



UNIVERSIDADE FEDERAL  
DO RIO DE JANEIRO



# Astroestatística

Aula 003 -- Prof. Walter Martins-Filho

# Estatística Descritiva

- Distribuições Estatísticas Paramétricas
  - Unidimensionais
  - Multidimensionais

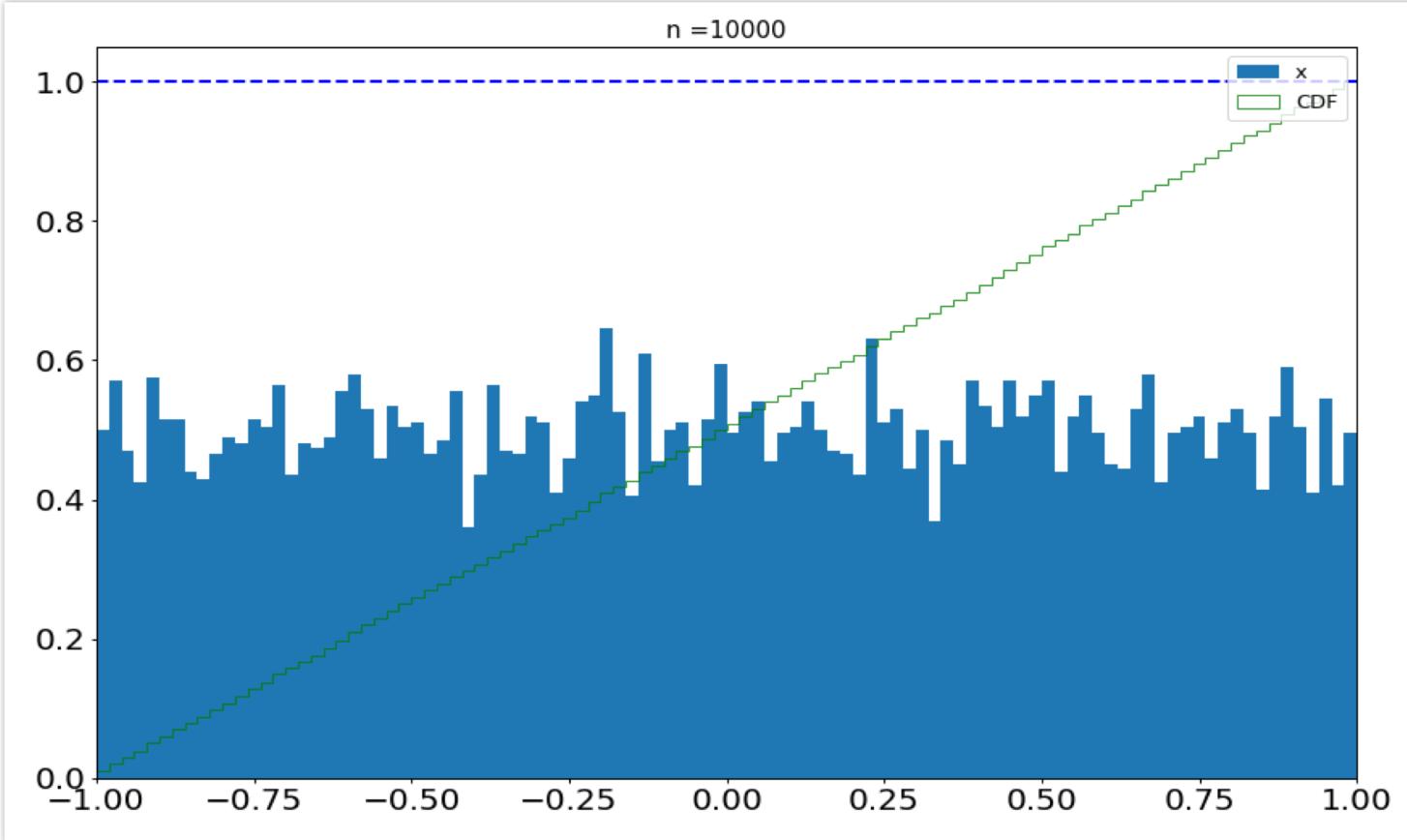
# Distribuições (úteis?)

- Por mais que qualquer função possa servir como base para definir distribuição estatística, o conjunto de distribuições que são usadas comumente são poucas

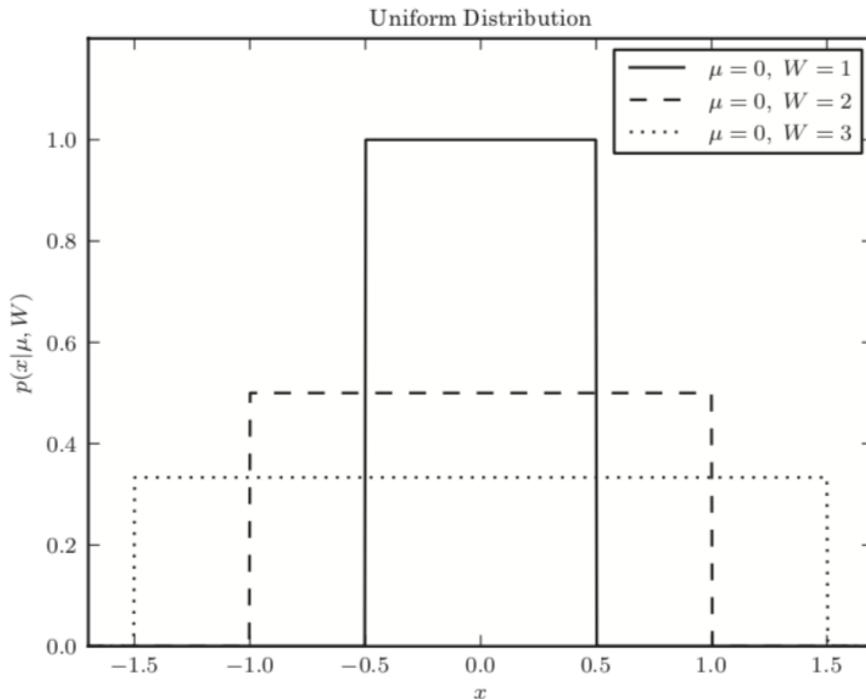
# Distribuição Uniforme

- O Distribuição em que qualquer valor pertencente a  $\{x_i\}$  possui a mesma probabilidade de ser sorteado.

$$p(x|\mu, W) = \frac{1}{W} \text{ se } |x - \mu| \leq \frac{W}{2}$$



Na prática, temos intervalo com valores dentro de uma caixa de largura  $W$ , que tendem a ter a mesma probabilidade conforme  $N \rightarrow \infty$



$$\sigma = \frac{W}{\sqrt{12}} \sim 0.3W$$

# Distribuição Uniforme

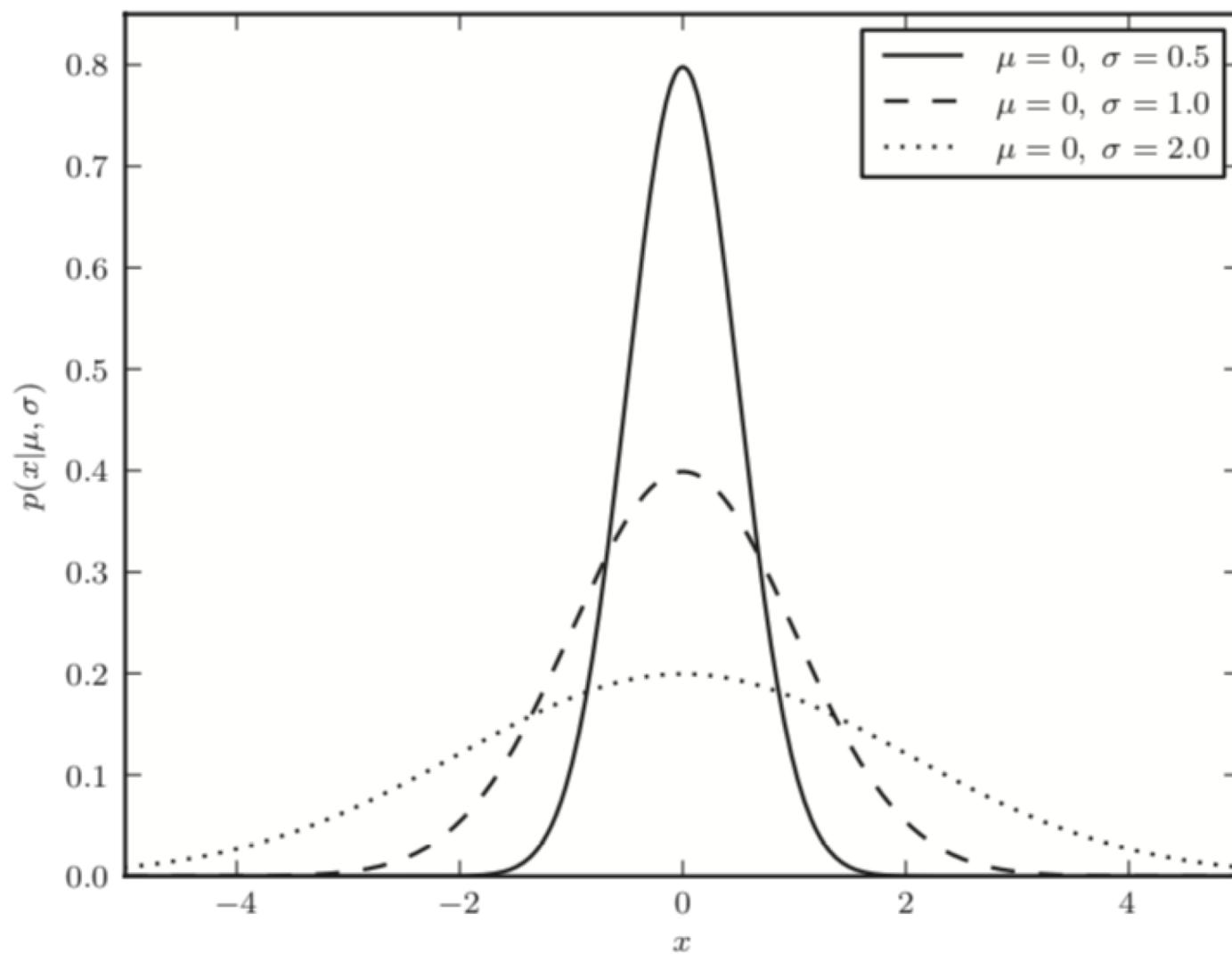
O desvio-padrão é quase 1/3 da largura da caixa.

