W&J Cap 4: Correlação e Associação

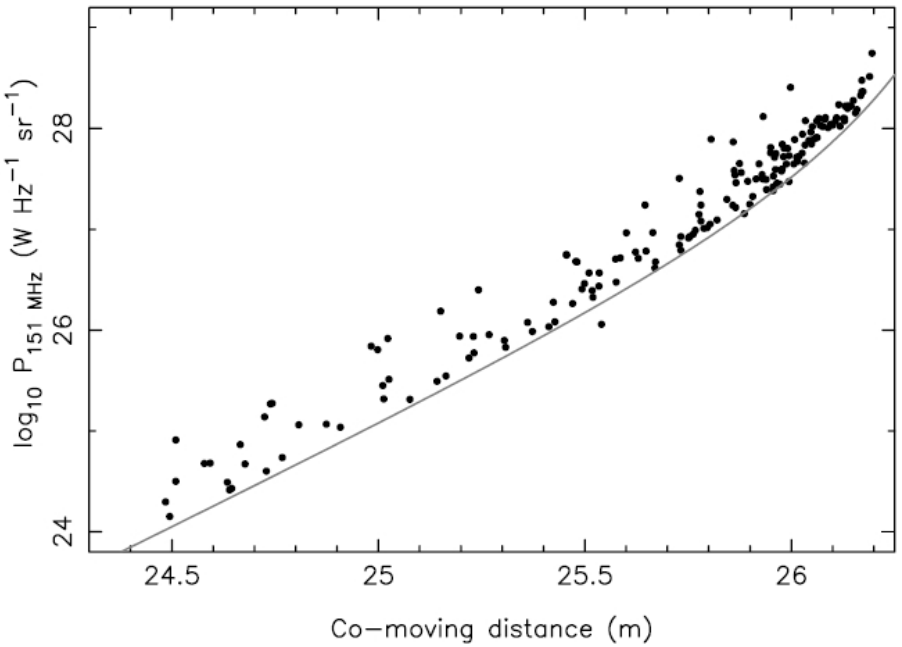
- Quando fazemos uma série de medidas é praticamente instintivo tentar correlacionar as observações com outros resultados. alguns dos motivos por trás desse 'instinto' são:
  (a) Checar se as medidas dos outros observadores são razoáveis.
  (b) Checar se as suas medidas são razoáveis.
  (c) Testar uma hipótese, possivelmente a razão inicial das observações.
  (d) Na ausência de qualquer hipótese ou conhecimento, ou na falta de algo melhor com os dados, ver se ele se correlaciona com algo na esperança de descobrir algo novo.

A Pescaria

Tomemos o último ponto primeiramente. Suponhamos que graficamos alguma coisa em função de alguma outra coisa, numa autentica "pescaria" nos dados. Existem vários perigos nesse tipo de expedição e devemos nos perguntar as seguintes questões:

1. Você consegue perceber alguma correlação com os olhos? Se não cálculos formais de estatística de correlação são, provavelmente, uma perda de tempo.

2. A correlação aparente pode se dever a um efeito de correlação? Veja, por exemplo a figura abaixo de Sandage (1972):

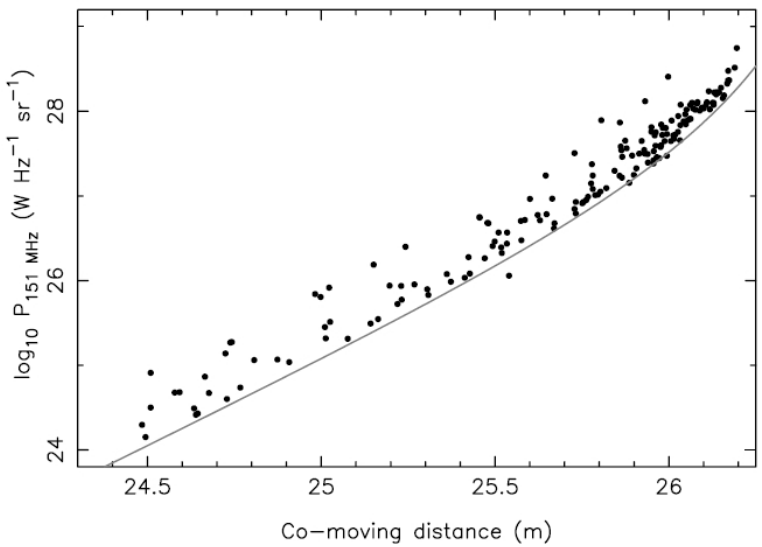


Luminosidades rádio versus módulo de distância:

$\mu = m - M = 5log(d(Mpc) + 25)$. A curva representa os limites das observações, impostos pela seleção em fluxo.

- Numa primeira vista a figura *prova* claramente a evolução em luminosidade de rádio-fontes, pois quão mais distante está o objeto mais brilhante ele é, não?
- Não. Isso está errado!

