



UNIVERSIDADE FEDERAL  
DO RIO DE JANEIRO



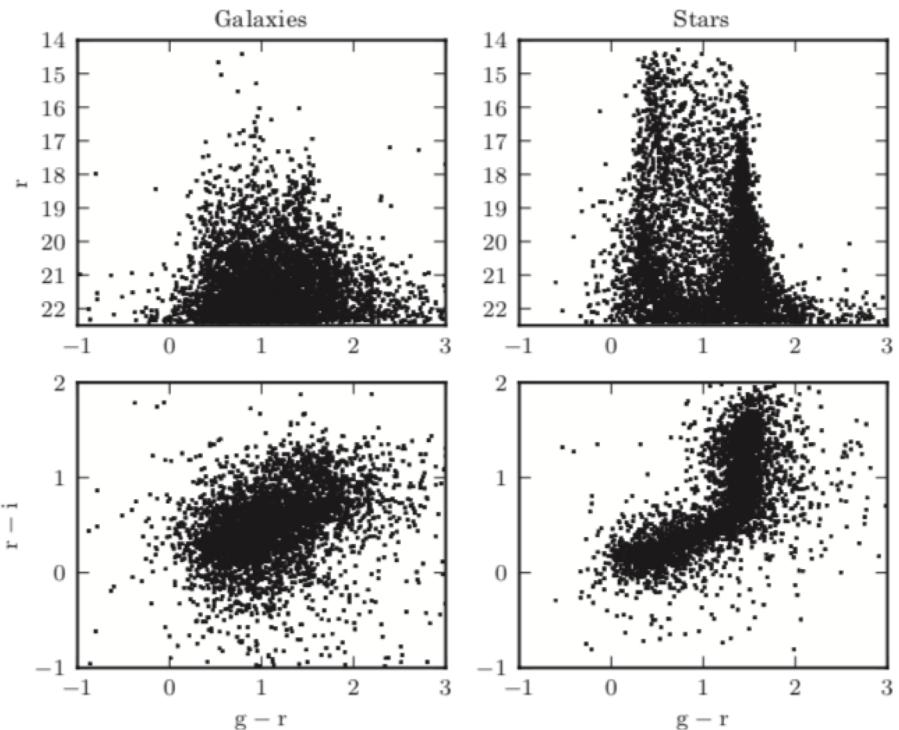
# Astroestatística

Aula 001 -- Prof. Walter Martins-Filho

# Introdução

O que é  
Astroestatística?

Dados SDSS entre -1 e 1 deg de declinação



# Página da Turma

- Repositório no GitHub
- Datas Importantes
- Sistema de Avaliação
- Links

## AstroestatisticaOVL362

Disciplina de Astroestatística lecionada no Observatório do Valongo/UFRJ em 2018B

[View On GitHub](#)



**Astroestatística – OVL362**

Disciplina de Astroestatística lecionada no Observatório do Valongo/UFRJ em 2018B

---

Esta disciplina visa cobrir a primeira parte da ênfase/especialização em Astronomia Computacional do atual currículo do curso de graduação em Astronomia da Universidade Federal do Rio de Janeiro. Alunos que tenham pré-requisito também podem cursar como disciplina eletiva.

A amenta do curso, disponibilizada pelo sistema da INTRANET-UFRJ abrange conteúdos adicionais, muitos que são cobridos também em outras disciplinas da mesma especialização.

A amenta original, pelo SICA/UFRJ é:

"Testes de hipótese paramétricos e não-paramétricos. Teste da Qui-quadrada e de Kolmogorov-Smirnov. ANOVA. Regressão linear. Método de máxima verossimilhança. Séries temporais astronômicas. Análise e decomposição de misturas multivariáveis. PCA. Estimação de densidades. Métodos de reamostragem (bootstrapping). Função de correlação. Wavelets. Análise Bayesiana. Introdução a redes neurais."