# Distribuição Normal

- Distribuição definida a partir da função Gaussiana

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Gaussian Distribution



# Distribuição Normal

- A distribuição normal possui algumas propriedades interessantes:
  - A convolução entre duas Gaussianas também é uma Gaussiana, logo, a distribuição final segue uma distribuição Normal
  - A média amostral e a variância amostral são independentes, o que permite estimá-los juntos de forma robusta.

# Problema: os erros não seguem uma distribuição Normal

- O teorema do Limite Central nos diz que a média dos resultados médios amostrais segue uma distribuição Normal.
- Problema ocorre que isso se baseia que a variância dos conjuntos amostrais  $k$ , onde  $\{x_i\}_k$ , sejam iguais. O que não necessariamente é verdadeiro!
- Ajuste linear, baseia-se da premissa do Teorema do Limite Central
- O Ajuste de Mínimos Quadrados torna-se inválido se a variância amostral dos conjuntos amostral não é idêntica.

# Distribuição Normal

- Outra característica da distribuição Normal é denominada “Information Content”
- Voltaremos neste assunto quando tratarmos na Parte 2 – Inferência Bayesiana.
- A distribuição cumulativa da distribuição Normal não é passível de ser calculada por funções elementais ...

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x' - \mu)^2}{2\sigma^2}\right) dx'$$

Definindo a função Erro Gaussiano:  $erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$

$$P(x|\mu, \sigma) = \frac{1}{2} \left( 1 \pm erf\left(\frac{|x - \mu|}{\sigma\sqrt{2}}\right) \right) \text{ se } sign(x) > 0 | x > \mu$$

# Distribuição Normal

CDF da distribuição Normal

# Exercício

- Mostre o histograma de fluxo, com a curva de Distribuição Normal e a curva da Distribuição Cumulativa no mesmo gráfico.

# Distribuição Normal

- Vamos assumir que queremos calcular a PDF dentro de determinado intervalo  $[a, b]$
- A PDF pode ser obtida por:

$$\int_a^b p(x|\mu, \sigma)dx = \int_a^\mu p(x|\mu, \sigma)dx + \int_\mu^b p(x|\mu, \sigma)dx$$
$$a = \mu - M\sigma \quad b = \mu + M\sigma \quad \left. \begin{array}{l} \text{Intervalo acaba tendo } erf(M/\sqrt{2}) \\ M = 1, 2, 3, \dots \rightarrow 0.682, 0.954, 0.997 \end{array} \right\}$$

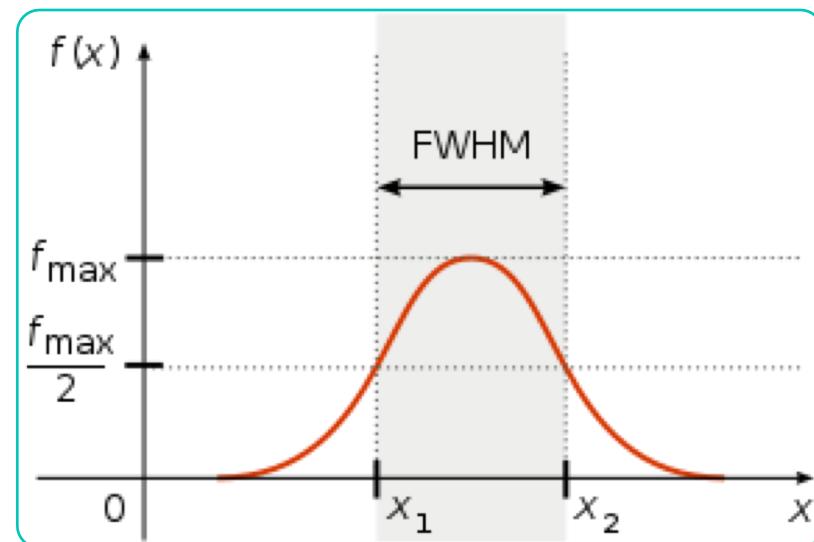
# Distribuição Normal

- A ideia anterior carrega a ideia de Intervalo de Confiança
- Se aplicamos no Intervalo Interquartil, teremos que:

$$p_{75} - p_{25} = \sigma 2\sqrt{2} \operatorname{erf}^{-1}(0.5) \approx 1.349\sigma$$

# FWHM

- FWHM, Full width at half maximum, em português, largura à meia altura, é nome dado a expressão/função que fornece a diferença entre extremos do intervalo o qual a distribuição de probabilidade decai pela metade do valor máximo:

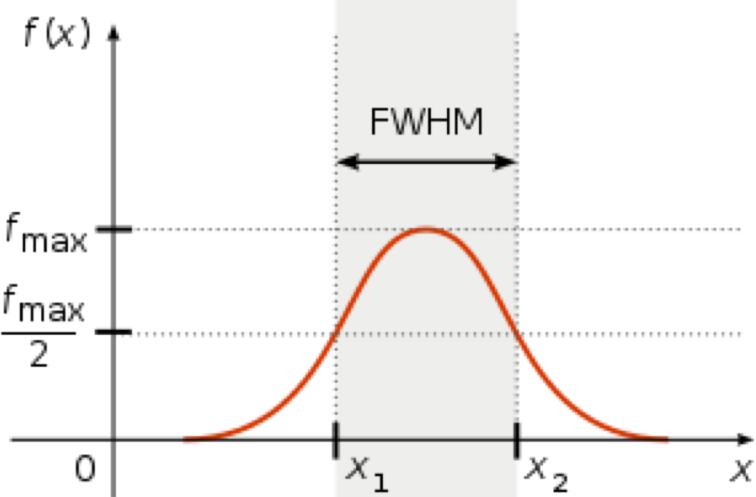


$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x' - \mu)^2}{2\sigma^2}\right) dx'$$

$$FWHM \equiv x_2 - x_1$$

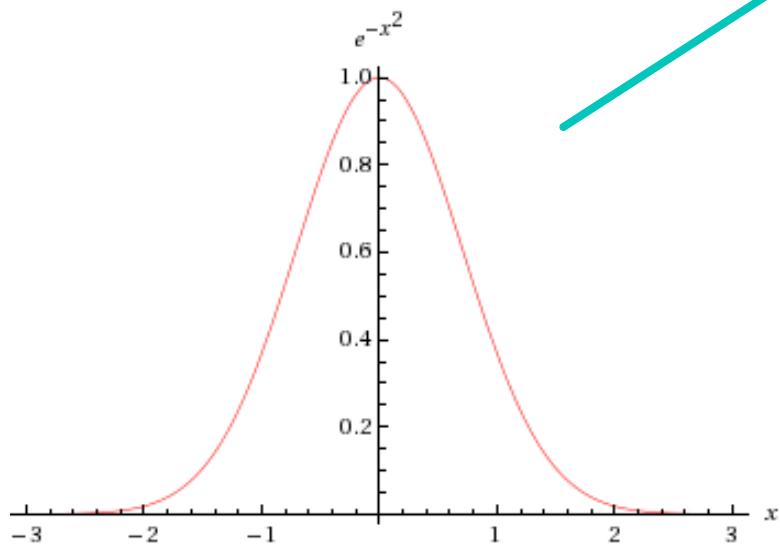
Por simetria, em distribuição padronizada,  $x_1 = -x_2$ . Vamos considerar formalismo geral. Temos que  $x_0$  ocorre quando  $p(x)|_{max/2}$ :