A Pescaria

- 

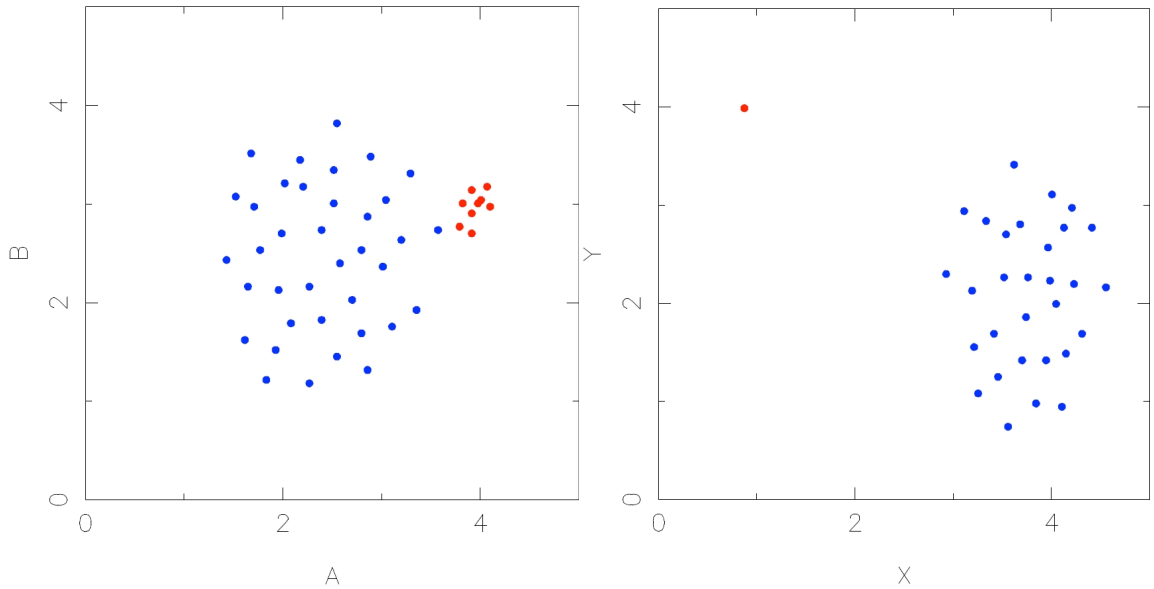
- Esse é o *viés de Malmquist em ação*

- O canto superior direito nunca poderia estar povoado por conta do limite de detecção. Mas e o superior esquerdo?
- Dada uma função de luminosidade seja menor para objetos cada vez mais brilhantes é pouco provável encontrar no pequeno volume do universo local objetos brilhantes.

A Pescaria

3. Se nós passamos pelo item (2), então podemos calcular formalmente a significância da correlação. Ademais, se há uma correlação, a reta ajustada faz sentido?

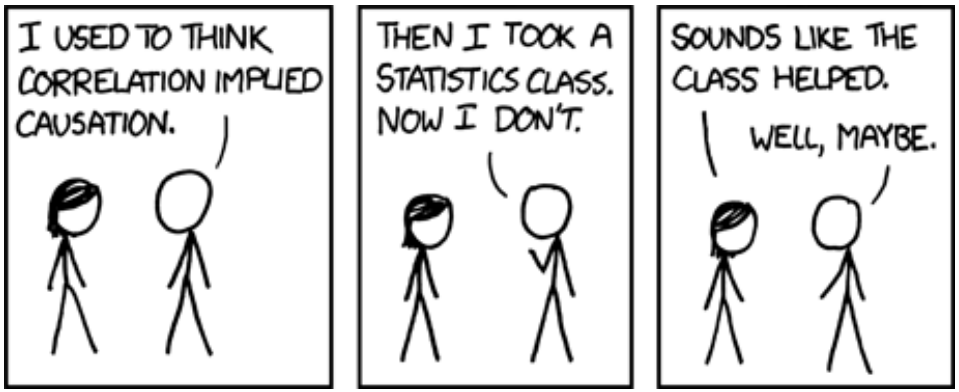
4. Se continuamos satisfeitos devemos retornar ao gráfico e nos perguntar se o resultado formal é realístico. Uma *regra do dedão*: se 10% dos pontos estão agrupados por si mesmos, ao cobri-los com o meu dedão eu destruo a correlação vista pelos meus olhos? Em caso positivo devemos duvidas do resultado, independentemente do resultado formal da análise. Cuidado em particular com gráficos como os de baixo, que sugerem efeitos de seleção, erros nos dados ou outra forma de "conspiração estatística".



Más correlações: em ambos casos testes formais indicam que existem correlações com alto grau de significância.

