



UNIVERSIDADE FEDERAL  
DO RIO DE JANEIRO



# Astroestatística

Aula 004 -- Prof. Walter Martins-Filho

# Estatística Descritiva

- Distribuições Estatísticas Paramétricas
  - Unidimensionais
  - Multidimensionais

Distribution	Parameters	$\bar{x}$	$q_{50}$	$x_m$	$\sigma$	$\sigma_G$	$\Sigma$	$K$
Gaussian	$\mu, \sigma$	$\mu$	$\mu$	$\mu$	$\sigma$	$\sigma$	0	0
Uniform	$\mu, W$	$\mu$	$\mu$	N/A	$W/\sqrt{12}$	$0.371W$	0	-1.2
Exponential	$\mu, \Delta$	$\mu$	$\mu$	$\mu$	$\sqrt{2}\Delta$	$1.028\Delta$	0	3
Poisson	$\mu$	$\mu$	$\mu - 1/3$	$\mu - 1$	$\sqrt{\mu}$	N/A	$1/\sqrt{\mu}$	$1/\mu$
Cauchy	$\mu, \gamma$	N/A	$\mu$	$\mu$	N/A	$1.483\gamma$	N/A	N/A
$\chi^2_{\text{dof}}$	$k$	1	$(1 - 2/9k)^3$	max $(0, 1 - 2/k)$	$\sqrt{2/k}$	N/A	$\sqrt{8/k}$	$12/k$

# Distribuições Multidimensionais

# Distribuições Bidimensionais

- Também chamadas de “Bivariate Distributions”
- $\{x_i, y_i\} \rightarrow h(x, y)dx dy$  Probabilidade entre  $x$  e  $x + dx$  e  $y$  com  $y + dy$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y)dx dy = 1$$

$$E[x_i^n] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i^n h(x_i | i = 1, 2) dx_1 dx_2$$

Por exemplo:

$$Var(x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^2 h(x, y) dx dy$$

## Analogamente

Os momentos seguem somatórios ou integrais duplas

# Covariância

- A covariância é definida como:

$$Cov(x, y) = V_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^2 (y - \mu_y)^2 h(x, y) dx dy$$

$$\sigma_x = \sqrt{Var(x)} = \sqrt{V_x}$$

$$\sigma_y = \sqrt{Var(y)} = \sqrt{V_y}$$

$$\sigma_{xy} = \sqrt{Var(xy)} = \sqrt{V_{xy}}$$

# Distribuições Marginais

- Em vez de integrar (ou somar) em todas as variáveis, podemos somente integrar em  $x$  ou  $y$  afim de obter uma margem da PDF para  $y$  ou  $x$ , respectivamente:

$$m(x) = \int_{-\infty}^{\infty} h(x, y) dy$$

# O que é independência?

- A independência entre as duas variáveis aleatória em distribuições bidimensionais ( e podemos extrapolar para casos multidimensionais) ocorre que qualquer intervalo imposto a alguma das variáveis não interfere no domínio (intervalo) da variável seguinte.

# Não-Correlacionada

- Se  $\sigma_{xy} = 0$ , consideramos que  $x$  e  $y$  não são correlacionadas entre si ("uncorrelated"), e podem ser tratadas separadamente
- Desta forma, as variáveis  $x$  e  $y$  são independentes e podemos escrever a função bidimensional como o produto de cada função de probabilidade:

