Correlation

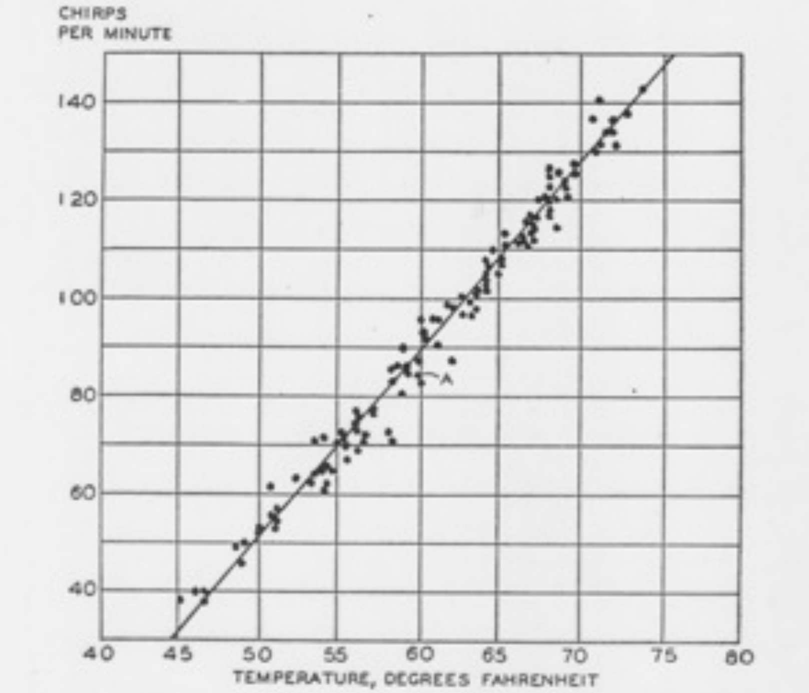


Chart 19.1. Temperature and Chirps per Minute of 115 Crickets. Data provided by Mr. Bert. E. Holmes.

When we last met ....

What are statistics? (Ultimate data-reducers)

- why do we use them (frequentist approach: compare with distributions)

- expectations values – how to relate statistics to probability-distribution parameters

- what do we demand from them?
  - efficiency
  - accuracy/consistency/closeness
  - robustness
  - lack of bias

- error analysis, error propagation

- combining distributions

- some statistics and their distributions - order statistics in particular

That Schechter Luminosity Function of last time ...

The error pointed out in the last lecture has survived though two editions of the book and >two generations of students.

The Schechter luminosity function is formally

$f(L)=K(L^*/L)^{\gamma}\exp(-L/L^*)$

Parameters K, L*, γ

So that here we should have written it in abbreviated form for the example as

$(x^*/x)^{\gamma}\exp(-x/x^*)$

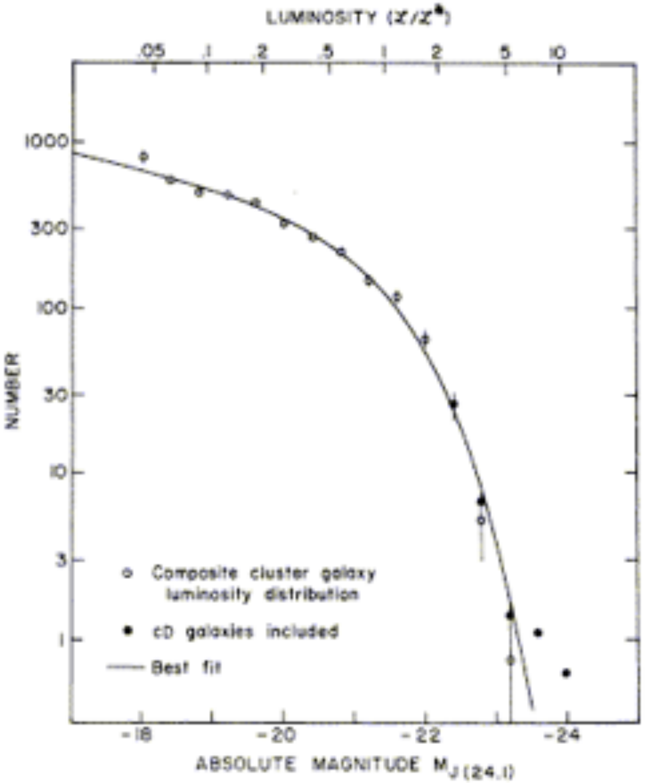
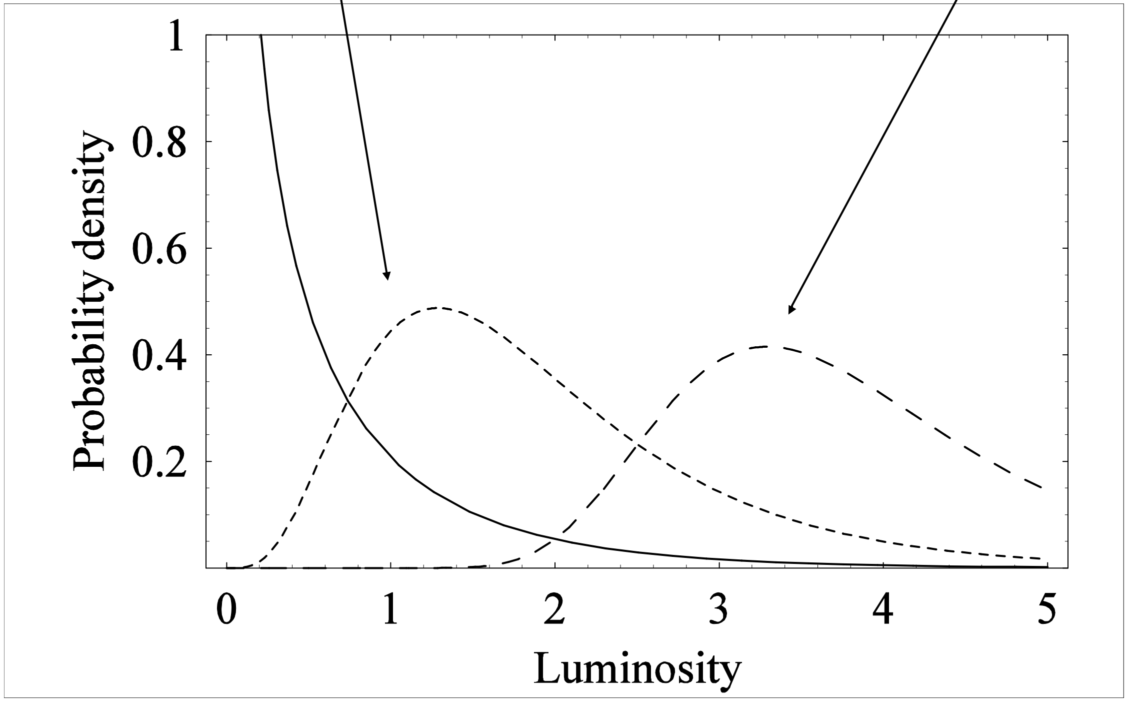


