Lecture 2 :
Searching for correlations

These lectures

- Lecture 1 : basic descriptive statistics
- Lecture 2 : searching for correlations
- Lecture 3 : hypothesis testing and model-fitting
- Lecture 4 : Bayesian inference

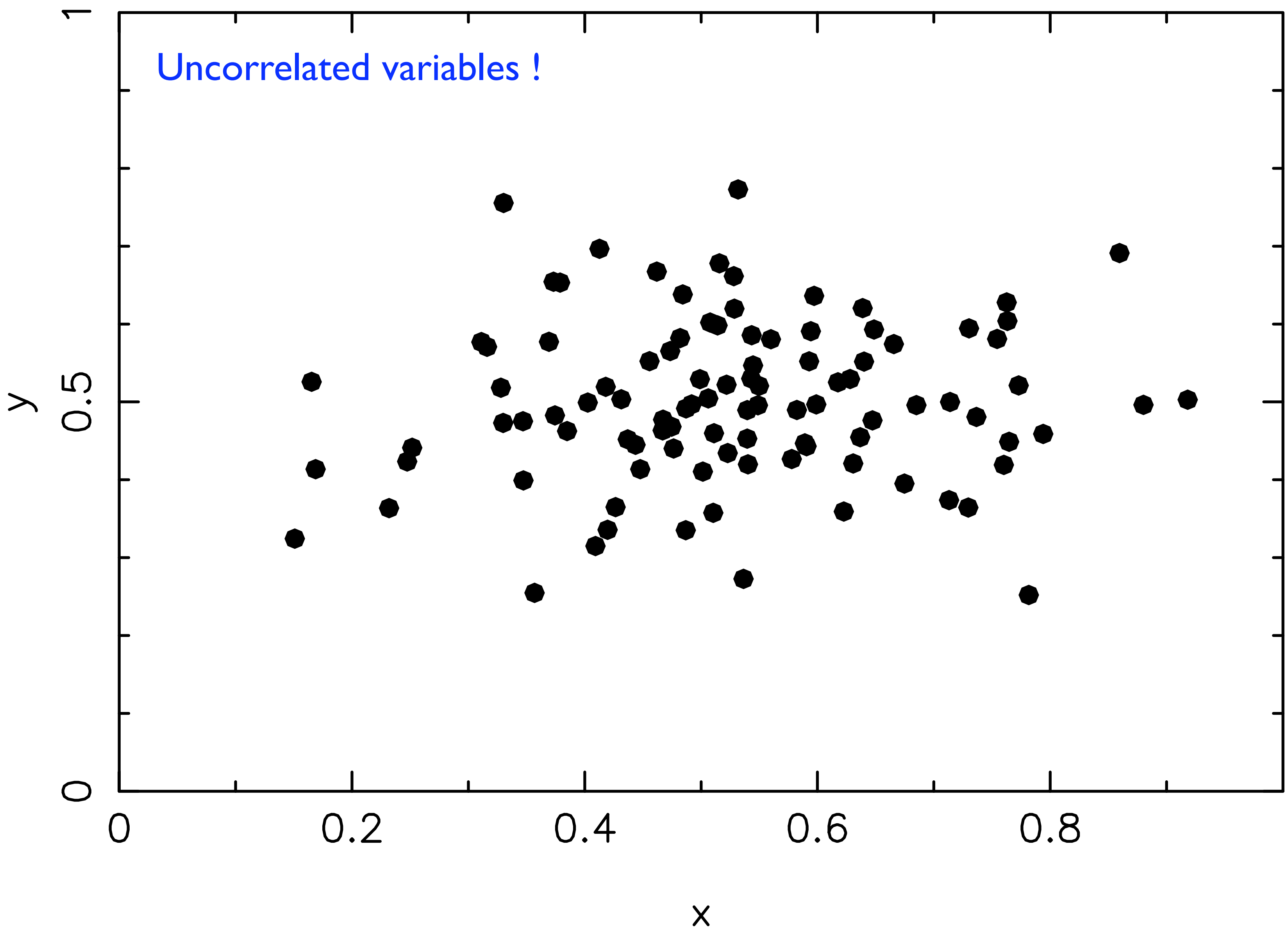
Lecture 2 : searching for correlations

- Correlation coefficient and its error
- How to quantify the significance of a correlation
- Bootstrap error estimates
- Non-parametric correlation tests
- Common pitfalls when searching for correlations
- Comparing two distributions

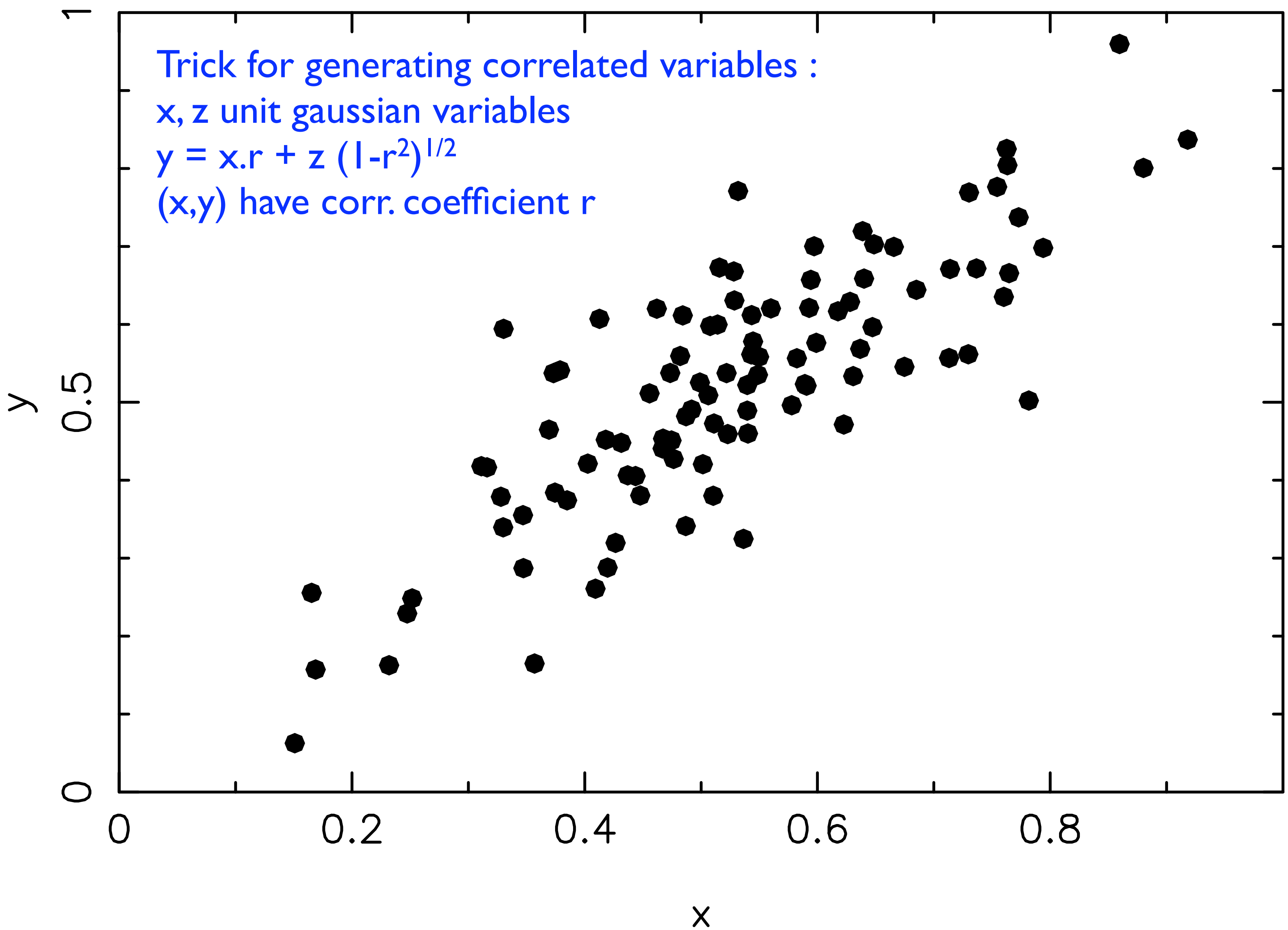
What is a correlation?