$$h(x, y) = h_x(x)h_y(y)$$

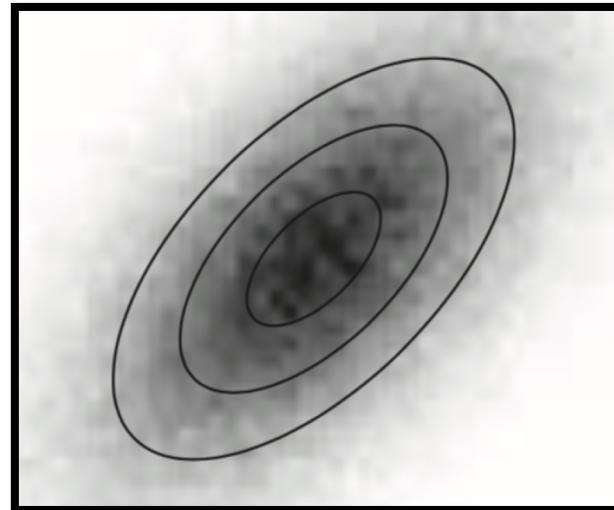
# Distribuição Bidimensional Normal

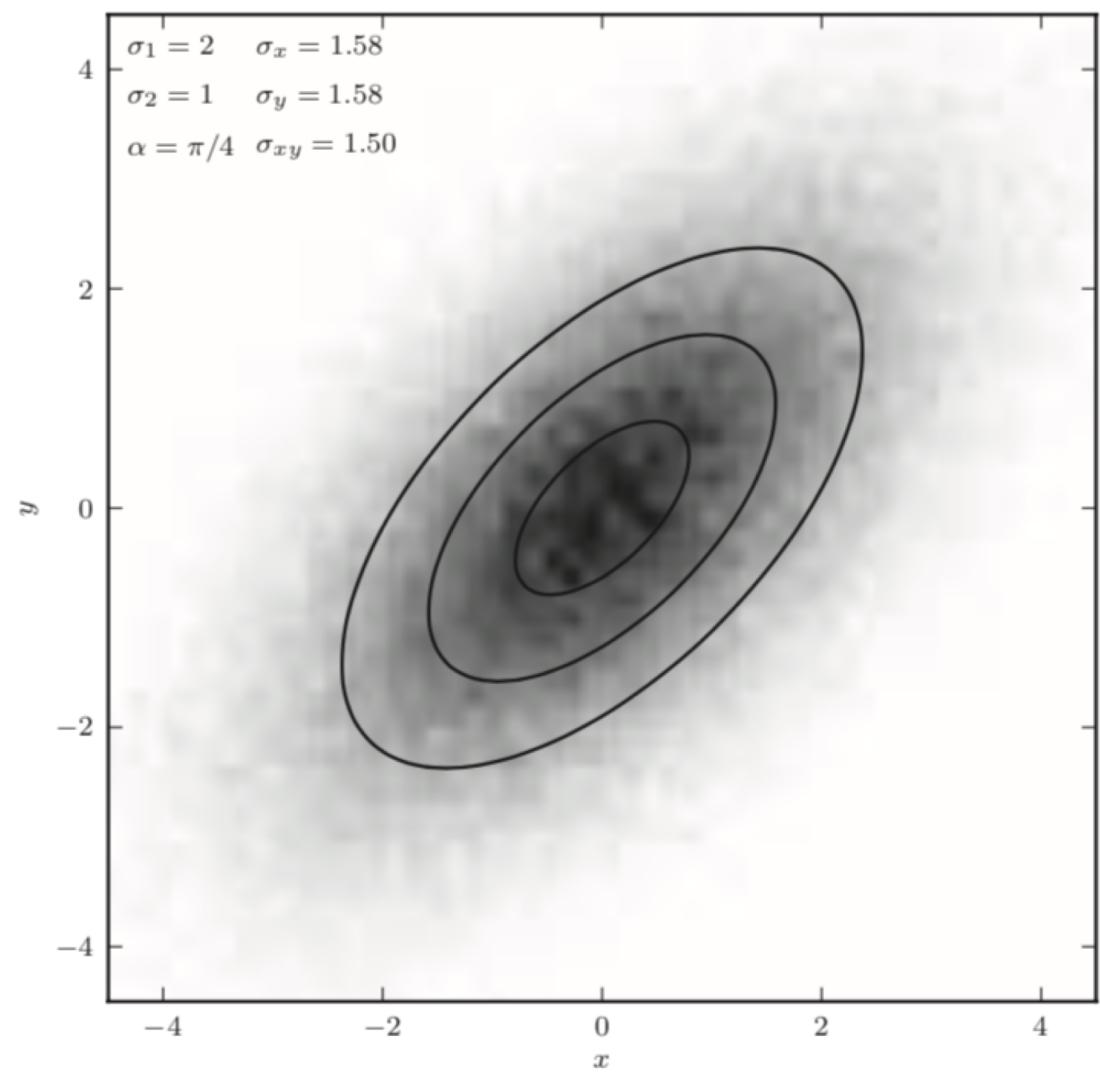
- Generalização da distribuição Normal, onde a distribuição bidimensional é descrita por uma Gaussiana 2D:

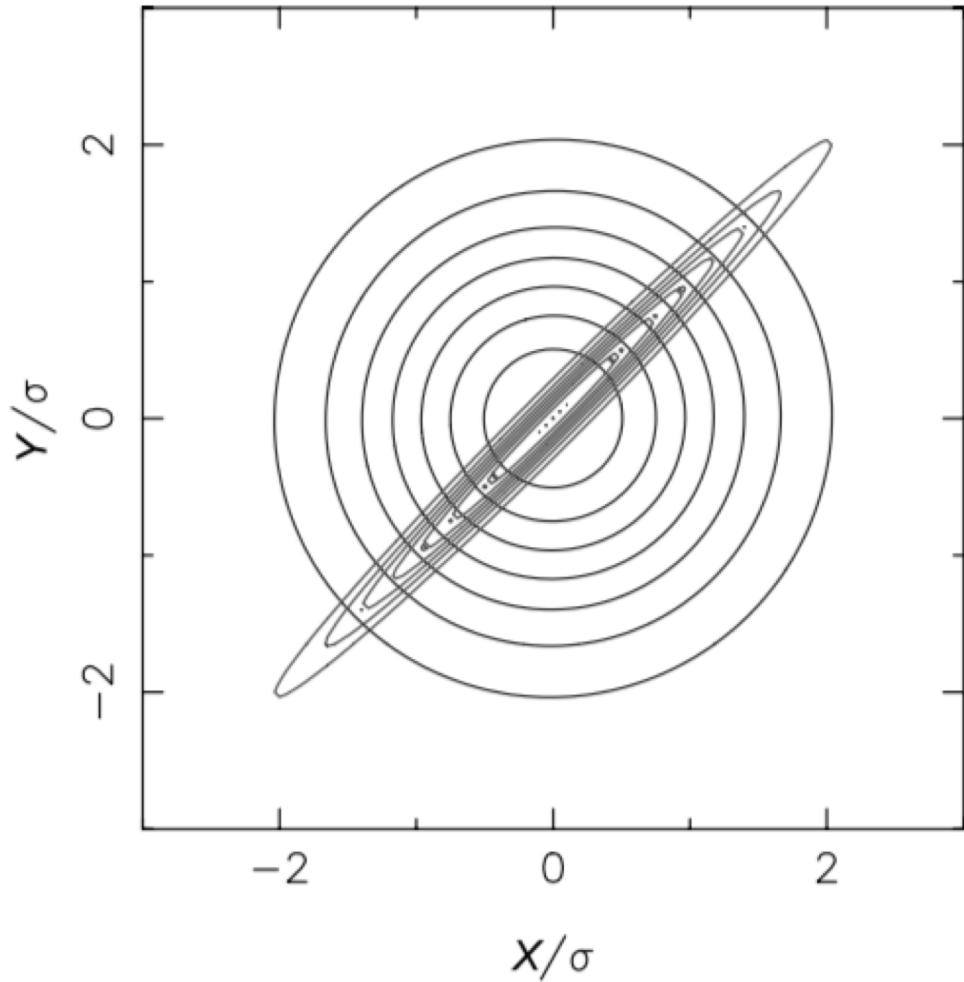
$$p(x|\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{z^2}{2(1-\rho^2)}\right)$$