FIG. 2.—Best fit of analytic expression to observed composite cluster galaxy luminosity distribution. Filled circles show the effect of including cD galaxies in composite.

Order statistics - Example 2

Here’s how the slide should have read:

Solid curve: A Schechter luminosity function ***$(x^*/x)^\gamma\exp[-x/x^*]$,*** a useful model for the luminosity function of field galaxies. Take ***$\gamma = 0.5$, $x^* = 1$***.



Max of 10 galaxies

Max of 100 galaxies

Correlation - why do we try it?

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.

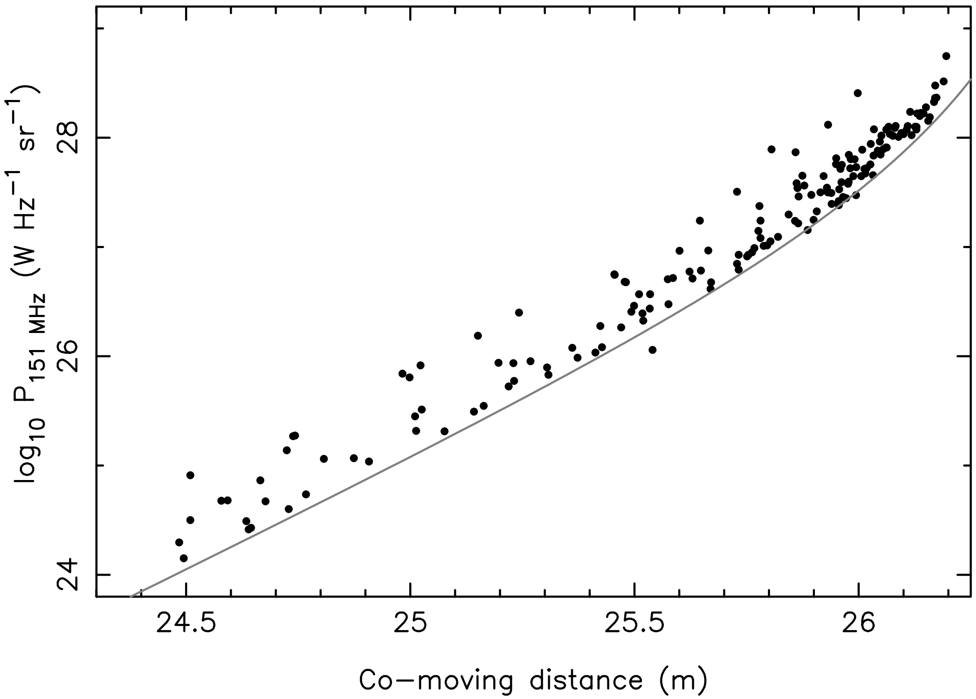
We are gonna do it - and we are going to fall into some deadly traps. We already have.

The fishing trip

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.



Radio luminosities of 3CR radio sources versus distance modulus

Still on the fishing trip ....

The plot **proves** luminosity evolution for radio sources? Are the more distant objects (at earlier epochs) clearly not the more powerful?

No! The sample is flux- (or apparent intensity) 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.

This is the only conclusion to be drawn from the diagram!

Still on the fishing trip ....

Astronomers produce many plots of this type, and say things like terms like ‘The lower right-hand region of the diagram is unpopulated because of the detection limit, but there is no reason why objects in the upper left-hand region should have escaped detection....’

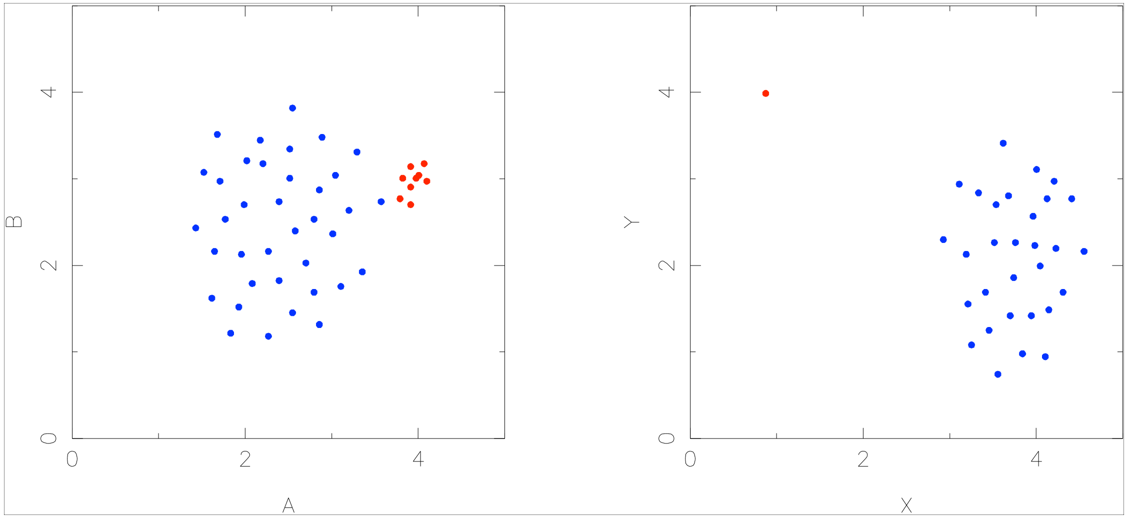
Nonsense – probabilities rule! There are only low-luminosity sources to be seen at low redshifts because there’s not enough volume to pick up the high-luminosity counterparts.

