]= −0.757013

The Mathematica function to compute this returns the same result:

In[•]:=
```
SpearmanRho[data[[All, 1]], data[[All, 2]]]
```

Out[•]= −0.757013

Degrees of freedom: df = n − 2 = 10 − 2 = 8

From the table we find a 5% significance level of our test for a r_s = 0.738 and a 2% significance level for r_s = 0.833. That means that the probability of the relationship you have found being a chance event is below 5 in a 100. You are more than 95% certain that your hypothesis is correct.

The exact result is:

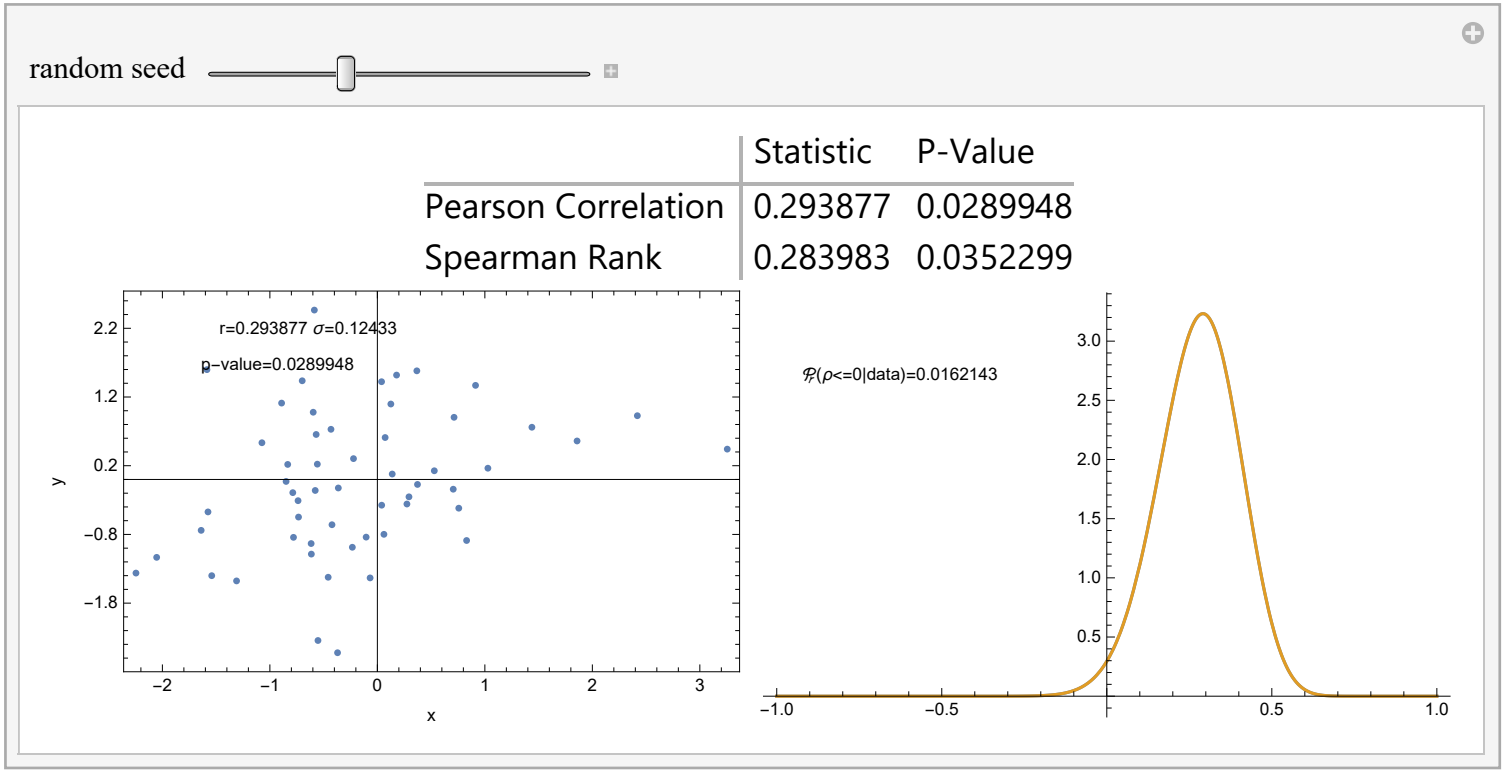
Out[•]=

	Statistic	P-Value
Spearman Rank	−0.757013	0.011239

Example - bad conclusion

Here we show a correlation at the notorious 2σ level, $r_s = 0.29$, $N = 55$. The hypothesis that the variables are unrelated $H_0 : \rho = 0$ is rejected at the 5 per cent level of significance. Here we have no idea of the underlying distributions; nor are we clear about the nature of the axes. The assumption of a bivariate Gaussian distribution would be rash in the extreme.

Out[•]=



HINT: The underlying true correlation coefficient $\rho = 0$!
All tests discard the probability that the data is uncorrelated!