$$z^2 = \frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} - 2\rho \frac{(x - \mu_x)(y - \mu_y)}{\sigma_x\sigma_y}$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$







$$\rho \neq 0$$

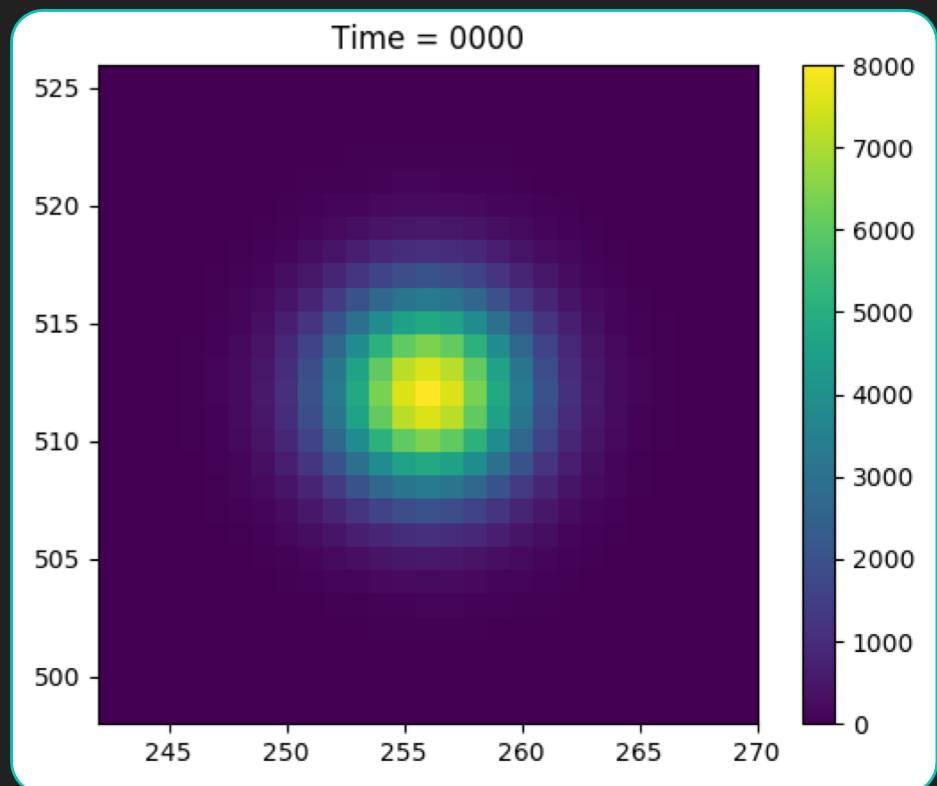
Aqui apresentamos as curvas de  $\rho = 0$  até  $\rho = 1$

No caso de  $\rho = 0$  (no caso,  $\rho = 0.01$ ), temos uma distribuição circular

No caso de  $\rho = 1$  (no caso,  $\rho = 0.999$ ), teremos dependência direta

# PSF: Point Spread Function

- A PSF de uma estrela pode ser aproximada por uma Normal bidimensional.
- Isto permite, dando a posição na grade do CCD, a amplitude (para desnormalizar), i.e., a quantidade máxima de fótons que devem cair na PSF na região central, podemos gerar objetos estelares artificiais
- Saber criar dados PSF sintéticos permite analisar, na prática, algum instrumento que se tenha em mãos ou algum código de redução/análise