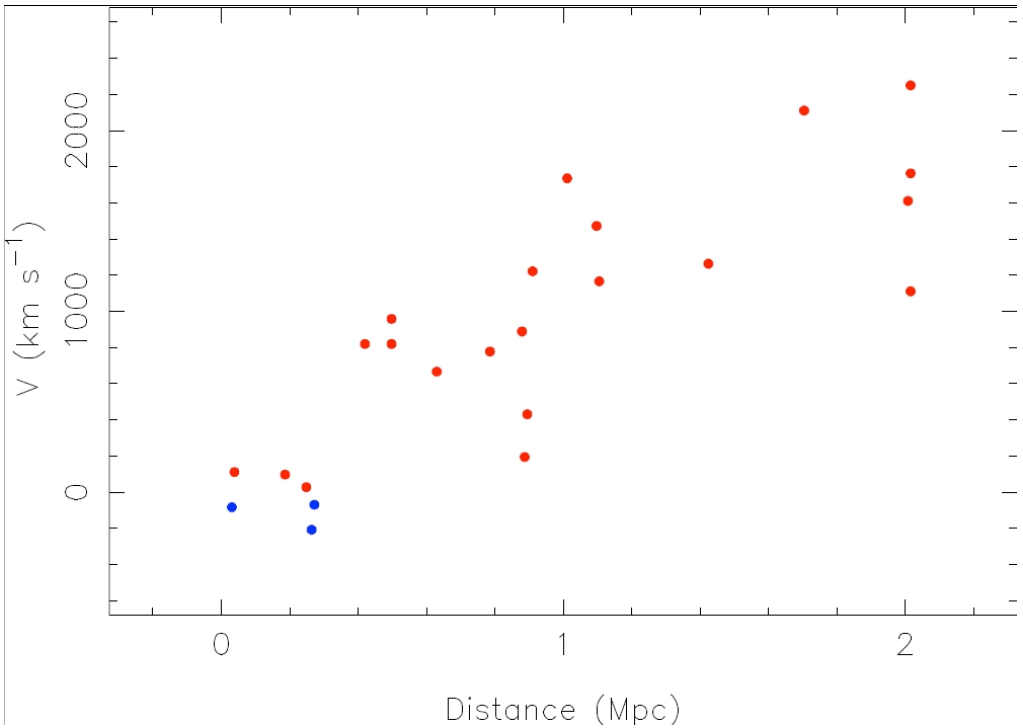
A Pescaria

6. Se nós *ainda* estamos confiantes, devemos lembrar que uma correlação não implica em conexão causal. Pode indicar apenas que ambas variáveis dependem de uma terceira variável. De fato existem métodos para se procurar por correlações entre variáveis aos quais se sabe de antemão que estão correlacionados a uma terceira variável. O problema, numa "pescaria" é conhecer ou mesmo desconfiar dessa terceira variável.



A Pescaria

- Por fim, isso tudo não deve ser um desencorajador. Veja a figura abaixo. Uma correlação relativamente pobre, mas nenhum agrupamento que pode ser rejeitado pela regra do dedão. De fato esse foi um dos primeiros diagramas de Hubble (Hubble 1936) que serviram para demonstrar a expansão do Universo.



Testando correlações

- Dados dois conjuntos de medidas (X_i, Y_i) como podemos testar formalmente se estes estão relacionados?

- A melhor forma de definir esse conceito de 'relação' é ao modelar os dados usando uma gaussiana bivariada com *coeficiente de correlação* ρ:

$$prob(x, y|\sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times$$

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

- Nesse modelo 'correlação'significa $\rho \neq 0$. Se a correlação for perfeita $\rho \to 1$ e for pequena $\rho \to 0$

Testando correlações

- O parâmetro ρ é o *coeficiente de correlação* e, nessa formulação, é dado por:

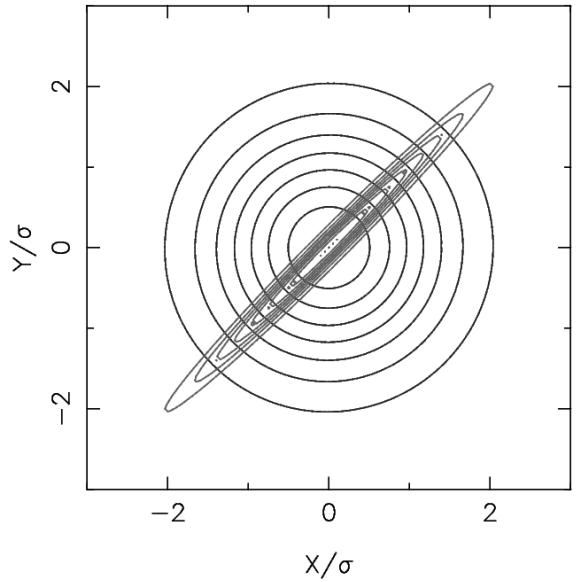
$$\rho = \frac{cov[x, y]}{\sigma_x \sigma_y},$$

onde $cov[x, y]$ é a covariância (seção 3.3.1) de x e y e σ_x e σ_y as variançcas.

