



UNIVERSIDADE FEDERAL
DO RIO DE JANEIRO



Astroestatística

Aula 002 -- Prof. Walter Martins-Filho

Estatística Descritiva

- Valor Esperado
- Momentos do Valor Esperado
- Percentils
- BIAS
- Teorema do Limite Central
- Outliers
- Estimadores Robustos

Valor Esperado

- Média, ou valor esperado, é resultado pela soma (ou integral) de cada valor possível para variável aleatória x , pesada pela sua probabilidade associada $p(x)$
- O Valor Esperado também é denominado *Primeiro Momento de Variável Aleatória*

$$E[x] = \sum_{i=0}^N x_i p(x_i)$$

$$E[x] = \int_{-\infty}^{\infty} x' p(x') dx$$

Momentos de Variável Aleatória

- O momento de uma variável aleatória é dado por:

$$\mu_n = E[x^n] = \int x^n p(x) dx$$

- O problema de Hausdorff
 - “Quais as condições para que se haja uma sequência de $n - th$ momentos ?”
 - Resposta de Hausdorff: Crescimento Monotônico, i.e., $\{\mu_n: n = 1,2,3, \dots\}$

Momentos de Variável Aleatória

- Desta forma, cada valor interessante para a variável aleatória passa a ser dado por momento.
- Sua PDF, é dada pela ordem zero:

$$\mu_0 = E[x^0] = \int x^0 p(x) dx = \int p(x) dx$$

Momentos de Variável Aleatória

- Por sua vez, o primeiro momento permite identificar qual seria valor mais provável pesado pelas probabilidades de saída, i.e., nossa MÉDIA, como falado no início:

$$\mu_1 = E[x^1] = \int x^1 p(x) dx = \int x p(x) dx$$

- O segundo momento identifica qual intervalo de possíveis valores de sair dada nossa distribuição de probabilidade, i.e., nossa medida de dispersão: VARIÂNCIA

$$\mu_2 = E[x^2] = \int x^2 p(x) dx$$

Momentos de Variável Aleatória

Exemplo: Distribuição Normal

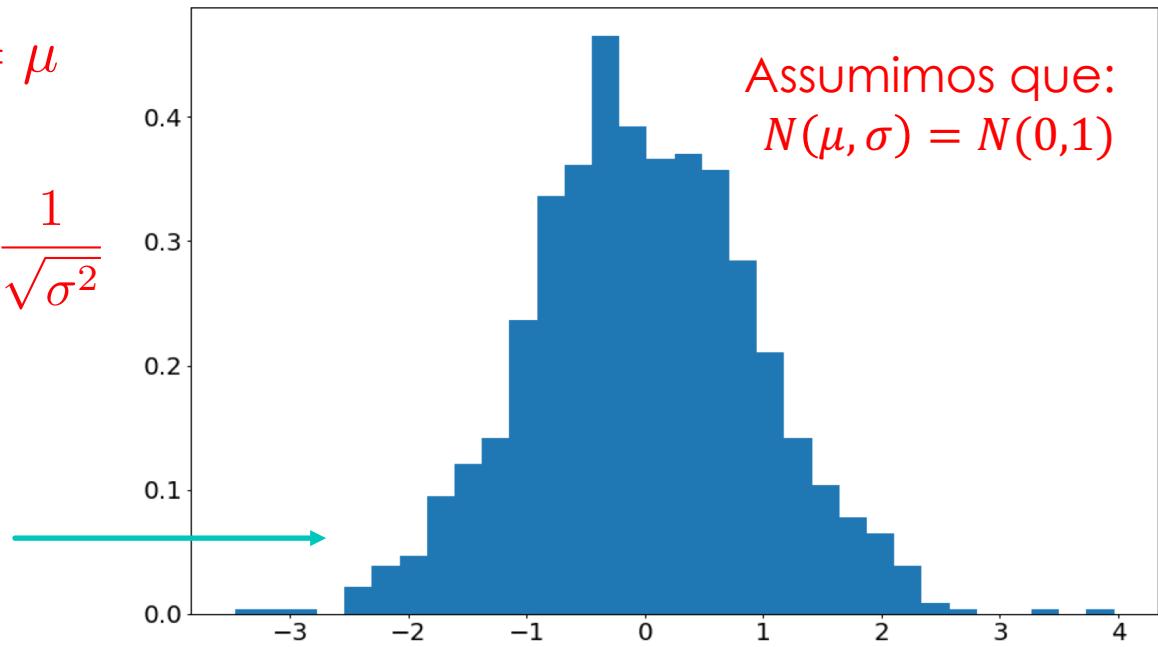
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$E[x] = \int_{-\infty}^{\infty} x p(x) dx = \mu$$

