Bloco 4

Estatística

Atualizado:
2012

Bibliografia

- Lena – Cap. 4 (parte) - só a inspiração...
- Wall & Jenkins, Practical statistics for astronomers
- Brandt – Statistical and computacional methods in data analysis
- Bevington – Data reduction and error analysis for the physical sciences
- Gnedenko & Khinchin - An elementary introduction to the theory of probability
- Dwass – Probability: theory and applications
- Saridis – Stochastic processes, estimation...

Link interessante:

- http://astrostatistics.psu.edu/

Tópicos

- Probabilidade
- Variáveis aleatórias
- Distribuições de probabilidade
- Ajuste de funções

Vale a pena lembrar que o conteúdo dessas anotações não tem rigor matemático. Algumas das referências citadas possuem essa característica e devem ser consultadas.

Introdução

- Observação
  - extração de parte da informação carregada pela radiação eletromagnética
  - seleção de informação depende da técnica utilizada
  - resultado: um sinal
- Processamento de sinais importante em várias áreas, além da astronomia
  - telecomunicações
  - controle de processos
  - medicina
- Ferramentas importantes para o processamento de sinais
  - probabilidade e estatística
  - transformada de Fourier

- Em uma observação astrofísica, ao se realizar uma medida, como em qualquer ciência experimental, temos como resultado um dado que deve ter sua:
  - avaliação correta e
  - exploração completa.
- Um experimento pode ter seu objetivo classificado como:
  - amostragem (*sampling*): para caracterização de uma população
  - teste de uma hipótese
  - estimativa de um parâmetro

"Nosso Problema"

Exemplo de teste de hipótese:

Um dado objeto astrofísico tem seu fluxo variável?

- Como testar essa hipótese? Como quantificar o resultado?

Realizar uma série de medidas de fluxo

Associar à série uma dispersão e compará-la com os valores individuais

Experimento

- Podemos pensar em um experimento (medida) como uma série de procedimentos que levam a uma ou mais quantidades = resultados
- Evento: ocorrência de um dado resultado
- Os resultados pode ser quantidades
  - contínuas
  - discretas
- A não-repetibilidade de um resultado pode ser causada, de modo geral, por:
  - acurácia finita
  - natureza da quantidade a ser medida

As contagens em um píxel de um CCD são contínuas ou discretas?

Nosso problema

- acurácia finita

Não podemos dizer se fluxo é variável.

E agora?

Grafico acima com erro pequeno?

Acurácia x precisão

- acurácia: quão próximo do valor real está a medida
- precisão: quão bem um dado valor foi determinado
  - o que chamamos normalmente de erro da medida está relacionada à sua precisão
- erros sistemáticos - afetam a acurácia:
  - diferença entre o valor real e a medida que podem ser reproduzidos
    - introduzidos por falhas de procedimento
  - tendências da medida
- erros aleatórios - afetam a precisão
  - medida da flutuação dos dados relacionada a precisão finita dos dados

Probabilidade

- A probabilidade, $P(\zeta)$, do evento ζ é um número positivo entre 0 e 1.
  - Quando se pensa em um único evento, a sua probabilidade só pode ser definida em <u>condições invariantes</u>

- Dois <u>eventos</u> são chamados <u>independentes</u>, ζ_1 e ζ_2, se a ocorrência de cada um deles é completamente desassociada da outra, isto é, a ocorrência de ζ_1 não interfere absolutamente na ocorrência de ζ_2. Nesse caso, a probabilidade dos dois eventos ocorrerem simultaneamente é dada pela multiplicação de suas probabilidades:

$$P(\zeta_1 \text{ e } \zeta_2) = P(\zeta_1) \cdot P(\zeta_2)$$

Exemplo

- Considere o lançamento de uma moeda com dois possíveis resultados, cara e coroa, de mesma probabilidade:

  ζ_1: cara

  ζ_2: coroa

  $P(\zeta_1) = P(\zeta_2) = 0.5$

Variáveis aleatórias ou estocásticas

- Ao se realizar um experimento, normalmente, não existe apenas 1 resultado possível, mas vários. Assim, podemos pensar em um conjunto de resultados possíveis para um dado experimento.
- Porém, esse conjunto ainda não define o comportamento de um experimento. É necessário saber quais são as probabilidades de um dado resultado ocorrer: Daí entra os conceitos de função de distribuição e de densidade de probabilidade.
- Por que um resultado não é sempre o mesmo? Porque existem condições que determinam o resultado que não conseguimos levar em conta ou controlar. Esses desvios podem ser expressos de uma maneira estatística.
- Assim, o comportamento de uma **variável aleatória** é definido por:
  - seu espaço amostral
  - sua distribuição de probabilidades

Definindo variável aleatória

- Suponha um processo que possa ter vários resultados, ζ, cada um deles associado a uma probabilidade $P(\zeta)$. A cada resultado pode ser associada uma dada quantidade real pela função $x(\zeta)$

A função $\mathbf{x}(\zeta)$ é uma variável aleatória

Note que, apesar do nome, **x**(ζ), não é uma variável!