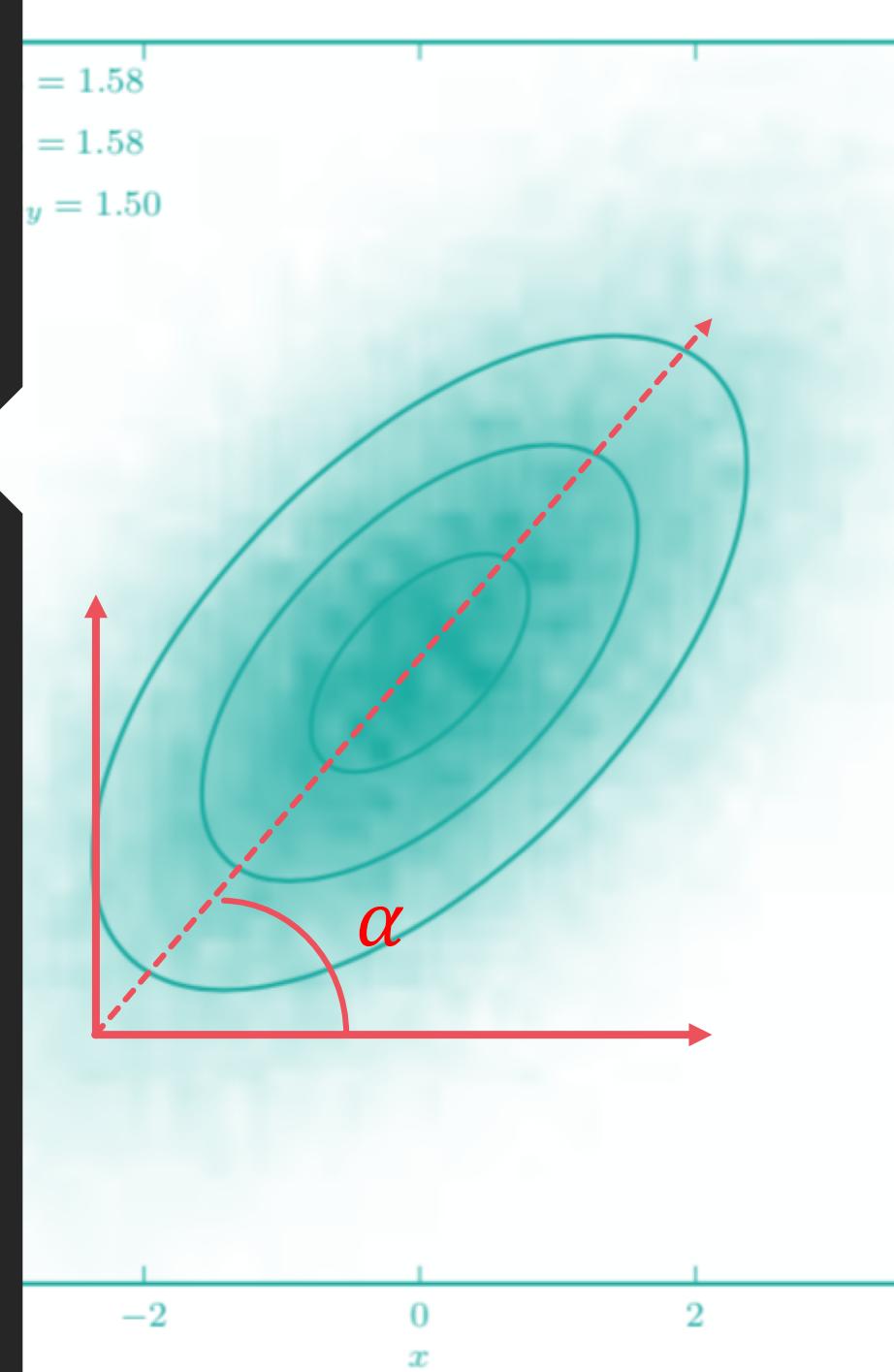


# Exercício

- Calcule a probabilidade marginal da distribuição bidimensional Normal em x e em y
- Crie uma PSF sintética assumindo que o máximo de contagens possíveis na grade seja de 5000 contagens. Faça plot 3D, e um mapa de cor como plot do slide 16. Assuma que  $\sigma_x = \sigma_y = 4$  e  $\sigma_{xy} = 0$
- Crie outra PSF artificial, porém assuma  $\sigma_{xy} \neq 0$ , escolha alguns valores em ordem sucessiva e verifique a mudança.

# Correlação da distribuição bidimensional Normal

- Para  $x$  e  $y$  não-correlacionadas em uma distribuição bidimensional Normal, teremos que  $\rho = 0$
- **Vamos considerar agora o caso em que  $\rho \neq 0$**
- Vamos chamar de  $\alpha$  o ângulo entre o eixo-x e o semi-eixo maior da elipse formada pela distribuição.



# Correlação da distribuição bidimensional Normal

- O ângulo  $\alpha$  é definido em  $-\frac{\pi}{2} \leq \alpha \leq \frac{\pi}{2}$

$$\tan 2\alpha = 2\rho \frac{\sigma_x \sigma_y}{\sigma_x^2 - \sigma_y^2} = 2 \frac{\sigma_{xy}}{\sigma_x^2 - \sigma_y^2}$$

Se fizermos uma rotação em  $\alpha$ :

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} x - \mu_x \\ y - \mu_y \end{bmatrix}$$

A correlação entre as duas variáveis desaparece.

$$P_1 = (x - \mu_x) \cos \alpha + (y - \mu_y) \sin \alpha$$

$$P_2 = -(x - \mu_x) \cos \alpha + (y - \mu_y) \sin \alpha$$

E os desvios-padrões tornam-se:

$$\sigma_{1,2}^2 = \frac{\sigma_x^2 + \sigma_y^2}{2} \pm \sqrt{\left( \frac{\sigma_x^2 - \sigma_y^2}{2} \right)^2 + \sigma_{xy}^2}$$

# Correlação da distribuição bidimensional Normal

- A rotação pelo ângulo  $\alpha$  transforma uma distribuição bidimensional com correlação para sistema de coordenadas aonde a distribuição conjunta de  $x$  e  $y$  é representada por duas distribuições unilaterais (unidimensionais) sem correlação para as coordenadas ( $P_1, P_2$ )
- Os novos eixos são denominados de “Eixos Principais” (*Principal Axis*) e dá origem ao tratamento chamado de Análise de Componentes Principais na análise de Dados.

# Estimador Robusto

- Quando trabalhando com dados reais, não estamos utilizando a população, logo, trabalhamos com  $p(x, y|\bar{x}, \bar{y}, s_x, s_y, s_{xy})$
- Podemos de forma análoga obter os eixos- $P_{1,2}$ , entretanto, como depende de valores amostrais, os eixos encontrados são só uma aproximação do eixo ideal. Outro fator é que outliers podem influenciar drasticamente essa medida

