DEPARTAMENTO DE FÍSICA NUCLEAR E ALTAS ENERGIAS

Tópicos Especiais de Física B IV: Introdução à análise de dados em FAE

Estatística básica – 1

PROFESSORES:

DILSON DE JESUS DAMIÃO

SANDRO FONSECA DE SOUZA

Estatística básica – 1

Está aula é baseada em um dos cursos de verão do CERN

Practical Statistics for Physicists

Louis Lyons/ Imperial College and Oxford

Livro de referência

Statistics for Nuclear and Particle Physicists, Cambridge University Press, 1986

J. H. Vuolo, Fundamentos da teoria de erros, 1996

V. Oguri, et. al., Estimativas e erros em experimentos de Física, 2013

Tópicos

1) Introdução
2) X^2
3) Estatística Frequentista e Bayesiana

Introdução

O que é estatística?

Probabilidade e estatística

Por que incertezas?

Incertezas sistematícas e estatísticas

Combinação de incertezas

Combinando dados de diferentes experimentos

Distribuições: Binomial, Poisson e Gaussiana

O que fazemos com estatística?

- **Determinação de parâmetros (valor esperado)**
  - Por exemplo, massa de partículas = 80 ± 2 GeV
- **Ajuste de dados / MC**
  - Os dados concordam com a teoria?
- **Teste de hipóteses**
  - Entre as teorias 1 e 2, qual é a mais adequada?
- **Nos ajuda a decidir**
  - Qual experimento devemos fazer a seguir?

FAE tem uma grande demanda de financiamento e tempo, então quanto mais tem se investe em estatística → melhor a informação dos dados.

Exemplo: Vamos jogar dado

Probabilidade

Temos que P(5) = 1/6, qual a P(5) 20 vezes em 100 tentativas?

Se não for tendencioso, qual a P(n #par em 100 tentativas)?

Teoria → Dados

Estatística

Tento 20 vezes o 5 em 100 tentativa, qual é P(5)?

Determinação de parâmetros

Se der 60 #par em 100 tentativas, isso é tendencioso?

Ajuste de dados

P(#par) = 2/3?

Teste de hipóteses

Dados → Teoria

Por que precisamos de incertezas?

- Interfere na conclusão dos nossos resultados
  - Pro exemplo: Resultado/Teoria = 0,970

Se 0,970 ± 0,050, dados compatíveis com a teoria

Se 0,970 ± 0,005, dados incompatíveis com a teoria

Se 0,970 ± 0,07, precisamos de um experimento melhor

Conhecem o experimento feito para testar a Relatividade Geral em Harwell na década de 60?

K. Hentschel, "Measurements of Gravitational Redshift Between 1959 and 1971", Annals of Science 53, pp. 269-295, 1996.

Incertezas sistemáticas + estatísticas

Veja o pêndulo por exemplo: $g = 4\pi^2 L / \tau^2$, $\tau = T/n$

- Estatísticas/Randômicas: acurácia imitada, tem resultados espalhados a cada repetição (método de estimativa) T, L
- Sistemáticas: Mais provável causar deslocamento ao invés de resultados espalhados T, L

Ao calibrar o instrumento Sistemática → Estatística

Existem mais sistemáticos: amplitude pequena, rigidez do fio, correção para g ao nível do mar, etc

Uma possibilidade de cancelar o sistemático dá-se ao fazer a razão de g em locais diferentes.

Apresentação de resultados

Apresentação de resultados: $g \pm \sigma_{esta} \pm \sigma_{sist}$

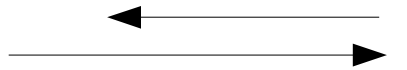
Ou com as incertezas combinadas em quadratura: $g \pm \sigma$

Pode-se também apresentar todas as incertezas sistemáticas separadamente, mas é muito raro. Isso é utilizado para ter acesso a correlação com outras medidas

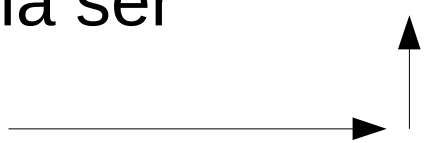
Combinação de incertezas

$$z = x - y$$

$$\delta_z = \delta_x - \delta_y [1] \qquad \sigma_z^2 = \sigma_x^2 - \sigma_y^2 [2]$$

1. [1] é para casos específicos 

   Também poderia ser 

   ou até mesmo  ?