$$E[x^2] = \int_{-\infty}^{\infty} x^2 p(x) dx = \frac{1}{\sqrt{\sigma^2}}$$

Em nosso exemplo:

$$E[x] = 0 \text{ e } E[x^2] = 1$$



Momentos de Variável Aleatória

- Mas a conta anterior para segundo momento somente funciona porque padronizamos (*standardized*) a distribuição de probabilidade (i.e., tornarmos a média igual a zero).
- Para caso geral, devemos remover esse BIAS criado nos momentos seguintes:

$$Var(x) = E[(x - \mu)^2] = E[x^2] - \mu^2$$

Momentos de Variável Aleatória

- A média e a variância NÃO NECESSARIAMENTE estão correlacionadas
- É importante notar que os momentos, por originarem-se de somatórios ou integrais, carregam características similares, como:
 - (1) O valor esperado de soma, ou diferença, de duas variáveis é a soma ou diferença dos valores esperados de cada um
 - (2) O valor esperado de uma $p(x) = cte$ é a própria cte .
 - etc

Momentos de Variável Aletatória

Assim, temos que:

$$E\left[\sum_{i=0}^N x_i\right] = \sum_{i=0}^N E[x_i]$$

$$E\left[\left(\sum_{i=0}^N x_i\right)^2\right] = \sum_{i=0, i=j}^N E[x_i x_j] + \sum_{i=0}^N \sum_{j=0, i \neq j}^N E[x_i x_j]$$

$$E\left[\left(\sum_{i=0}^N x_i\right)^2\right] = \sum_{i=0}^N Var(x_i) + \sum_{i=0}^N \sum_{j=0, i \neq j}^N Cov(x_i, x_j)$$

Definimos a covariância como:

$$Cov(x_i, x_j) = E[(x_i - E[x_i])(x_j - E[x_j])]$$

Covariância

- Na prática, a COVARIÂNCIA avalia a mutual dependência entre x_i e x_j
- SE x_i e x_j são independentes entre si, $Cov(x_i, x_j) = 0$
- Entretanto, o CONTRÁRIO NÃO É VERDADEIRO, existe a possibilidade de que duas variáveis dependentes entre si gerem uma covariância nula.

Caso Homocedástico

- Vamos assumir que todas as variáveis em $\{x_i\}$ possuem a mesma variância σ^2
- Se $\{x_i\}$ são independentes, então, o primeiro e segundo momentos tornam-se:

$$\mu = E[x] = \frac{1}{n} \sum_{i=0}^n x_i$$

$$Var(x) = \frac{1}{n^2} \sum_{i=0}^n Var(x_i) = \frac{\sigma^2}{n}$$

A variância final é a média aritmética das variâncias!

O nosso conhecido *mean square variation*, ou STD, **standart deviation!**

A forma padronizada

- Com o STD, podemos padronizar a nossa distribuição e com isto compará-las com outras distribuições também padronizadas:

$$x_{STD} = \frac{x - \mu}{\sigma}$$

- Distribuições padronizadas possuem média ZERO e variância UM.