Dado isto, a amenta de conteúdo é:

1. Parte 1:

<https://waltersmartinsf.github.io/AstroestatisticaOVL362/>

# Conteúdo da disciplina

- Parte 1:
  - Análise Exploratória de Dados
  - Inferência Clássica ou Fquentista (ou Frequencista)
- Parte 2:
  - Inferência Bayesiana
  - Regressão

# Sistema de Avaliação

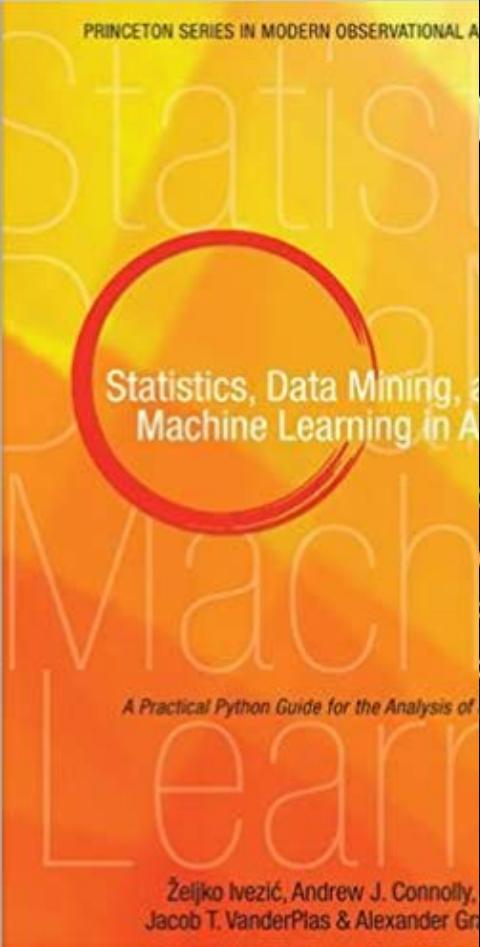
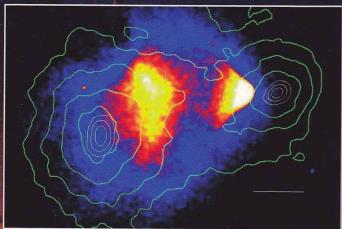
- 2 provas
  - P1: conteúdo da parte 1
  - P2: conteúdo da parte 2
- Prova Final
- Segunda-chamada
- Datas encontram-se no site da turma

$$NF = \frac{P1 + P2}{2}$$

$$MF = \frac{NF + PF}{2}$$

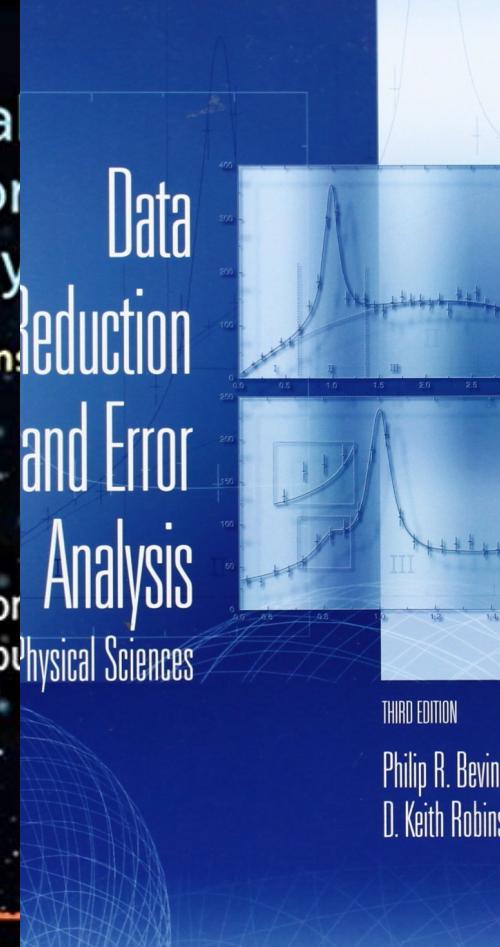
## Statistical Methods for Astronomers SECOND EDITION