- O coeficiente de correlação pode ser estimado por:

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

onde r é conhecido como coeficiente de correlação de Pearson.



Testando correlações

- Assumindo, sem perda de generalidade, que as médias são zero, os contornos da figura caem de um fator $1/e$, a partir do máximo quando:

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

ou, em notação matricial, quando:

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

- A inversa da matriz central é conhecida como *matriz de covariância* ou *matriz de erro*:

$$C = \begin{pmatrix} \frac{1}{\sigma_x^2} & cov(x,y) \\ cov(x,y) & \frac{1}{\sigma_y^2} \end{pmatrix}$$

- Os elementos não-diagonais da matriz podem ser estimados por:

$$\frac{1}{N-1}\overline{(x_i-\overline{X_i})((x_j-\overline{X_j})}.$$

Testando correlações

- Essa matriz é particularmente valiosa para para cálculos de propagação de erros, mas tem outras várias aplicações, como em PCA (Análise de Componentes Principais)e no modelamento via máxima verossimilhança.

- A gaussiana multi-variada é um exemplo da classe (e o mais familiar) de distribuições multi-variadas que dependem apenas do vetor $\vec{x}$ via a chamada forma quadrática:

$$\vec{x}^T C \vec{x}.$$

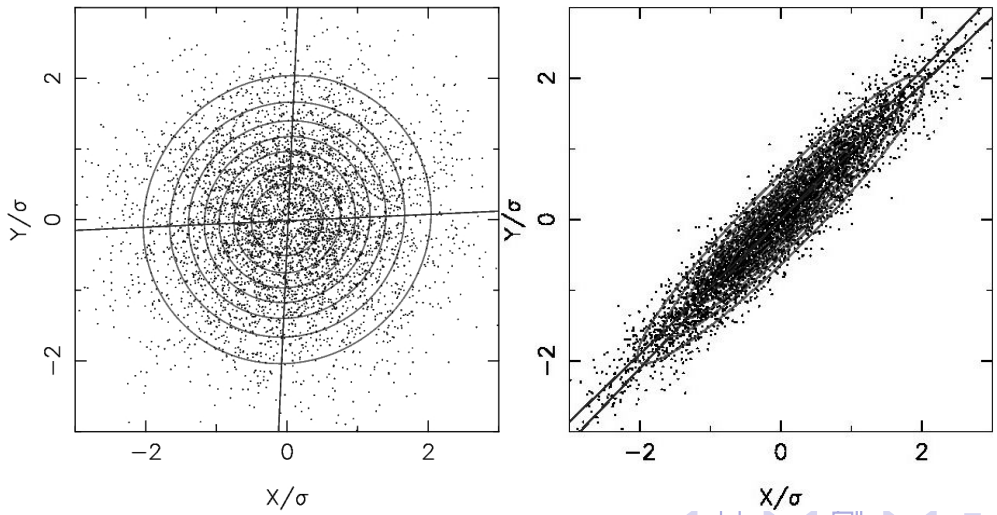
- A forma quadrática define uma elipse, de modo que toda essa classe produz contornos de equiprobabilidade elípticos.

Testando correlações

- Uma caso especial de geração de números aleatórios envolve gaussianas multivariadas com parâmetros σ_i e ρ_i. Isso é crucial para testar rotinas de ajuste de modelos e, graças à discussão das matrizes de erros, simples de formular:
  1. Monte a matriz de erro ou covariança (para o caso bivariado a matriz de erros é: $e_{1,1} = \sigma_x^2$, $e_{2,1} = e_{1,2} = cov[x, y] = \rho\sigma_x\sigma_y$ e $e_{2,2} = \sigma_y^2$, como vimos).
  2. Encontre os auto-valores e auto-vetores da matriz de covariança.
  3. Combine os autovetores, na matriz de transformação T que diagonaliza a matriz de covariança.
  4. Sorteie pares gaussianos (x', y'), não-correlacionados, com variançcas iguais aos dois auto-valores. Compute os pares (x, y) de acordo com:

$$\begin{pmatrix} x \\ y \end{pmatrix} = [T] \quad \begin{matrix} x' \\ \end{matrix} \quad \begin{matrix} y' \\ \end{matrix}$$

  5. Com isso é possível gerar os pontos dos gráficos abaixo:



Testando correlações: modo bayesiano

- Esse modelo gaussiano é bastante específico e não pode ser usado em qualquer situação para testar correlações. Para resolver esse problema (e vários outros) sempre há dois caminhos: bayesiano e não-bayesiano. Comecemos pelo primeiro.