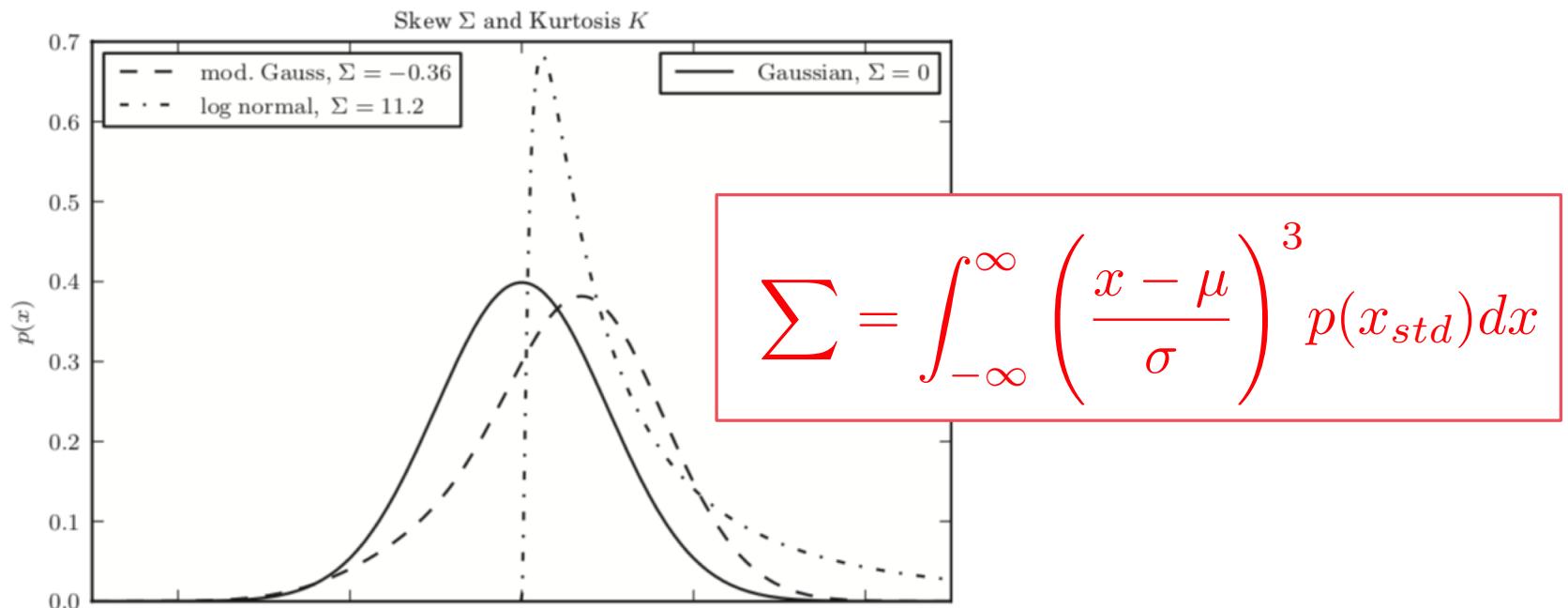
SKEWNESS: terceiro momento

- O skewness é o nome dado ao terceiro momento de uma variável aleatória.
- O skewness é definido na literatura com a letra Σ
- Ele provém informação se a distribuição é simétrica, tende para lado direito ou lado esquerdo do valor esperado, a média
- O terceiro e quarto momentos de variáveis aleatórias são correspondentes a FORMA que a distribuição possui.
- Distribuições com um dos lados mais alongado do que outro, baseado no valor médio, possuem valores positivos**
- Distribuições que possuem dois lados idênticos, possuem valor nulo**
- Distribuições com um dos lados menor do que outro, baseado na média, possuem valor negativo**

$$\Sigma > 0$$

$$\Sigma = 0$$

$$\Sigma < 0$$



SKEWNESS: terceiro momento

O skewness é formalmente posto em termos da distribuição padronizada:

Kurtosis: quarto momento

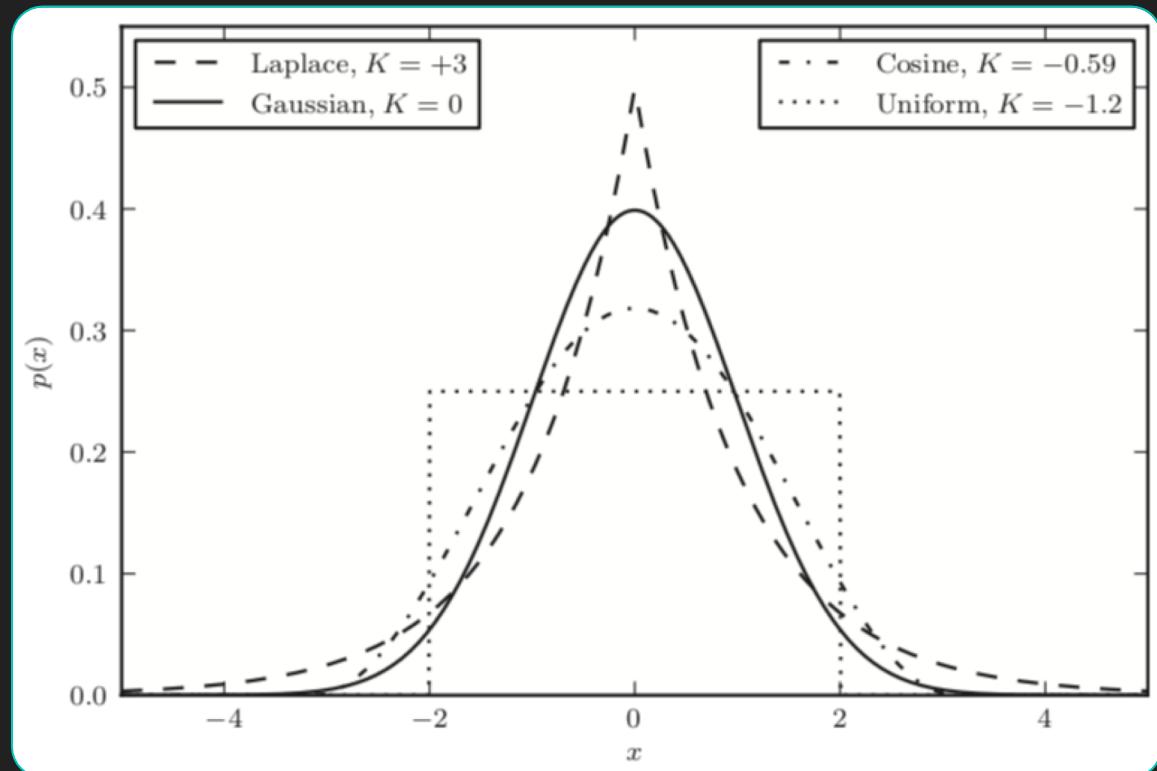
- Kurtosis é definido como quarto momento de uma variável aleatória

$$K = \int_{-\infty}^{\infty} \left(\frac{x - \mu}{\sigma} \right)^4 p(x_{std}) dx$$

- Em geral, o termo Kurtosis é definido baseado na distribuição Normal, i.e., no perfil da função Gaussiana.
- O quarto momento permite identificar o quanto suave é a distribuição. Em outras palavras, se existe pico.