2. $$\sigma_z^2 = \overline{\delta_x^2} + \overline{\delta_y^2} - 2\overline{\delta_x \delta_y}$$
   $$= \sigma_x^2 - \sigma_y^2$$

Combinação de incertezas

3. O cálculo da média é o suficiente: N medidas $x_i \pm \sigma$

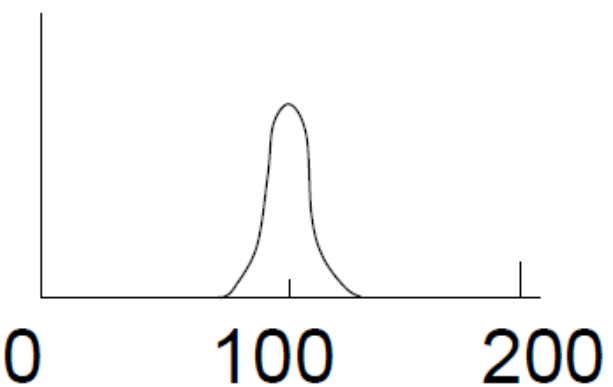
   [1] $x_i \pm \sigma$ ou [2] $x_i \pm \sigma/\sqrt{N}$?

4. Vamos jogar moeda

   Caso tire cara = 0 e coroa = 2 (1±1)

   Depois de 100 jogadas,

   [1] 100 ± 100 ou [2] 100 ± 10 ?



Prob (0 ou 200) = $(1/2)^{99} \sim 10^{-30}$

Compare com a idade do universo $\sim 10^{18}$ segundos

Propagação de erros para diferentes funções

- **Ver capítulo 4 de V. Oguri, et. al., Estimativas e erros em experimentos de Física, 2013**

Em geral: $u = f(x, y)$

$$\sigma_{\bar{u}}^2 = \left(\frac{\partial f}{\partial x}\right)^2\Bigg|_{(\bar{x},\bar{y})} \sigma_{\bar{x}}^2 + \left(\frac{\partial f}{\partial y}\right)^2\Bigg|_{(\bar{x},\bar{y})} \sigma_{\bar{y}}^2 + \frac{2}{N}\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\Bigg|_{(\bar{x},\bar{y})} \sigma_{xy}$$

Propagação de erros para diferentes funções

- **Ver capítulo 4 de V. Oguri, et. al., Estimativas e erros em experimentos de Física, 2013**

$$\bar{u} = f(\bar{x}, \bar{y})$$

i) $u = x \pm y \longrightarrow \sigma_{\bar{u}} = \sqrt{\sigma_{\bar{x}}^2 + \sigma_{\bar{y}}^2 \pm 2r\sigma_{\bar{x}}\sigma_{\bar{y}}}$

ii) $u = xy$

ou

$u = x/y$ $\longrightarrow \dfrac{\sigma_{\bar{u}}}{|\bar{u}|} = \sqrt{\left(\dfrac{\sigma_{\bar{x}}}{\bar{x}}\right)^2 + \left(\dfrac{\sigma_{\bar{y}}}{\bar{y}}\right)^2 \pm 2r\left(\dfrac{\sigma_{\bar{x}}}{\bar{x}}\right)\left(\dfrac{\sigma_{\bar{y}}}{\bar{y}}\right)}$

Combinação de resultados

- **Ver capítulo 4 de V. Oguri, et. al., Estimativas e erros em experimentos de Física, 2013**

$$\bar{x} = \frac{\sum_{i=1}^{N} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}$$

$$\frac{1}{\sigma_{\bar{x}}^2} = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$$

ou

$$\sigma_{\bar{x}} = \frac{1}{\sqrt{\sum_{i=1}^{N} \frac{1}{\sigma_i^2}}}$$

Diferença entre média e adição

Suponha uma ilha isolada com número de habitantes constante. Quantas pessoas são casadas?