- O método bayesiano usa o teorema da Bayes para extrair a probabilidade da distribuição para ρ a partir da verossimilhança dos dados e *a prioris* razoáveis.

- Isso foi feito por Jeffreys (1961), para o modelo da gaussiana bivariada que obteve o posterior total, $p(\rho, \sigma, \sigma_y, \mu_x, \mu_y|\text{dados})$ e marginalizou sobre os parâmetros irrelevantes para obter:

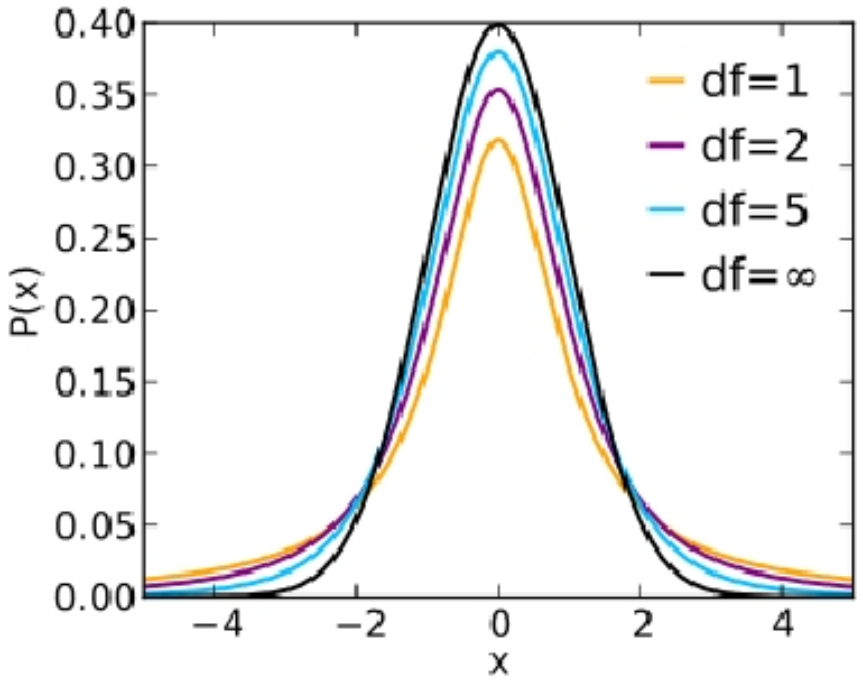
$$prob(\rho|\text{dados}) \propto \frac{(1-\rho^2)^{\frac{N-1}{2}}}{(1-r\rho)^{N-\frac{3}{2}}} \left(1 + \frac{1}{N-1/2}\frac{1+r\rho}{8} + ...\right)$$

- Com os dados (X_i, Y_i) calcula-se r e, dado em intervalo de interesse, calcula-se $prob(\rho)$.

Distribuição t de Student

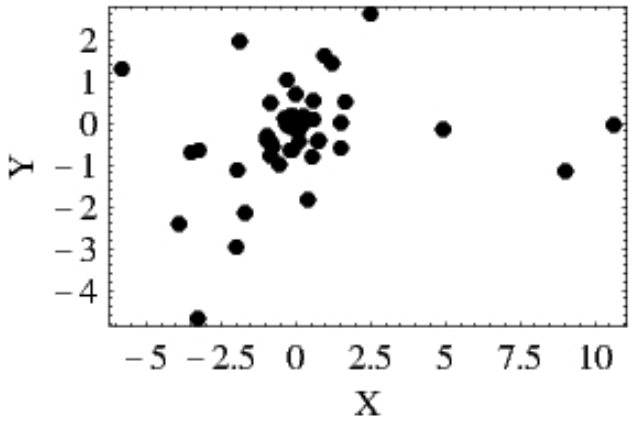
- $$f(t,\nu) = \Gamma[(\nu+1)/2] \times \frac{(1+t^2)^{[(\nu+1)/2]}}{\sqrt{\pi\nu}\Gamma(\nu/2)}$$

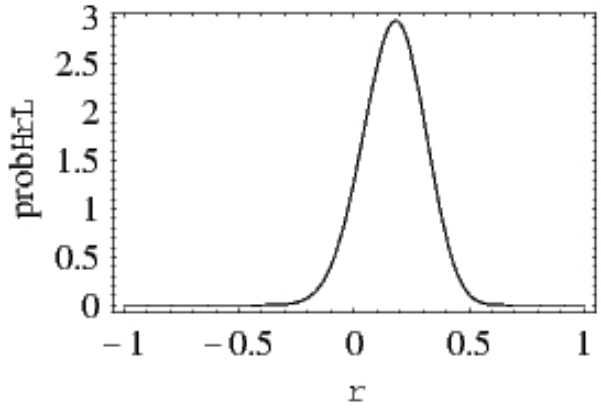
  Distribuição usada para comparar médias: Dados nx_i pontos gerados a partir de uma distribuição Normal com parâmetros (μ, σ), então $t = (\bar{x}_s - \mu)/(\sigma_S/\sqrt{N})$ é distribuído conforme $f(t,\nu)$, onde ν, é o número de 'graus de liberdade': $\nu = n - 1$.