Partial Correlation

A "hidden third variable" can be dealt with (provided that its influence is recognized in the first place) by **partial correlation** in which the partial correlation between two variables is considered by nullifying the effects of the third variable upon the variables being considered.

Consider a sample of N objects for which the parameters x_1, x_2 and x_3 have been measured. The **first order partial correlation coefficient** between variables x_1 and x_2 is

Out[ ]//TraditionalForm=

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}$$

If there are four variables, then the **second-order partial correlation coefficient** is

Out[ ]//TraditionalForm=

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{\left(1 - r_{14.3}^2\right)\left(1 - r_{24.3}^2\right)}}$$

Partial Correlation

where the correlation is being examined between x_1 and x_2 with x_3 and x_4 held constant.

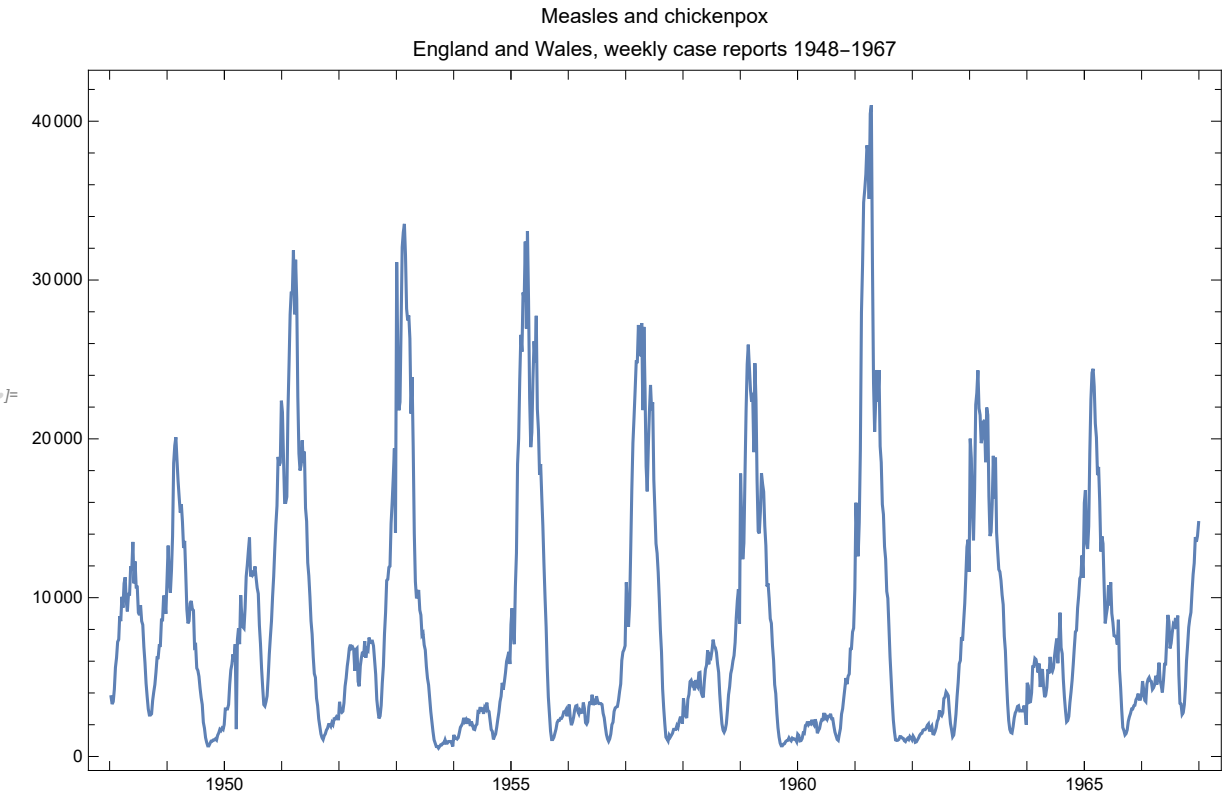
And so forth for higher-order partial correlations between more than four variables, with the **standard error of the partial correlation coefficients** being given by