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

Still fishing ....

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?**



Suspect correlations: in each case formal calculation will indicate that a correlation exists to a high degree of significance.

The fishing trip continues ....

5. If **still** confident, remember that

a correlation does not prove a causal connection. Examples:

¤ The price of fish in Billingsgate Market and the size of feet in China.

¤ Number of violent crimes in cities versus number of churches.

¤ The quality of student handwriting versus their height.

¤ Stock market prices and the sunspot cycle.

¤ In World War II, bombing accuracy was far greater when enemy fighter planes were present.

¤ Cigarette smoking versus lung cancer.

¤ Health versus alcohol intake.

1. Lurking third variables

2. Similar time scales

3. Causal connection

There are ways of searching for intrinsic correlation between variables when they are known to depend mutually upon a third variable. But '**known**'?????

Wilkinson & Pickett: *The Spirit Level*

`Correlations’ show that higher income inequality correlates with higher crime rate, higher infant mortality, lower life expectancy, worse gender inequality, lower education standards, higher obesity rates……

Figure 5a: Wilkinson and Pickett’s plot of inequality against homicide rates[33]

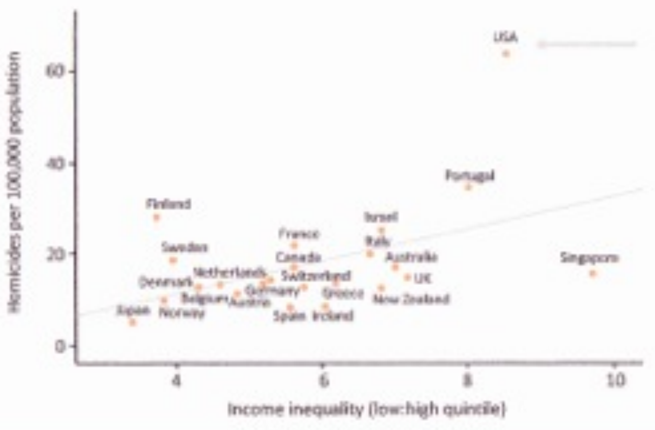
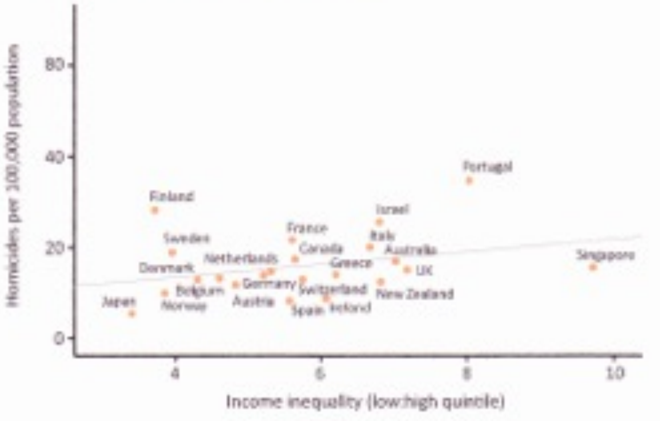


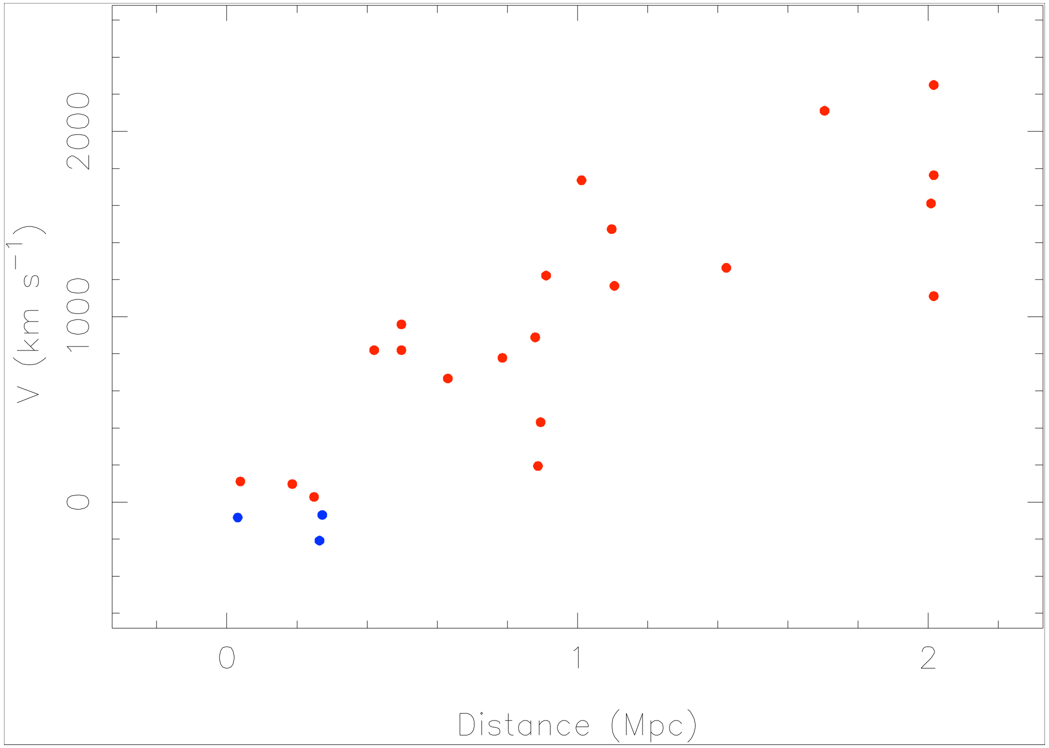
Figure 5b: Wilkinson and Pickett’s plot of inequality against homicide rates, excluding the USA



Critique by Peter Saunders: ***Beware of False Prophets*** shows that it is (statistical) garbage. The ‘correlations’ are false or of no significance.The data are selective. There have been other critical reviews and books along the same lines.