É uma função.

Evento

- Definimos como um **evento** um subconjunto de resultados dentro do espaço amostral. A um evento pode ser associado uma probabilidade, de acordo com as probabilidades dos resultados (eventos elementares)

- Exemplos de eventos:
  - { $X(\zeta) < 0{,}74$ }
  - 5 caras consecutivas

Densidade de probabilidade

- A densidade de probabilidade (de uma variável aleatória), f(x), é tal que a probabilidade, P, da v.a. $x(\zeta)$ estar entre dois valores x_1 e x_2 é:

$$P=\int_{x_1}^{x_2} f(x)\,dx$$

- Se pensarmos em eventos discretos (não contínuos), a distribuição de probabilidade pode ser expressa a partir das probabilidades, p_i, de cada evento, ζ_i, como:

$$f(x)=\sum_i p_i\,\delta(x-x_i)$$

Probabilidade cumulativa

- **A probabilidade cumulativa**, F(x), ou função de distribuição é a probabilidade da variável aleatória $x(\zeta)$ possuir um valor menor que x:

$$F(x)=\int_{-\infty}^{x} f(u)\,du$$

- A função F(x) é tal que:
  - $F(-\infty) = 0$
  - $F(\infty) = 1$ (normalizada)
  - F(x) é uma função crescente
  - F(x) é contínua à direita

Descrição de uma variável aleatória

- Considere uma função, H, dependente de uma variável aleatória f(x).
  - H[f(x)] é também uma variável aleatória
  - e possui uma distribuição de probabilidade que a descreve
- Podemos pensar em uma medida como um procedimento para determinar a variável aleatória que descreve a obtenção de uma grandeza física que é inicialmente desconhecida
- Assim, é adequado ter meios para descrever a variável aleatória, mesmo que parcialmente, a partir das medidas
- Entretanto, vamos inicialmente introduzir algumas definições formais que dependem do conhecimento da variável aleatória

Algumas definições

- A média, ou valor médio ou valor esperado, η, de uma v.a. contínua é:

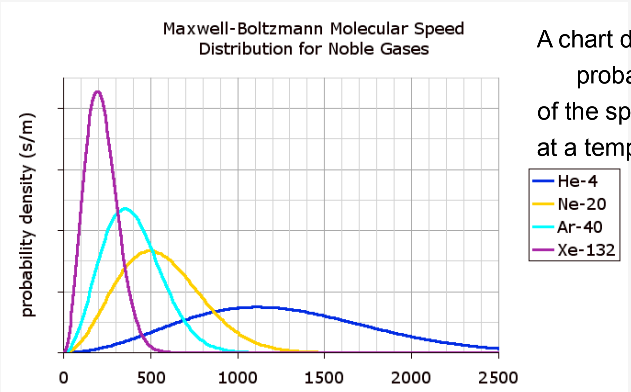
$$\eta = E\{x\} = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

- Se a v.a. é discreta, o valor médio é:

$$\eta = E\{x\} = \sum_n x_n p_n$$

- A moda é o valor mais provável, sendo o valor x onde f(x) é máxima
- A mediana, x_m, é definida como:

$$P(x \le x_m) = F(x_m) = 1/2$$



Distribuição

A chart displaying the speed probability density functions of the speeds of a few noble gases at a temperature of 298.15K (25 C).

Wikipedia

$$f(v) = 4\pi \left(\frac{m}{2\pi k T}\right)^{3/2} v^2 \exp\left[\frac{-m v^2}{2 k T}\right]$$

Velocidade + provável

$$v_p = \sqrt{\frac{2kT}{m}}$$

Velocidade média

$$\langle v \rangle = \sqrt{\frac{8kT}{\pi m}}$$

- A variância, σ^2, é uma medida da concentração da distribuição de probabilidade em torno da média:

$$\sigma^2 = E\{(x-\eta)^2\} = \int_{-\infty}^{\infty} (x-\eta)^2 f(x)\, dx$$

$$\sigma^2 = E\{x^2\} - E\{x\}^2$$

$$\sigma^2 = \sum_n (x_n - \eta)^2 p_n$$

  - Existem outras quantidades que medem a dispersão de uma v.a., como o desvio médio, por exemplo.