Número de homens casados = 100 ± 5 k

Número de mulheres casadas = 80 ± 30 k

Total = 180 ± 30 k

Média = 99 ± 5 k

Total = 198 ± 10 k

Concepção teóricas adicionais (inquestionáveis) melhoram a precisão da resposta

Distribuição binomial

Número N fixo de ensaios independentes

Podendo ter somente dois resultados: "sucesso" / "fracasso"

Qual é a probabilidade s de sucessos?

Exemplos de experimentos binomiais:

Jogue o dados 100 vezes. Sucesso = "6". Qual a probabilidade de termos 0, 1,. . . , 49, 50, . . . 100 sucessos?

A eficiência da reconstrução de traços = 98%. Para 500 traços, probabilidade que 490, 491, ...499 , 500

A distribuição angular é 1 + 0,7 cos θ? Qual a probabilidade de ter 52/70 eventos com cos θ > 0 ?

Distribuição binomial

$$P = \frac{N!}{(N-s)!s!} p^s (1-p)^{N-s}$$

Número esperado de sucessos $= \sum sP = Np$

Variância do número de sucessos $= Np(1-p)$

Se p ~ 0, variância ~ NP

Se p ~ 1, variância ~ N(1-p)

Distribuição binomial

Exemplo: Considere que numa grande rede de computadores, em 60% dos dias ocorre alguma falha. Construir a distribuição de probabilidades para a variável aleatória X = número de dias com falhas na rede, considerando o período de observação de três dias. (Suponha independência.)

N = 3 p = 0,6 1 - p = 0, 4

$$P = \frac{N!}{(N-s)!s!} p^s (1-p)^{N-s}$$

Distribuição binomial

Exemplo: N = 3 p = 0,6 1 - p = 0, 4

$$P = \frac{3!}{(3-s)!s!} 0,6^s (0,4)^{N-s}$$

$$P(S=0) = \frac{3!}{(3-0)!0!} 0,6^0 (1-0,6)^{3-0} = 1.0,6^0.0,4^3 = 0,064$$

$$P(S=1) = \frac{3!}{(3-1)!1!} 0,6^1 (1-0,6)^{3-1} = 3.0,6^1.0,4^2 = 0,288$$

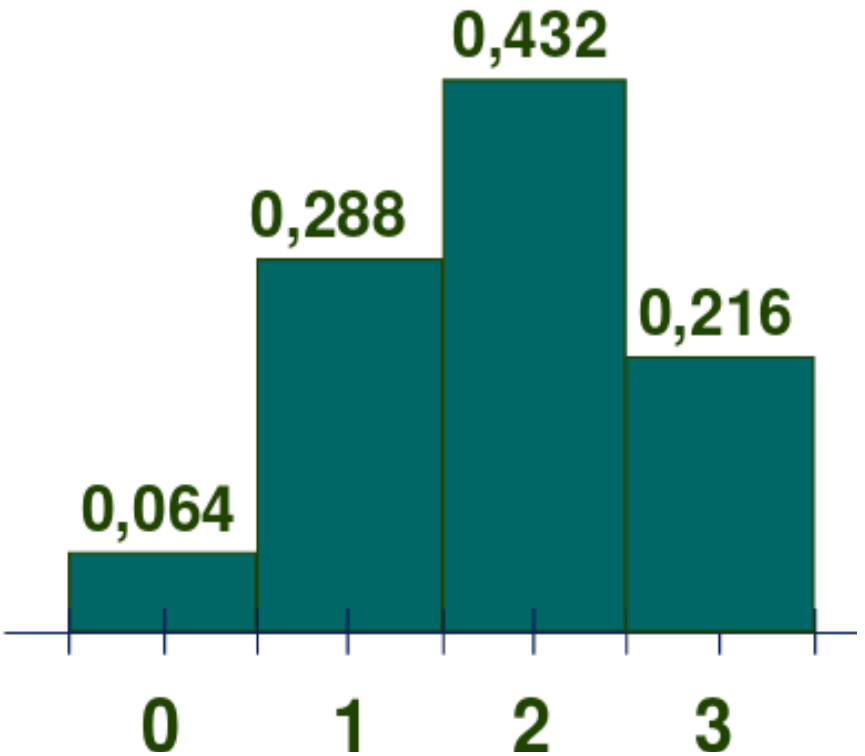
$$P(S=2) = \frac{3!}{(3-2)!2!} 0,6^2 (1-0,6)^{3-2} = 3.0,6^2.0,4^1 = 0,432$$