The end of the fishing trip - big fish are out there

Don’t get too discouraged by all the foregoing. Consider the example figure, a ragged correlation if ever there was one, although there are no nasty groupings of the type rejected by the Rule of Thumb.



An early Hubble diagram (Hubble 1936); recession velocities of a sample of 24 galaxies versus distance measure.

Correlation - the standard model

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**:

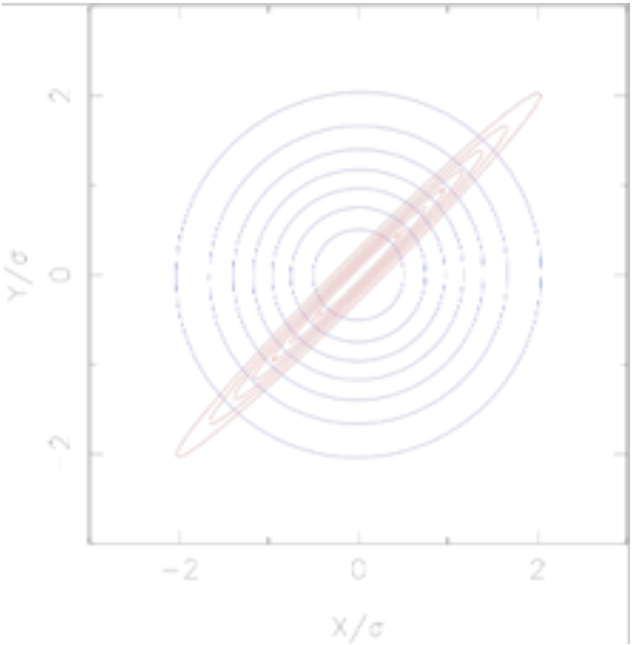
$$\begin{aligned}\mathrm{prob}(x, y \mid \sigma_x, \sigma_y, \rho) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\times \exp\left(\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x\sigma_y}\right)\right).\end{aligned}$$

Correlation - the standard model, continued

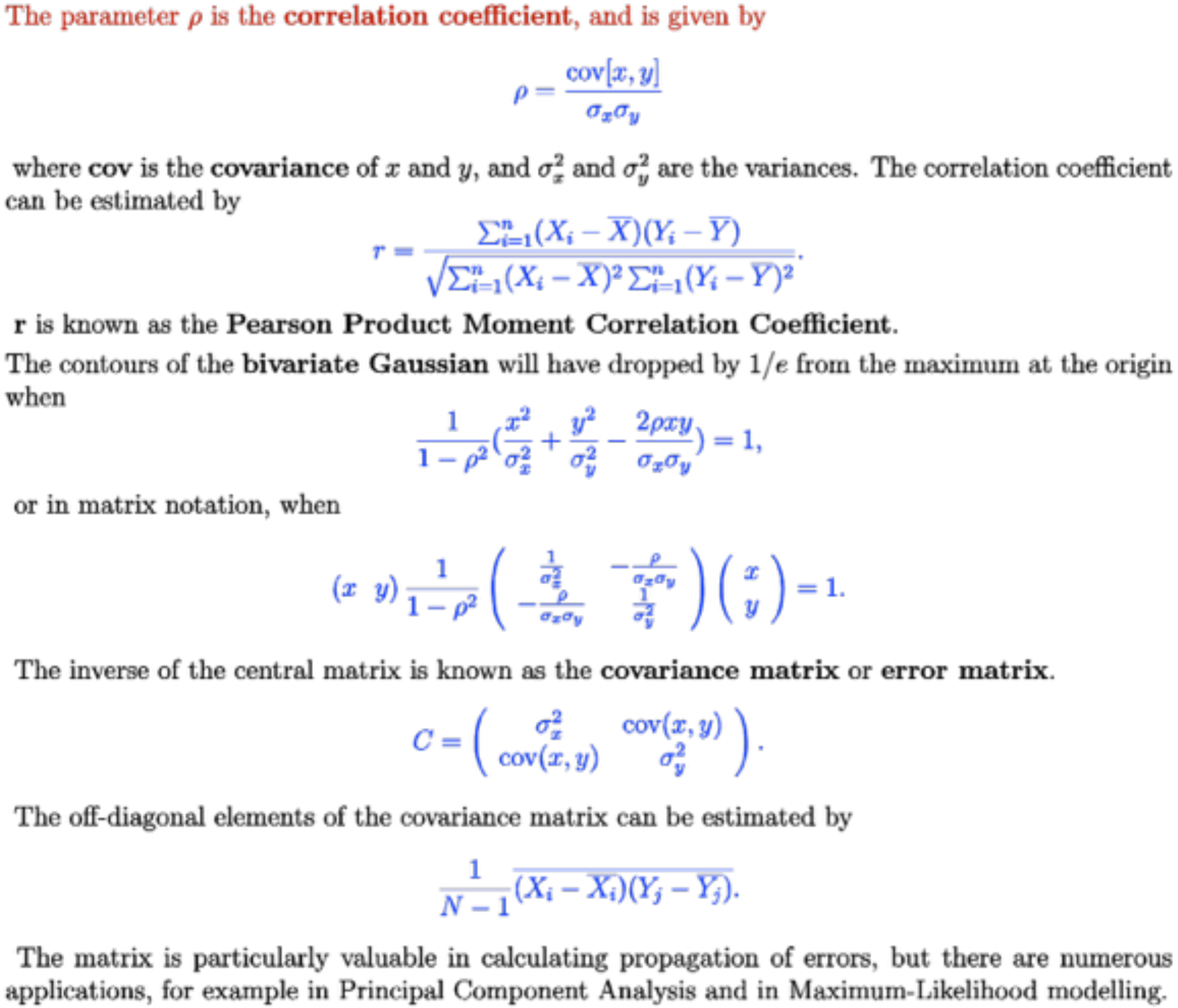
$$\begin{aligned}\mathrm{prob}(x, y \mid \sigma_x, \sigma_y, \rho) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \\ &\times \exp\left(\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}-\frac{2\rho x y}{\sigma_x\sigma_y}\right)\right).\end{aligned}$$

This model is so well developed that **‘correlation’** and **‘ρ ≠ 0’** are nearly synonymous; if **ρ** → **0** there is little correlation, while if **ρ** → **1** the correlation is perfect.