$$e^{-(x_0 - \mu)^2 / (2\sigma^2)} = \frac{1}{2} p(x|\mu, \sigma) \Big|_{max}$$



Para simplificar, só usamos a parte exponencial em distribuição padronizada, i.e.,  $p(\mu) = 1$

$$e^{-(x_0 - \mu)^2 / (2\sigma^2)} = \frac{1}{2} p(\mu) = \frac{1}{2}$$



Reorganizando os termos:

$$(x_0 - \mu)^2 = 2\sigma^2 \ln 2$$

$$x_0 = \mu \pm \sigma \sqrt{2 \ln 2}$$

$$\rightarrow M = \sqrt{2 \ln 2}$$

# FWHM

- Logo, assumindo  $M = \sqrt{2 \ln 2}$ , teremos dentro do intervalo a distribuição aonde os extremos correspondem a metade do valor máximo.
- Desta forma, o FWHM é definido como:

$$FWHM \equiv x_2 - x_1 = 2\sigma\sqrt{2 \ln 2} \approx 2.3548\sigma$$

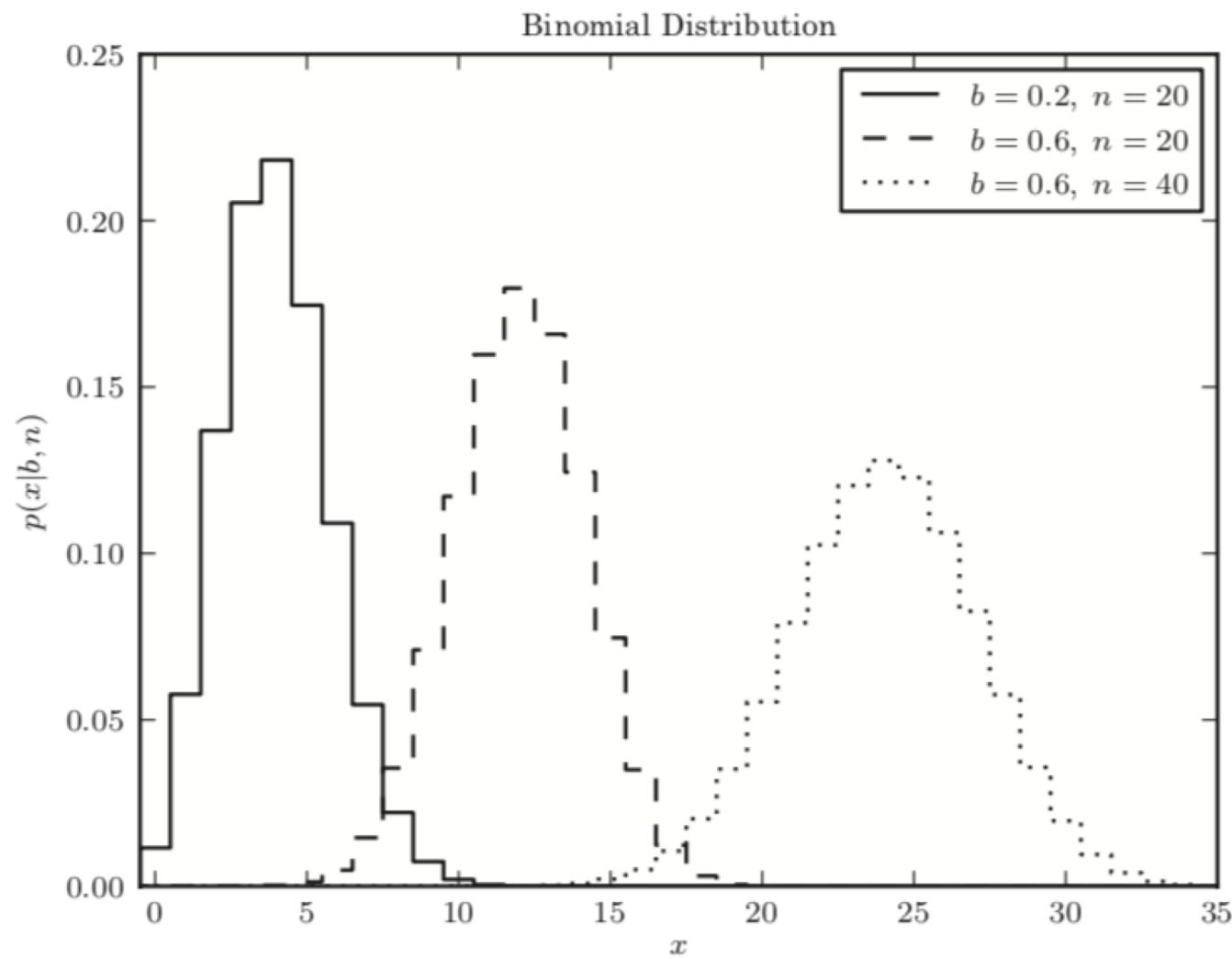
# *The born of log-normal*

- Quando faz-se uma transformação por uma função exponencial ou potência, a consequência é uma distribuição log-normal
- A distribuição log-normal surge quando o resultado final é produto de vários valores positivos, e o Teorema do Limite Central aplica-se somente para a soma dos log
- Por mais que os momentos também variem, o uso de percentis se mantém o mesmo: a mediana da primeira distribuição é a mediana da distribuição seguinte.

# Distribuição Binomial

- Também chamada de distribuição de Bernoulli
- A distribuição binomial é a distribuição que descreve uma variável que pode obter somente dois valores.
- Assumimos que a probabilidade de sucesso é  $p = b$ , sendo assim, a probabilidade da outra condição ocorrer, pelo axioma de Kolmogorov é  $p = 1 - b$
- A distribuição é dada por:

$$p(k|b, N) = \frac{N!}{k!(N-k)!} b^k (1-b)^{N-k}$$



# Aglomerados de Galáxias

- Temos amostra de 100 aglomerados de galáxias (novo survey)
- Por um estudo anterior, a quantidade de aglomerados com galáxia central dominante é de 10%.
- Selecionamos aleatoriamente 30 destas para estudo de raio—  
X
- Quantos destes aglomerados esperamos que tenham uma galáxia central dominante?

# Analizando

- Temos caso em que existe duas chances: ter ou não ter centro dominante identificado por raio—X
- Um estudo anterior diz que da população que se conhece, 10% deles devem ter uma galáxia central dominante
- Probabilidade:
  - Sucesso (ter centro dominante)  $\rightarrow p=0.10$
  - Pelo axioma de Kolmogorov  $\rightarrow 1-p = 0.9$

Isto são características de uma distribuição binomial.

# Aglomerados de Galáxias

- Assim, a probabilidade de se encontrar  $n$  aglomerados com galáxia central dominante numa amostra de 30 aglomerados, baseados em estudos anteriores, será de:

$$p(n) = \binom{30}{n} 0.1^n 0.9^{30-n}$$

# Aglomerados de Galáxias

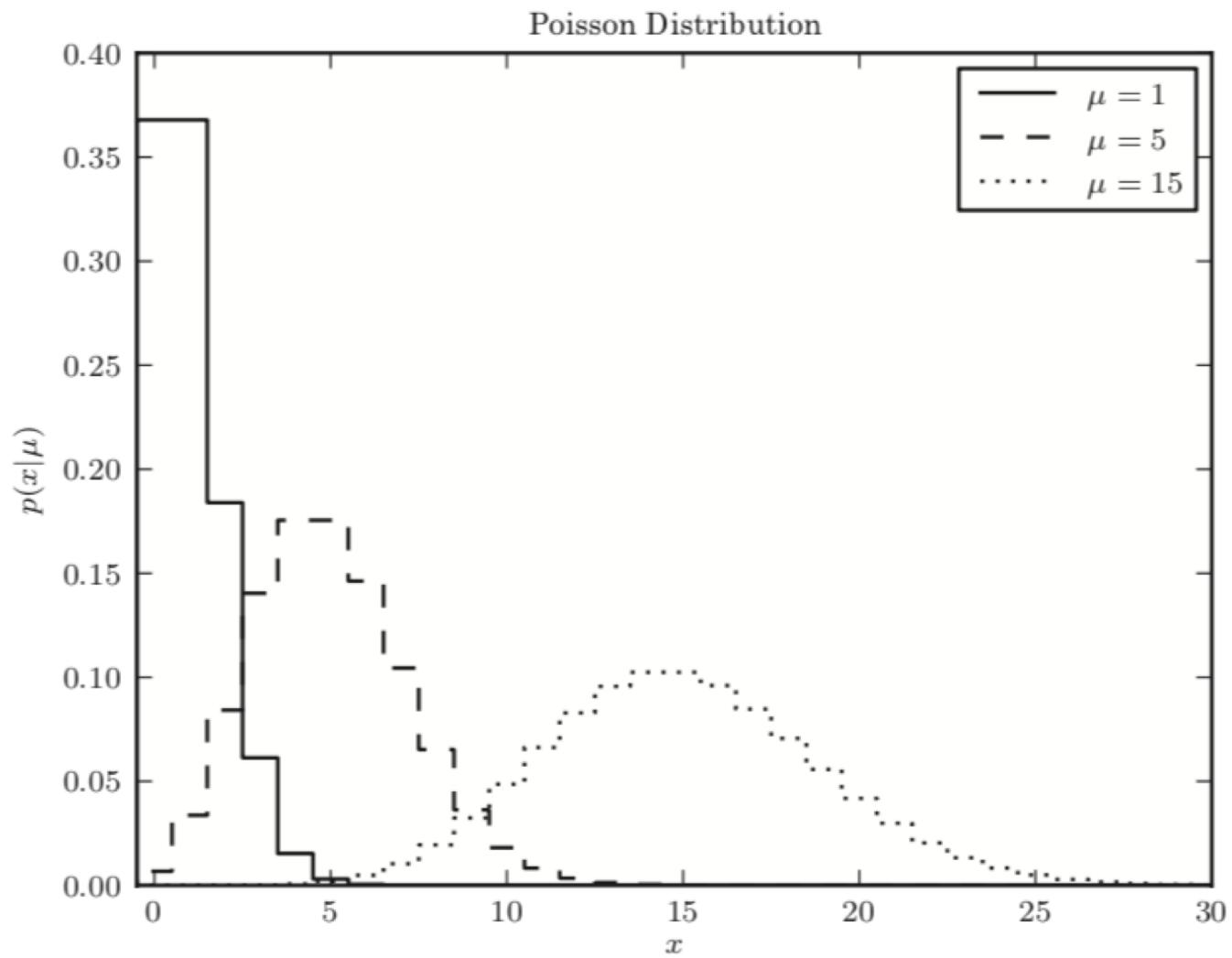
- Analisamos aleatoriamente 30 objetos da amostra inicial.
- Porém, encontramos que 10 destes possuem região central ativa, i.e., 33%
- Pela probabilidade assumida anteriormente  $p(n = 10) = 0.01$ 
  - Logo, se encontramos 10 objetos, as chances disso ocorrer são 1%.

O que você conclui com isto?

# Distribuição de Poisson

- A distribuição de Poisson é caso especial da distribuição binomial.
- No caso de  $N \rightarrow \infty$ , então as chances de sucesso, ou probabilidade de sucesso, tornam-se  $p = k/N$ , e se mantém fixada pela quantidade de sucessos  $k$  (ou TAXA de SUCESSOS).
- Assim, o número de sucessos  $k$  gera média  $\mu = pN$

$$p(k|\mu) = \frac{1}{k!} \mu^k e^{-\mu}$$



# Distribuição de Poisson

- A distribuição de Poisson é importante em astronomia pois descreve a distribuição de fótons contados dentro de um intervalo.

# Fótons no CCD

- A contagem de fótons no CCD DURANTE UMA INTEGRAÇÃO numérica, tem as seguintes características:
  - (1) Fótons são detectados ou não detectados
  - (2) A detecção ocorre durante intervalo de tempo  $t$
  - (3) A chegada, e por conseguinte a detecção ou não, para cada fóton é majoritariamente independente. (casos especiais acontecem devido fótons seguirem distribuição de Bose-Einstein, mas que é irrelevante ao CCD em sua escala de tempo)
  - Fótons detectados (TAXA de SUCESSO) chegam com taxa temporal  $\lambda$

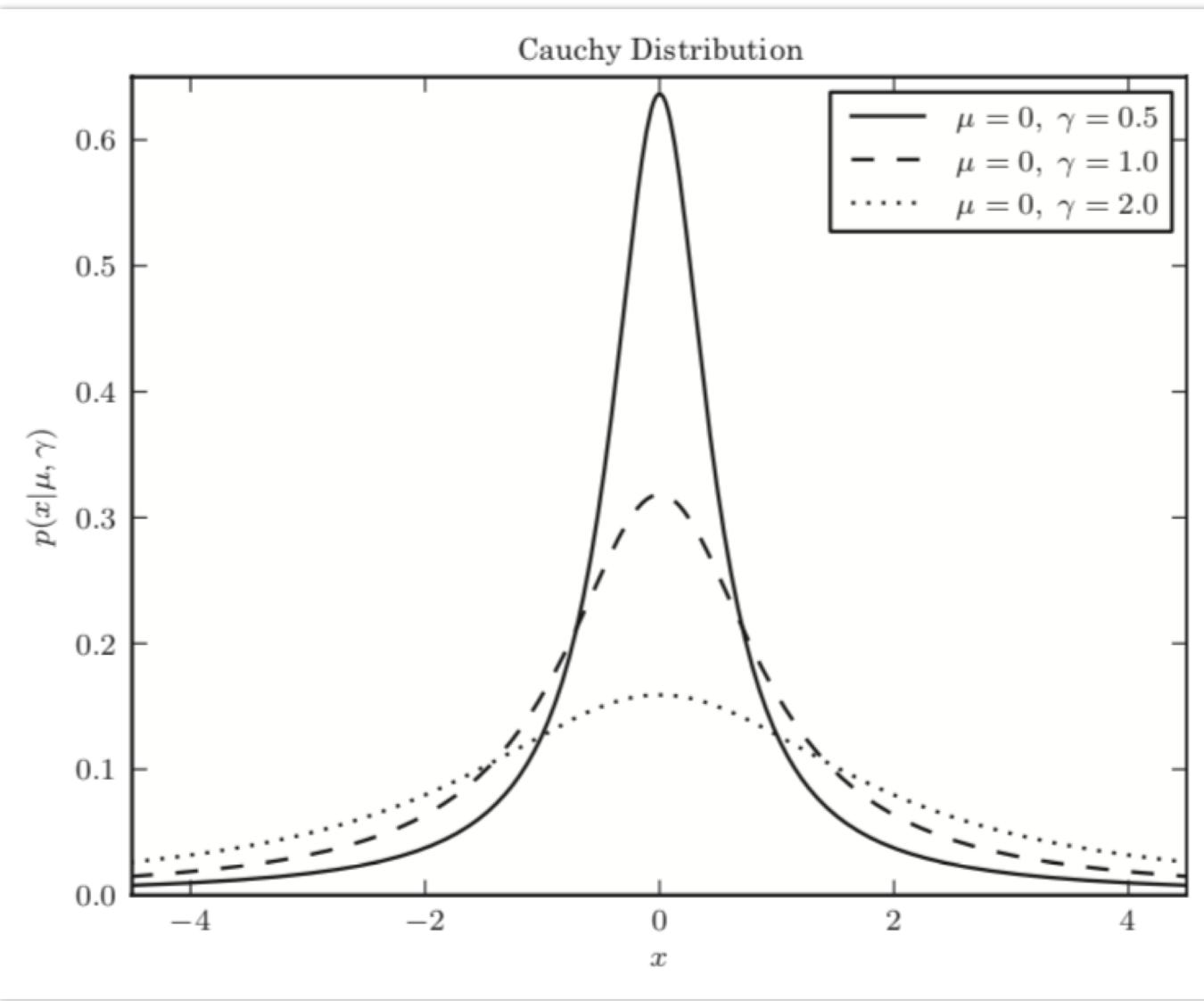
# Fótons no CCD

- O valor médio de fótons que chegam é dado por  $\mu = \lambda t$
- A flutuação segue então  $\sigma = \sqrt{\mu}$
- Todas as Características anteriores permite identificar que, em placa CCD, a distribuição de chegada de fótons segue distribuição de Poisson durante a integração da observação.