- Desvio padrão, σ, ou desvio quadrático médio:

$$\sigma = +\sqrt{\sigma^2}$$

- O desvio padrão possui a mesma unidade da variável aleatória. Essa característica faz com que ele possa ser utilizado como uma estimativa do erro associado ao valor esperado

Momentos

- O momento em torno da origem de ordem k, λ_k, de uma v.a. é dado por:

$$\lambda_k = E\{x^k\} = \int_{-\infty}^{\infty} x^k f(x)\,dx$$

- Podemos também definir o momento central que é dado por:

$$\mu_k = E\{(x-\eta)^k\} = \int_{-\infty}^{\infty} (x-\eta)^k f(x)\,dx$$

- Se conhecemos todos os momentos, conhecemos a distribuição f(x)

- O momento em torno da origem de ordem 1 é a média
- Alguns momentos centrais tem nomes específicos:
  - segunda ordem: variância
  - terceira ordem: *skewness*
    - mede a simetria da distribuição a partir da média
    - valores positivos – concentração à direita
  - quarta ordem: *kurtosis*
    - mede a concentração de probabilidade em torno da média

Qual o valor dos momentos centrais de ordem 0 e 1?

Distribuição de mais de uma variável aleatória

- Vamos considerar uma variável aleatória, y, que é composta por n variáveis aleatórias de modo que podemos escrever sua distribuição cumulativa de densidade como:
  - $F(x_1, x_2, \ldots, x_n)$
- A densidade de probabilidade pode ser escrita como:

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \ldots \partial x_n} F(x_1, x_2, \ldots, x_n)$$

- Se as probabilidades associadas a cada v.a. são independentes, as v.a. também serão independentes. Nesse caso, temos que:

$$f(x_1, x_2, \ldots, x_n) = g(x_1)\, g(x_2) \ldots g(x_n)$$

- Podemos definir os momentos em torno da origem nesse caso como:
  - $\lambda_{k1\,k2\,\ldots\,kn} = E\{ x_1^{k1}\ x_2^{k2} \ldots x_n^{kn} \}$

- E os momentos centrais como:
  - $\mu_{k1\,k2\,\ldots\,kn} = E\{ (x_1 - \langle x_1\rangle)^{k1} (x_2 - \langle x_2\rangle)^{k2} \ldots (x_n - \langle x_n\rangle)^{kn} \}$

- O valor esperado pode ser expresso como:

$$E\{H(\vec{x})\} = \int H(\vec{x})\, f(\vec{x})\, d\vec{x}\,,$$
$$onde$$
$$\vec{x} = (x_1, x_2, \ldots, x_n)$$

Alguns momentos específicos

- Médias
  - $\lambda_{10\ldots0} = E\{ x_1^1\ x_2^0 \ldots x_n^0 \} = E\{ x_1\} = \langle x_1\rangle$
  - $\lambda_{01\ldots0} = E\{ x_1^0\ x_2^1\ {}_x \ldots x_n^0 \} = E\{ x_2\} = \langle x_2\rangle$
  - ...
- Variâncias
  - $\mu_{200\ldots0} = E\{ (x_1 - \langle x_1\rangle)^2 \} = s^2(x_1)$
  - $m_{020\ldots0} = E\{ (x_2 - \langle x_2\rangle)^2 \} = s^2(x_2)$
  - ...
- Covariâncias
  - $c_{ij} = cov(x_i,x_j) = E\{ (x_i - \langle x_i\rangle) (x_j - \langle x_j\rangle) \}$

Lei de propagação de erros

- Uma variável aleatória composta por v.a. independentes possui as covariâncias nulas.
- Se uma variável aleatória y é uma combinação linear de variáveis aleatórias independentes, podemos definir a lei de propagação de erros:

$$\sigma_y^2 = \sum_{j=i}^{n} \left(\frac{\partial y}{\partial x_j} \right)^2 \sigma_{xj}^2$$