Left: linear contours of the bGpd. Near circular: **ρ = 0.01**, little connection between **x** and **y;** highly elliptical: **ρ = 0.99**, strong correlation between **x** and **y.** Negative values of **ρ** reverse the tilt: ‘**anticorrelation**’.



Correlation - standard model, continued even more



The parameter ρ is the **correlation coefficient**, and is given by

$$\rho = \frac{\mathrm{cov}[x,y]}{\sigma_x \sigma_y}$$

where **cov** is the **covariance** of x and y, and σ_x^2 and σ_y^2 are the variances. The correlation coefficient 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.**

The contours of the **bivariate Gaussian** will have dropped by $1/e$ from the maximum at the origin when

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

or in matrix notation, when

$$(x \quad y)\frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\sigma_x^2} & -\frac{\rho}{\sigma_x\sigma_y} \\ -\frac{\rho}{\sigma_x\sigma_y} & \frac{1}{\sigma_y^2} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = 1.$$

The inverse of the central matrix is known as the **covariance matrix** or **error matrix**.

$$C = \begin{pmatrix} \sigma_x^2 & \mathrm{cov}(x,y) \\ \mathrm{cov}(x,y) & \sigma_y^2 \end{pmatrix}.$$

The off-diagonal elements of the covariance matrix can be estimated by

$$\frac{1}{N-1}\overline{(X_i - \overline{X_i})(Y_j - \overline{Y_j})}.$$

The matrix is particularly valuable in calculating propagation of errors, but there are numerous applications, for example in Principal Component Analysis and in Maximum-Likelihood modelling.

Before the tests, let's play random numbers again

How to generate numbers obeying a bivariate (or even multivariate) Gaussian, with given $\boldsymbol{\sigma_i}$ and $\boldsymbol{\rho_i}$?

Following the discussion of error matrices, it's quite simple to formulate:

1. Set up the error matrix and determine the covariance matrix from it. (For the bivariate case, the error matrix is $\mathbf{e_{1,1} = \sigma_x{}^2}$**,** $\mathbf{e_{2,1} = e_{1,2} = cov[x,y] = \rho\, \sigma_x\, \sigma_y}$**,** $\mathbf{e_{2,2} = \sigma_y{}^2}$, as we have seen.)

2. Find the **eigenvalues** and **eigenvectors** of the covariance matrix.

3. Combine the eigenvectors, the column vectors, into the transformation matrix **T**, the matrix that diagonalizes the covariance matrix.

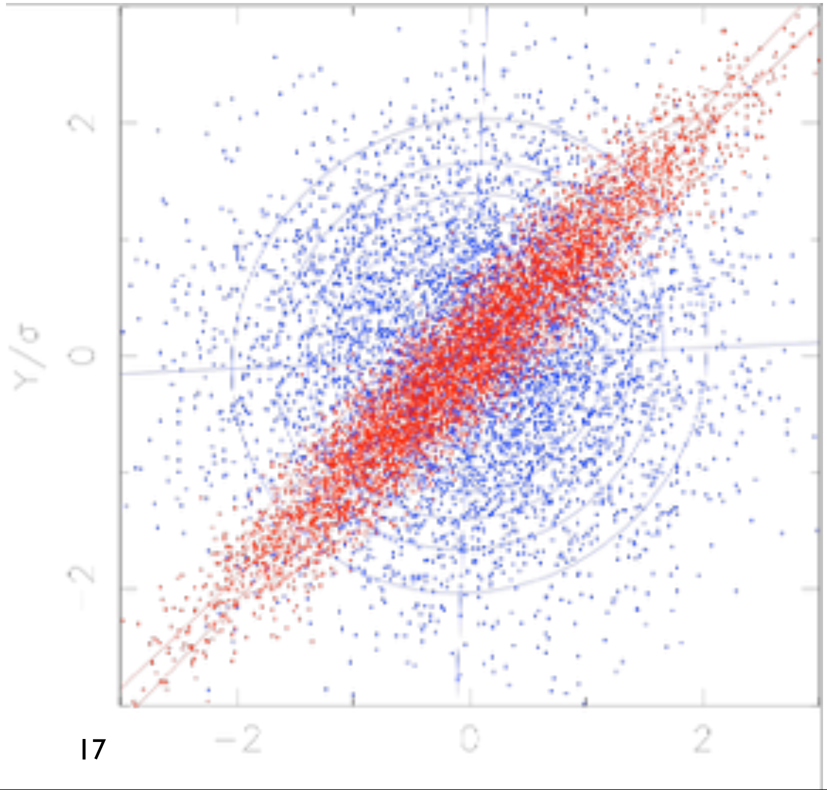
Playing random numbers, continued

We have got there! Now draw (X',Y') Gaussian pairs, uncorrelated, with variances equal to the two eigenvalues.

Compute the (X,Y) pairs according to

$$\begin{pmatrix} X \\ Y \end{pmatrix} = [T] \begin{pmatrix} X' \\ Y' \end{pmatrix}$$

The points in the figure were obtained in this manner, with ρ = 0.05 and 0.95, 5000 points each.



Formal testing - what are we doing?

1. For bivariate data, what we really want to know is whether or not **ρ = 0.**

2. Using the bivariate Gaussian is a very specific model.

3. A Gaussian is assumed - it allows only two variances, and assumes that both **x** and **y** are random variables.

4. $\boldsymbol{\sigma}_{\mathbf{x}}$ and $\boldsymbol{\sigma}_{\mathbf{y}}$ include both the errors in the data, and their intrinsic scatter -- all presumed Gaussian.

5. Does not apply 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.

CAUTION! Are your data right for this testing process?