- Two variables are correlated if they share a statistical dependence / relationship
- For example, measurements of temperature at noon and 1pm every day are correlated, because they both lie consistently above the mean daily temperature
- Correlations between variables are important because they indicate some underlying physical relationship between those variables

What is a correlation?

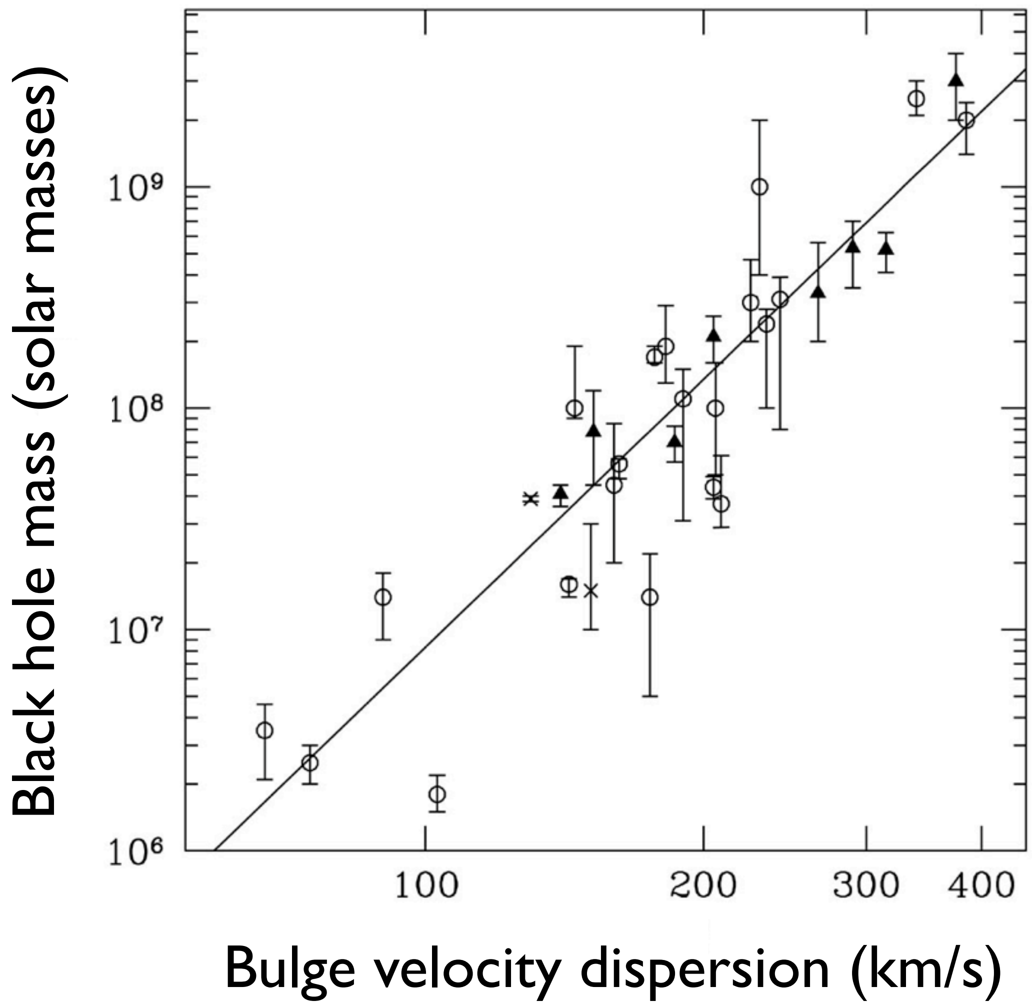


What is a correlation?



What is a correlation?

- Example in astronomy : black-hole / bulge relation



Correlation coefficient

- Describes the strength of the correlation between (x, y)
- Means : (μ_x, μ_y)
- Standard deviations : (σ_x, σ_y)
- Definition of correlation coefficient :

$$\rho = \frac{< (x - \mu_x)(y - \mu_y) >}{\sigma_x \sigma_y} = \frac{< xy > - \mu_x \mu_y}{\sigma_x \sigma_y}$$

$$< xy >= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy \, P(x, y) \, dx \, dy$$

Correlation coefficient

$$\boxed{\rho = \frac{< (x-\mu_x)(y-\mu_y) >}{\sigma_x\,\sigma_y} = \frac{< xy > - \mu_x\,\mu_y}{\sigma_x\,\sigma_y}}$$

- No correlation [P(x,y) separable into f(x) g(y)] :

$$< xy > = < x >< y > = \mu_x\,\mu_y \qquad \rho = 0$$

- Complete correlation :

$$y = C\,x \qquad \rho = +1$$

- Complete anti-correlation :

$$y = -C\,x \qquad \rho = -1$$

- Possible range is $-1 \leq \rho \leq +1$

Lies, damn lies and statistics

BBC NEWS

Watch One-Minute World News

News Front Page

Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
Business
Health
Medical notes
Science & Environment

Last Updated: Tuesday, 22 July, 2003, 10:28 GMT 11:28 UK

E-mail this to a friend   Printable version

Eating pizza 'cuts cancer risk'

Italian researchers say eating pizza could protect against cancer.

Pizzas are covered with a potentially protective tomato sauce

Researchers claim eating pizza regularly reduced the risk of developing oesophageal cancer by 59%.

The risk of developing colon cancer also fell by 26% and mouth cancer by 34%, they claimed.

Example 3

Why was this poor statistics? Correlation is not the same as causation. Other dietary or lifestyle habits could be a third variable! [... more examples later ...]