# Distribuição de Cauchy

- Também chamada de distribuição Lorentziana
- Definida pela média  $\mu$  e pelo fator de escala  $\gamma$
- Parecida com uma Gaussiana, mas suas bordas decaem para  $|x|$  grandes, seguindo uma lei potencial de  $x^{-2}$

$$p(x|\mu, \gamma) = \frac{1}{\pi\gamma} \left( \frac{\gamma^2}{\gamma^2 + (x - \mu)^2} \right)$$

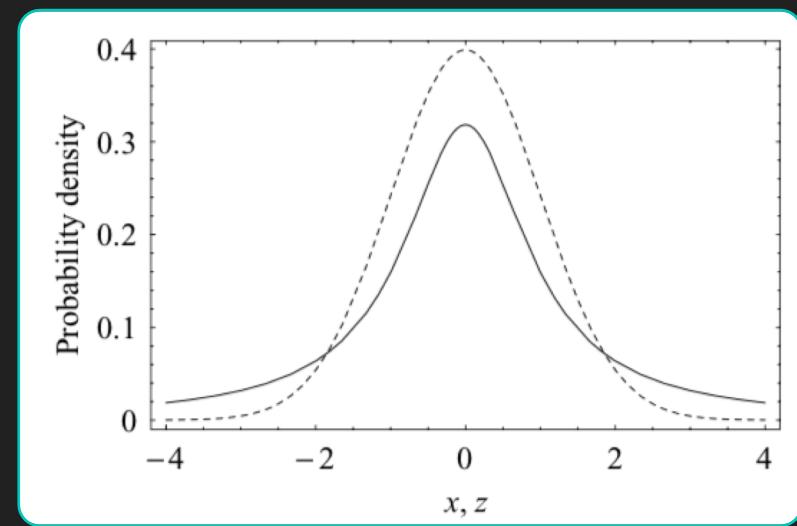


# Distribuição de Cauchy

- A distribuição de Cauchy é usada para caracterizar também a PSF de objetos astronômicos ou linhas de absorção e emissão.
- Contudo, por experiência, ela é melhor aplicada quando queremos caracterizar psf nas colunas (eixo-y) quando se tem o espectro astronômico espalhado no eixo-x em vez de aplicação direta para ajuste PSF fotométrico.
- **Característica útil:** A razão de duas variáveis aleatórias  $\{x, y\}$ , segue uma distribuição de Cauchy, quando  $p(x)$  e  $p(y)$  estão padronizadas

# Quebrando paradigma na fotometria: Baixo S/N

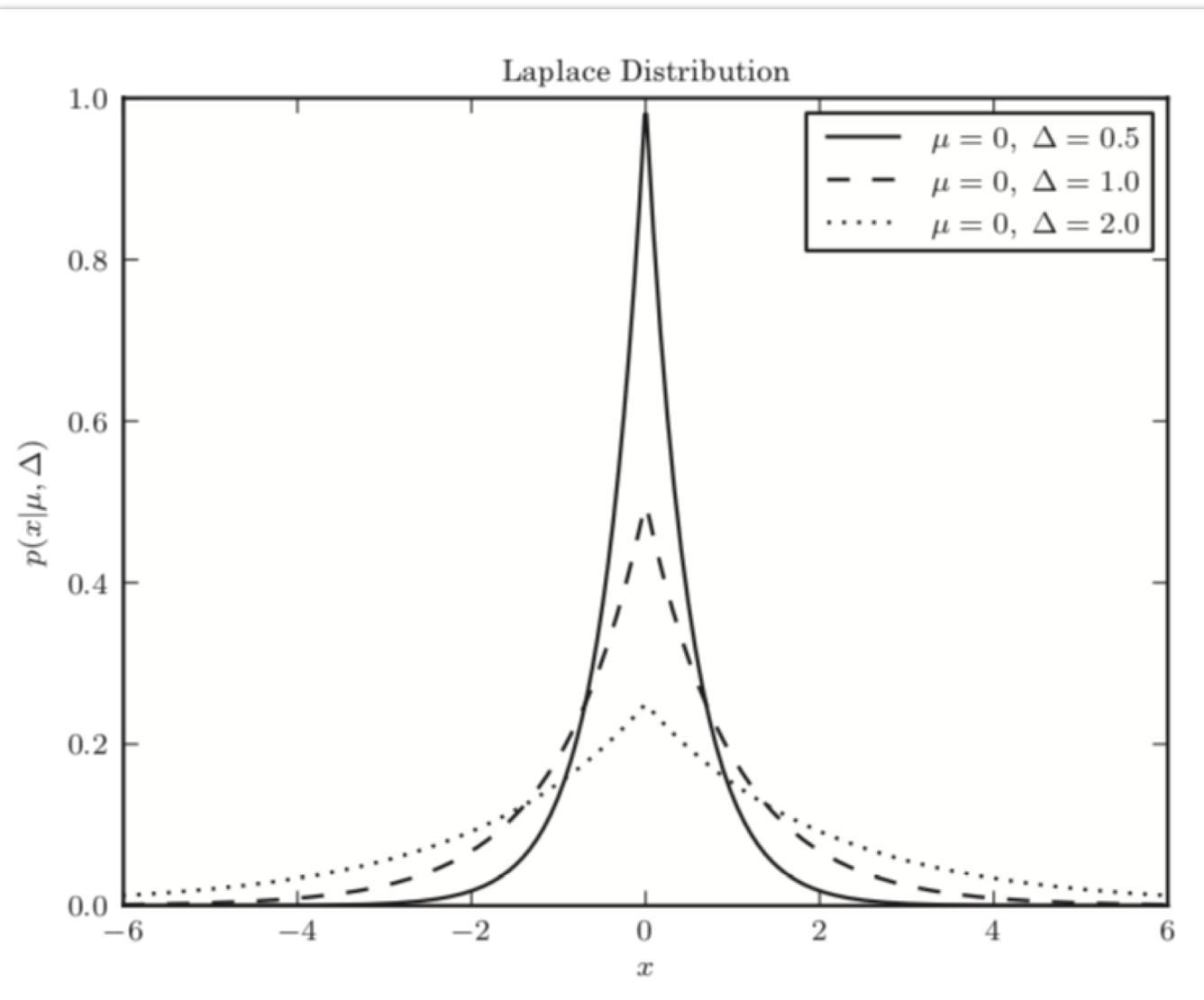
- S/N → duas variáveis que consideraremos aleatórias  $\{S_i\}$  e  $\{N_i\}$
- O caso de se ter outliers em distribuição normal causa mudanças nas regiões das “asas” das gaussianas
- Esse é maior motivo se ter uma distribuição de Cauchy explicando PSF aonde tenha-se baixo S/N
- O que também explica porque acabo experimentando uma distribuição de Cauchy em espectroscopia de exoplanetas (low S/N)



# Distribuição de Laplace

- A distribuição de Laplace também é chamada de distribuição exponencial
- Essa distribuição normalmente é definida para  $x > 0$ , entretanto, o formalismo permite aplicar para  $x \in \mathbb{R}$
- No caso geral, ela é denominada de “exponencial dobrada” (em inglês, *double exponential*)

$$p(x|\mu, \Delta) = \frac{1}{2\Delta} e^{-\frac{|x-\mu|}{\Delta}}$$



# Fótons no CCD (parte 2)

- Diferente de contabilizar os fótons durante um intervalo de tempo  $t$  (i.e., integrada no tempo), vamos considerar os fótons que chegam do mesmo objeto em instantes sucessivos
- A distribuição de Laplace descreve eventos físicos que acontecem em instantes consequentes no tempo, o qual possuem taxa constante,  $\eta$ , e independentes (cada objeto astrotípico emite independentemente de outro objeto astrotípico em seu FoV)
- O número de eventos (fótons chegando no detector) do objeto astrotípico segue então distribuição de Laplace

# Fótons no CCD (parte 2)

- Desta forma, os fótons de um objeto astrofísico, que chegam no detector em instantes sucessivos, durante intervalo de tempo  $T$ , são descritos como:

$$p(t|\eta) = \frac{1}{\eta} e^{-\frac{t}{\eta}} \quad \mu = \frac{T}{\eta}$$

# Distribuição $\chi^2$

- Motivação: Se obtemos  $\{x_i\}$  de uma distribuição Normal, e definirmos  $z$  como nova variável usando a padronização da distribuição, teremos que soma dos quadrados de  $z_i$  seguirá uma distribuição  $\chi^2$
- A distribuição  $\chi^2$  é interessante, como veremos durante a parte de inferência, pois permite avaliar modelos de comportamento para dados.