J. V. Wall and C. R. Jenkins



## Modern Statistical Methods for Astronomy With R Applications

Eric D. Feigelson  
G. Jogesh Babu



# Livros-Textos

# Linguagens de Programação

- Python
  - Python é linguagem de programação geral.
  - Atualmente, o uso de pacotes como PANDAS permite trabalhar com grande volumes de dados e forma fácil
  - Grande uso dos astrônomos como todo.
- R
  - Linguagem e/ou Programa com objetivo principal de tratar estatisticamente projetos científicos e empresariais
  - Existe comunidade astronômica, focada em estatística, usa comumente este programa.

# *Data Mining vs Machine Learning*

- Data Mining
  - Técnicas para analisar e descrever dados estruturados
- Machine Learning
  - Técnicas que permitem interpretar dados baseados em modelos de comportamento (agrupamento, classes, etc)

# Astroestatística: data mining e machine learning aplicada à astronomia

- Nesta disciplina: veremos em grande parte data mining. Ou seja, iremos analisar dados astronômicos baseados na estrutura de catálogos e com isto, interpretá-los.

# Exemplo: espaço de fase raio-temperaturaestelar

- “As medidas de tamanho e temperaturas de grupo de estrelas apresentam uma sequência bem definida no diagrama tamanho-temperatura. Algumas estrelas encontram-se fora dessa região.”

Qual ponto de vista por data mining?  
Qual ponto de vista por machine learning?

# Exemplo: espaço de fase raio-temperaturaestelar

Visão Dados estruturados

- Baseado na estrutura dos dados obtidos, podemos QUANTITATIVAMENTE descrever o que faz parte desta sequência e o que não faz parte.

# Exemplo: espaço de fase raio-temperaturaestelar

## Visão Modelo de comportamento

- Um modelo de estrutura estelar pode predizer qual é a temperatura e tamanho de estrelas presentes na Sequência Principal. Com isto, podemos rejeitar objetos que não correspondem a este modelo.

# Análise Exploratória de Dados

- Abordagem de apresentar o conjunto de dados estruturados afim de obter conhecimento sobre os mesmo e interpretá-los da forma mais fácil possível.

# Probabilidade

$$\{x_i\}, i = 1,..N$$

- $x$ : medida escalar
- O experimento: obter  $N$ -vezes a medida escalar
- $\{x_i\}$ : conjunto amostral
- $x$  pode ser número real, possuir valor discreto, ser tag de classe ou mesmo, podemos não obter valores, NaN

# Distribuição de $x$

- Como estimar a distribuição de  $x$ , que denominaremos  $h(x)$ , onde os valores de  $\{x_i\}$  são obtidos?
- $h(x)$  distribuição populacional
- $h(x)dx$  probabilidade entre  $x$  e  $x + dx$
- $h(x)dx$  é denominada Função de Densidade de Probabilidade (pdf)