Algumas distribuições de probabilidade importantes

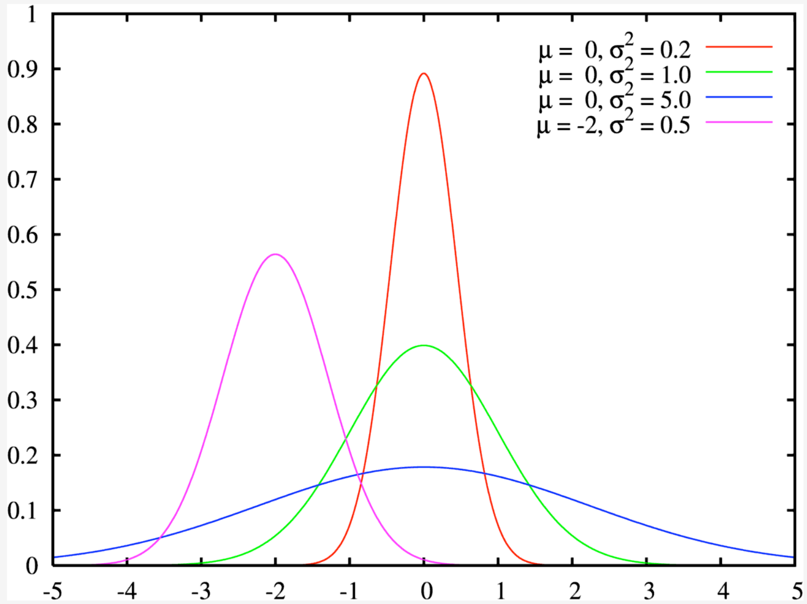
- **<u>Distribuição normal ou Gaussiana</u>**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\eta)^2}{2\sigma^2} \right]$$

  - Se $\eta = 0$, a distribuição é dita <u>centrada</u>
  - Função cumulativa de probabilidade:

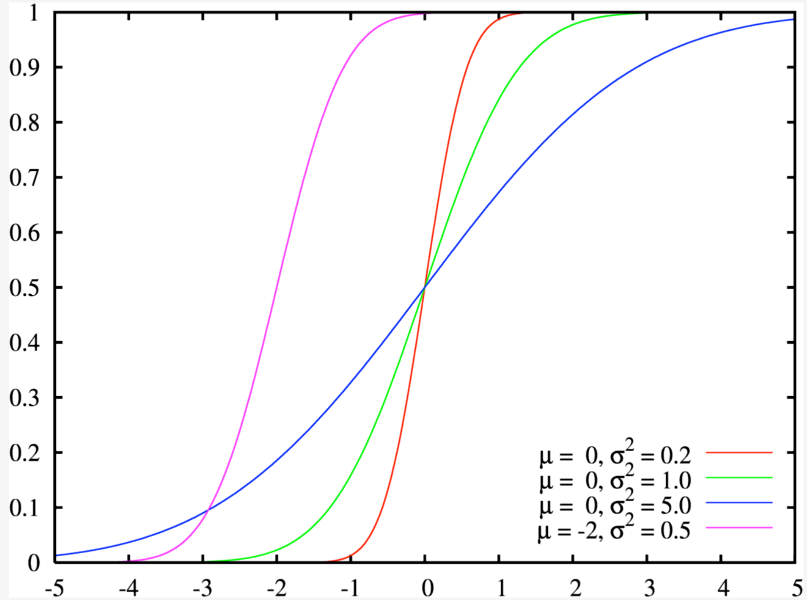
$$F(x) = \int_{-\infty}^{x} f_g(u)\,du = \frac{1}{2}\left\{ 1 + erf\left(\frac{x-\eta}{\sigma\sqrt{2}} \right) \right\}$$
$$onde$$
$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\,dt$$

Funçao de distribuição de probabilidade Gaussiana

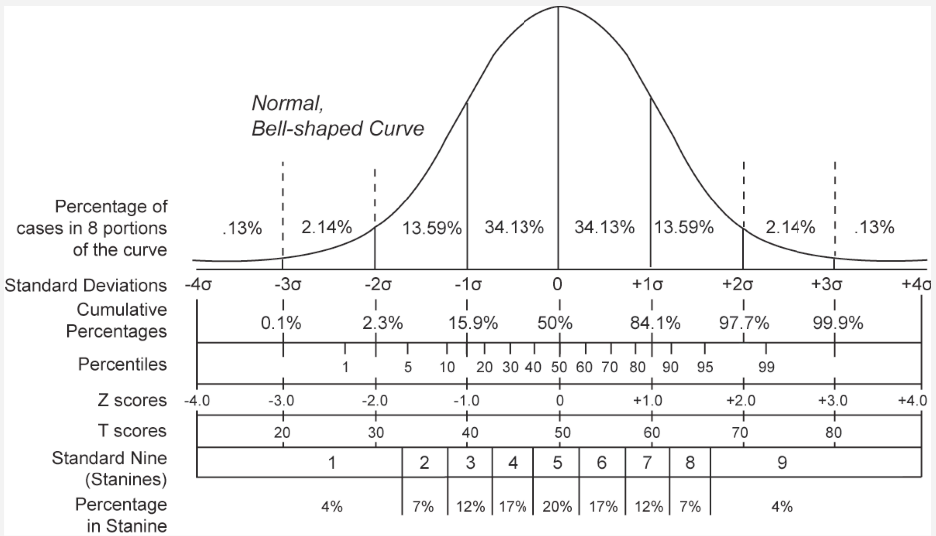


Wikipedia

Funçao cumulativa de probabilidade Gaussiana



Wikipedia



Wikipedia

Qual a probabilidade de uma medida estar mais distante que 3 ς da média?