Estimating the correlation coefficient

- We can estimate the Pearson product-moment correlation coefficient as :

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

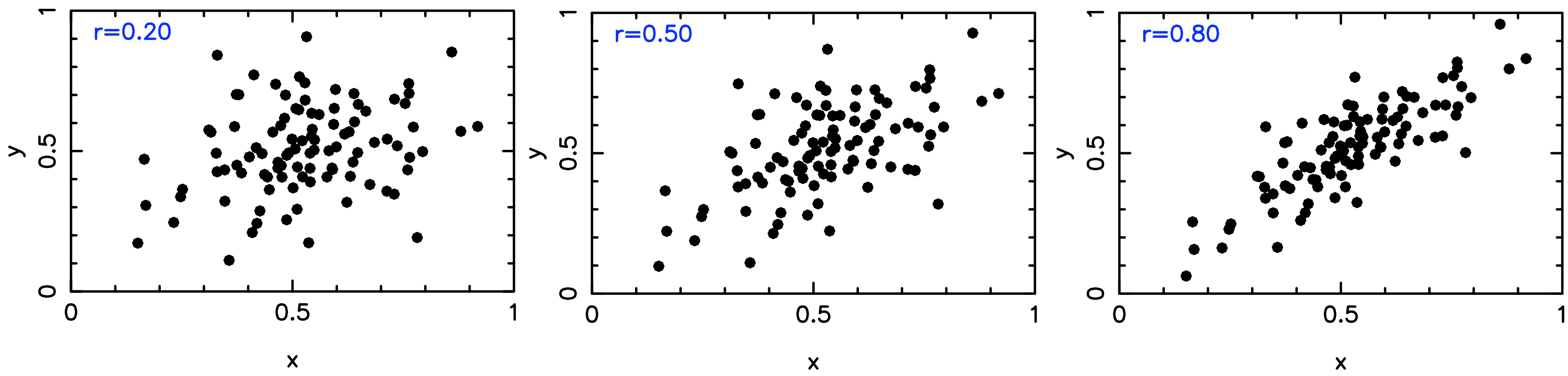
$$r = \frac{\sum_{i=1}^{N} x_i\, y_i - N\,\overline{x}\,\overline{y}}{(N-1)\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}} \quad \text{c.f.} \quad \rho = \frac{< xy > - \mu_x\, \mu_y}{\sigma_x\, \sigma_y}$$

- The possible range of values is

$$-1 \leq r \leq +1$$

Estimating the correlation coefficient

- If the correlation is statistically significant :
  - $0 < |r| < 0.3$ is a "weak correlation"
  - $0.3 < |r| < 0.7$ is a "moderate correlation"
  - $0.7 < |r| < 1.0$ is a "strong correlation"



Estimating the correlation coefficient

- Assumption : (x,y) are drawn from a bivariate Gaussian distribution about an underlying linear relation :

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

- If this model is true, then the uncertainty in the measured value of r is

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

Estimating the correlation coefficient

- Example 1 : who discovered the distance-redshift relation, Lemaitre (1927) or Hubble (1929)?

arXiv:1106.3928 etc.

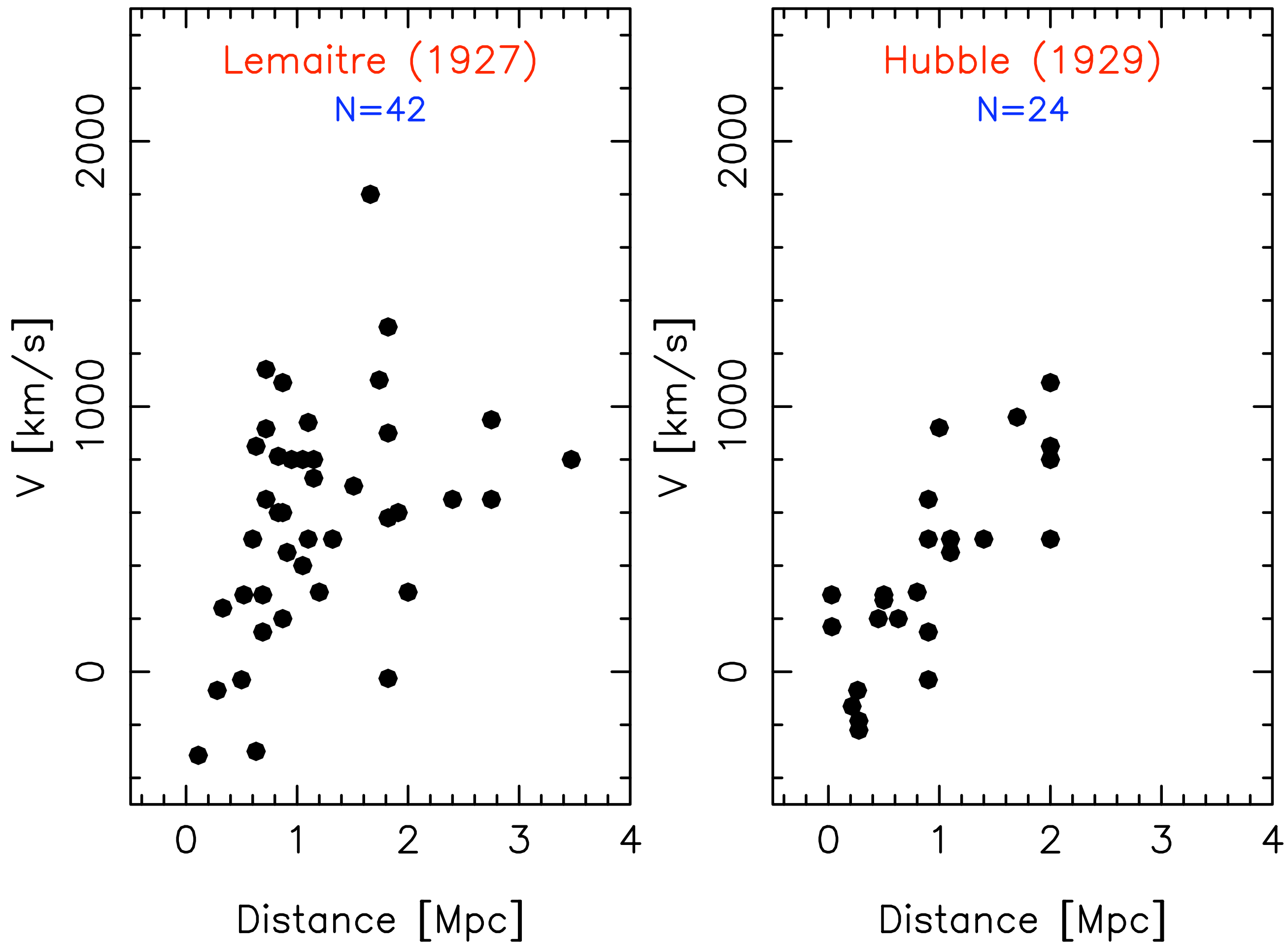
"A Hubble Eclipse: Lemaître and Censorship"

David L. Block, School of Computational & Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa.

Abstract. One of the greatest discoveries of modern times is that of the expanding Universe, almost invariably attributed to Hubble (1929). What is not widely known is that the original treatise by Lemaître (1927) contained a rich fusion of both theory and of observation. The French paper was meticulously censored when printed in English - all discussions of radial velocities and distances (and the very first empirical determination of "H") were omitted. Fascinating insights are gleaned from a letter recently found in the Lemaître archives. An appeal is made for a Lemaître Telescope, to honour the discoverer of the expanding universe.

Estimating the correlation coefficient

- Example 1 : who discovered the distance-redshift relation, Lemaitre (1927) or Hubble (1929)?



Estimating the correlation coefficient

- Part (a) : For each dataset, find the Pearson product-moment correlation coefficient and its error
- Lemaitre : r = 0.38 +/- 0.15
- Hubble : r = 0.79 +/- 0.13

Is a correlation significant?