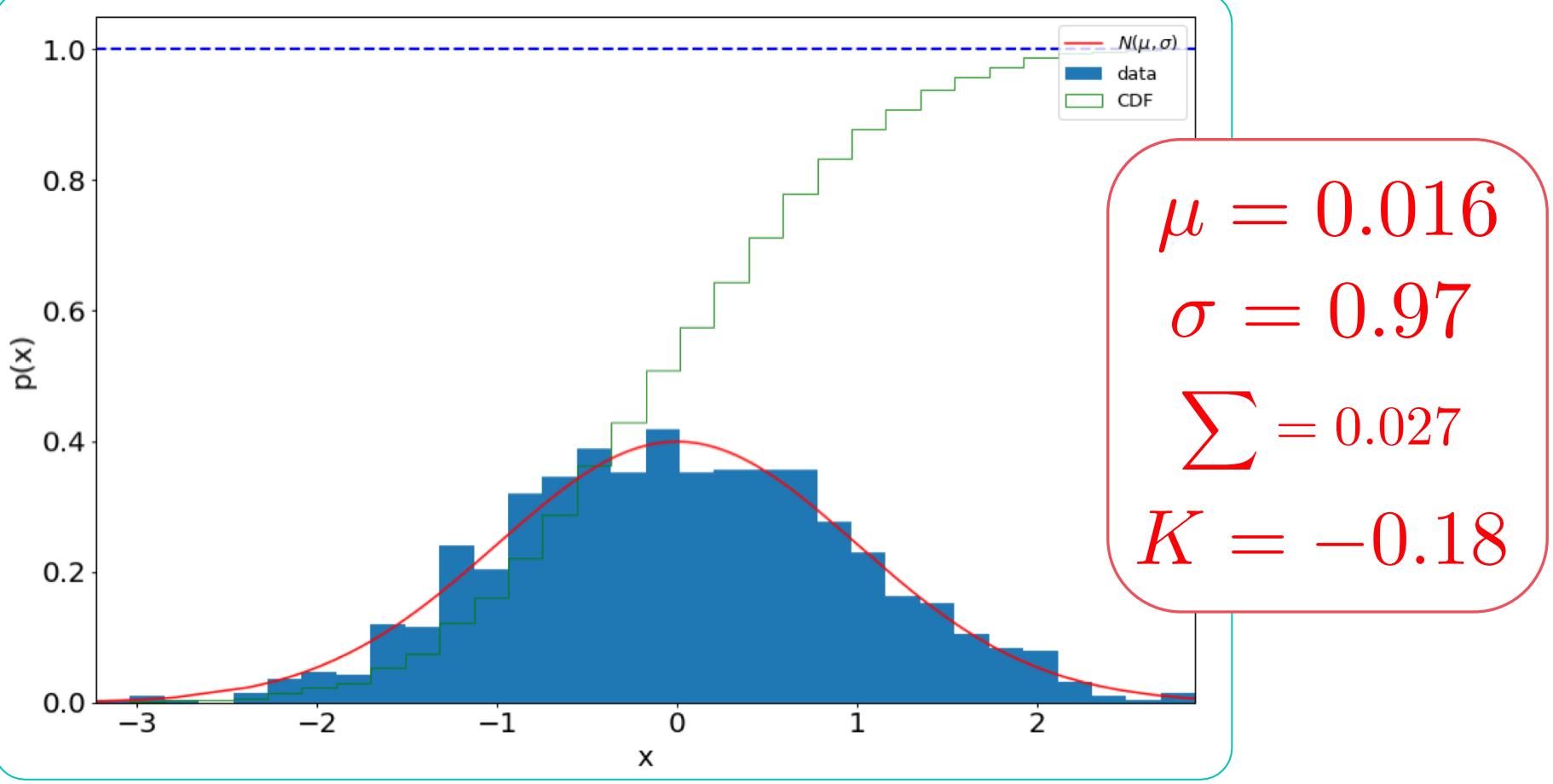
Kurtosis: quarto momento

- Se $K > 0$: temos um pico
- Se $K = 0$: temos perfil gaussiano
- Se $K < 0$: perfil da distribuição começa a tornar-se um platô, i.e., uma distribuição uniforme dentro de um determinado intervalo



Mais momentos (?)

- Da forma como se define, podemos obter cada vez mais momentos a partir de nossa variável aleatória x e de sua distribuição de probabilidade $p(x)$
- Entretanto, nem todos momentos são de alguma forma úteis na prática em astronomia.
- É normal, usarmos somente os primeiros 5 momentos:
 - Ordem zero: nossa PDF
 - Ordem um: a média
 - Ordem dois: a variância
 - Ordem três: o skewness
 - Ordem quadro: o kurtosis



Exemplo: Gaussiana

- Neste caso, em que $n = 1000$, de $N(\mu, \sigma) \sim N(0, 1)$

Exercício

- Usando Python ou R. Crie de forma sintética, dados para o fluxo de uma estrela.
 - Assuma que o fluxo segue uma distribuição normal $p(F) = N(F_\mu, \sigma_F)$. assumindo valores para F_μ e σ_F . O mais comum é centrada em 1 e variância 1.
 - Crie histograma, procure a função que ajuste uma distribuição normal. Verifique se retorna a média e variância que você assumiu previamente.
 - Calcule as magnitudes, obtenha a distribuição de magnitudes.
 - Encontre as rotinas no Python OU R que calcule os momentos e coloque eles no gráfico do histograma, assim como exemplo da Gaussiana anteriormente comentado.
 - Inclua a Distribuição Cumulativa
 - Apresente seus resultados na próxima AULA em folha impressa com seu gráfico.

Percentils

- Os percentis são proporção nas CDFs
- Eles indicam o valor que x assume, quando a CDF, a distribuição cumulativa, atinge certa quantidade.
- Os percentis mais usados são:
 - 25% $\rightarrow p_{25}$
 - 50% $\rightarrow p_{50}$ também denominado mediana
 - 75% $\rightarrow p_{75}$

Mean Square Errors

- Mean square errors é nome dado a classe de estimadores que usam a variância e um formalismo para bias presente na amostra afim de identificar possíveis discrepâncias entre o conjunto amostral e a população:

$$MSE = Var(x) - bias^2$$

- O “bias” é definido como valor esperado da diferença entre estimador e valor verdadeiro. Na prática, não conhecemos valor verdadeiro, e MSE passa a ser usado com valor de possível modelo.