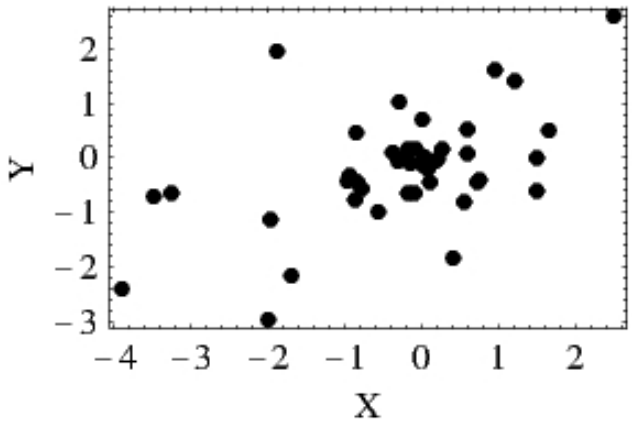
- 

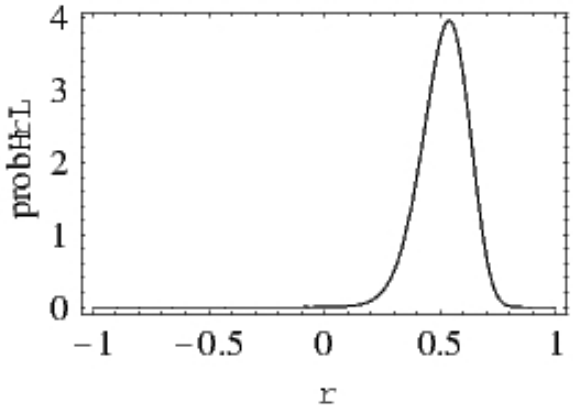
  Distribuição t de Student com 1, 2 e 5 graus de liberdade e a distribuição normal (infinitos graus de liberdade).

Exemplo W&J P. 77



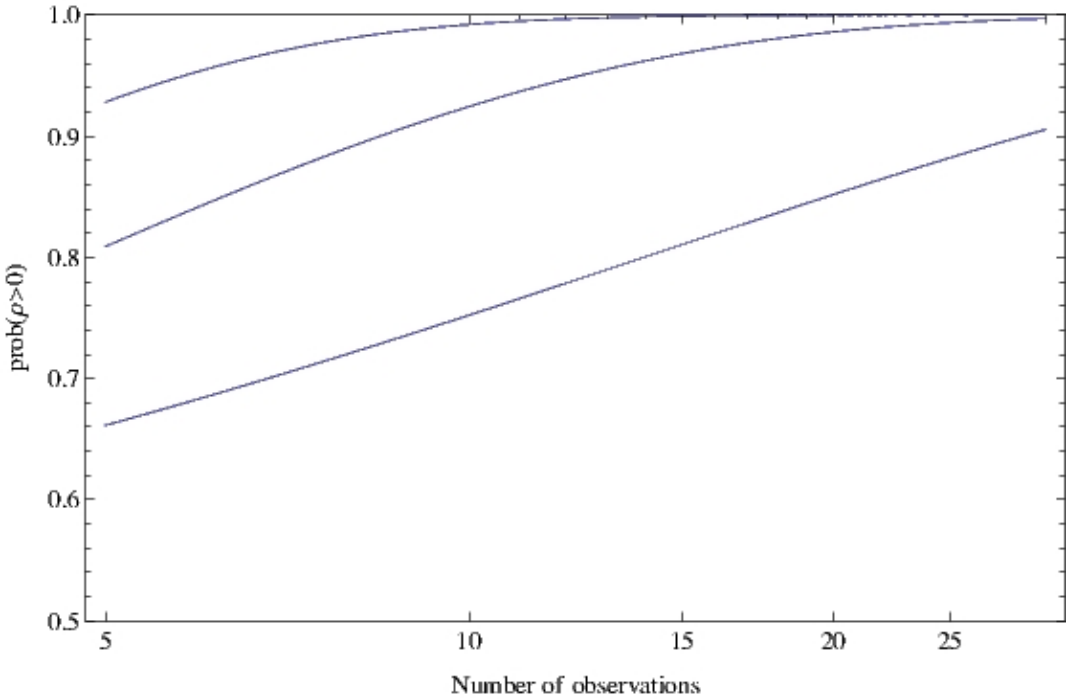






Painéis Superiores 50 pontos gerados usando uma função t de student bivariada com coeficiente de correlação verdadeiro 0.5. **Painéis inferiores** O mesmo após uma remoção de pontos que distam mais de 4σ da média.