- What is the probability of obtaining the measured value of r if the true correlation is zero? (also depends on N)

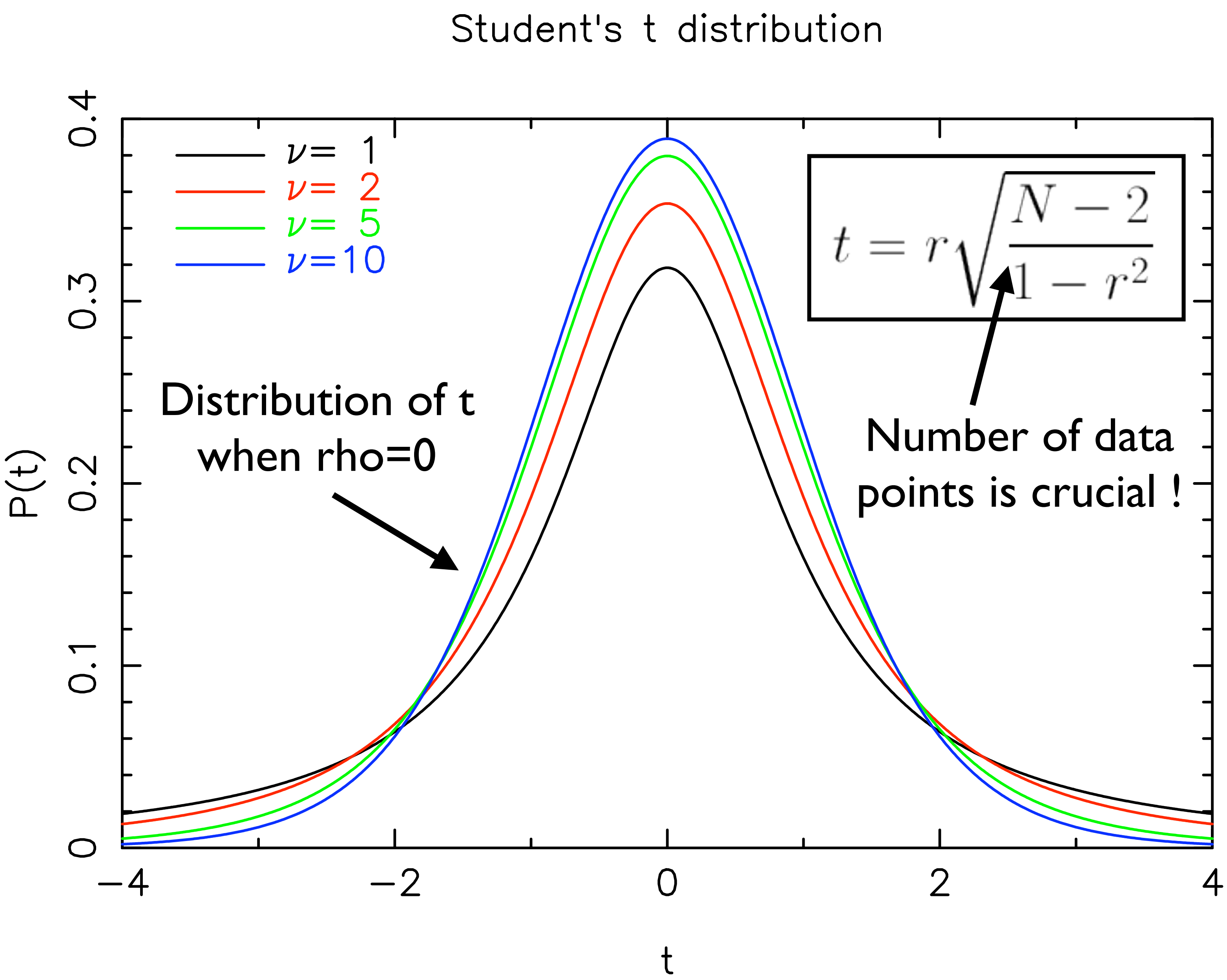
- In order to determine whether the correlation is significant, calculate

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

- This obeys the Student's t probability distribution with number of degrees of freedom $\nu = N - 2$

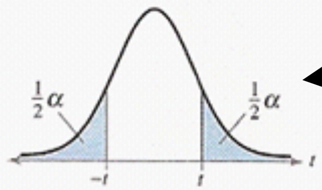
- Consult tables (2-tailed test) with these two numbers

Is a correlation significant?



Is a correlation significant?

- Standard tables list the critical values that t must exceed, as a function of nu, for the hypothesis that the two variables are unrelated to be rejected at a particular level of statistical significance (e.g. 95%, 99%)



Two-tailed test

"Two-tailed test" : correlation or anti-correlation

d.f.	One tail, α / Two tails, α	0.25 / 0.50	0.10 / 0.20	0.05 / 0.10	0.025 / 0.05	0.01 / 0.02	0.005 / 0.01
1		1.000	3.078	6.314	12.706	31.821	63.657
2		.816	1.886	2.920	4.303	6.965	9.925
3		.765	1.638	2.353	3.182	4.541	5.841
4		.741	1.533	2.132	2.776	3.747	4.604
5		.727	1.476	2.015	2.571	3.365	4.032
6		.718	1.440	1.943	2.447	3.143	3.707
7		.711	1.415	1.895	2.365	2.998	3.499
8		.706	1.397	1.860	2.306	2.896	3.355
9		.703	1.383	1.833	2.262	2.821	3.250
10		.700	1.372	1.812	2.228	2.764	3.169
11		.697	1.363	1.796	2.201	2.718	3.106
12		.695	1.356	1.782	2.179	2.681	3.055
13		.694	1.350	1.771	2.160	2.650	3.012
14		.692	1.345	1.761	2.145	2.624	2.977
15		.691	1.341	1.753	2.131	2.602	2.947
16		.690	1.337	1.746	2.120	2.583	2.921
17		.689	1.333	1.740	2.110	2.567	2.898

Is a correlation significant?

- Part (b) : Determine the statistical significance of the correlation

- Lemaitre : t=2.60 , nu=40, prob=1.3e-2 (2.5 sigma)

- Hubble : t=6.03 , nu=22 , prob=4.5e-6 (4.6 sigma)

Linear regression line

- The regression line is the linear fit that minimizes the sum of the squares of the y-residuals