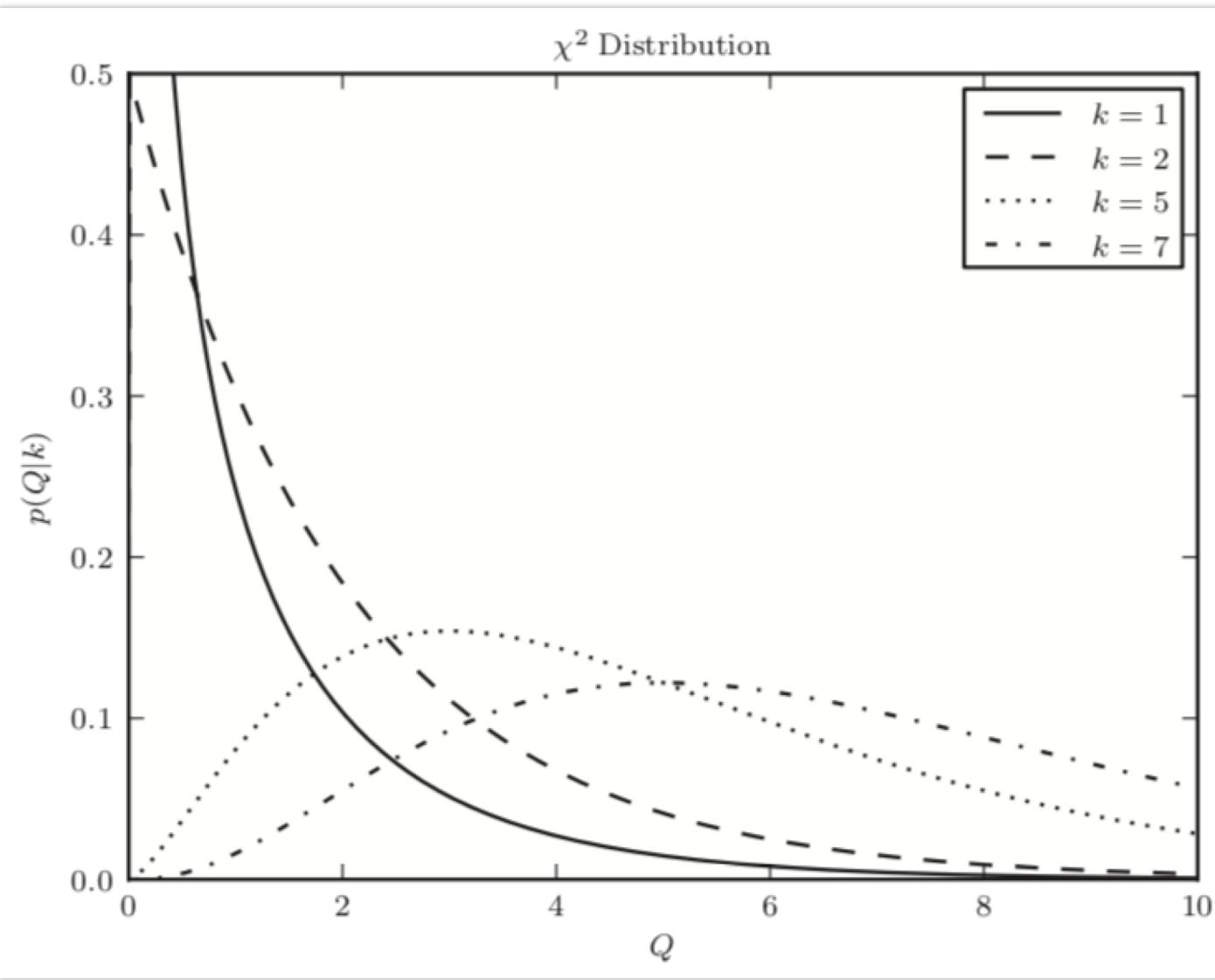
# Distribuição $\chi^2$

- $N$ -pontos coletados/observados
- $k = N$  graus de liberdade

$$p(Q|k) \equiv \chi^2(Q|k) = \frac{1}{2^{k/2}\Gamma(k/2)} Q^{\frac{k}{2}-1} e^{-Q/2} \text{ se } Q > 0$$

Onde

$$z_i = \frac{x_i - \mu}{\sigma} \quad Q = \sum_{i=0}^k z_i^2 \quad \Gamma(\xi) \equiv (\xi - 1)! \text{ se } \xi > 0$$



# Distribuição $\chi^2_{dof}$

- $\chi^2$  per degree of freedom
- Podemos também definir a distribuição  $\chi^2$  normalizada pelos graus de liberdade  $k$

$$\chi^2_{dof}(Q|k) \equiv \chi^2\left(\frac{Q}{k} \middle| k\right)$$

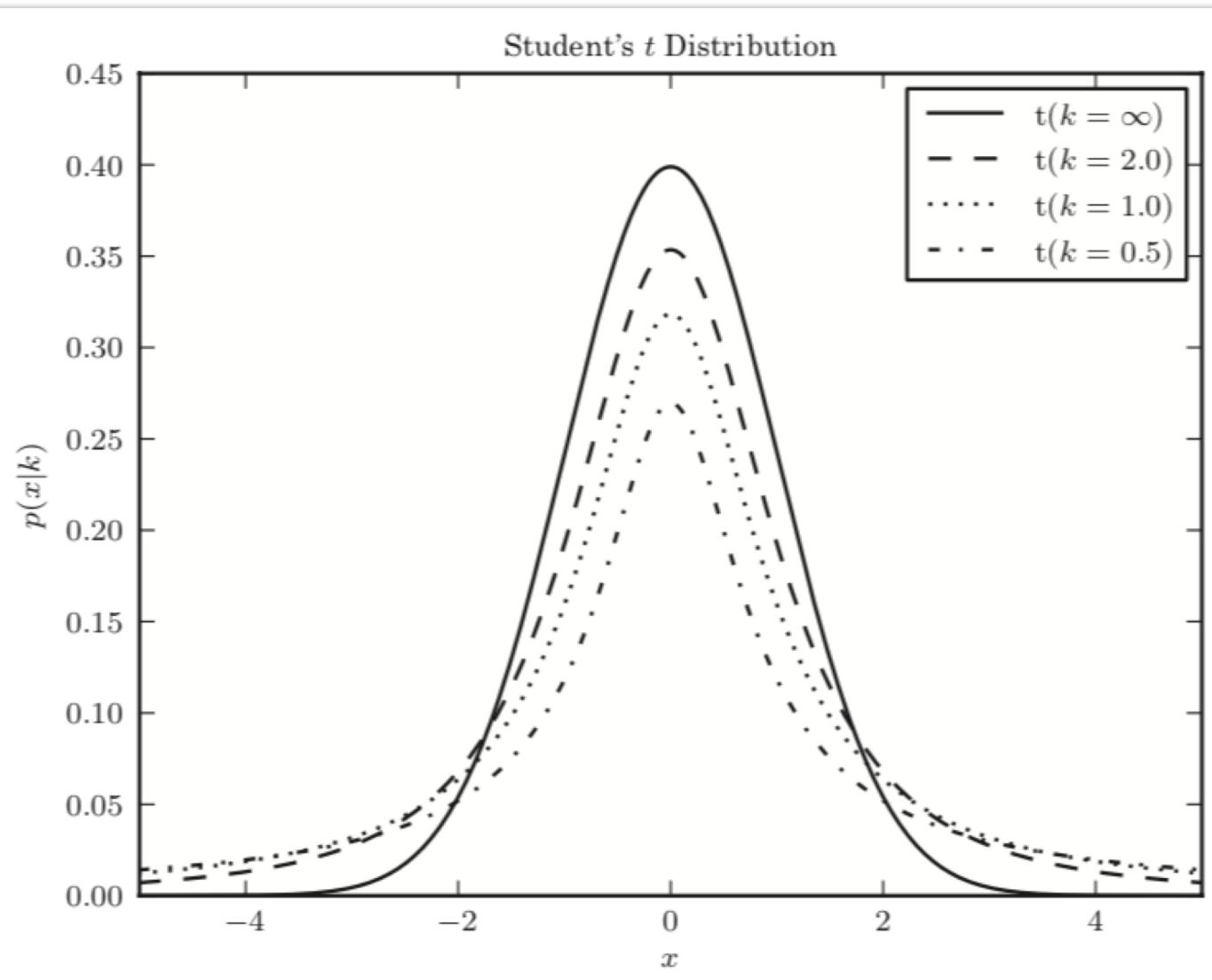
# Distribuição $\chi^2_{dof}$

- O valor médio da  $\chi^2_{dof}$  é 1
- A soma  $Q$  é muito suscetível a outliers e em geral deve ser lidado este problema antes de se calcular  $Q$

# Distribuição *t*-Student

- Distribuição publicada por William Gosset (1908), e desenvolvida na cervejaria Guinness Breweries (aberta desde 1759)

$$p(x|k) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k} \Gamma(\frac{k}{2})} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$



# Distribuição $F$ de Fisher

- A distribuição  $F$  de Fisher é útil pois descreve a razão entre duas  $\chi^2_{dof}$  variáveis independentes com graus de liberdade  $d_1$  e  $d_2$ , respectivamente
- Assim, se  $x_1$  e  $x_2$  provém da mesma distribuição populacional, a proporção entre ambas é uma distribuição  $F$  de Fisher