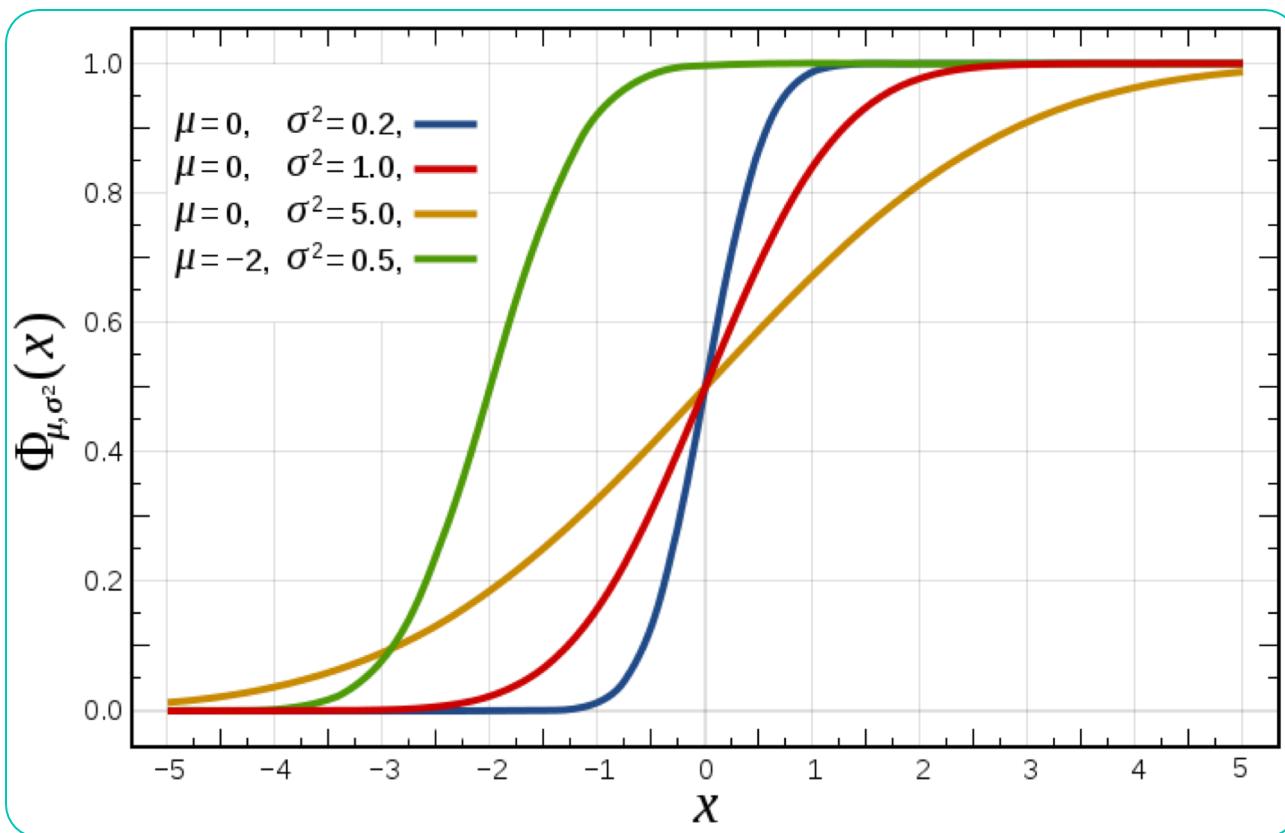
# PDF em termos astronômicos

- A distribuição de densidade de probabilidade é denominada nos artigos de astronomia como:
  - “*differential distribution function*”
  - “*probability function*”

# Função de Distribuição Cumulativa

- A integral de  $h(x)$  é denominada de Função de Distribuição Cumulativa e representa toda a probabilidade acumulada até certo valor  $x$ :

$$H(x) = \int_{-\infty}^x h(x')dx'$$



Exemplo CDF

○ Distribuição Normal  $N(\mu, \sigma^2)$

# População vs. Amostra

- Nem sempre conhecemos toda a população, mesmo que soubermos quais são todos os possíveis resultados para o valor medido  $x$ , não temos a ideia da frequência de cada resultado
- Amostra: os resultados que conseguimos obter e organizar de forma estruturada de  $x$
- População: todos os resultados existentes de  $x$

# População vs. Amostra

	População	Amostra
PDF	$h(x)$	$f(x)$
CDF	$H(x)$	$F(x)$

Qual tamanho necessário da amostra para que ela represente a população?

# Tamanho da amostra e erro amostral

- Tamanho amostral é geralmente determinado por 2 métodos:
  - (1) Por estimativa da média populacional
  - (2) Por estimativa da proporção populacional
- São problemas tratados na inferência clássica.