Out[•]//TraditionalForm=

$$\sigma_{r_{12.34\mathrm{m}}} = \frac{1 - r^2_{12.34\mathrm{m}}}{\sqrt{\mathrm{N} - m}}$$

Autocorrelation

Out[ ]=

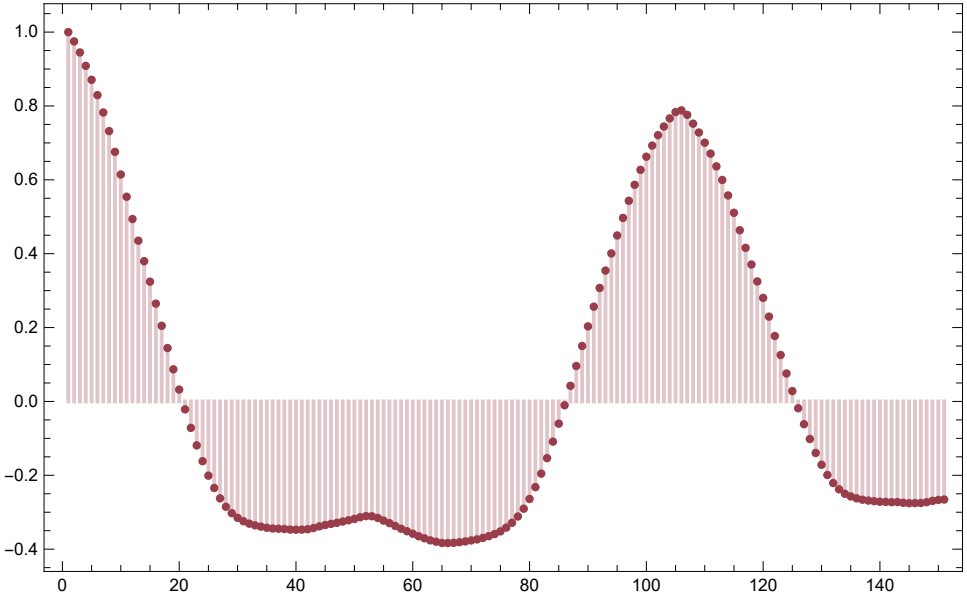


Autocorrelation is a measure of the internal correlation within a time series.

For example, for the measles time series you might ask: "take an arbitrary spot in time, on the average what does the time series look like in four weeks time, compared to now? " In other words, you are asking a question about how strong is the internal association within the series at a period of four weeks? That association could be very strong and positive (i.e. the series in four weeks is similar to now), or be very strong and negative (i.e. the series in four weeks is very dissimilar to now) or there could be a weak or no relationship (i.e. there is no identifiable association).

Autocorrelation

Out[ ]=



We can see that for time lag of a few weeks the auto-correlation is still very high, meaning that if we have a high in measles cases in a given week, it is very likely that the situation is similar in the coming weeks. There is an anti-correlation for a time lag of ab out one year, i.e. if we have many measles cases this year, we will have a measles low next year. And there seems to be a two year periodicity of measles infections indicated by the peak of the auto-correlation function at a time lag of ~105 weeks.