- With intercept [y = a + b x] :

$$b = \frac{\sum_{i=1}^{N} x_i\, y_i - N\, \overline{x}\, \overline{y}}{(N-1)\, \sigma_x^2} = r\, \frac{\sigma_y}{\sigma_x}$$

$$a = \mu_y - b\, \mu_x$$

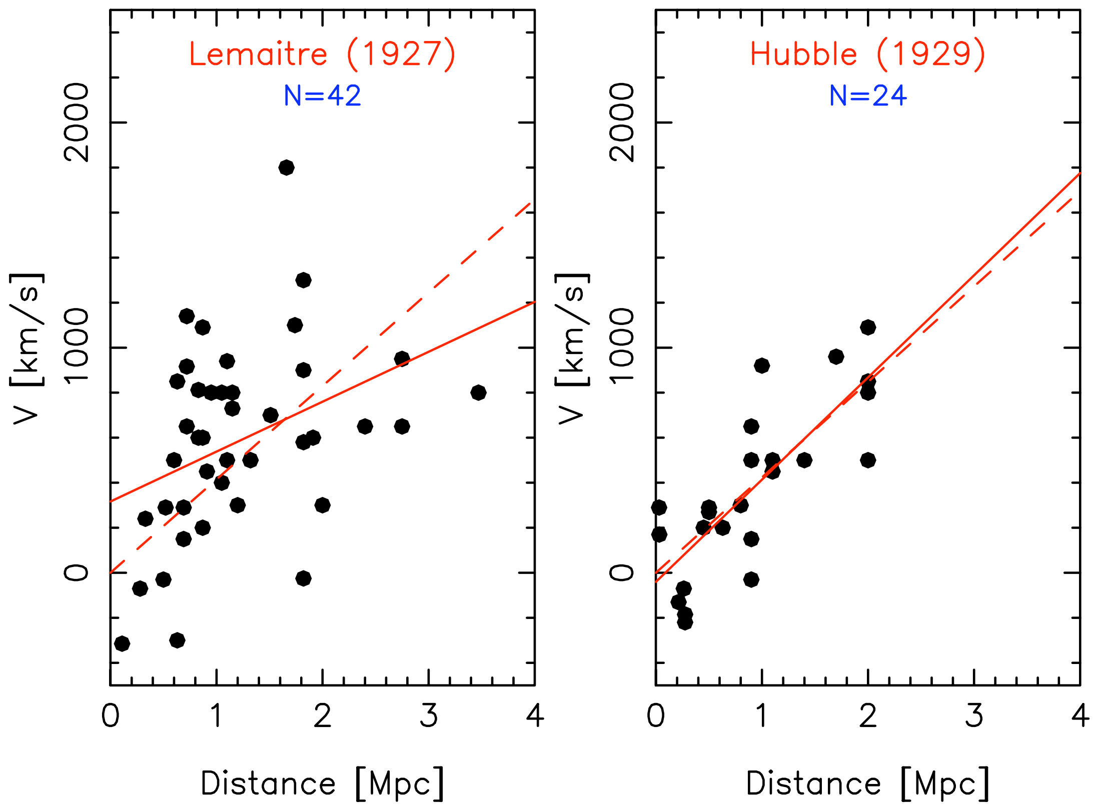
- Without intercept [y = b x] :

$$b = \frac{\sum_{i=1}^{N} x_i\, y_i}{\sum_{i=1}^{N} x_i^2}$$

Linear regression line

- Part (c) : Determine linear least-squares regression lines of the form V = HD and v = H'D + C

- Lemaitre : H=414.9 , H'=221.7 , C=316.5
  Hubble : H=423.9 , H'=453.9 , C=-40.4



Linear regression line

- Aside : Hubble and Lemaitre both found values of H_0 ~ 420 km/s/Mpc with independent techniques ! How could they both be wrong? [Example of statistical bias]

- Today would probably indicate confirmation bias, but Hubble didn't even cite Lemaitre's result!

- Lemaitre : assumed galaxy apparent magnitude was standard candle - scuppered by Malmquist bias

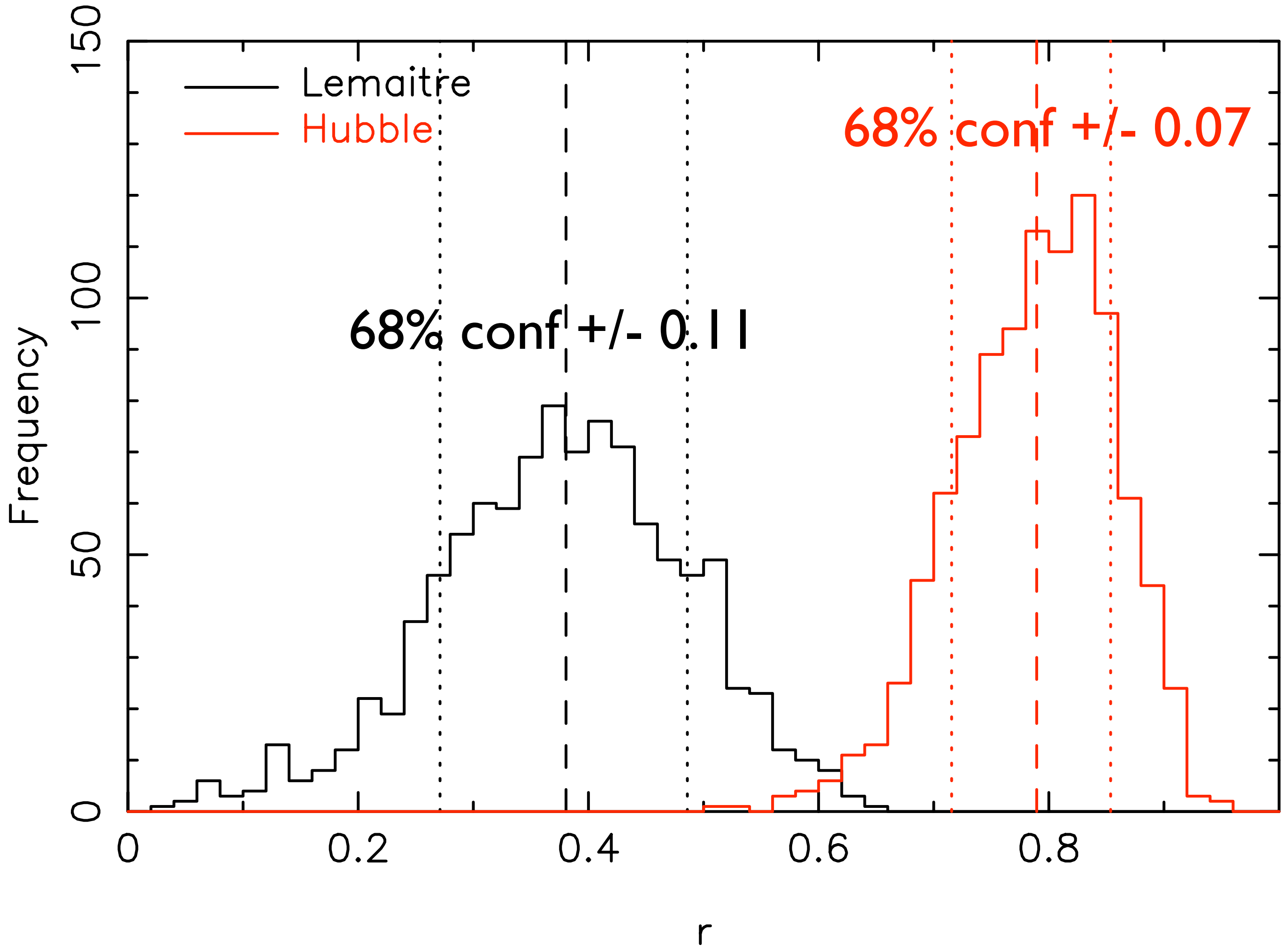
- Hubble : used "brightest stars" as standard candles, but could not distinguish brightest star from HII region (systematic error bias due to aperture effect)

Bootstrap errors and probabilities

- Bootstrap statistics allow us to determine parameter errors and probability distributions using just the data
- If we have N data points, repeatedly draw at random samples of N points (with replacement)
- Re-compute the parameter of interest for each bootstrap sample
- The distribution of the re-computed parameters estimates the uncertainty in the measurement from the original sample