$$P(S=3) = \frac{3!}{(3-3)!2!} 0,6^3 (1-0,6)^{3-3} = 1.0,6^3.0,4^0 = 0,216$$

Distribuição binomial

Exemplo: N = 3 p = 0,6 1 – p = 0, 4

x	p(x)
0	0,064
1	0,288
2	0,432
3	0,216
Total	1



Distribuição binomial

Estatística: Estime p e σ_p tendo s (e N)?

$p = s/N$

$\sigma_p^2 = 1 / N \; s/N \; (1 - s/N)$

Casos limite:

- p = const., N → ∞: Binomial → Gaussiana

  Contínua →∞

  - $\mu = N p$, $\sigma_p^2 = N p (1-p)$
- N → ∞, p → 0, Np = const.: Binomial → Poisson
  - $\mu = N p$, $\sigma_p^2 = N p$

Distribuição de Poisson

Probabilidade de N eventos independentes ocorrerem num tempo t contínuo com uma taxa constante.

Exemplos: eventos in bin de histogramas (lembre do limite da Binomial)

Distribuição de Poisson

Probabilidade de N eventos independentes ocorrerem num tempo t contínuo com uma taxa constante.

Limite da Binomial

$$P(X = x) \approx \binom{n}{x} p^x (1-p)^{n-x}$$

$$n \mapsto \infty$$
$$p \mapsto 0$$
$$n\,p \mapsto \lambda > 0$$

$$P(X = x) \longrightarrow \frac{\lambda t^x e^{-\lambda t}}{x!} \qquad (x = 0, 1, 2, \ldots)$$

Distribuição de Poisson

As probabilidade de uma distribuição de Poisson:

$$P_x = \frac{e^{-\lambda t}\lambda t^x}{x!} = e^{-\mu}\mu^x/x!$$

$$< n >= t = \mu$$

$$\sigma_n^2 = \mu \rightarrow \mathbf{x} \pm \sqrt{\mathbf{x}}$$

Binomial

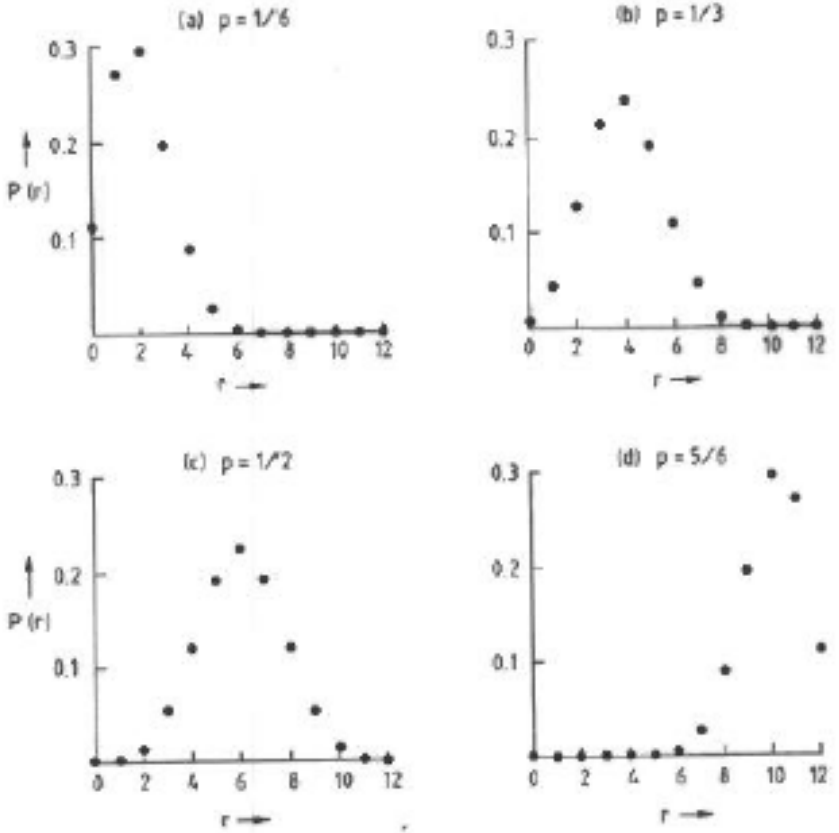


Fig. A3.1 The probabilities $P(r)$, according to the binomial distribution, for r successes out of 12 independent trials, when the probability p of success in an individual trial is as specified in the diagram. As the expected number of successes is $12p$, the peak of the