$$\tan 2\alpha = 2 \frac{s_x s_y}{s_x^2 - s_y^2} r$$

$$Cov(u, w) = 0$$

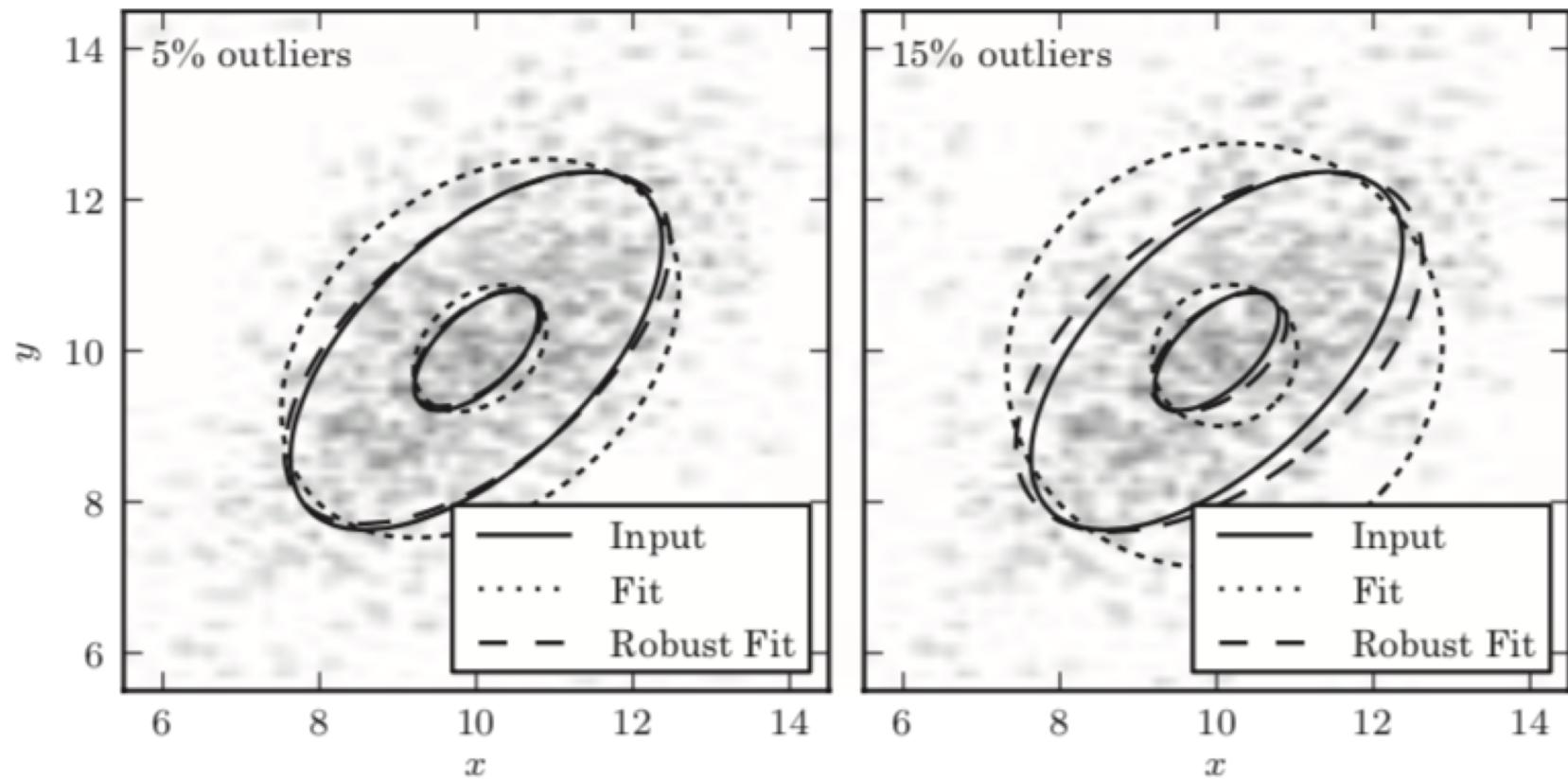
Assumindo a rotação por  $\alpha$ :  $(x, y) \rightarrow (u, w)$

$$u = \frac{\sqrt{2}}{2} \left( \frac{x}{\sigma_x} + \frac{y}{\sigma_y} \right)$$

$$\rho = \frac{V_u - V_w}{V_u + V_w}$$

$$w = \frac{\sqrt{2}}{2} \left( \frac{x}{\sigma_x} - \frac{y}{\sigma_y} \right)$$

Podemos usar  $\sigma_G$  para estimar um desvio robusto para elipsoides baseado nos percentis



# Exemplo: elipsoide de velocidades para estrelas

- Vamos assumir que obtemos dados de velocidades de estrelas para região da galáxia. Mas queremos determinar aquelas que fazem parte do Halo e as que fazem parte do Disco.
- Dois grupos estelares, podem contaminar a projeção no plano das velocidades galácticas das estrelas amostradas. Isto se deve a terem origem em grupos cinemáticos diferentes
- A melhor forma de identificar é avaliar mediana e os percentis, pois permite separar dois grupos cinemáticos

# Distribuição Normal multidimensional

- Vamos assumir mais dimensões do que duas
- A Gaussiana de  $N$ -dimensões é denominada “*Multivariate Gaussian*”
- Em vez de considerar nomes para cada variável aleatória, iremos considerar vetor  $\mathbf{x}$  de variáveis.

$$\mathbf{x} = (x_1 \ x_2 \ x_3 \ \cdots \ x_m)$$

# Reformulando nosso vetor

- Porém, cada variável dentro do vetor  $\vec{x}$  pode assumir  $N$ -valores possíveis, pois cada  $x_m$  é uma amostra
- Assim, iremos considerar índice  $i$  para amostra, e índice  $j$  para variável aleatória:

$$\mathbf{x} = (x_i^j | j = 1, \dots, m | i = 1, \dots, N)$$

# Reorganizando função normal em termos matriciais

- Então, se temos posição média geral, iremos considerar a diferença entre  $\mathbf{x} - \bar{\mathbf{x}}$ , onde  $\bar{\mathbf{x}}$  é vetor de valores médios:

$$\bar{\mathbf{x}} = (\mu_1 \ \mu_2 \ \cdots \ \mu_m)$$

$$p(\mathbf{x}|I) = \frac{1}{(2\pi)^{M/2} \sqrt{\det(\mathbf{C})}} \exp \left( -\frac{1}{2} \mathbf{x}^T \mathbf{H} \mathbf{x} \right)$$

Sendo:

$$p(\mathbf{x}|I) = \frac{1}{(2\pi)^{M/2} \sqrt{\det(\mathbf{C})}} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{H} \mathbf{x}\right)$$

$\mathbf{C}$  é a matriz de covariância

$\mathbf{H}$  é a inversa da matriz de covariância, i.e.,  $\mathbf{H} = \mathbf{C}^{-1}$

$\mathbf{x}$  é o vetor de variáveis aleatórias

$\mathbf{x}^T$  é a transposta do vetor  $\mathbf{x}$

Cada elemento da matriz de covariância é definido como:

$$\mathbf{C} = [C_{kj}] \quad |C_{kj} = \int_{-\infty}^{\infty} x^k x^j p(\mathbf{x}|I) d^M x$$