BIAS

- Devido a sempre estarmos usando dados de uma amostra, devemos tomar cuidado que os momentos calculados anteriormente podem não corresponder aos verdadeiros momentos da população.
- Algumas correções foram sugeridas com passar do tempo, afim de que não se crie BIAS, uma variação que se acumule ao longo de toda análise de sua distribuição
- O BIAS mais comum é a média se desviar por algum valor:
 $\mu_{populacao} = \mu - cte$

$$\bar{x} = \frac{1}{N} \sum_{i=0}^N x_i$$

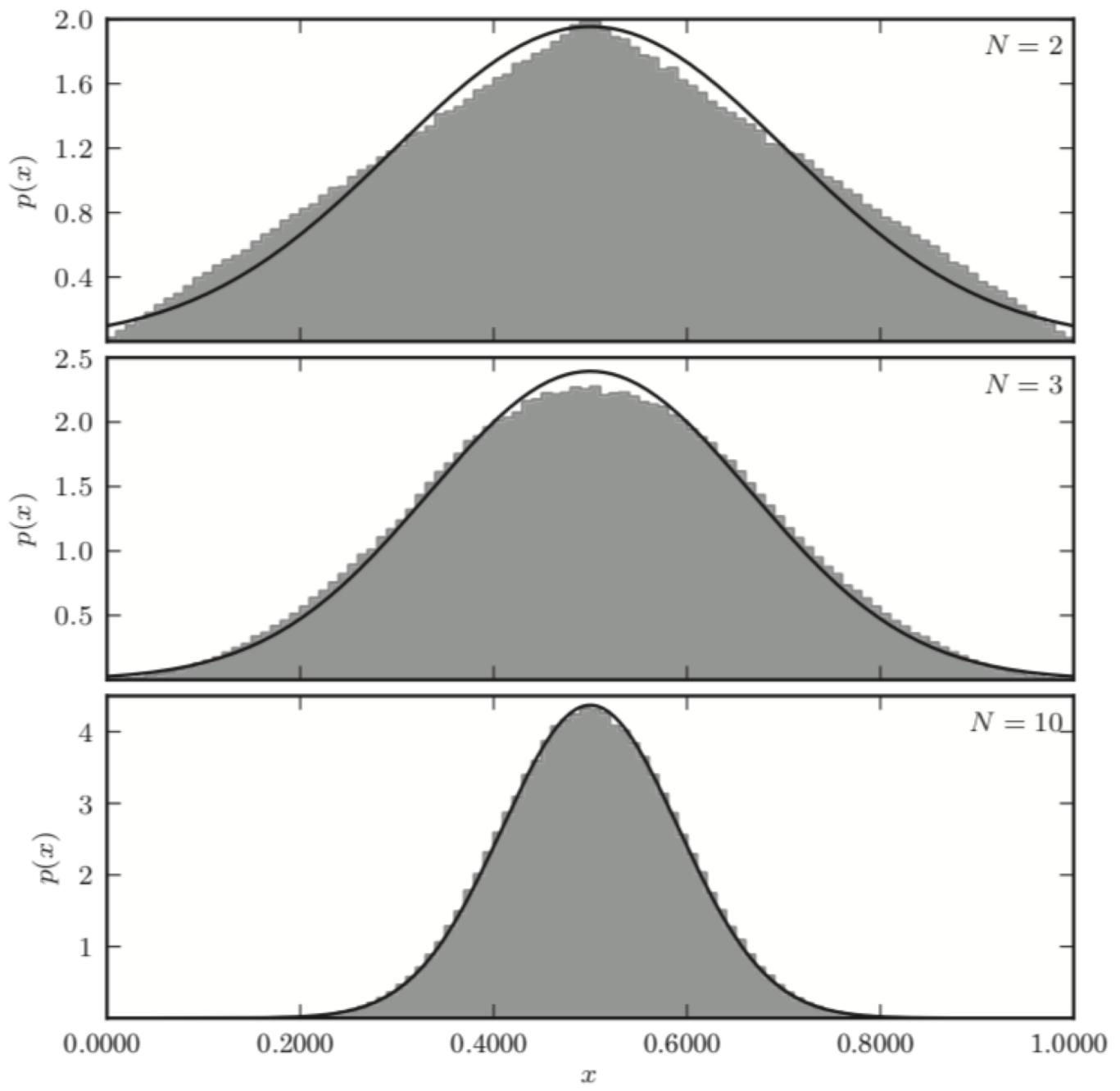
$$s = \sqrt{\frac{1}{N-1} \sum_{i=0}^N (x_i - \bar{x})^2}$$