distribution moves to the right as p increases. The RMS width of the distribution is $\sqrt{12p(1-p)}$ and hence is largest for $p = \frac{1}{2}$. Since the chance of success in the $p = \frac{1}{6}$ case is equal to that of failure for $p = \frac{5}{6}$, the diagrams (a) and (d) are mirror images of each other. Similarly the $p = \frac{1}{2}$ situation shown in (c) is symmetric about $r = 6$ successes.

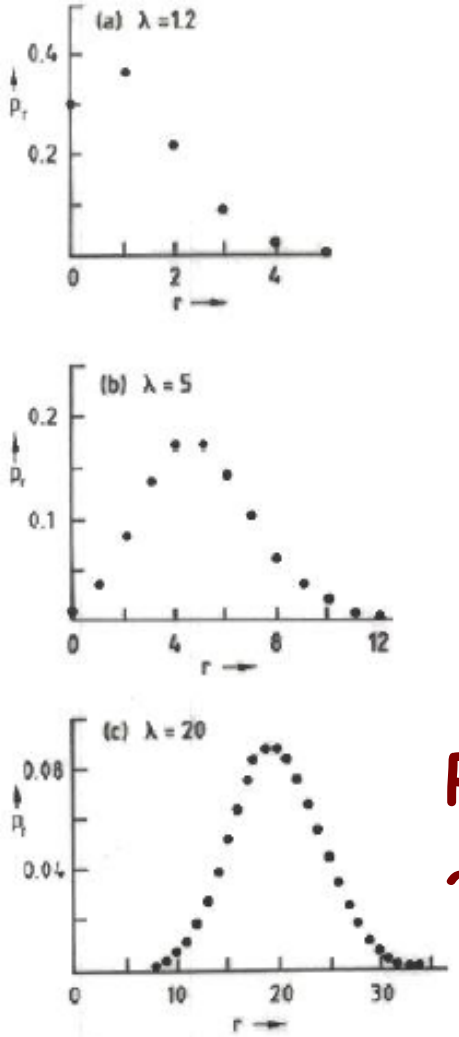
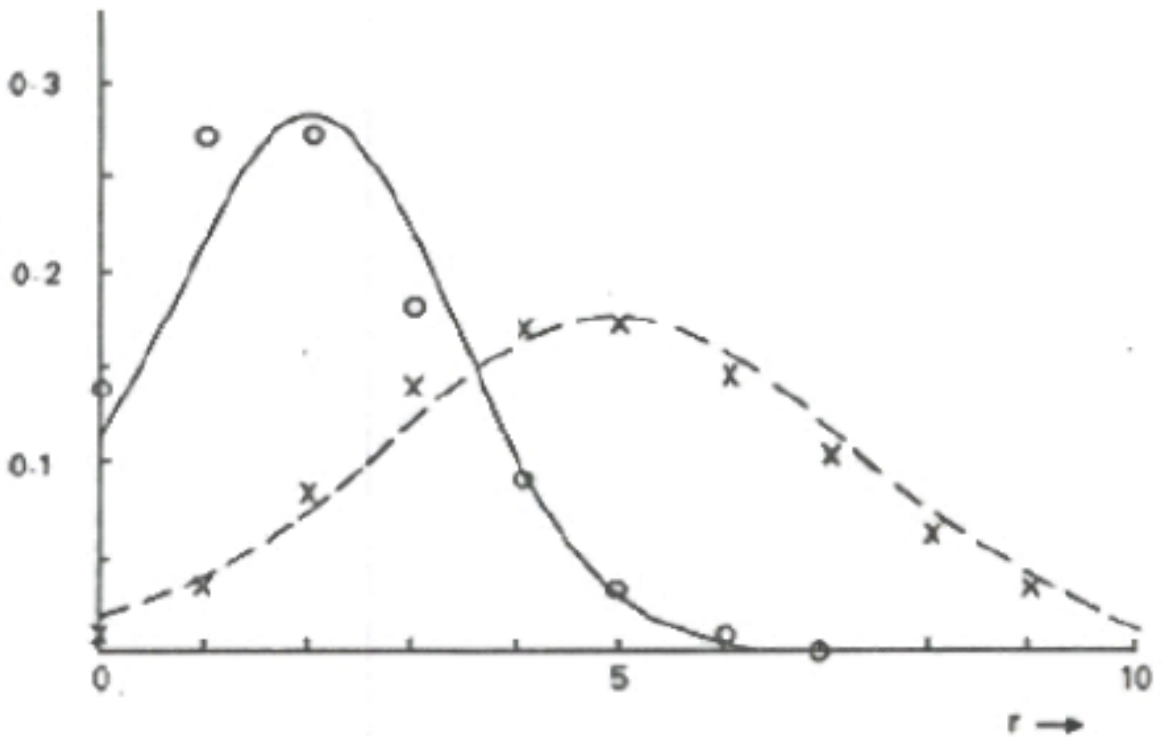