Estamos aplicando a ideia do segundo momento entre as várias variáveis aleatórias do vetor  $\mathbf{x}$

O argumento da exponencial é definido por:

$$\mathbf{H} = \mathbf{C}^{-1}$$

$$\mathbf{x}^T \mathbf{H} \mathbf{x} = \sum_{k=1}^M \sum_{j=1}^M H_{kj} x^k x^j$$

No caso bidimensional:

$$\mathbf{x}^T \mathbf{H} \mathbf{x} = -\frac{z^2}{2(1 - \rho^2)} \quad \det(\mathbf{C}) = \sigma_x^2 \sigma_y^2 - \sigma_{xy}^2 = 1 - \rho^2$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad z^2 = \frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} - 2\rho \frac{(x - \mu_x)(y - \mu_y)}{\sigma_x \sigma_y}$$

## Ganhando noção de matriz de covariância

Vamos rever o termo  $H$ , dentro da exponencial para o caso bidimensional:

$$\mathbf{x}^T \mathbf{H} \mathbf{x} = \sum_{k=1}^M \sum_{j=1}^M H_{kj} x^k x^j \quad \longrightarrow \quad \text{Covariância amostral}$$

$$H = C^{-1} \quad \mathbf{C} = Cov(X, Y) = \frac{1}{N} \sum_{i=1}^N (x - \mu_x)(y - \mu_y)$$

$$\mathbf{C}(\mathbf{x}, \mathbf{y}) = \frac{1}{N} [x_1 - \mu_1 \ x_2 - \mu_2 \ \dots \ x_N - \mu_N] \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \dots \\ y_N - \mu_N \end{bmatrix}$$

Se tivermos uma matriz com mais de 2 elementos

$x_1$	$x_2$	$x_3$	$x_4$	$x_5$
$x_{11}$	$x_{12}$		$x_{41}$	$x_{51}$
$x_{21}$		...		
		...		
		...		
				$x_{75}$

$$\mathbf{C} = \begin{bmatrix} Cov(x_1, x_1), Cov(x_1, x_2), Cov(x_1, x_3), Cov(x_1, x_4), Cov(x_1, x_5) \\ Cov(x_2, x_1), Cov(x_2, x_2), Cov(x_2, x_3), Cov(x_2, x_4), Cov(x_2, x_5) \\ Cov(x_3, x_1), Cov(x_3, x_2), Cov(x_3, x_3), Cov(x_3, x_4), Cov(x_3, x_5) \\ Cov(x_4, x_1), Cov(x_4, x_2), Cov(x_4, x_3), Cov(x_4, x_4), Cov(x_4, x_5) \\ Cov(x_5, x_1), Cov(x_5, x_2), Cov(x_5, x_3), Cov(x_5, x_4), Cov(x_5, x_5) \end{bmatrix}$$

Na situação de mesma variável:

$$\mathbf{C}(\mathbf{x}, \mathbf{x}) = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_x)^2 = \sigma_x^2$$

Temos que:

$$\mathbf{C}(\mathbf{x}, \mathbf{y}) = \sigma_{x,y}^2$$

Assim, qual propriedade da matriz de covariância ajuda em seu cálculo?

$$\mathbf{C} = \begin{bmatrix} Cov(x_1, x_1), Cov(x_1, x_2), Cov(x_1, x_3), Cov(x_1, x_4), Cov(x_1, x_5) \\ Cov(x_2, x_1), Cov(x_2, x_2), Cov(x_2, x_3), Cov(x_2, x_4), Cov(x_2, x_5) \\ Cov(x_3, x_1), Cov(x_3, x_2), Cov(x_3, x_3), Cov(x_3, x_4), Cov(x_3, x_5) \\ Cov(x_4, x_1), Cov(x_4, x_2), Cov(x_4, x_3), Cov(x_4, x_4), Cov(x_4, x_5) \\ Cov(x_5, x_1), Cov(x_5, x_2), Cov(x_5, x_3), Cov(x_5, x_4), Cov(x_5, x_5) \end{bmatrix}$$

Nossa matriz de covariância é quadrada, com lado superior espelhado com inferior! Nossa diagonal principal são as variâncias de cada coluna em nossa tabela de dados.

$$\mathbf{C} = \begin{bmatrix} \sigma_{1,1}^2, \sigma_{1,2}^2, \sigma_{1,3}^2, \sigma_{1,4}^2, \sigma_{1,5}^2 \\ \sigma_{2,1}^2, \sigma_{2,2}^2, \sigma_{2,3}^2, \sigma_{2,4}^2, \sigma_{x,5}^2 \\ \sigma_{3,1}^2, \sigma_{3,2}^2, \sigma_{3,3}^2, \sigma_{3,4}^2, \sigma_{x,5}^2 \\ \sigma_{4,1}^2, \sigma_{4,2}^2, \sigma_{4,3}^2, \sigma_{4,4}^2, \sigma_{x,5}^2 \\ \sigma_{5,1}^2, \sigma_{5,y}^2, \sigma_{5,3}^2, \sigma_{5,4}^2, \sigma_{5,5}^2 \end{bmatrix}$$

Voltando a nossa Normal bidimensional, teremos que

$$p(x|\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy}) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{z^2}{2(1-\rho^2)}\right)$$

$$z^2 = \frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} - 2\rho \frac{(x - \mu_x)(y - \mu_y)}{\sigma_x\sigma_y}$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

$$\frac{1}{2(1-\rho^2)} \times 2\rho \frac{(x - \mu_x)(y - \mu_y)}{\sigma_x\sigma_y}$$