O teorema do limite central ou teorema de Lyapunov

- Dada uma série de variáveis aleatórias, x1, x2, x3, ..., xi, ..., xn, identicamente distribuídas e estatisticamente independentes (com desvio padrão não nulo), a variável aleatória representada pela soma das v.a. acima tende à distribuição normal quando n tende a infinito.

Distribuição de Poisson

- A distribuição de probabilidades de Poisson (discreta) rege eventos que ocorrem a uma dada taxa temporal média que é independente do tempo decorrido desde o último evento. Ela fornece a probabilidade que ocorra um dado número de eventos em um intervalo fixo de tempo.
- Essa estatística rege a detecção de fótons
- É uma distribuição de eventos discretos

$$P(k)=e^{-\lambda}\frac{\lambda^k}{k!}$$

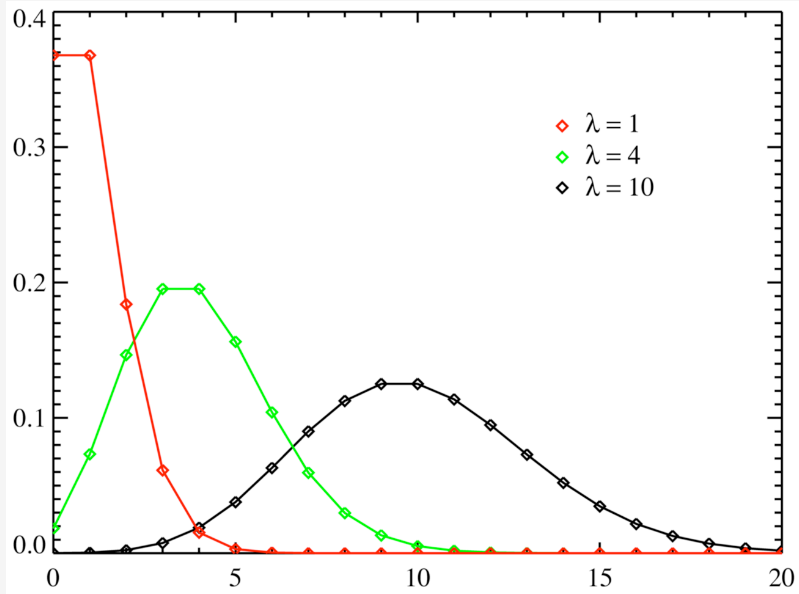
λ: taxa de eventos

$$f(x)=e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\delta(x-k)$$

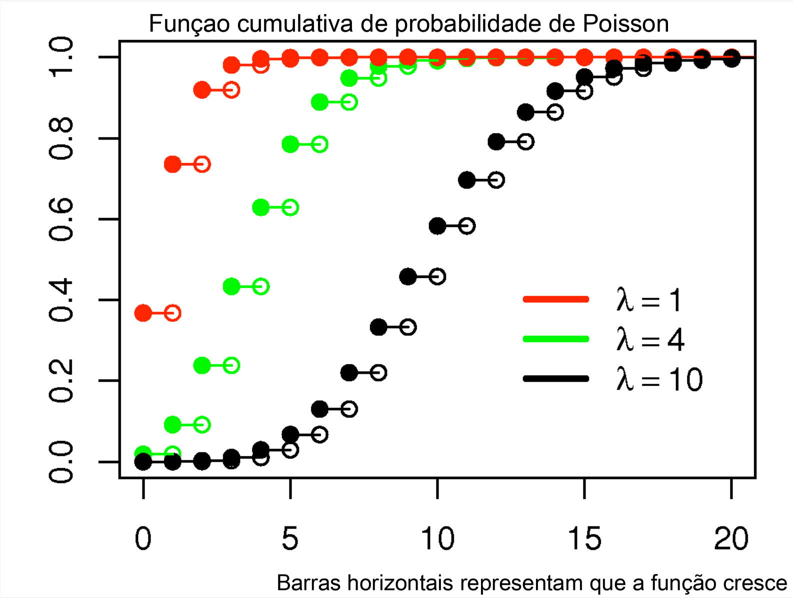
k: número de eventos em um intervalo de tempo fixo