Autocorrelation

Formally, the auto-covariance describes the variance between values of a stochastic process $(X_t)_{t\in T}$ at different times

Out[•]//TraditionalForm=

$$\gamma(\text{t1},\ \text{t2}) = \text{Cov}(X_{\text{t1}},\ X_{\text{t2}}) = \text{E}((X_{\text{t1}} - \mu_{\text{t1}})\,(X_{\text{t2}} - \mu_{\text{t2}}))$$

The auto-correlation is defined as the normalized auto-covariance

Out[•]//TraditionalForm=

$$\rho(\text{t1},\ \text{t2}) == \frac{\gamma[\text{t1},\ \text{t2}]}{\sigma_{\text{t1}}\,\sigma_{\text{t2}}} \quad \text{, with} \quad -1 \le \rho(\text{t1},\ \text{t2}) \le +1$$

In signal theory, the auto-correlation function is used to correlate a signal with itself at various time shifts τ, called lag.

Out[•]//TraditionalForm=

$$\Psi(\tau) = \lim_{T\to\infty} \frac{1}{2\,T} \int_{-T}^{T} x(t)\,x(t+\tau)\,dt$$

PCA - Principal Component Analysis

1) Get some data and subtract the mean

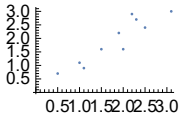
Out[269]=

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.
2.3	2.7
2.	1.6
1	1.1
1.5	1.6
1.1	0.9

DataAdjusted =

$x-\bar{x}$	$y-\bar{y}$
0.69	0.49
−1.31	−1.21
0.39	0.99
0.09	0.29
1.29	1.09
0.49	0.79
0.19	−0.31
−0.81	−0.81
−0.31	−0.31
−0.71	−1.01



PCA - Principal Component Analysis

2) Calculate the covariance matrix

```
Variance[dataAdjusted[[All, 1]]]
Variance[dataAdjusted[[All, 2]]]
```

```
0.616556
```

```
0.716556
```

$$\frac{\text{dataAdjusted[[All, 1]].dataAdjusted[[All, 2]]}}{10 - 1}$$

```
0.615444
```

In[270]:=
```
cov = Covariance[dataAdjusted]; cov // MatrixForm
```

Out[270]//MatrixForm=

$$\begin{pmatrix} 0.616556 & 0.615444 \\ 0.615444 & 0.716556 \end{pmatrix}$$

PCA - Principal Component Analysis

3) Calculate the eigenvalues and eigenvectors of the covariance matrix

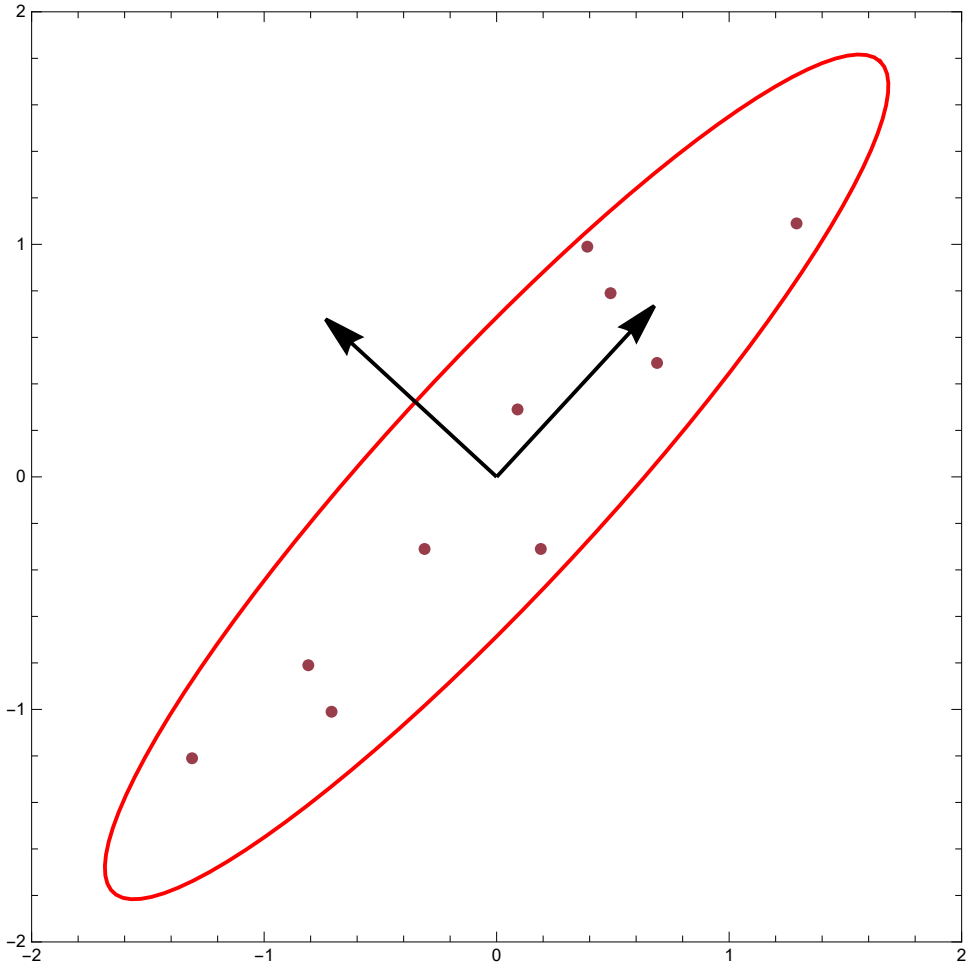
In[271]:=
```
eval = Eigenvalues[cov]; MatrixForm[eval]
evec = Eigenvectors[cov]; MatrixForm[evec]
```