Correlation testing - the Bayesian way

Uses 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; thus we have to integrate these ‘nuisance variables’ out of the full posterior probability **prob(ρ**,**σ_x**,**σ_y, μ_x,μ_y | data).**

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

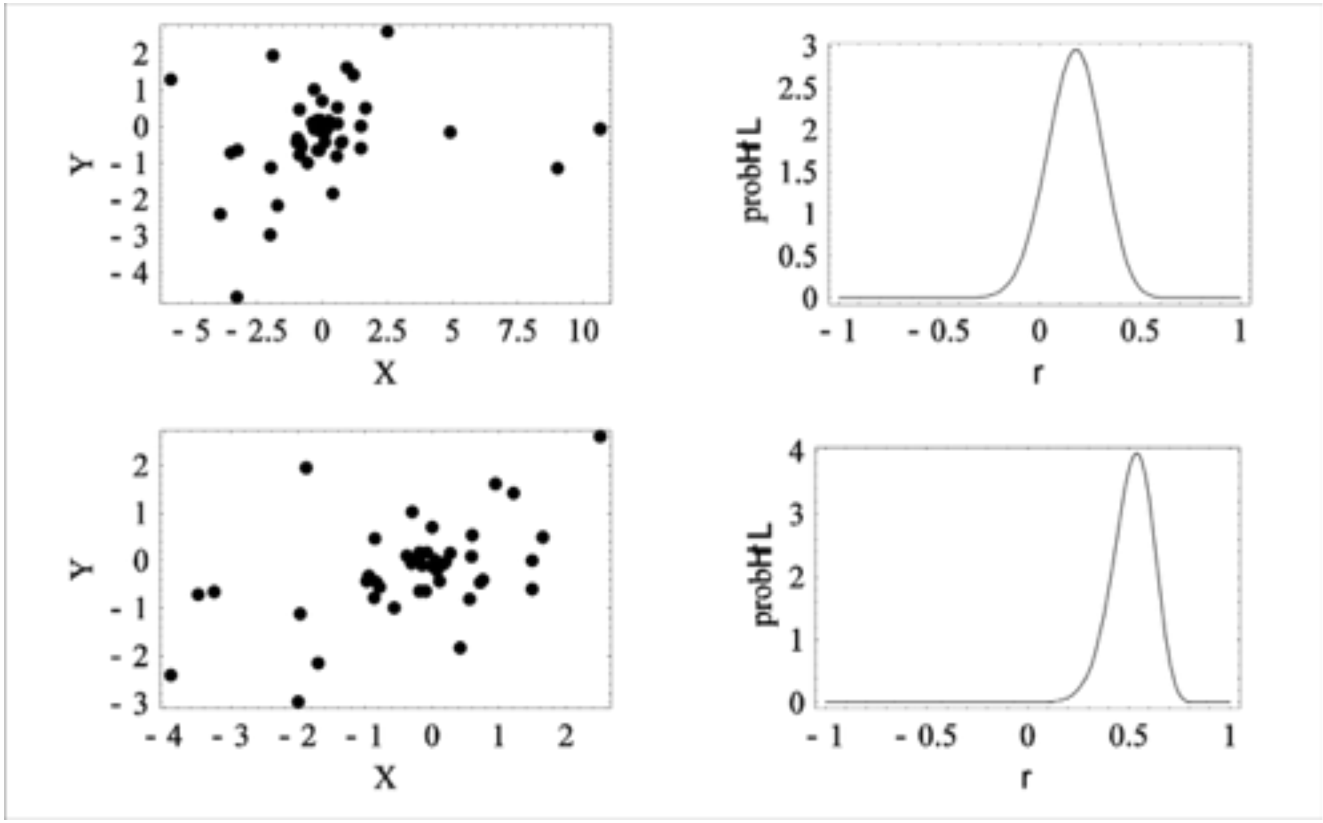
$$\mathrm{prob}((\rho \mid \ \mathrm{data}) \propto \frac{(1-\rho^2)^{\frac{n-1}{2}}}{(1-\rho r)^{n-\frac{3}{2}}}(1+\frac{1}{n-1/2}\frac{1+r\rho}{8}+\ldots)$$

The Bayesian test for correlation is thus simple: compute **r** from the **(X_i,Y_i),** and calculate prob(**ρ**) for the range of interest.

That’s it – and we have, as always with Bayes, what we really want to know.

Bayesian correlation testing - Example 1

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.
Top panels: rotten result! Then remove the outliers **> 4σ. Better!**



50 X_i,Y_i chosen at random from a bivariate Gaussian with ρ = 0.5, outliers added. The Jeffreys probability distribution of correlation coefficient is shown, peaking at around 0.2 for the upper panel. The data have been restricted to ±4σ in the lower panel; the distribution now peaks at 0.44.

Bayesian correlation testing - more

Given this probability distribution for $\boldsymbol{\rho}$, we can answer questions like ‘what is the probability that $\boldsymbol{\rho} > 0.5$?' or ‘what is the probability that $\boldsymbol{\rho}$ from data set A is bigger than $\boldsymbol{\rho}$ from data set B?’.

The utility of the Bayesian approach is not that prior information is accurately incorporated, but rather that we get an answer to the question we really want to ask.