# Problema de $f(x)$

- Quando os erros de  $f(x)$  são muito grandes (quão grandes?),  $f(x)$  não tenderá a  $h(x)$ , mesmo quando  $N \rightarrow \infty$ .
- $f(x)$  sempre será modelo de  $h(x)$

Como estimar  $h(x)$ ?

# Paramétrico vs. Não-Paramétrico

- Metodologia Paramétrica
  - Conhecemos uma função analítica  $f(x)$  que descreve a probabilidade de  $\{x_i\}$
- Metodologia Não-Paramétrica
  - Desconhecemos uma função analítica, mas podemos ajustar comportamento afim de descrever  $\{x_i\}$ . Comportamento este que pode ser a combinação de várias funções de probabilidade analíticas conhecidas.

# Axiomas de Kolmogorov

- Probabilidade de qualquer evento é sempre maior ou igual a zero
- A probabilidade do conjunto de todos os possíveis resultados de  $x$  é igual a 1.
- Se cada resultado encontrado em  $\{x_i\}$  não possuir dependência entre si, i.e., sejam disjuntos, então a probabilidade devida a união dos resultados (também chamados de eventos) em  $\{x_i\}$  é a soma de cada probabilidade individual de  $x_i$ ,  $i = 1, \dots, N$

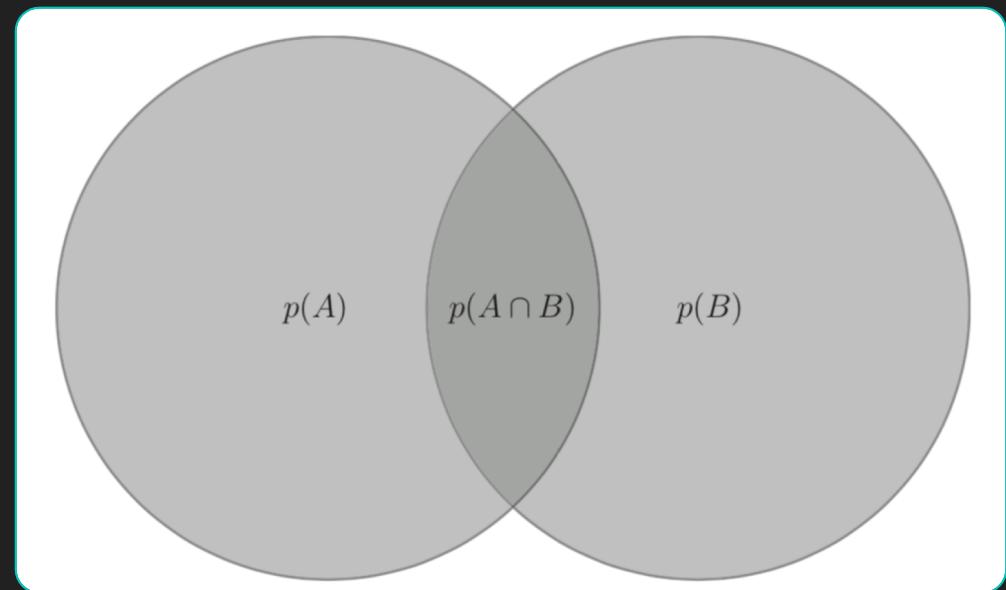