Out[271]//MatrixForm=

$$\begin{pmatrix} 1.28403 \\ 0.0490834 \end{pmatrix}$$

Out[272]//MatrixForm=

$$\begin{pmatrix} 0.677873 & 0.735179 \\ -0.735179 & 0.677873 \end{pmatrix}$$


2
1
0
−1
−2
−2
−1
0
1
2

PCA - Principal Component Analysis

4) Choose components and from feature vector

Sort the eigenvectors by their eigenvalue, highest to lowest. This gives you the components in order of significance. The eigenvector with the **highest eigenvalue** is the **principal component** of the data set. In our example, the eigenvector with the larges eigenvalue was the one that pointed down the middle of the data. It is the most significant relationship between the data dimensions.

Form the feature vector, which is a matrix of vectors, formed from the eigenvectors that you want to keep from the list of eigenvectors.

Out[ ]//TraditionalForm=

$$\text{feature vector} = (\text{eig}_1, \text{eig}_2, \ldots, \text{eig}_p)$$

Given our example, we can either form a feature vector using both eigenvectors or we can choose to leave out the smaller, less significant component:

$$\begin{pmatrix} 0.677873 & -0.735179 \\ 0.735179 & 0.677873 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0.677873 \\ 0.735179 \end{pmatrix}$$

PCA - Principal Component Analysis

5) Derive the new data set

Once we have chosen the components (eigenvectors) that we wish to keep in our data and formed a feature vector, we simply take the transpose of the vector and multiply it on the left of the original data set, transposed.

Out[•]//TraditionalForm=

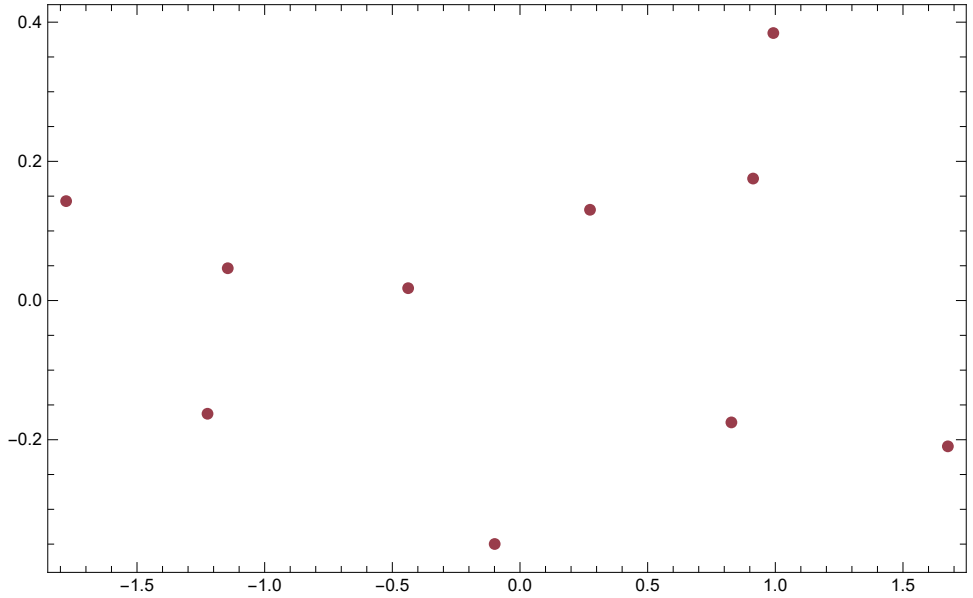
$$\text{final data} = \text{RowFeatureVector} \times \text{RowDataAdjusted}$$