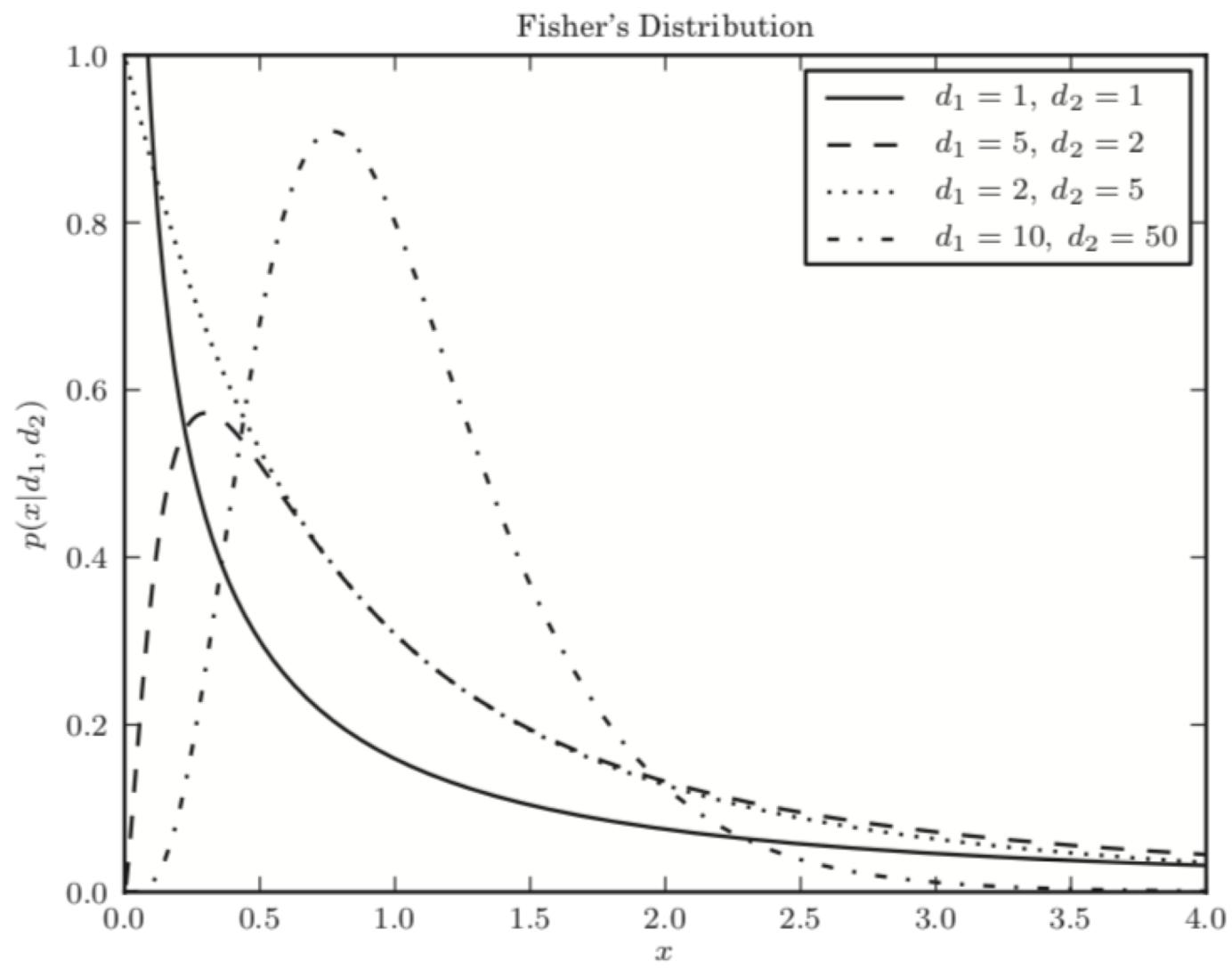
# Distribuição $F$ de Fisher

- A distribuição SOMENTE está definida para  $0 \leq x$ ,  $d_1 > 0$  e  $d_2 > 0$
- $C$  é a constante de normalização
- $B(x|\xi_1, \xi_2)$  é a Função Beta

$$p(x|d_1, d_2) = C \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}} x^{\frac{d_1}{2}-1}$$

$$C = \frac{1}{B(d_1/2, d_2/2)} \left(\frac{d_1}{d_2}\right)^{d_1/2}$$

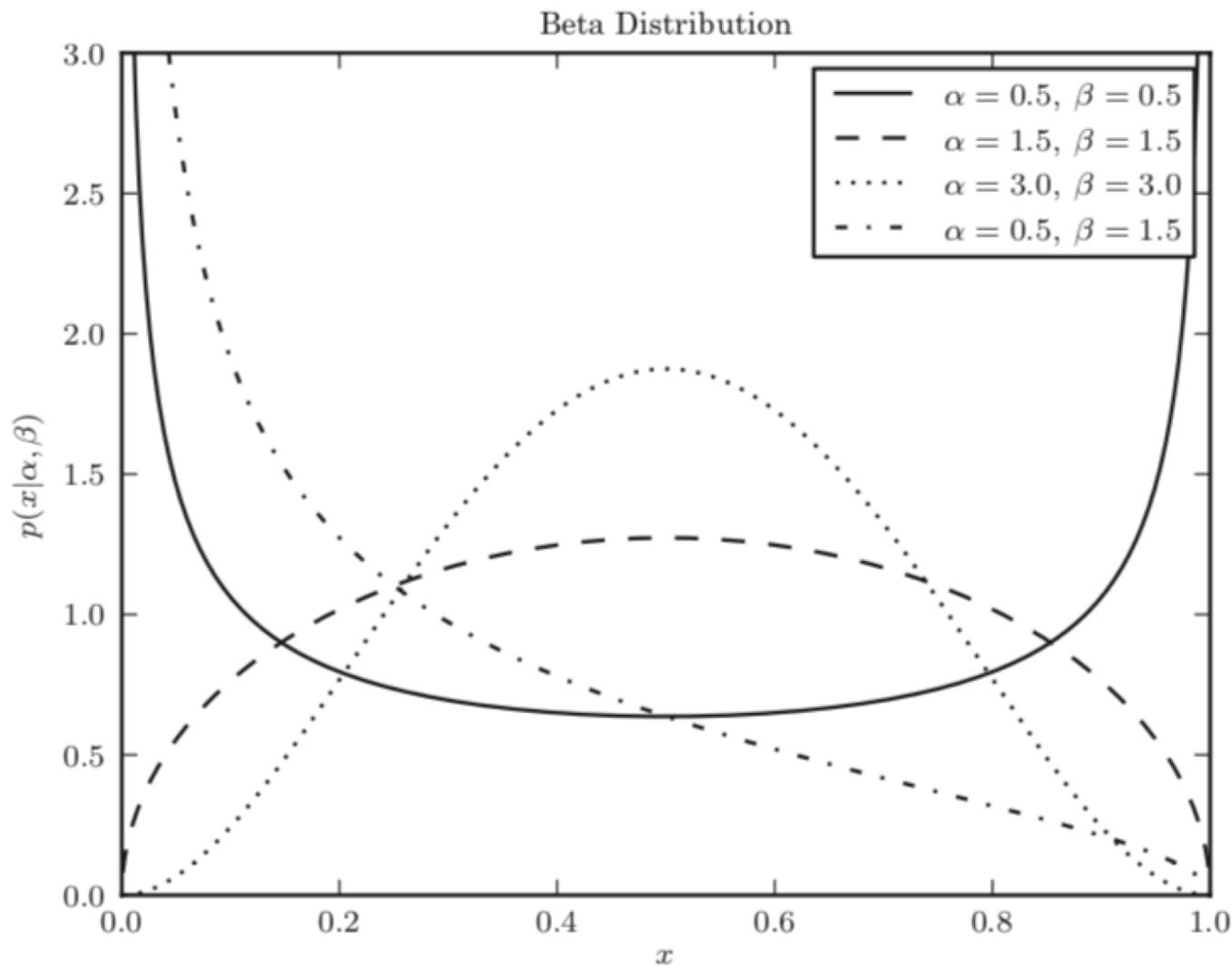
$$B(x|\xi_1, \xi_2) = \int_0^1 x^{\xi_1-1} (1-x)^{\xi_2-1} dx$$



# Distribuição Beta

- A distribuição Beta é definida entre  $0 \leq x \leq 1$ , aonde  $\alpha > 0$  e  $\beta > 0$
- O fato de os termos  $(\alpha, \beta)$  possam definir diferentes formatos para distribuição permite que seja útil para Análises de Inferência Bayesiana