$$0 \leq p(x)$$

$$p(\Omega) = 1$$

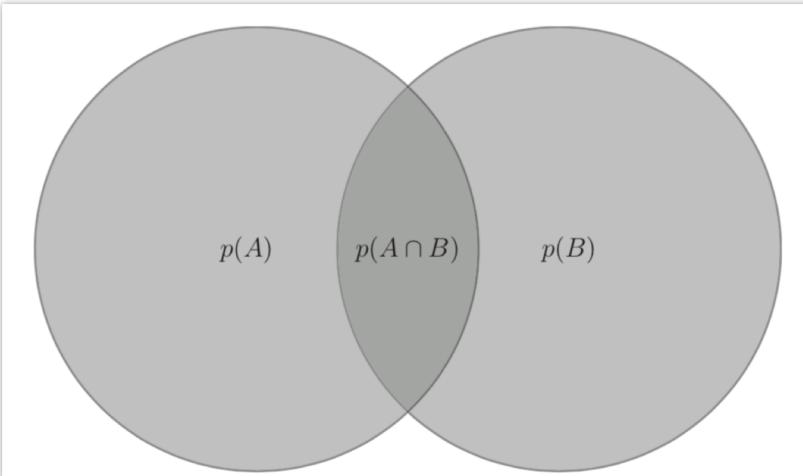
$$p(U_{i=1}^{\infty} x_i) = \sum_{i=1}^{\infty} p(x_i)$$

# União e Diagrama de Venn

- Vamos considerer que  $\{x_i\} = A \cup B$
- Assim a probabilidade da união é dada por:



$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

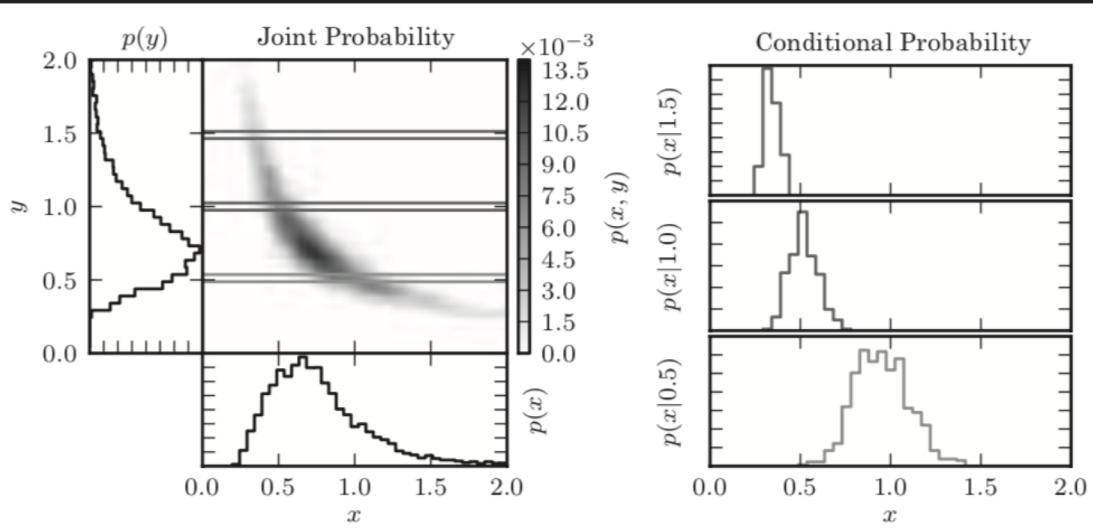


$$\frac{p(A \cap B)}{p(B)} = p(A|B)$$

$$p(A \cap B) = p(A|B)p(B) = p(B|A)p(A)$$

## A probabilidade de intersecção

Vamos pensar em termos da proporção de  $p(A \cap B)$  por  $p(B)$



$$p(x, y) = p(x)p(y) ?$$

# “Joint Probability”

- “random variables” → valores medidas que estão sujeitos a pequenas variações aleatórias
- Variáveis aleatórias: discretas e contínuas
- Duas variáveis aleatórias  $x$  e  $y$  são independentes, se e somente se, a probabilidade conjunta delas é o produto das probabilidades

# Teorema de Bayes

- Vamos assumir que  $x$  e  $y$  não são independentes entre si.

$$p(x|y) = p(x \cap y)p(y) = p(y \cap x)p(x) = p(x, y)$$

Como seria o diagrama de Venn para esta situação?