where "**RowFeatureVector**" is the matrix with the eigenvectors in columns transposed, so that the eigenvectors are now in the rows with the most significant eigenvector at the top, "**RowDataAdjusted**" is the mean adjusted data transposed, i.e. the data items are in each column, with each row holding a separate dimension.

PCA - Principal Component Analysis

5) Derive the new data set

eig_1	eig_2
0.82797	−0.175115
−1.77758	0.142857
0.992197	0.384375
0.27421	0.130417
1.6758	−0.209498
0.912949	0.175282
−0.0991094	−0.349825
−1.14457	0.0464173
−0.438046	0.0177646
−1.22382	−0.162675



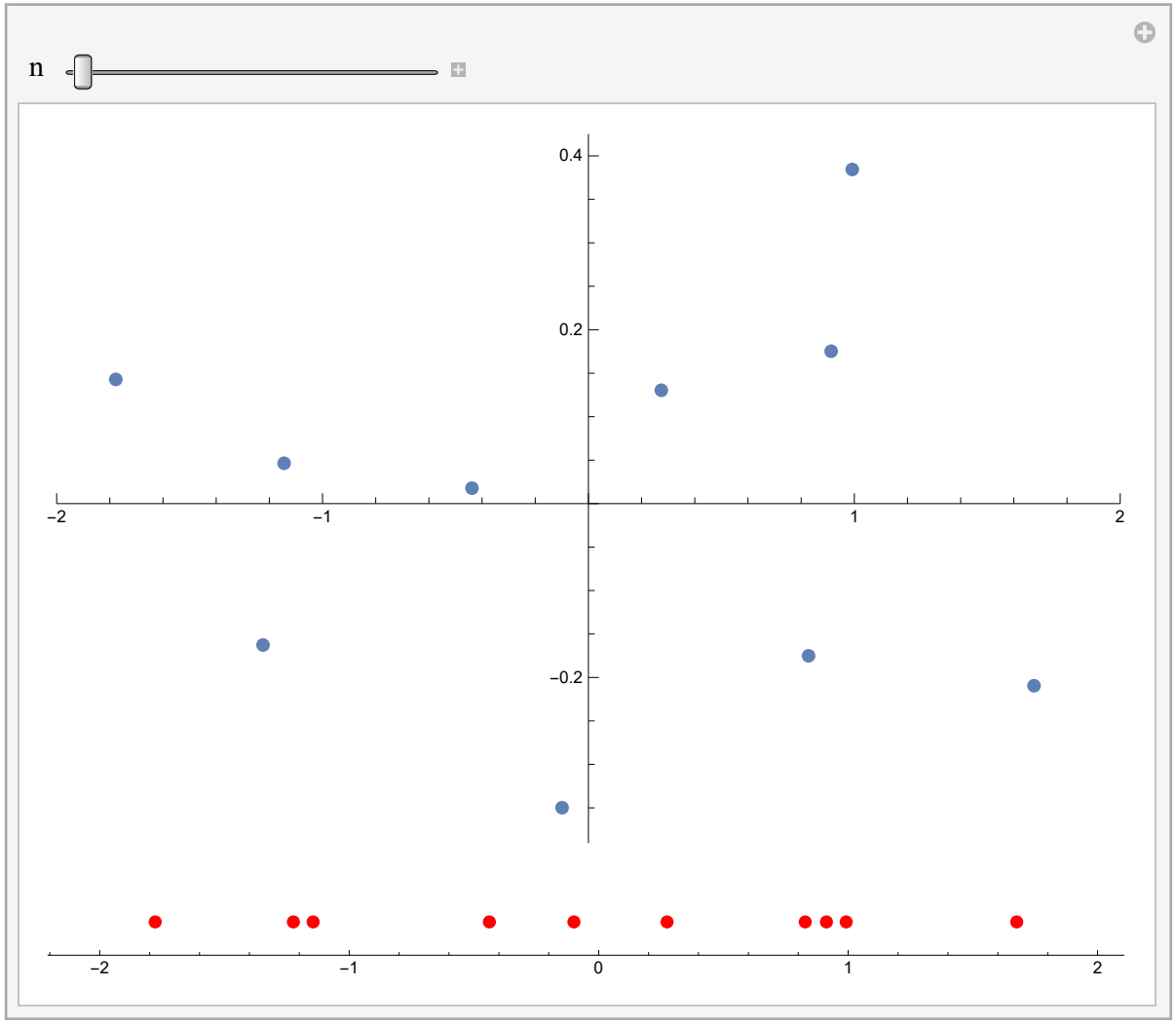
We changed our data from being in terms of the axes *x* and *y* to being in terms of our two eigenvectors. In the case of when the new data set has reduced dimensionality, i.e. we have left some of the eigenvectors out, the new data is only in terms of the vectors that we decided to keep.