Fig. A4.1 Poisson distributions for different values of the parameter λ. (a) $\lambda = 1.2$; (b) $\lambda = 5.0$; (c) $\lambda = 20.0$. P_r is the probability of observing r events. (Note the different scales on the three figures.) For each value of λ, the mean of the distribution is at λ, and the RMS width is $\sqrt{\lambda}$. As λ increases above about 5, the distributions look more and more like Gaussians.

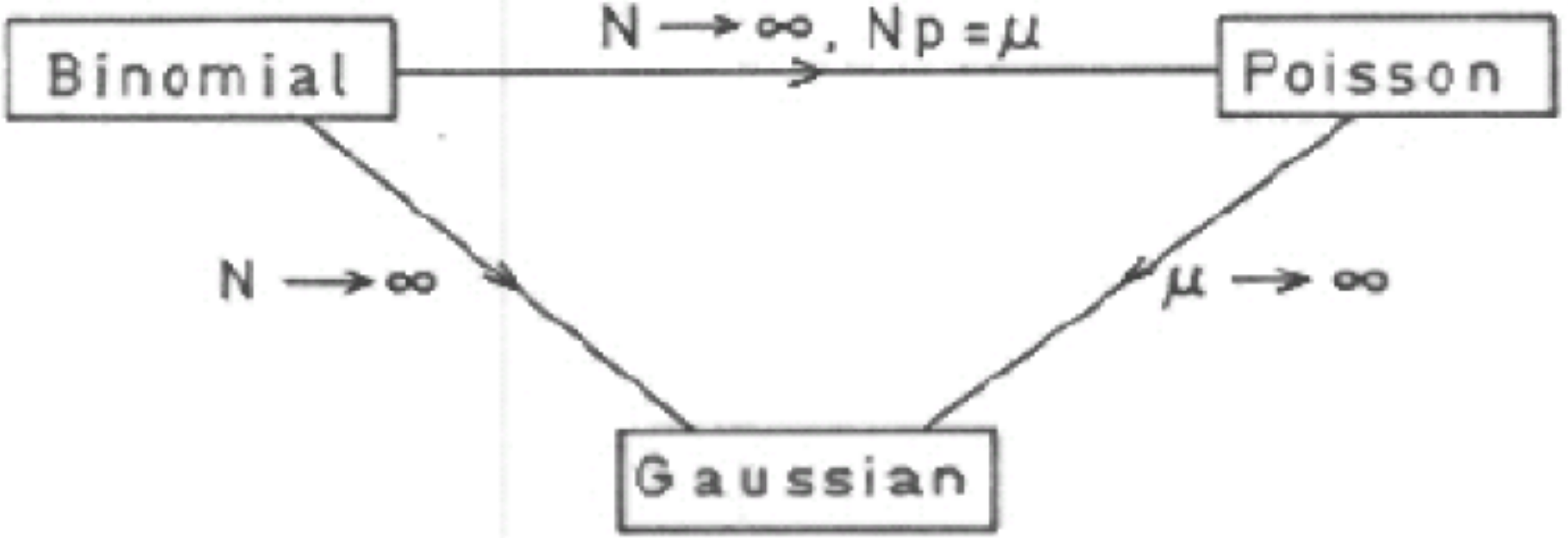
Poisson, ~ gaussiana

Relevante para o melhor acordo do ajuste



Poisson

Gaussian

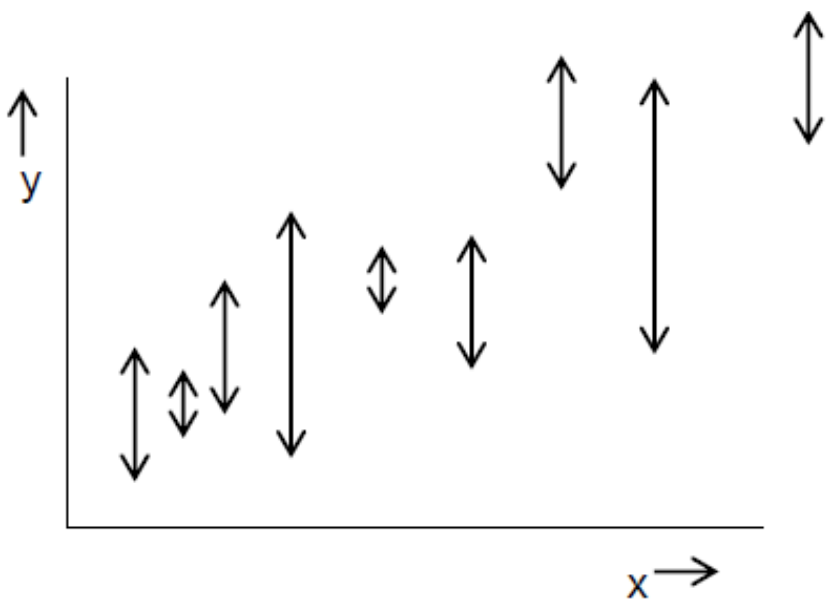

Binomial
N → ∞, Np = μ
Poisson
N → ∞
μ → ∞
Gaussian

Ajuste de funções

Vamos discutir o problema de obter a melhor descrição dos dados em termos de alguma teoria, que possuem parâmetros cujos valores não são conhecidos inicialmente.