Média: λ Variância: λ

Funçao de distribuição de probabilidade de Poisson



Wikipedia



Barras horizontais representam que a função cresce aos saltos, pois a função de distribuição de probabilidade só existe para valores inteiros

Wikipedia

Freqüência

- A probabilidade de um dado evento não é normalmente conhecida.
- Muitas vezes a medida tem como objetivo a determinação da probabilidade associada a um evento
  - de modo mais geral, um experimento pode fornecer a distribuição de probabilidade dos resultados possíveis
- Nesse caso, pode ser útil a definição de freqüência, h, do evento j:
  - h é a fração de ocorrências do evento j com relação ao número total de experimentos
  - h é uma variável aleatória da qual podemos definir média e variância
- É comum considerar a freqüência h como uma estimativa da probabilidade real do evento j

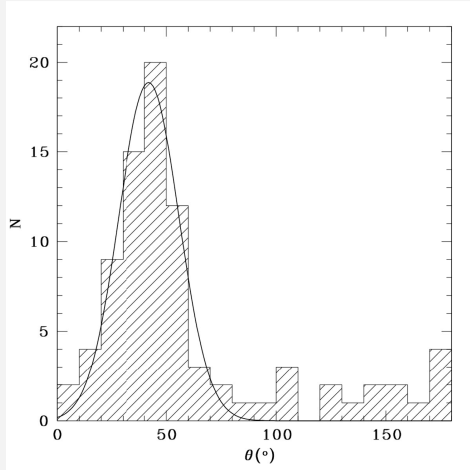
Lei dos grandes números

- A freqüência de um evento – fração do número de ocorrências com relação ao número total de experimentos, h - tende ao valor real da probabilidade, p, com o aumento do número de repetições do evento
  - É possível demonstrar que:
    - a média de h tende a p
    - a variância da média diminui com o número total de experimentos

Amostragem

- considere uma série de medidas com o objetivo de determinar a distribuição de probabilidade que descreve o conjunto de resultados
- a densidade de probabilidade pode ser estimada pela distribuição de freqüências que é fornecida pelos dados

Número de objetos com um dado valor do ângulo de polarização



- População: conjunto de todos os resultados possíveis de um experimento
  - pode ser finito ou infinito
  - discreto ou contínuo
- Amostra: sub-conjunto da população que resulta de uma série de experimentos
  - podemos considerar que os resultados da amostra seguem um distribuiçao de probabilidade da variável aleatória que os descrevem
- Amostragem aleatória
  - elementos das amostras, eventos, são independentes
  - cada evento é amostrado pela mesma variável aleatória e portanto pela mesma densidade de probabilidade

- Uma função da amostra é uma variável aleatória e é chamada de estatística
- <u>Exemplo</u>: considere uma amostra de n resultados, z_i: $x_1(z_1)$, $x_2(z_2)$, $x_3(z_3)$, ... , $x_n(z_n)$
  - média da amostra

$$\bar{x} = \frac{1}{n}\sum_i x_i$$

Estimativa de parâmetros

- Suponhamos um experimento cuja forma funcional da densidade de probabilidade dos resultados, f(x), seja conhecida, mas cujos parâmetros não conhecemos.
- Esses parâmetros podem ser estimados pela amostra
  - Exemplo: aproximação da emissão de uma estrela por um corpo negro.
    - Consideramos que a densidade de fluxo segue a lei de Planck, mas não conhecemos a temperatura. A temperatura pode ser estimada pelo índice de cor.
- Se o valor estimado baseia-se em uma amostra, ele é uma estatística, e especificamente um estimador, S

$$S = S(x_1, x_2, x_3, \ldots, x_n)$$

Estimadores

- Estimador sem viés
  - seu valor independe do tamanho da amostra

$$E\{S(x_1, x_2, x_3, \ldots, x_n)\} = \lambda, \qquad \forall\, n$$

- Estimador consistente
  - sua dispersão tende a zero para uma amostra infinita

$$\lim_{n\to\infty} \sigma(S) = 0$$

Amostragem de população

- População:
  - N elementos
    - $y_1, y_2, y_3, \ldots, y_N$
- Amostra
  - n elementos
    - $x_1, x_2, x_3, \ldots, x_n$
- Média da amostra