PCA - Principal Component Analysis

5) Derive the new data set

Out[•]=

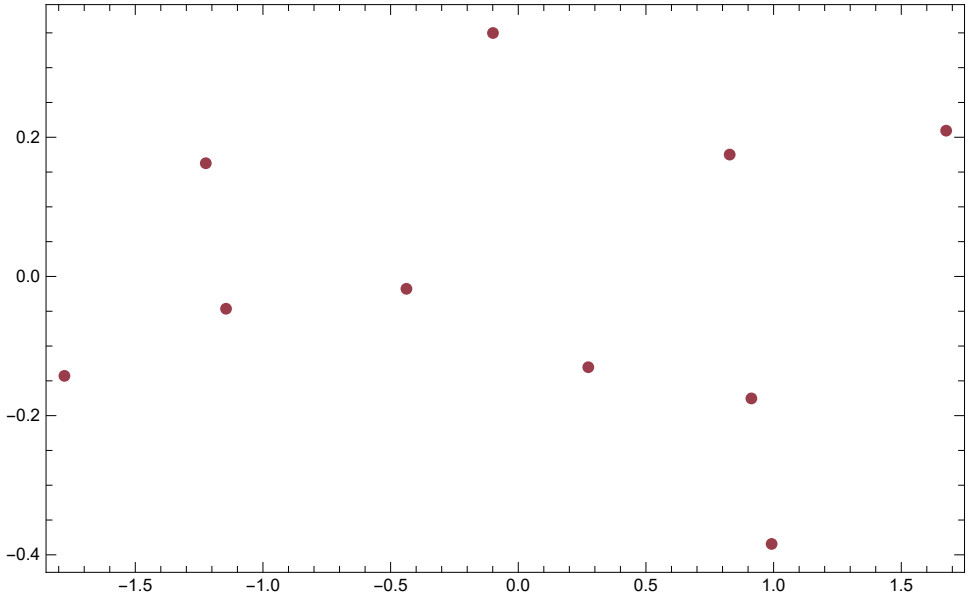
```
0.82797
-1.77758
0.992197
0.27421
1.6758
0.912949
-0.0991094
-1.14457
-0.438046
-1.22382
```



```
PrincipalComponents[dataAdjusted]
```

```
{{0.82797, 0.175115}, {-1.77758, -0.142857}, {0.992197, -0.384375}, {0.27421, -0.130417}, {1.6758, 0.209498},
 {0.912949, -0.175282}, {-0.0991094, 0.349825}, {-1.14457, -0.0464173}, {-0.438046, -0.0177646}, {-1.22382, 0.162675}}
```

```
ListPlot[PrincipalComponents[dataAdjusted]]
```



Init