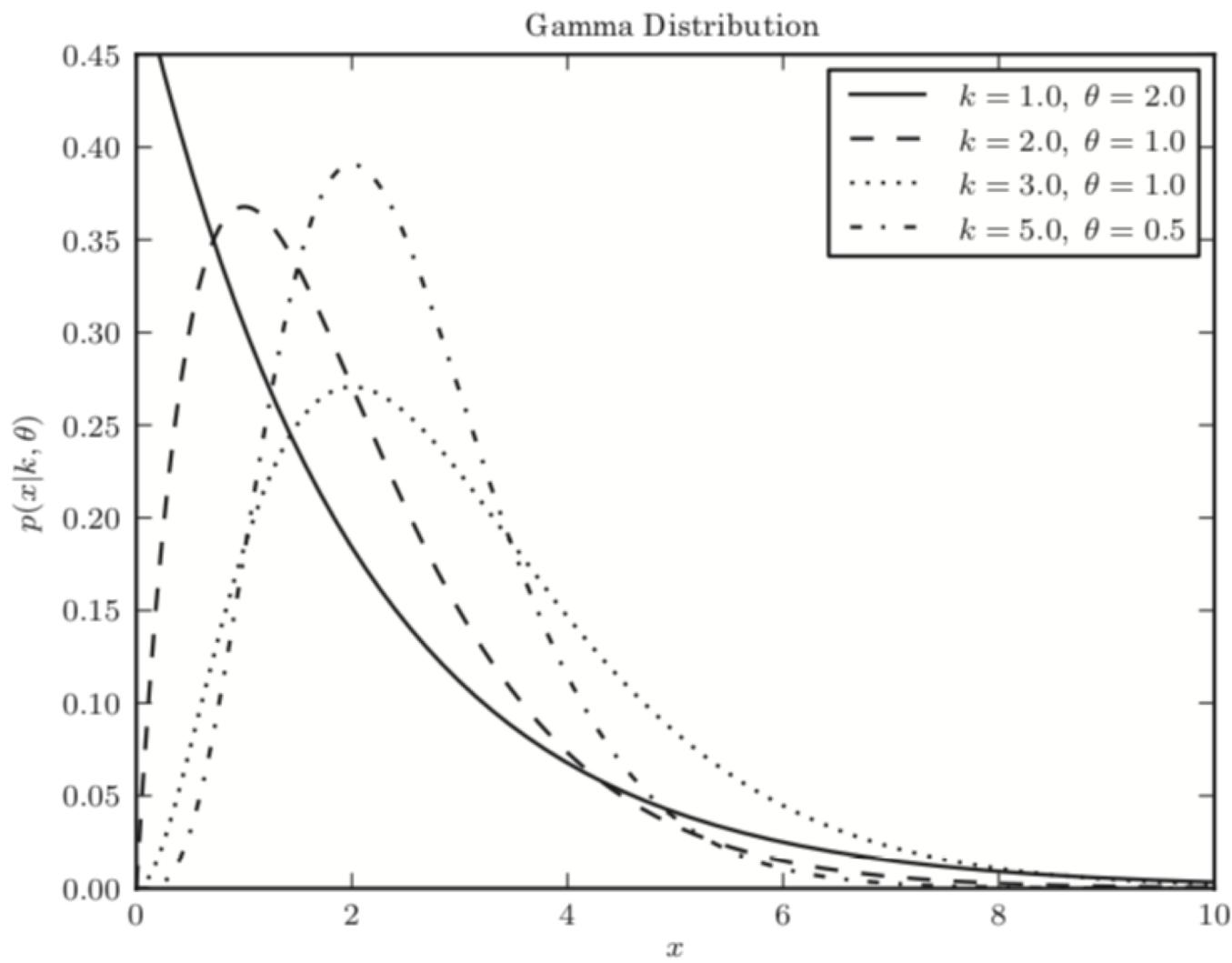
$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$



# Distribuição Gamma

- A distribuição Gamma é útil pois ela permite caracterizar funções *Prior* em Análises Bayesianas.
- A distribuição é definida em  $0 < x < \infty$
- 2 parâmetros:
  - (1) parâmetro de forma  $k$
  - (2) parâmetro de escala  $\theta$

$$p(x|k, \theta) = \frac{1}{\theta^k} \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)}$$

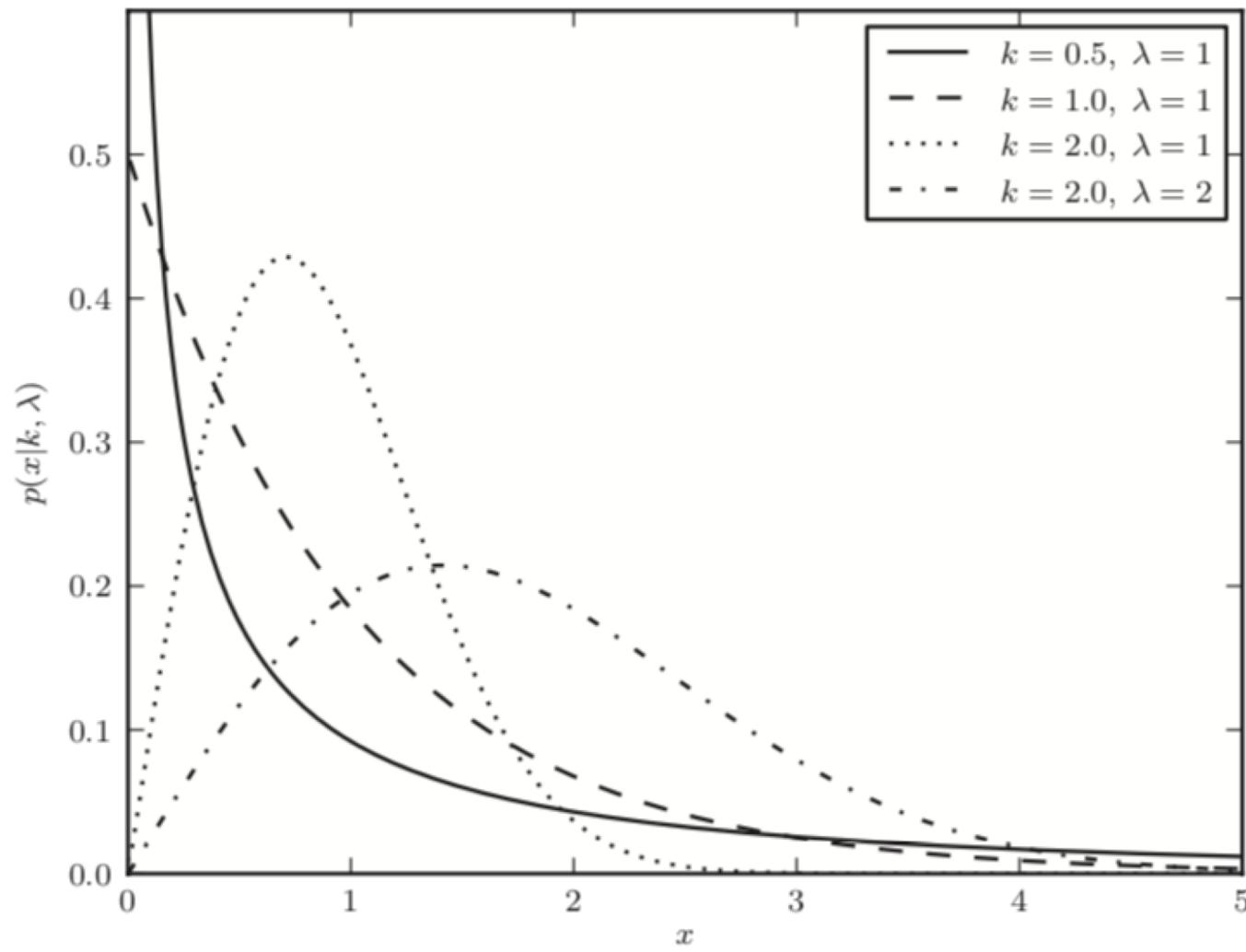


# Distribuição Weibull

- A distribuição de Weibull foi desenvolvida para descrever eventos físicos em que (1) taxa variável de processos que gerem falhas aleatórias (**quebras de corpos físicos, como asteroides colidindo**), (2) distribuição de valores extremos, (3) **distribuição do tamanho de partículas**
- Definida para  $0 \leq x$
- 2 parâmetros:
  - (1) parâmetro de forma  $k$
  - (2) parâmetro de escala  $\lambda$

$$p(x|k, \lambda) = \frac{k}{\lambda} \left( \frac{x}{\lambda} \right)^{k-1} e^{-(x/\lambda)^k}$$

### Weibull Distribution



## Exercício:

- Usando as simulações de distribuições normais, apresente os intervalos interquartis e FWHM de cada um.
- Escreva todos os momentos (de ordem Zero a Quarta) das distribuições unidimensionais (exceto da distribuição de Cauchy). Pode deixar na forma de integral.
- Calcule o intervalo interquartil da distribuição de Laplace

# Referências

- Livros-Textos
- <http://mathworld.wolfram.com/GaussianFunction.html>
- <http://mathworld.wolfram.com/FullWidthatHalfMaximum.html>
- <http://mathworld.wolfram.com/Studentst-Distribution.html>