$$\bar{x} = \frac{1}{n}\sum_i x_i$$

  - A média da amostra possue como valor esperado a média da população

$$E\{\bar{x}\} = \bar{y}$$

- Variância da amostra

$$s^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$$

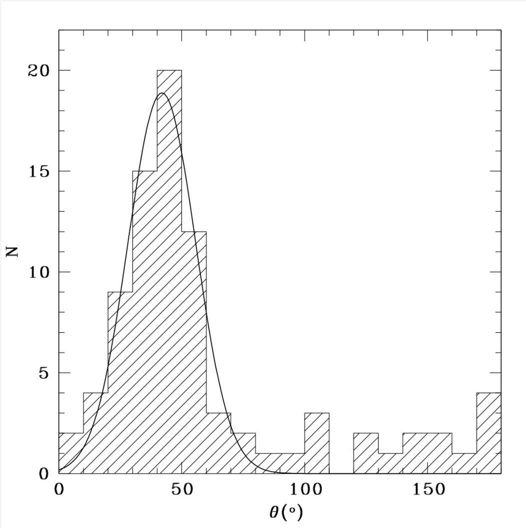
  - A variância da amostra possue como valor esperado a variância da população

$$E\{s^2\} = \sigma^2(y)$$

- Variância da média da amostra

$$\sigma^2(\bar{x}) = \frac{1}{n} s^2$$

A média é um estimador consistente?



A distribuição ao lado foi ajustada a uma Gaussiana.

A média da amostra é igual ao valor médio da Gaussiana?

Você poderia estimar visualmente a moda da distribuição observada?

Root mean square (RMS)
Média quadrática

$$RMS = \sqrt{\frac{1}{n}\sum_i^n (x_i)^2}$$

- A média quadrática é útil para termos uma noção do valor médio do valor <u>absoluto</u> de uma grandeza, independentemente de seu sinal

Medidas com diferentes precisões

- Os estimadores de uma amostra citados até agora consideram que todas as medidas possuem o mesmo erro. Mas, esse pode não ser o caso.
- Considere uma série de medidas (x_1, x_2, ... ,x_n), as quais podemos associar diferentes distribuições de probabilidade que possuem mesma média h, mas diferentes variâncias (s^2_1, s^2_2, ... ,s^2_n)
- Nesse caso, a média da amostra e sua variância são dadas por:

$$\bar{x} = \frac{\sum_i^n x_i/\sigma_i^2}{\sum_i^n 1/\sigma_i^2} \qquad \sigma^2(\bar{x}) = \left\{\sum_i^n 1/\sigma_i^2\right\}^{-1}$$

Estatística

- Uma quantidade calculada a partir dos valores de uma amostra é chamada de estatística
- Assim, um estimador é uma estatística
  - a média é uma estatística
  - a variância é uma estatística
  - etc

Hipótese nula

- De modo geral, a hipótese nula é a hipótese inicialmente assumida.
- É comum realizar testes estatísticos para estimar se ela é válida ou não.
  - isso se faz pelo cálculo de uma *estatística* específica para o teste desejado
- A essa *estatística* costuma se associar um grau de confiança que a hipótese nula seja válida, dado pela probabilidade da estatística ocorrer

Graus de liberdade (wiki)

- O número de graus de liberdade é o número de valores no cálculo final de uma estatítica que são livres para variar
- Estimates of statistical parameters can be based upon different amounts of information or data. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom (df). In general, the degrees of freedom of an estimate is equal to the number of independent scores that go into the estimate minus the number of parameters estimated as intermediate steps in the estimation of the parameter itself.

  - No ajuste de uma reta a partir de dois pontos, qual o grau de liberdade? E a partir de três pontos?

Teste de χ^2 da qualidade do ajuste

Teste de hipótese

- O teste de χ2 é usado para comparar um conjunto de dados, Di, aos valores esperados (um modelo, por exemplo), Mi

  Nesse caso, podemos definir a estatística do teste de χ2

$$\chi^2=\sum_i^n \frac{(D_i-M_i)^2}{M_i}$$

  - cuja função de densidade é:

$$f(\chi^2)=k(\chi^2)^{\lambda-1} e^{\frac{-\chi^2}{2}}$$

$$onde$$

$$\lambda=\frac{n}{2}$$

n: graus de liberdade

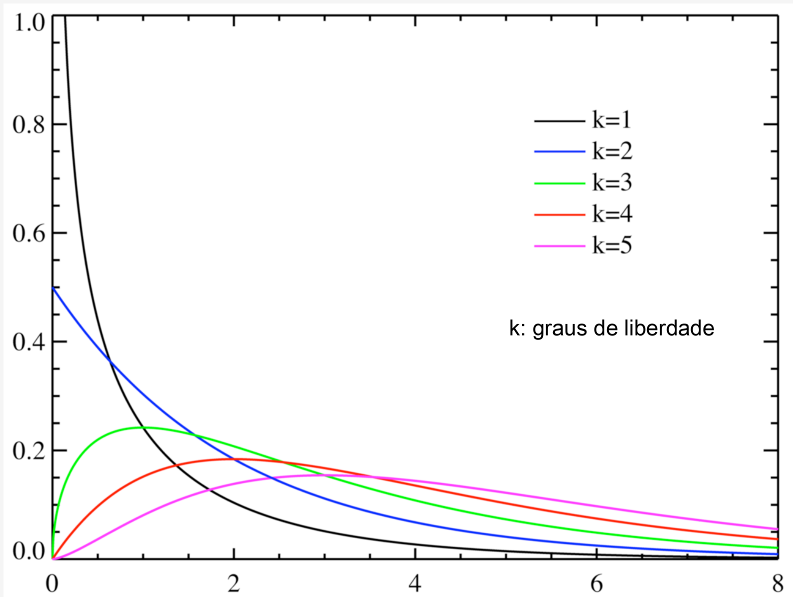
$$k=\frac{1}{\Gamma(\lambda)2^{\lambda}}$$

- Como testar a hipótese nula, isto é, que os dados sejam bem representados pelo modelo?
- Em primeiro lugar, você deve definir com qual grau de confiança você quer testar a hipótese.
  - O valor normalmente usado é 5% ou menos. Isto é, o X2 é maior que o valor medido em apenas 5% (ou menos) dos casos se a hipótese nula é verdadeira.
- O problema define qual o grau de liberdade.
- A seguir, você deve comparar qual o valor obtido com o X2 com o esperado para o grau de confiança definido. Se o X2 for maior, você descarta a hipótese nula com 5% de confiança.
- Vide a seguir alguns valores da probabilidade de X2 como função do grau de liberdade.

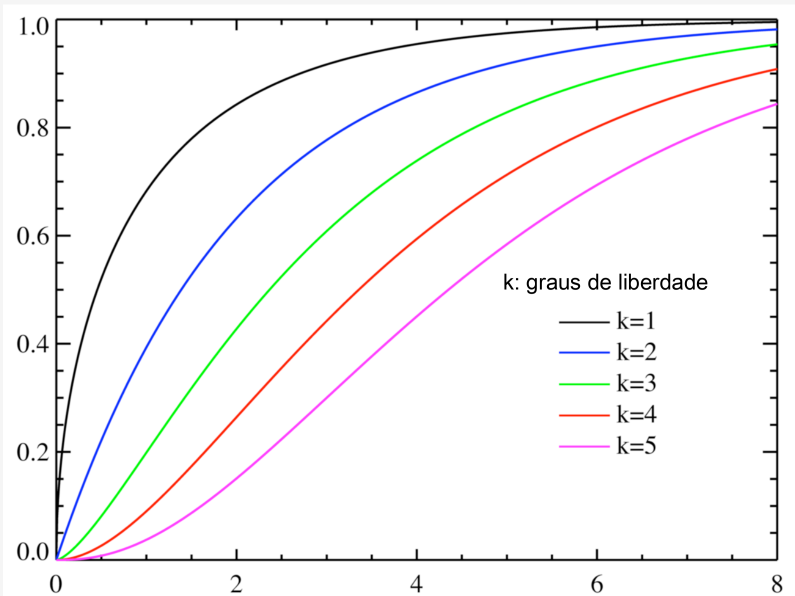
Degrees of freedom (df)	χ^2 value [11]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
	Nonsignificant								Significant		

http://en.wikipedia.org/wiki/Chi-squared_distribution

Densidade de probabilidade de χ^2



Funçao cumulativa de probabilidade de χ^2



Método dos mínimos quadrados

- Uma medida, y_j, pode ser considerada como:

$$y_j = x + \epsilon_j$$

erro

quantidade desconhecida

- Podemos determinar ε_j de modo que:

$$\sum_j \epsilon_j = \sum_j (x - y_j)^2 = mínimo$$

Aplicação do método dos mínimos quadrados

- Realização de n medidas do valor x representados por y_j, com cada medida possuindo variância estimada de σ_j^2
- x é uma função conhecida de n_p parâmetros desconhecidos
- Podemos definir:

$$M = \sum_{j}^{n} \frac{(x - y_j)^2}{\sigma_j^2}$$

- M segue uma distribuição de probabilidade de χ^2
  - ⇨ o número de graus de liberdade é $(n - n_p)$
- Além disso, a minimização de M permite um método para determinar os parâmetros n_p