Amostra

- Afim de não interpretar erroneamente, iremos usar a partir de agora os termos \bar{x} para a média amostral e s para seu desvio padrão.
- O fator $N - 1$ é conhecido como Correção de Bessel, tornando a variância amostral com menor bias

Teorema do Limite Central

- “Quando maior for tamanho da amostra, a distribuição amostral da média tende a ser descrita por uma distribuição normal.”
- NOTE que:
 - Não é a distribuição de $\{x_i\}$ que tende a uma distribuição normal!
 - É a distribuição de \bar{x} que tende a ser uma distribuição normal

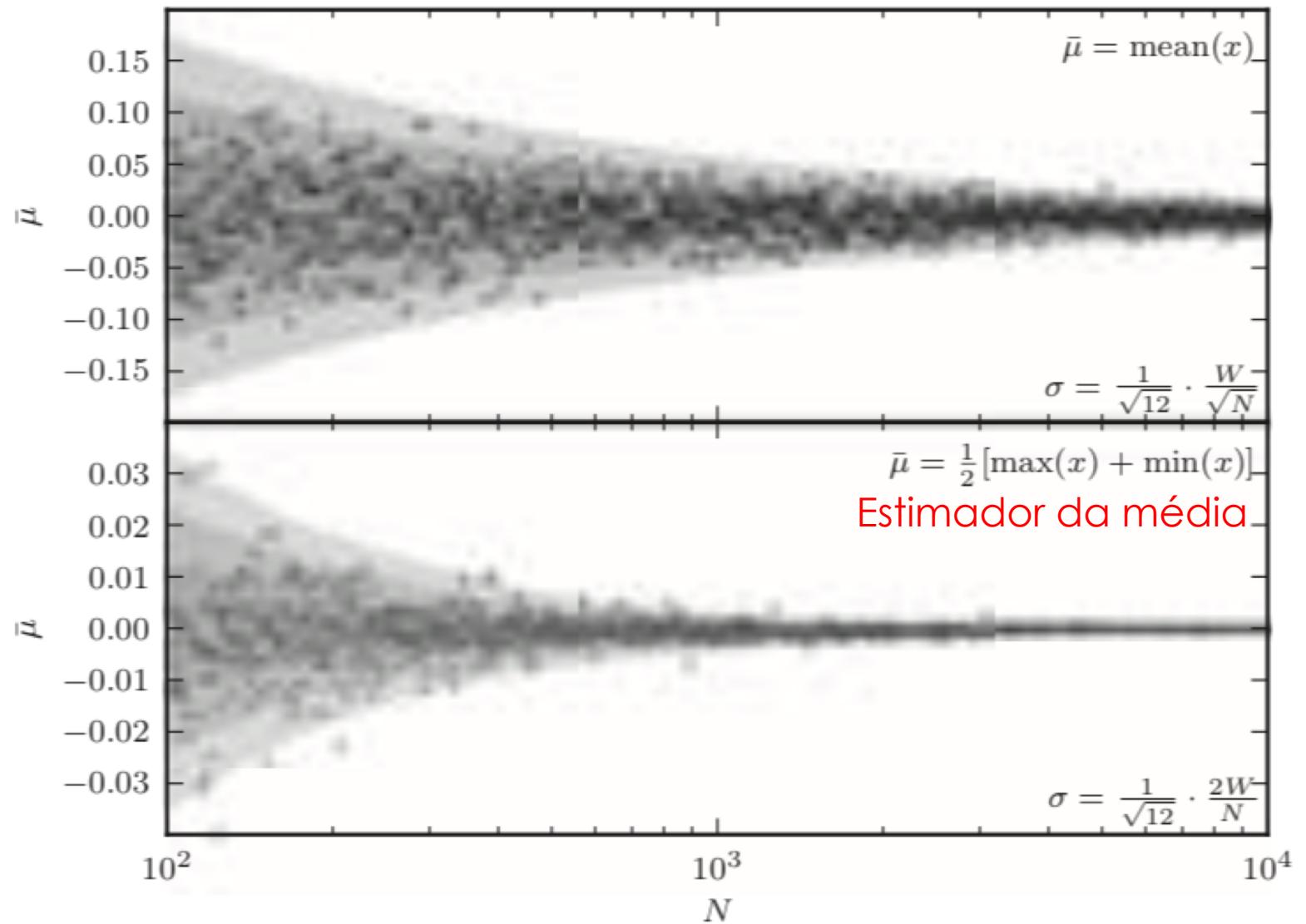


$$\mu = \frac{[\max(x_i) - \min(x_i)]}{2}$$

$$\tilde{W} = [\max(x_i) - \min(x_i)] \frac{N}{N - 2}$$

Estimadores de média e largura

- Assumindo o Teorema do Limite Central, e várias amostras sucessivas, podemos estimar a média e a largura (ou desvio-padrão) da população como:



Desvio-padrão da Média

- Vamos assumir que $N > 10$
- \bar{x} e s , pelo Teorema do Limite Central, a média e a variância seguiram distribuições normais
- Vamos assumir que as diversas distribuições $\{x_i\}_k$ onde $k = 1, \dots, N$ são as amostras coletadas de x , possuem o MESMO desvio padrão. Temos que:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Definimos que este seja o desvio-padrão da média

Desvio-padrão da distribuição de s

- Aplicando a ideia da Correção de Bessel junto ao Teorema do Limite Central, o desvio-padrão s será dado por:

$$\sigma_s = \frac{s}{\sqrt{2(N-1)}} = \frac{1}{\sqrt{2}} \sqrt{\frac{N}{N-1}} \sigma_{\bar{x}}$$

Exercício

- Crie k distribuições vindas de uma distribuição normal de média zero e variância 1 (em Python ou R), com $N > 10$
- Calcule a média e desvio-padrão de cada uma dessas k -distribuições. Faça um histograma da média e do desvio-padrão.