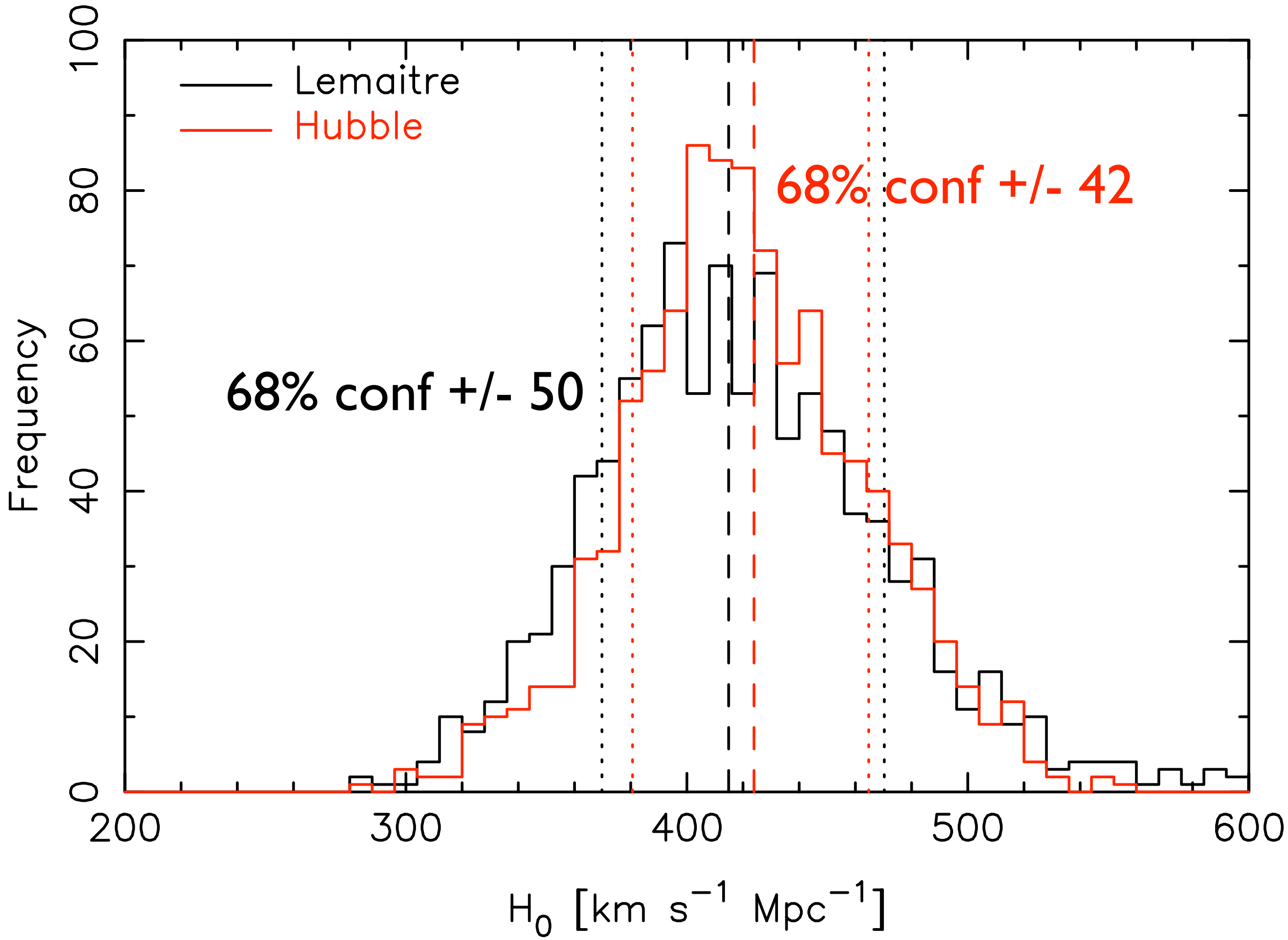
Bootstrap errors and probabilities

- Applied to our example : create 1000 bootstrap samples and measure the correlation coefficients



Bootstrap errors and probabilities

- Applied to our example : create 1000 bootstrap samples and do a linear regression fit



Non-parametric correlation coefficient

- If we do not want to assume (x,y) are drawn from a bivariate Gaussian we can use a non-parametric correlation test
- Let (X_i, Y_i) be the rank of (x_i, y_i) in the overall order such that $1 \leq (X_i, Y_i) \leq N$
- Find Spearman rank correlation coefficient

$$r_s = 1 - 6\frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{N^3 - N}$$

- Compare to standard tables with $\nu = N - 2$

D [Mpc]	Rank
0.03	1.0
0.03	2.0
0.21	3.0
0.26	4.0
0.28	5.5
0.28	5.5
0.45	7.0
0.50	8.5
0.50	8.5
0.63	10.0
0.80	11.0
0.90	13.5
0.90	13.5
0.90	13.5
0.90	13.5
1.00	16.0
1.10	17.5
1.10	17.5
1.40	19.0
1.70	20.0
2.00	22.5
2.00	22.5
2.00	22.5
2.00	22.5

Non-parametric correlation coefficient

- Standard tables list the critical values that r_s must exceed, as a function of nu, for the hypothesis that the two variables are unrelated to be rejected at a particular level of statistical significance (e.g. 95%, 99%)