- Um outro uso interessante dessa distribuição é o calculo da probabilidade de ρ ser positivo em função do número de pontos. Isso tem a ver com a quantidade de dados necessária para detectarmos correlações com confiança.



Probabilidade de ρ ser positivo como função do tamanho da amostra N para os seguintes valores de r 0,25 (curva de baixo), 0,50 e 0,75 (curva de cima).

- Isso nos lembra que uma das características mais interessantes do modo bayesiano é que, como este produz ao final uma PDF para todos os valores razoáveis de ρ ele dá resposta a várias perguntas possíveis, como a feita acima. O modo clássico, geralmente produz apenas um valor (o máximo do posterior) e a significância associada.

Testando correlações: modo clássico

- No modo clássico a abordagem começa supondo ρ como uma quantidade fixa, não uma variável sobre a qual associamos uma distribuição de probabilidades. Essa abordagem, portanto chega na probabilidade dos *dados* dados ρ e a hipótese de fundo da gaussiana bivariada. O resultado, obtido por Fisher em 1944 é:

$$prob(r|\rho, H) \propto \frac{(1-\rho^2)^{(N-1)/2}\,(1-r^2)^{(N-4)/2}}{(1-r\rho)^{N-3/2}}\left(1+\frac{1}{N-1/2}\,\frac{1+r\rho}{8}+...\right)$$

- O que pode ser feito com essa resposta? O procedimento padrão é escolher uma "hipótese nula" simples, como $\rho = 0$ e, então, computar a probabilidade, sob a hipótese nula de r ter o valor atual ou ser ainda maior.

- Se esse valor é muito pequeno, sente-se que a hipótese nula é improvável.

Testando correlações: modo clássico

- O teste paramétrico padrão é tentar rejeitar a hipótese nula ($\rho = 0$), e isso é feito ao computar-se$r$. Note que $-1 < r < 1$ e $r = 0$ significa sem correlação.

- Para testar a significância da correlação (valor não-zero para r), calcule:

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

  que obedece a uma distribuição probabilística t de Student com $N-2$ graus de liberdade (isso é feito apenas para que se possam usar as tabelas de t).

- Nesse momento estamos entrando no domínio do *teste de hipóteses* algo que faremos mais sistematicamente mais adiante. O procedimento é o que se segue.

Testando correlações: modo clássico

- Consulte tabelas de valores críticos de t (por exemplo Tabla B.3 de W&J).

Table B.3 *Critical values of Student's t distribution*

	Level of significance for one-tailed test					
	0.100	0.050	0.025	0.010	0.005	0.0005
	Level of significance for two-tailed test					
	0.200	0.100	0.050	0.020	0.010	0.001
$\nu = 1$	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587

- Se o valor de t for maior que o valor crítico correspondente da probabilidade (teste bilateral ou *two-tailed*) então a hipótese nula (ausência de correlação) foi rejeitada no nível de significância especificada

- Esse nível de significância (usualmente 5% ou 1%) é a máxima probabilidade de rejeitarmos a hipótese nula sendo esta verdadeira, que consideramos aceitável.

Testando correlações: modo clássico

- Essa abordagem provavelmente não responde à questão de quão correlacionados estão os pontos.

- Esse teste é usado amplamente e é formalmente poderoso, mas tem um série de restrições: *i*) a relação entre as variáveis tem que ser linear; *ii*) os dados precisam ser gerados a partir de uma distribuição Normal; *iii*) os dados não podem apresentar agrupamentos.

- Existe o teste F para lidar com a não-linearidade, mas para evitarmos esses problemas a melhor alternativa é usar um teste não-paramétrico.

Testando correlações: modo clássico não-paramétrico

- O mais conhecido teste não-paramétrico consiste em computar o coeficiente de correlação de Spearman:

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

onde existem N pares de dados e os N valores de cada uma das duas variáveis são reordenadas de modo que (X_i, Y_i) representa as posições do i-ésimo par, $1 < X_i < N$, $1 < Y_i < N$.

- O intervalo de valores é $0 < r_s < 1$ e um valor alto significa alta probabilidade de correlação. Para saber a significância do resultado é necessário recorrer-se a tabelas (B.5 do W&J), que é aplicável para $4 \geq N \geq 30$.