- Verifique se o desvio-padrão da distribuições da médias e dos desvio-padrões de N seguem as ideias anteriores. Calcule usando a fórmula e calcule usando uma rotina no seu programa.
- Faça um gráfico de convergência como no slide-28.

“Real data have outliers”

- Para casos reais, distribuições podem ter valores considerados outliers.
- Outliers são valores tão discrepantes que estão acima ou abaixo de percentis p_5 ou p_{95} na distribuição cumulativa
- Entretanto, os momentos calculados anteriormente são facilmente influenciados pelos valores outliers na amostra.

Mediana: um estimador robusto

- A mediana, por não derivar-se de uma média com termos em proporção a sua probabilidade, não é dependente de valores considerados outliers.
- Devido a essa não-influência, a mediana torna-se ótimo estimador de parâmetros de escala e localização dentro da distribuição de valores $\{x_i\}$

Estudo de caso: Raios Cósmicos

- Exemplo:

- $\{x_i\}$: são as distribuições de contagem em placa CCD
- Raios cósmicos são eliminados combinando imagens tiradas em instantes diferentes
- Cada pixel serve como conjunto amostral, a combinação pela mediana permite remover qualquer outlier que tenha caído em determinado pixel, i.e., desta forma, removendo os raios cósmicos

Intervalo Interquartil

- Enquanto a mediana serve como estimador mais robusto do que a média para localização na distribuição, o intervalo interquartil é estimador mais robusto para escala
- Intervalo interquartil é definido pela região compreendida pelos percentis $p_{75} - p_{25}$
- O intervalor interquartil é re-normalizado como σ_G para se relacionar a uma perfeita Gaussiana:

$$\sigma_G = 0.7413(p_{75} - p_{25})$$

Referências

- Livros-textos
- WolframAlpha Web
- <https://stackoverflow.com/questions/20011122/fitting-a-normal-distribution-to-1d-data>
- <https://stackoverflow.com/questions/9378420/how-to-plot-cdf-in-matplotlib-in-python>
- <https://www3.nd.edu/~rwilliam/stats1/x21.pdf>