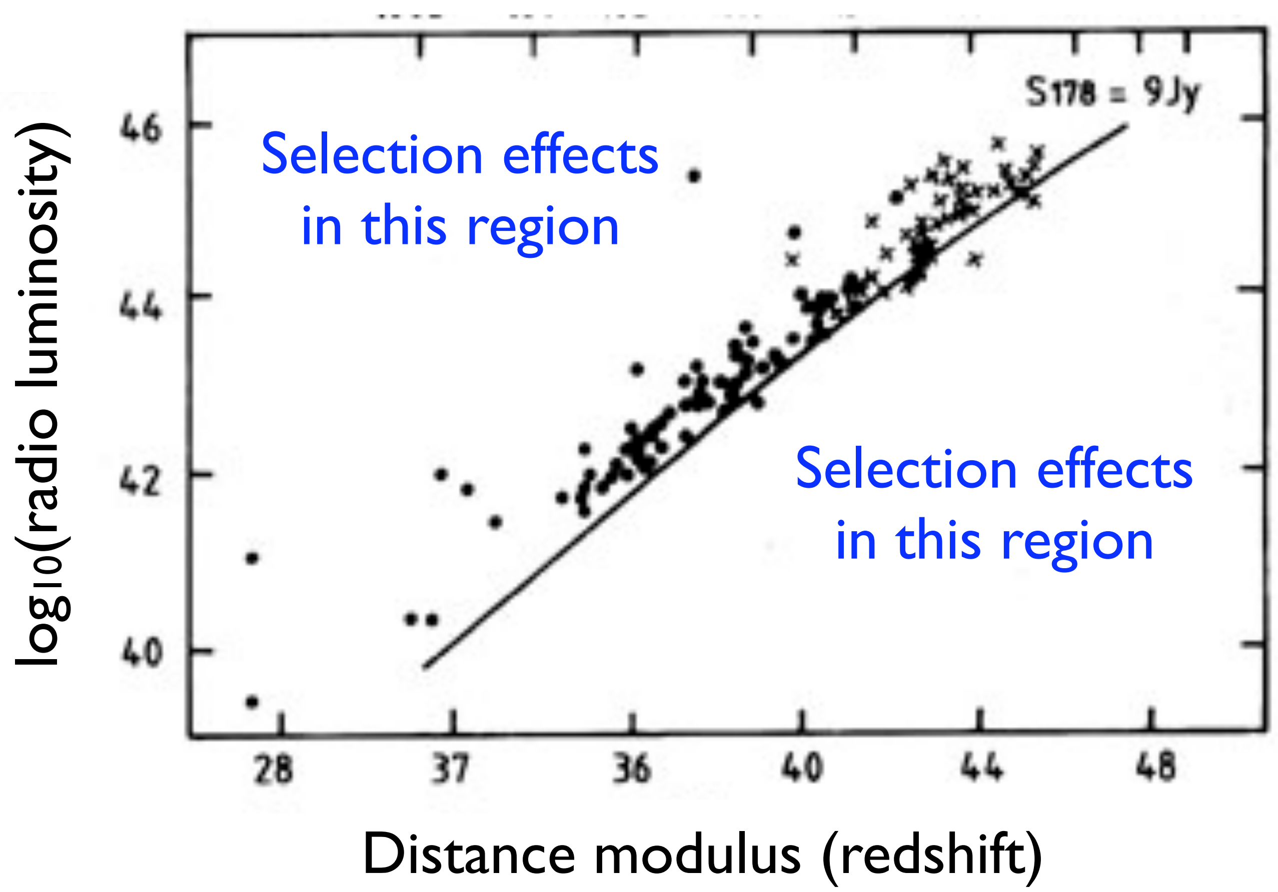
	Level of Significance for Two-Tailed Test			
df	.10	.05	.02	.01
5	.900	1.000	1.000	—
6	.829	.886	.943	1.000
7	.714	.786	.893	.929
8	.643	.738	.833	.881
9	.600	.683	.783	.833
10	.564	.648	.746	.794
12	.506	.591	.712	.777
14	.456	.544	.645	.715
16	.425	.506	.601	.665
18	.399	.475	.564	.625
20	.377	.450	.534	.591
22	.359	.428	.508	.562
24	.343	.409	.485	.537
26	.329	.392	.465	.515
28	.317	.377	.448	.496
30	.306	.364	.432	.478

Non-parametric correlation coefficient

- Part (e) : Determine the Spearman rank cross-correlation coefficient and its statistical significance
- Lemaitre : r_s=0.42 , nu=40 , prob=6.0e-3 (2.8 sigma)
- Hubble : r_s=0.80 , nu=22 , prob=3.4e-6 (4.7 sigma)
- [Results very similar to Pearson product-moment correlation coefficient, but fewer assumptions]

Issues with correlations

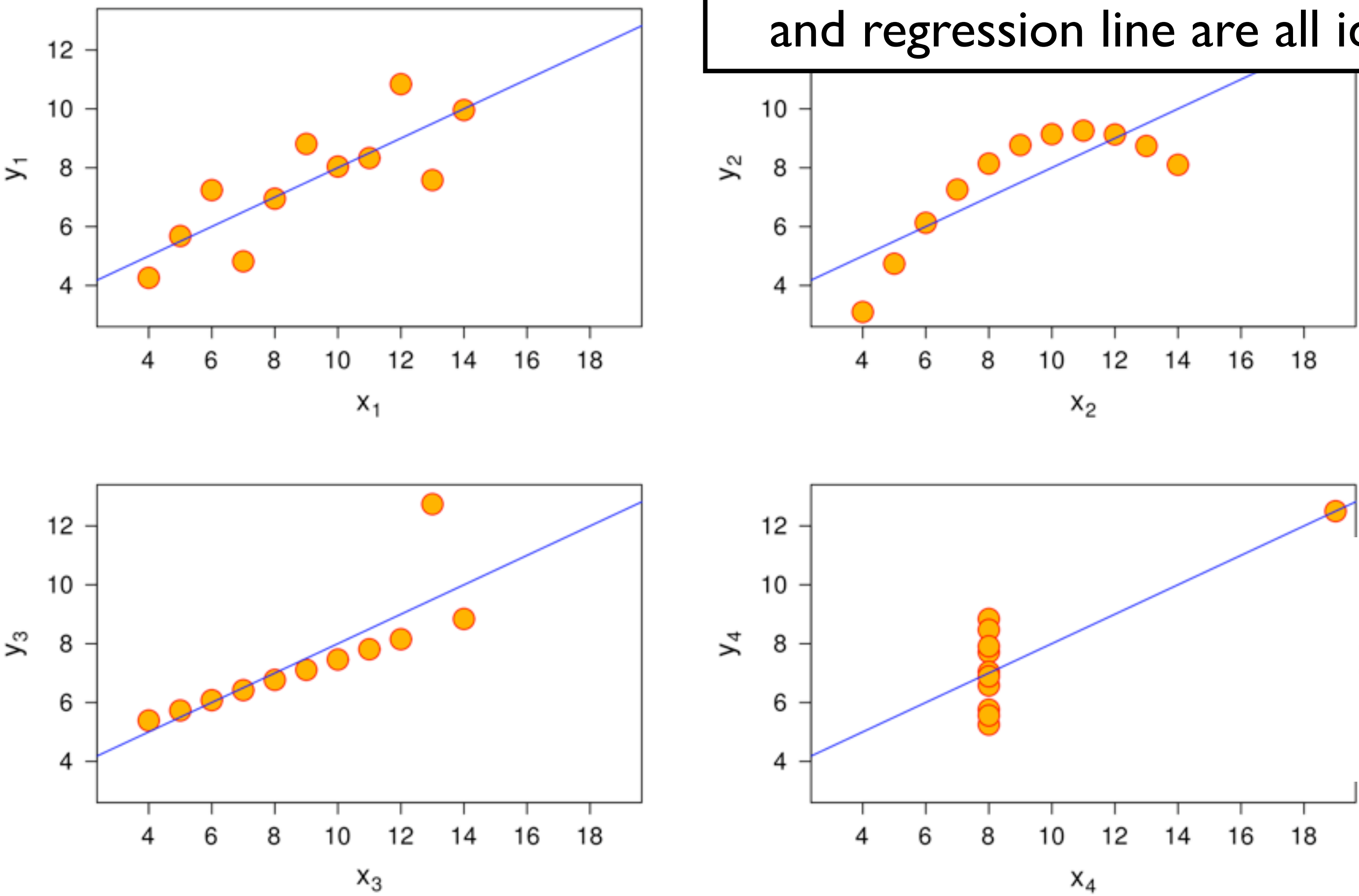
- Selection effects leading to spurious correlations, for example Malmquist bias



Issues with correlations

- Is the correlation driven by a small number of outliers, so is not robust?

Mean, variance, correlation coefficient and regression line are all identical



“Rule of thumb”

Issues with correlations

- Correlation does not necessarily imply causation
- Sleeping in your shoes causes you to wake up with a headache! [third variable - you were probably having a few drinks the night before...]
- Ice cream sales cause drowning! [third variable - hotter weather means more people are at the beach...]
- Having grey hair causes cancer! [third variable - age...]
- Obesity causes global warming! [third variable - everyone is getting richer...]