- Caso $N > 30$ calcula-se

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

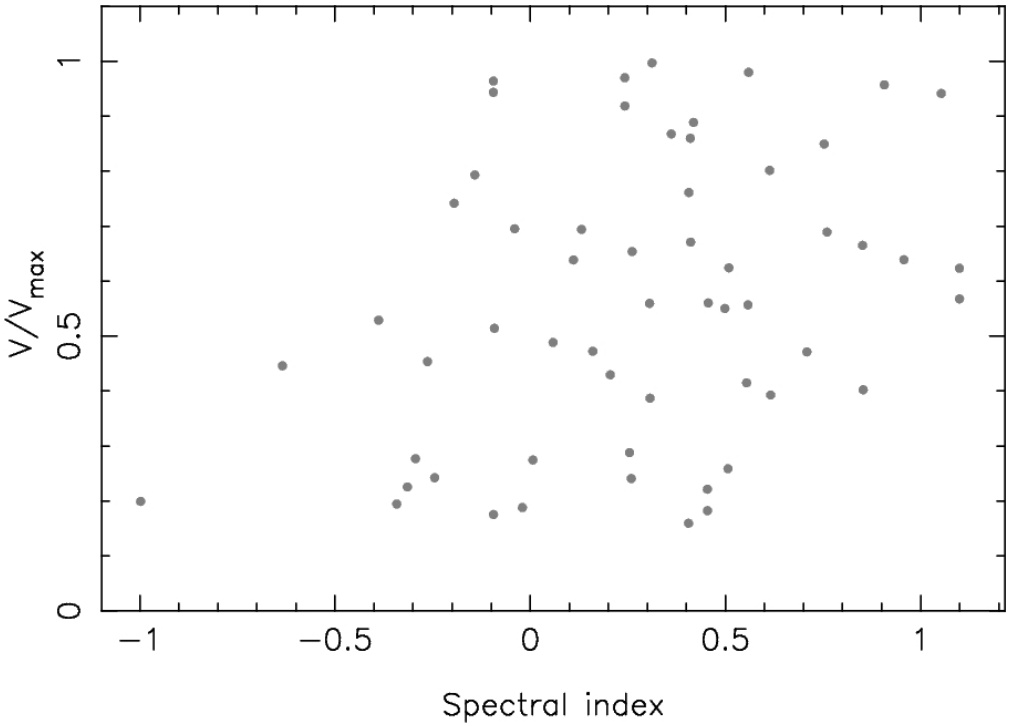
que é uma estatística cuja distribuição tenda à distribuição t com $N - 2$ graus de liberdade.

- Como o uso de r_s se compara com o uso de r, o mais poderoso teste de correlação paramétrico? Bastante bem: com eficiência de 91%. A moral da história aqui é: quando em dúvida, muito pouco se perde ao usar um teste não-paramétrico.

Testando correlações: modo clássico não-paramétrico

Exemplo W&J P. 81

- Na figura abaixo se vê uma correlação com significância de 2σ. Obtém-se $r_s = 0.28$, $N = 55$ e a hipótese de qua ambas variáveis não está correlacionadas é rejeitada ao nível de 5% de significância.



- Aqui, não temos ideia da distribuição subjacente, nem temos claro o que são os eixos. A hipótese de uma gaussiana bivariada seria precipitada ao extremo.

Testando correlações: testes bayesianos *versus* não-bayesianos

- Ainda que os testes não-paramétricos contornem alguns dos problemas que envolvem a abordagem não-gaussiana mas não tocam na questão fundamental: qual é a questão real?

- Por outro lado o modo bayesiano, cujo forte é fornecer respostas para todo tipo de perguntas, força a utilização de um modelo.

- Há pouca diferença na prática entre o teste de Fischer e os resultados da distribuição de Jeffreys.

Correlações parciais

- É possível lidar com a questão da 'terceira variável à espreita' (uma vez que suspeita-se da existência desta) através da *correlação parcial*., na qual correlação 'parcial' entre duas variáveis é considerada ao se remover o efeito da terceira (ou quarta ou mais) variável sobre as variáveis em consideração.

- Correlação parcial é uma quase ciência em si mesma. Ela é coberta nas formas paramétrica e não-paramétrica por Maclin(1982), Siegel & Castellan (1988) e Stuart & Ord (1994).

- Na forma paramétrica, considere uma amostra de N objetos para os quais os parâmetros x_1, x_2 e x_3 foram medidos. O *coeficiente de correlação parcial de primeira ordem* entre as variáveis x_1 e x_2 é:

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

onde r é o coeficiente de correlação de Pearson definido na equação 1.

Correlações parciais

- No caso de existirem quatro variáveis, então o *coeficiente de correlação parcial de segunda ordem* é:

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

  onde a correlação examinada é a entre x_1 e x_2, enquanto que x_3 e x_4 são mantidas constantes. (E assim por diante para ordens maiores.)

- A significância vem do teste t de Student como anteriormente:

$$t = r_{12.34\ldots m} \sqrt{\frac{N - m}{1 - r_{12.34\ldots m}^2}},$$

  que se comporta como uma função t com $N - m$ graus de liberdade e onde m é o número de variáveis envolvidas.

- Exemplo W&J P. 84

- Fim da aula 5