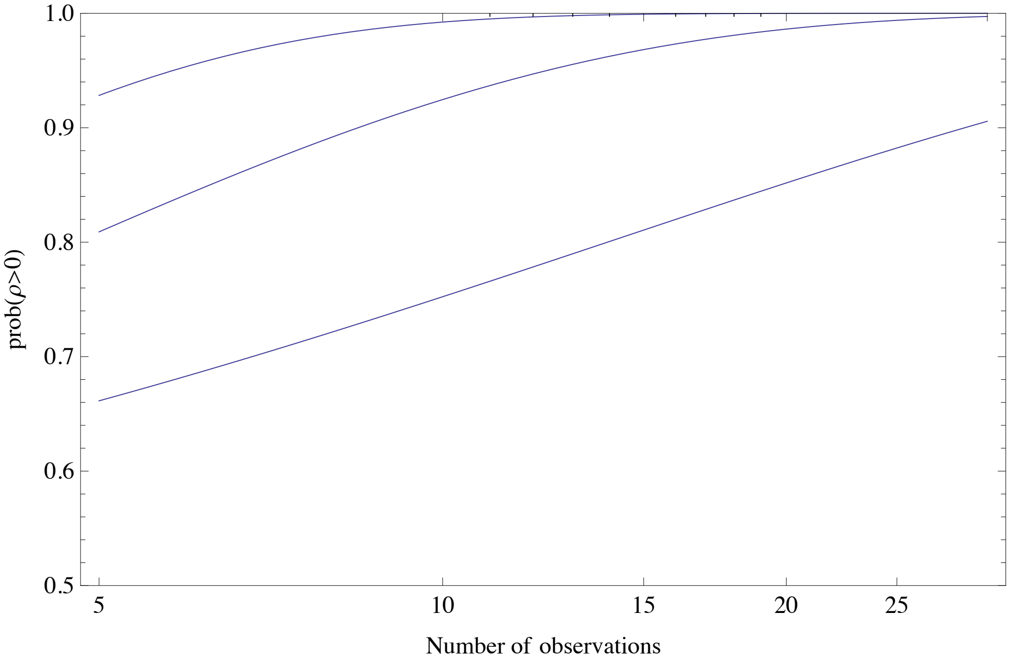
Jeffreys used a uniform prior for $\boldsymbol{\rho}$ - not obviously justifiable, and certainly not correct if $\boldsymbol{\rho}$ is close to 1 or -1. But **in these cases a statistical test is a waste of time** anyway.

Bayesian correlation testing - Example 2

As a second example, something generally useful: 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.

Figure right: the probability of ρ being positive as a function of sample size, for ρ=0.25 (lowest curve), 0.5 and 0.75 (top curve).



Correlation testing: Classical approach

ρ now taken to be a fixed quantity -

…and we are about to find the probability of the **data**, given **ρ** (+ hypothesis of bivariate Gaussian).

The result (Fisher 1944) is:

$$\mathrm{prob}(r \mid \rho, H) \propto \frac{(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}}{(1-\rho r)^{n-3/2}}\left(1+\frac{1}{n-1/2}\frac{1+r\rho}{8}+\ldots\right)$$

The standard parametric test is to attempt to reject the null hypothesis that **ρ** = 0:

Correlation testing: Classical approach

The standard parametric test is to attempt to reject the null hypothesis that **ρ** = 0:

1. Compute **r**. Note **-1 < r < 1**; **r=0** for no correlation, and the standard deviation in r is

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

2. Compute the probability, under this hypothesis, of **r** being this big or bigger.
If this probability is 'very small' we may conclude that the null hypothesis is unlikely.

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

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

which obeys the probability distribution of the **'Students' t statistic** with **N-2** degrees of freedom. (The transformation simply allows us to use tables of **t**.)

4. Check the **table of critical values** for **t**; if **t** exceeds that corresponding to a critical value of the probability (two-tailed test), then the hypothesis that the variables are unrelated can be rejected at the specified level of significance. This level of significance (say 1%, or 5%) is the **maximum probability** which we are willing to risk in deciding to reject the null hypothesis (no correlation) **when it is in fact true**.

Correlation testing, classical approach - 2

☻ This approach probably has not answered the question!

☻ We embark on this sort of investigation when it is apparent that the data contain correlations; we merely want some justification by knowing `how much'.

☻ The inclusion in the test of values of **unobserved values of r** 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 would …..?)
- 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.

Correlation testing: classical, non-parametric

☺ However, to circumvent the problems it is far better to go to **non-parametric tests**. These permit additional tests on data which are not numerically defined (binned data, or ranked data), so that in some instances **they may be the only alternative.**

The best known non-parametric test for correlation:

1. For the N data pairs of **(X_i,Y_i),** make rank tables of **X_i** and **Y_i** such that **(XR_i,YR_i)** pairs represent the ranks for the **i^{th}** pair, **$1 < XR_i < N$, $1 < YR_i < N$**.

2. Compute the **Spearman Rank Correlation Coefficient**:

$$r_s = 1 - 6\frac{\sum^{N}(XR_i - YR_i)^2}{N^3 - N}$$

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 **$4 \leq N \leq 30$**. If **r_s** exceeds an appropriate critical value, the hypothesis that the variables are unrelated is **rejected** at that level of significance.

Correlation testing: classical, non-parametric

4. If **N** exceeds 30, compute

$$t_r = r_s\sqrt{\frac{N-2}{1-r_s^2}},$$

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

In comparison with **r**, $\mathbf{r_s}$ has an efficiency of 91%. Pretty good.

Take-away message – if in doubt, **go non-parametric.**

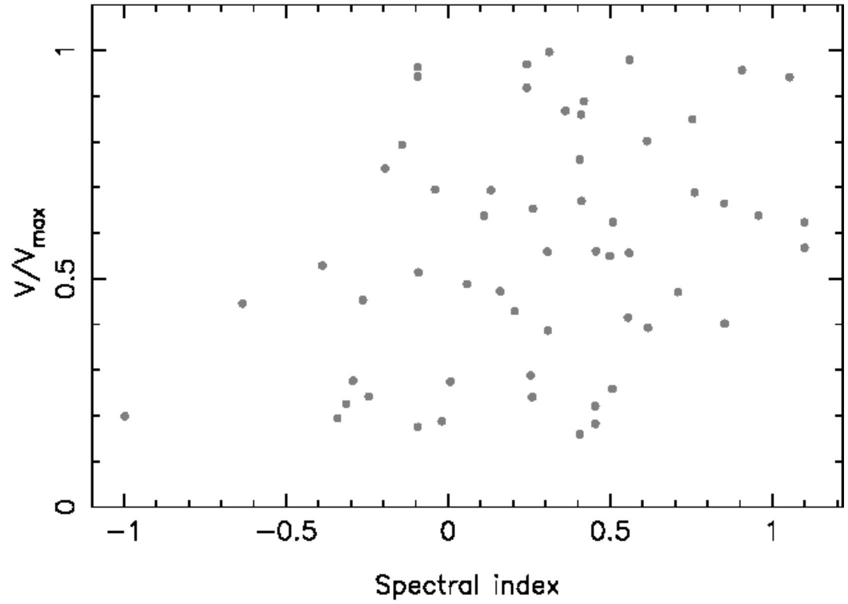
Classical corr testing: Non-Parametric Example

A `correlation' at the notorious **2σ** level is shown. Here, $\mathbf{r_s = 0.28}$, **N = 55**, and the hypothesis that the variables are unrelated is rejected at the **5% 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 crazy, especially in view of a uniformly-filled Universe producing a $\mathbf{V/V_{max}}$ statistic **uniformly** distributed between **0 and 1** (Schmidt 1968).