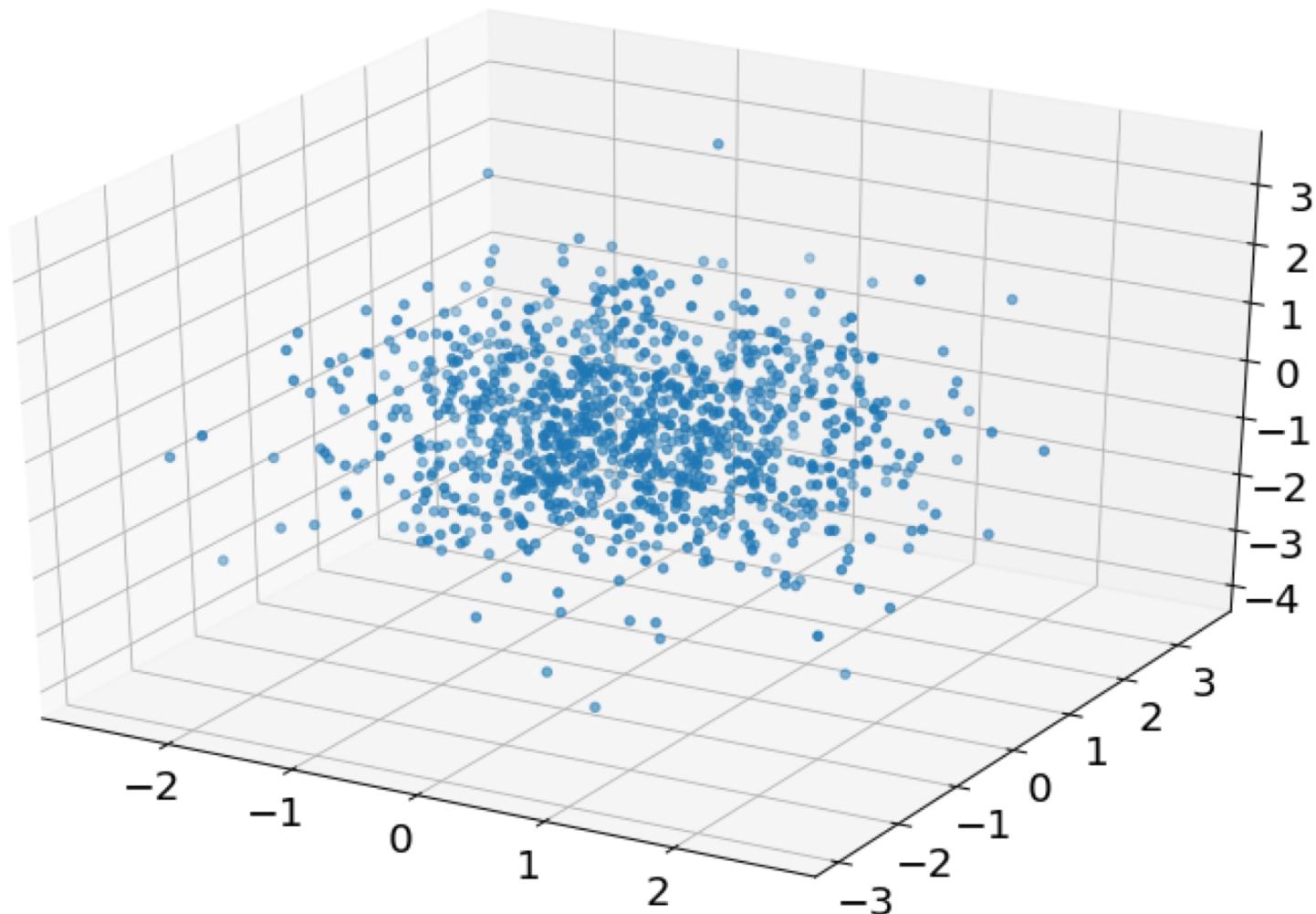
$$H(x, y) = \frac{\rho}{(1 - \rho^2)} \times \frac{1}{\sigma_x\sigma_y}$$

Como provar:  
 $H = C^{-1} = \frac{1}{\det(C)} \times \text{Cofator}(C)^T$

# Exercícios

- Prove os termos de Covariância e fator da exponencial para caso de uma Distribuição Normal de 3-dimensões. Chame cada dimensão como considere mais compreensível:
  - $x^j, j = 1, 2, 3$ ; ou
  - $\mathbf{x} = (x, y, z)$
- Em que condições podemos por  $p(x, y, z) = p(x)p(y)p(z)$ , cada uma sendo uma distribuição gaussiana? Prove formalmente (i.e., assuma condição e demonstre que nesta condição  $p(x, y, z)$  cai no caso assumido).

Exemplo de distribuição Normal em 3D para caso de independência entre x,y, e z  $\rightarrow N(x, y, z | \mu^j = 0, \sigma^j = 1, \forall j, N = 1000)$



# Coeficientes de Correlação

Pearson

Spearman

Kendall

- A ideia dos coeficientes de correlação é criar estimativa válida da interdependência entre duas amostras, já que desconhecemos a real interdependência populacional

# Coeficiente de Pearson

- O coeficiente de Pearson dá uma ideia se duas amostras  $\{x_i\}$  e  $\{y_i\}$ , de TAMANHOS IGUAIS  $N$ , são diretamente, inversamente ou não possuem correlação alguma (i.e.,  $r = 0$ )
- $-1 \leq r \leq 1$

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

# Coeficiente de Pearson

- Quando as duas amostras são independentes e vieram de distribuições Normais, os valores do coeficiente  $r$  segue uma distribuição  $t$ -Student para os diversos tamanhos amostrais considerados:

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

- É interessante perceber que conhecendo valor específico para  $r$ , podemos associar uma probabilidade de correlação, dada a distribuição  $t$  – Student. I.e., no caso de  $r > 0.72$  e  $N = 10$ , a probabilidade associada é de  $p(r > 0.72) = 0.99$ , ou seja, temos 99% de certeza que existe correlação direta entre  $x$  e  $y$ .