Lies, damn lies and statistics

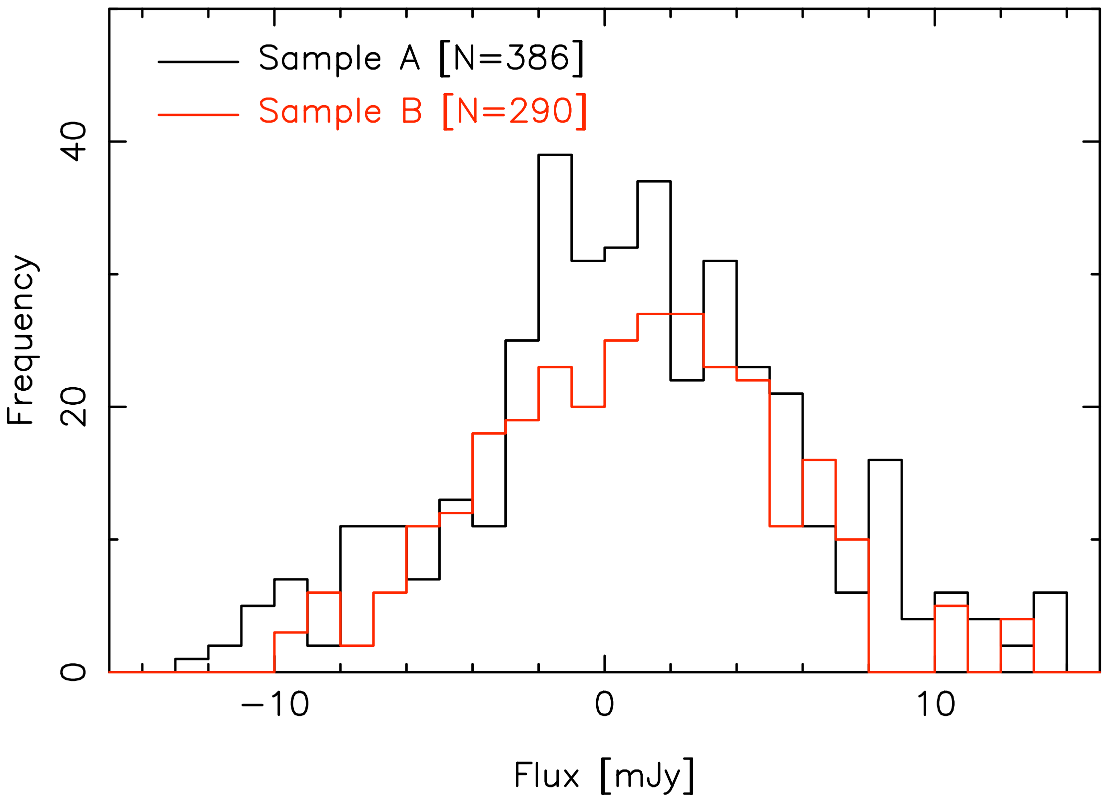
Example 4

"We don't accept the idea that there are harmful agents in tobacco"
[Phillip Morris, 1964]

Why was this poor statistics? Cigarette companies were attempting to invoke a "third variable". But correlation does sometimes imply causation, if it can be demonstrated by independent lines of evidence

Comparing two distributions

- Are two samples drawn from the same distribution?
- Example 2 : samples of flux densities measured at random positions [A] and galaxy positions [B]



Comparing two distributions

- Part (a) : Are their means consistent?

- Calculate t statistic and no. of degrees of freedom :

$$t = \frac{|\mu_x - \mu_y|}{\sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}}} \qquad \nu = \frac{\left(\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}\right)^2}{\frac{\sigma_x^4}{N_x^2(N_x-1)} + \frac{\sigma_y^4}{N_y^2(N_y-1)}}$$

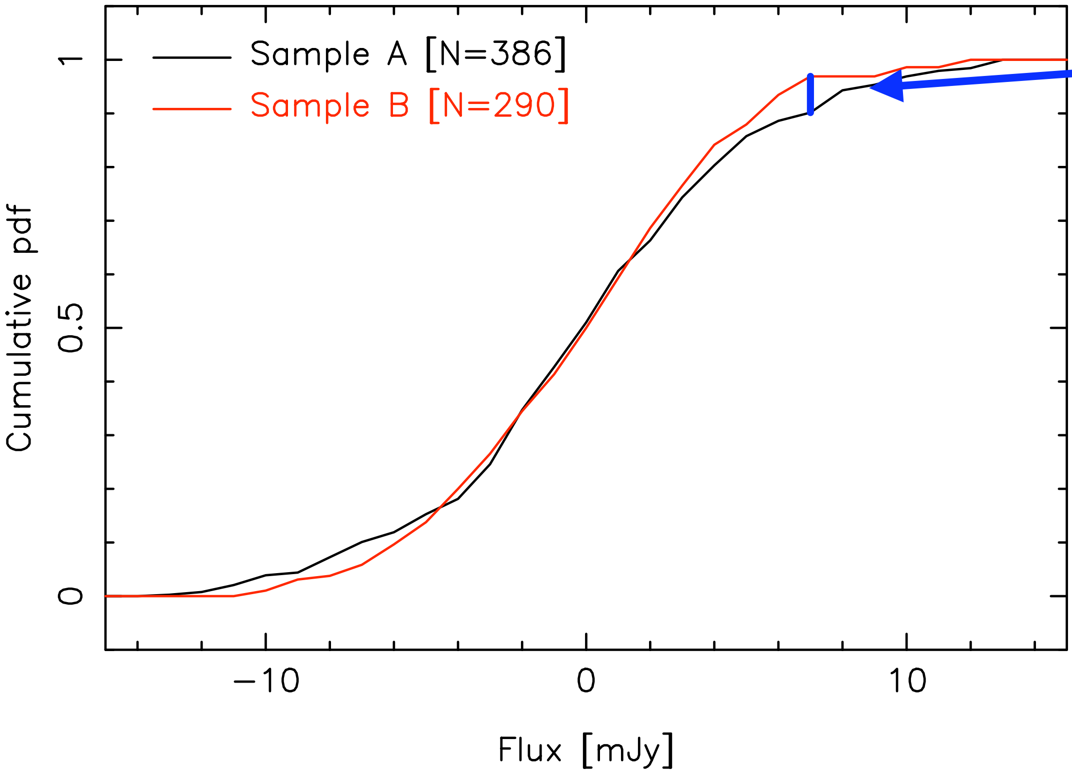
- Compare to Student's t distribution

- t = 0.31, nu = 661.5, prob of consistency = 0.76

- Small print : assumes (x,y) are normally-distributed populations

Comparing two distributions

- Part (b) : Are the full distributions consistent?
- The Kolmogorov-Smirnov test considers the maximum value of the absolute difference between the cumulative probability distributions



Max diff = D = 0.067

$$\nu = \frac{N_x N_y}{N_x + N_y} = 165.6$$

$$Q = \left(\sqrt{\nu} + 0.12 + \frac{0.11}{\sqrt{\nu}}\right) D = 0.876$$

Prob(Q) = 0.427