V/V_{max} as a function of high-frequency spectral index for a sample of radio quasars selected from the Parkes 2.7-GHz survey.



Correlation testing : DIY via permutations

The permutation test:

1. We have data **(X_1,Y_1), (X_2,Y_2), ….** and we wish to test the null hypothesis that **x** and **y** are uncorrelated.

2. If we have some **home-made test statistic** $\mathscr{B}$, we can calculate its distribution, on the assumption of the null hypothesis, by simply calculating its value for many permutations of the **x**'s amongst the **y**'s.

3. For any reasonable data set there will be far more possible permutations than we can reasonably explore; => choose a **random set** to give an estimate of the distribution of $\mathscr{B}$.

Correlation testing : DIY via permutations (2)

4. If it turns out that the observed value of $\mathscr{B}$ is very improbable under the null hypothesis, we may be interested in **estimating the distribution for non-zero correlation.** This is the route to useful Bayesian analysis, of the kind we described for the product-moment coefficient **ρ**. Here the bootstrap method is ideal.

These methods can be used to derive distributions of statistics like **Spearman's** or **Kendall's correlation coefficients** in cases when a correlation is apparently present.

Correlation testing : comments

1. The **non-parametric tests** circumvent some of the issues involved in the non-Bayesian approach, but they have no bearing on the fundamental issue – **what was the real question**?

2. But as ever, the Bayesian approach, strong in answering the real question, forces reliance on a **model**.

3. There is very little difference between the **Fisher test** and results from **Jeffreys distribution**. We can show this with random Gaussian data with a correlation of zero. We use the **r-distribution** to find the probability of **r** being as large or larger than we observe, on the hypothesis that **ρ=0**. If this probability is small, the test is hinting at the possibility that the correlation is actually positive Therefore we compare with the probability from the Jeffreys distribution that **ρ** is positive. Now we expect the probability from **ρ** to be large; and in fact we can see, either from simulations or from the algebraic form of the distributions, that **the sum of these two probabilities is always → 1**.

Interpreting the standard Fisher test (illegally!) to be telling us the chance that **ρ** is positive, actually works very well!

Correlation found! Now what?

First question: what's the law relating the variables?

We rush off and fit 'regression lines', often by Least Squares.

But recognize that we're now model-fitting. There is a crucial distinction.

In the model fitting (coming later) :

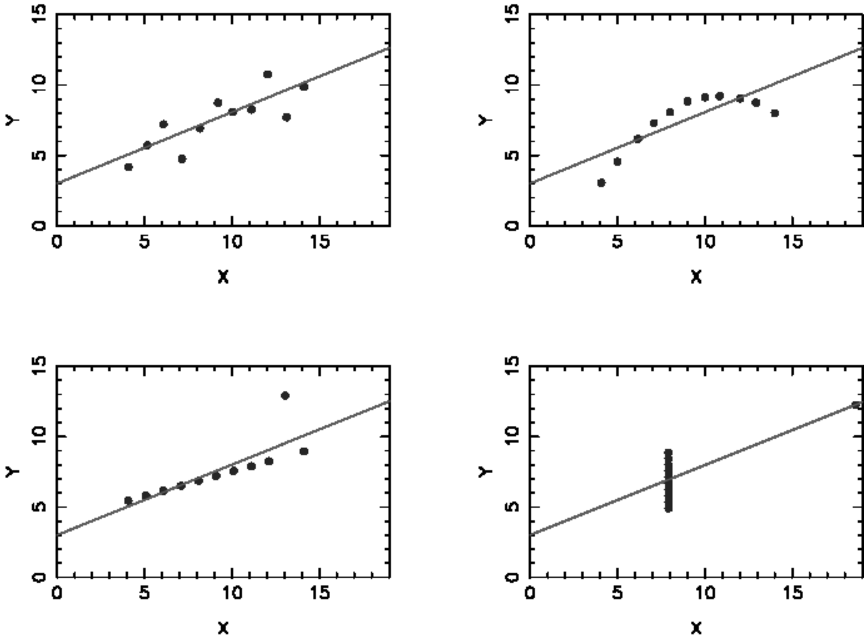
- Are there better quantities to minimize than the squares of deviations?
- What errors result on the regression-line parameters?
- Why should the relation be linear?
- What are we trying to find out?

Example: If we have found a correlation between **x** and **y**, which variable is dependent; do we want to know (**x** on **y**) or (**y** on **x**)? The coefficients are generally completely different.

As an argument against blind application of correlation testing, consider the example of Anscombe's (1973) famous quartet:

Anscombe’s Quartet: correlation vs independence

Anscombe’s quartet: 4 fictitious sets of 11 (Xi,Yi), each with the same <X>, <Y>, identical coefficients of correlation, regression lines, residuals in Y, estimated standard errors on slope, and covariance matrices.



Note:

1. In ¾ cases, the points are clearly related; they are far from independent but still show only indifferent quality of correlation. At upper right, choice of the ‘right’ relation would result in a perfect fit.

2. **X** independent of **Y** means **prob(X,Y)=prob(X)prob(Y)**, or **prob (X|Y)=prob(X). X** correlated with **Y** means **prob(X,Y) ≠ prob(X)prob(Y)** in a way such as to give **r ≠ 0.** We can have **prob(X,Y) ≠ prob(X)prob(Y)** AND **r = 0**, example: Union Jack.