# Coeficiente de Pearson

- Quando as duas amostras  $\{x_i\}$  e  $\{y_i\}$  possuem algum dependência, ou melhor, quando desconhecemos formalmente a função  $p(x, y)$  da população que originou, podemos considerar a estimativa da probabilidade do coeficiente de correlação  $r$  pela distribuição de Fisher:

$$F(r) = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right)$$

No mesmo exemplo anterior, se  $N = 10$  e  $r > 0.72$ , teremos  $F(r > 0.72) = 0.8$ . Ou seja, 80% de chances de as duas amostras possuírem correlação direta.

# Coeficiente de Pearson

- Existem dois problemas com uso do coeficiente de Pearson:
  - (1) o coeficiente desconsiderar erros de  $\{x_i\}$  e  $\{y_i\}$
  - (2) Extremamente sensível a outliers. Quando ocorre isto, o uso de testes não-paramétricos para correlação é aconselhado.

# Testes de Correlação Não- paramétricos

- Teste do coeficiente de Spearman
- Teste do coeficiente de Kendall

# Coeficiente de Spearman

- Mais utilizado na literatura porque é mais simples de calcular
- Baseado no conceito de “rank”  $R_i^x$
- O coeficiente de Spearman é definido semelhante ao Pearson, porém usando os valores de rank  $R_i^x$ :

$$r_s = \frac{\sum_{i=1}^N (R_i^x - \bar{R}^x)(R_i^y - \bar{R}^y)}{\sqrt{\sum_{i=1}^N (R_i^x - \bar{R}^x)^2} \sqrt{\sum_{i=1}^N (R_i^y - \bar{R}^y)^2}}$$

# Conceito de rank

- Coloque os dados em ordem ascendente, i.e.,  $x_i < x_{i+1}$
- O índice  $i$  dos dados organizados é denominado rank  $R_i^x$ 
  - Propriedades

$$\sum_{i=1}^N R_i = \frac{N(N + 1)}{2}$$

$$\sum_{i=1}^N (R_i)^2 = \frac{N(N + 1)(2N + 1)}{6}$$

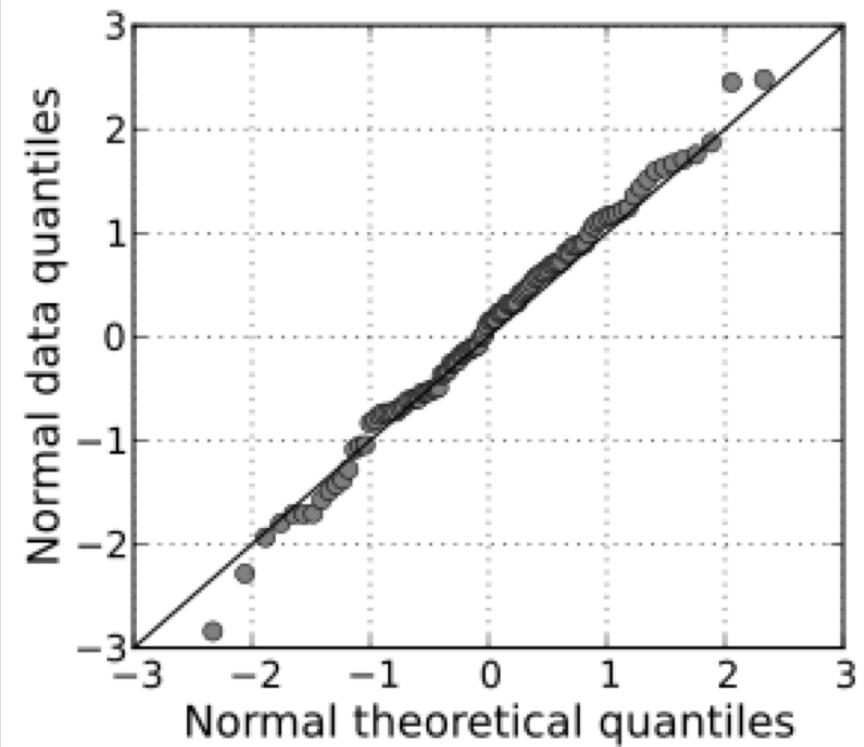
# Coeficiente de Spearman

- Uma forma alternativa de reescrever o coeficiente de Spearman é:

$$r_s = 1 - \frac{6}{N(N^2 - 1)} \sum_{i=1}^N (R_i^x - R_i^y)^2$$

- O qual pode ser derivada utilizando as duas propriedades anteriores

# Exemplo: QQ-plot



# Coeficiente de Kendall

- Kendall por sua vez baseia-se na comparação dos ranks
- Se duas distribuições não tiverem correlação, então, seus os ranks  $R_j^x = R_j^y$  e  $R_k^x = R_k^y$ , para qualquer  $j, k$  irão produzir **pares concordantes**:
  - Pares concordantes:  $(x_j - x_k)(y_j - y_k) > 0$
  - Pares discordantes:  $(x_j - x_k)(y_j - y_k) < 0$
- Desta forma, concordância refere-se que as duas diferenças tenham mesmo sinal.

# Coeficiente de Kendall

- O Desta forma, o coeficiente de Kendal é definido pela quantidade de pares concordantes,  $N_c$ , e de pares discordantes,  $N_d$ :

$$\tau = 2 \frac{N_c - N_d}{N(N - 1)}$$

- Onde  $-1 \leq \tau \leq 1$

# Coeficiente de Kendall

- O coeficiente de Kendall pode ser compreendido como a diferença de probabilidade de que os dados tenham a mesma ordem pela probabilidade de que não a tenham.
- Quando as variáveis  $\{x_i\}$  e  $\{y_i\}$  sejam independentes, então podemos aproximar  $\tau$  por uma distribuição Normal
- Quando existe dependência, as distribuições de  $\tau$  e  $r_s$  tornam-se difíceis de estimar, recorrendo a casos particulares.

Como trabalhar os coeficientes de Kendall e Spearson?

# Referências

- Livros-textos
- <http://mathworld.wolfram.com/BivariateNormalDistribution.html>