“joint probability” quando existe dependência

# Teorema de Bayes

- Assim, se conhecermos a proporção de  $x$  com relação ao conjunto universo  $\Omega$ , e  $x$  seja dependente de  $y$  em alguma parcela/proporção, então podemos estimar  $y$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

# Teorema de Bayes

- Dado que conhecemos  $p(y|x)$  e  $p(x)$  podemos expressar a probabilidade de  $y$  de forma contínua (lei da probabilidade total) como:

$$p(y) = \int p(y|x)p(x)dx$$

# Teorema de Bayes

- Combinando  $p(x|y)$  com a lei da probabilidade total, temos que:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}$$

Na situação de caso discreto, a integral passa a ser somatório para valores em  $\{x_i\}$

# Transformação de Variáveis Aleatórias

- “Qualquer função de uma variável aleatória é também uma variável aleatória”
- $x \rightarrow y(x) \mid y = \Phi(x) \rightarrow x = \Phi^{-1}(y)$
- Assim, podemos derivar a probabilidade  $p(y)$  conhecendo  $p(x)$  e como a função inversa mapeia  $x$  em termos de  $y$ . Em outras palavras, fazemos convolução de  $p(x)$ :

$$p(y) = p[\Phi^{-1}(y)] \left| \frac{d\Phi^{-1}(y)}{dy} \right|$$

Prova. Vamos iniciar nossa prova pela CDF de  $x$ :

$$H(x) = \int_{-\infty}^x h(x')dx'$$

$$CDF(y) = P(Y < y) = P(g(X) < y) = P(X < g^{-1}(y)) = H(g^{-1}(y))$$

Desta forma temos que:

$$CDF(y) = \int_{-\infty}^{x=g^{-1}(y)} h(x')dx'$$

$$\int_{-\infty}^y p(y')dy' = \int_{-\infty}^{x=g^{-1}(y)} h(x')dx' \quad \text{arrow} \quad (p(y')dy') \Big|_{y'=y} = (h(x')dx') \Big|_{x=g^{-1}(y)}$$

$$p(y) = h(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

O módulo é definido pois  $p(y) > 0$ .

# Exemplo

$$y = \Phi(x) = e^x \rightarrow x = \Phi^{-1}(y) = \ln(y)$$

$$p(x) = \begin{cases} 1 & \text{se } 0 \leq x \leq 1 \\ 0 & \text{se } x < 0 \text{ e } x > 1 \end{cases}$$

$$p(y) = 1 \times \frac{d \ln(y)}{dy} = \frac{1}{y} \text{ se } 1 \leq y \leq e$$

Logo, uma distribuição com certo comportamento, neste caso uniforme, não leva a também uma distribuição uniforme.

# Transformações de Variáveis aleatórias

- Para características de distribuições cumulativas, como a mediana (50% dos seus dados acima e abaixo daquele valor) não mudam com transformações monotônicas
- De forma semelhante, podemos estimar a distribuição dos erros, baseada na propagação por Série de Taylor:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!} (x - a)^n$$

$$y_0 = \Phi(x_0) \rightarrow \sigma_y = \left| \frac{d\Phi(x)}{dx} \right|_0 \sigma_x$$

## Exemplo: Fluxo vs. Magnitudes

$$m \equiv -2.5 \log \left( \frac{F}{F_0} \right)$$

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

$$F = \Phi^{-1}(m) = F_0 10^{-m/2.5}$$

$$p(m) = p(F) \frac{d}{dm} \left( F_0 10^{-m/2.5} \right)$$

# Exemplo: Fluxo vs. Magnitudes

Incluindo os erros:

$$\sigma_m = \left| \frac{d}{dF} \left( -2.5 \log \left( \frac{F}{F_0} \right) \right) \right| \sigma_F$$

# Exemplo: Fluxo vs. Magnitudes

$$p(m) = p(F) \frac{d}{dm} \left( F_0 10^{-m/2.5} \right)$$

$$u \equiv -m/2.5 \rightarrow \frac{d}{dm} F_0 10^{-m/2.5} = F_0 \frac{d}{dm} 10^u$$

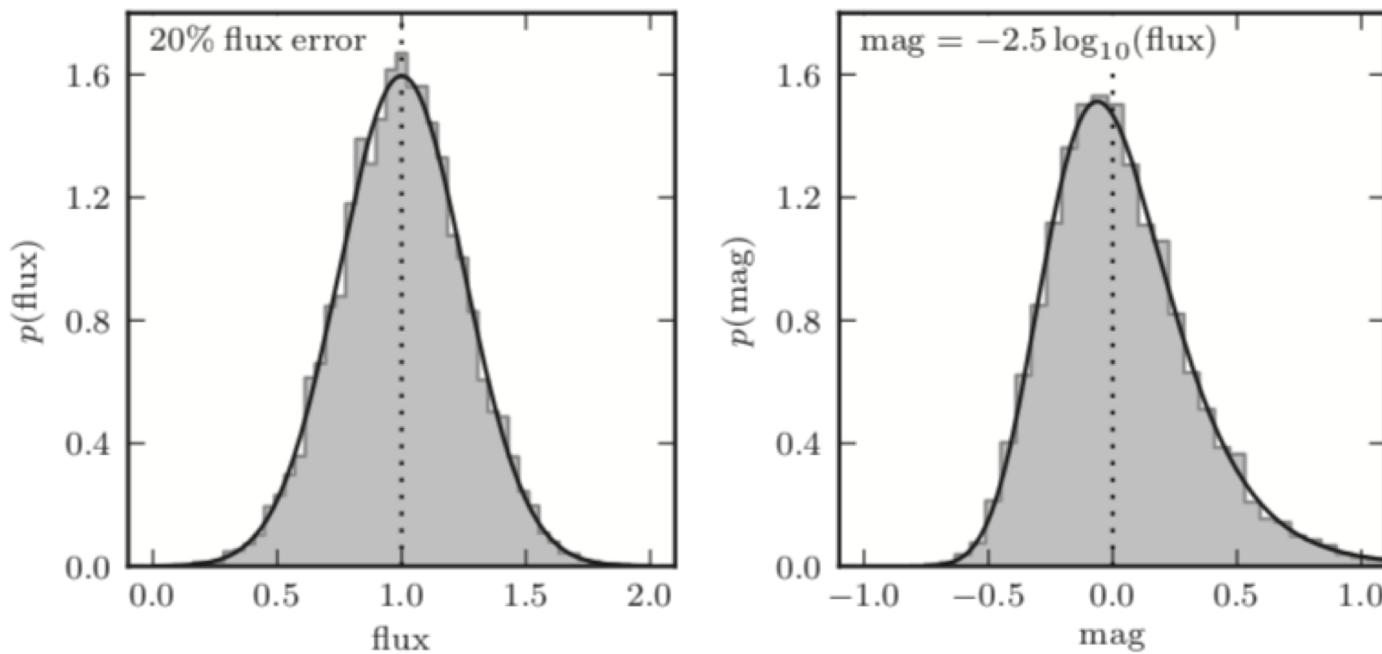
$$F_0 \frac{d}{dm} 10^u = F_0 \ln 10 \times 10^u \times \frac{du}{dm}$$

$$p(m) = p(F) \left| -\frac{F_0 \ln 10}{2.5} \times 10^{-m/2.5} \right|$$

Convolução entre Gaussiana e uma função potência

# Exemplo: Fluxo vs. Magnitudes

Se tivermos menos do que  $3\sigma$  de precisão, a  $p(m)$  também será uma distribuição Normal. Mas com  $5\sigma$  de confiança,  $p(m)$  será uma distribuição log-Normal, com um dos lados decaindo mais lentamente do que outro.



# Referências

- Livros-Textos
- <https://oscarbonilla.com/2009/05/visualizing-bayes-theorem/>