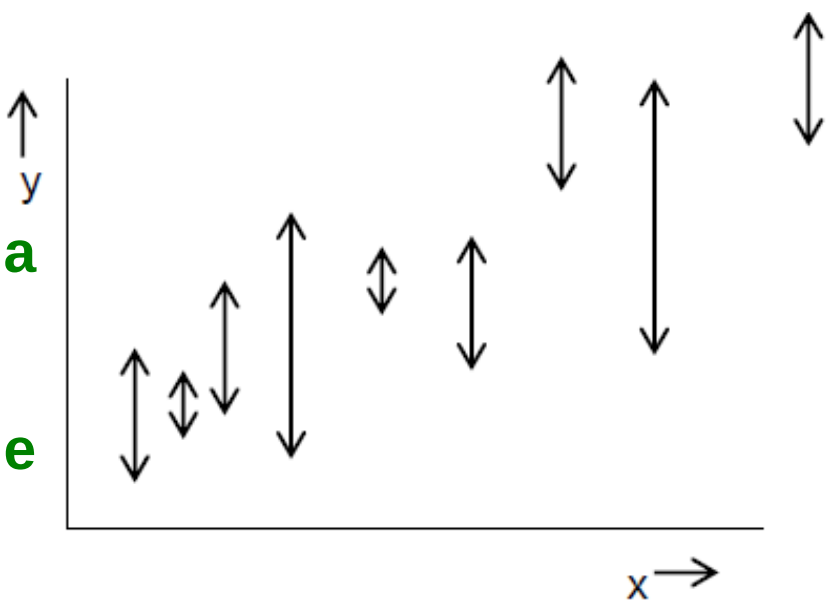
Dados: $\{x_i, y_i \pm \sigma_i\}$

Teoria : $y = ax + b$



Ajuste de funções

Vamos discutir o problema de obter a melhor descrição dos dados em termos de alguma teoria, que possuem parâmetros cujos valores não são conhecidos inicialmente.



1) Os dados são consistentes com a teoria? Concordância do ajuste

2) Quais sãos os coeficientes angular e linear? Determinação de parâmetros

Esse método não é único e pode ser utilizado com outras funções!

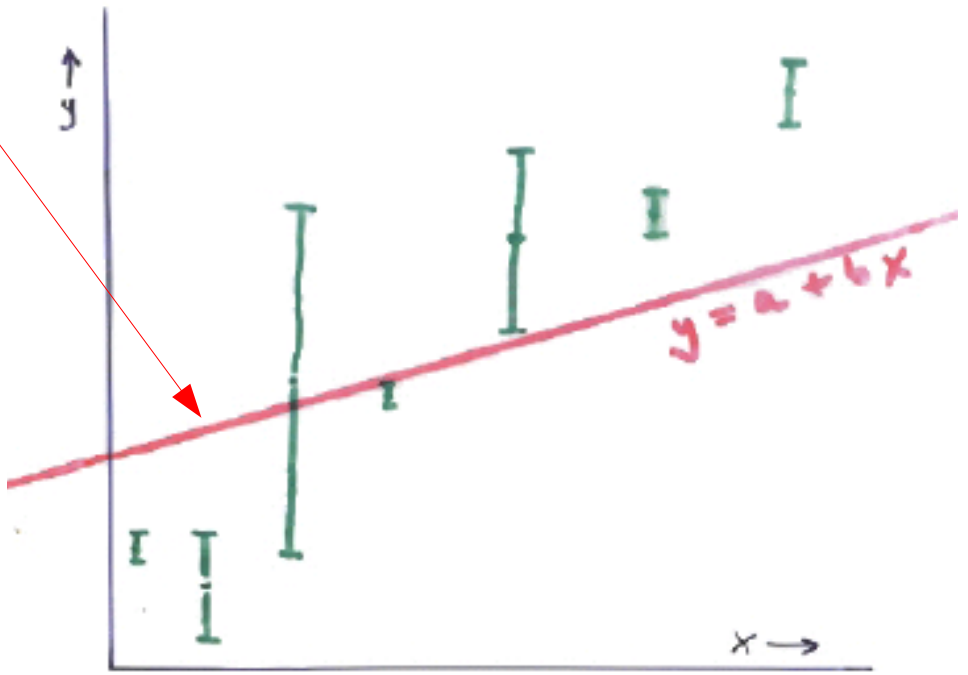
Ajuste de funções

Esse é o melhor ajuste possível?

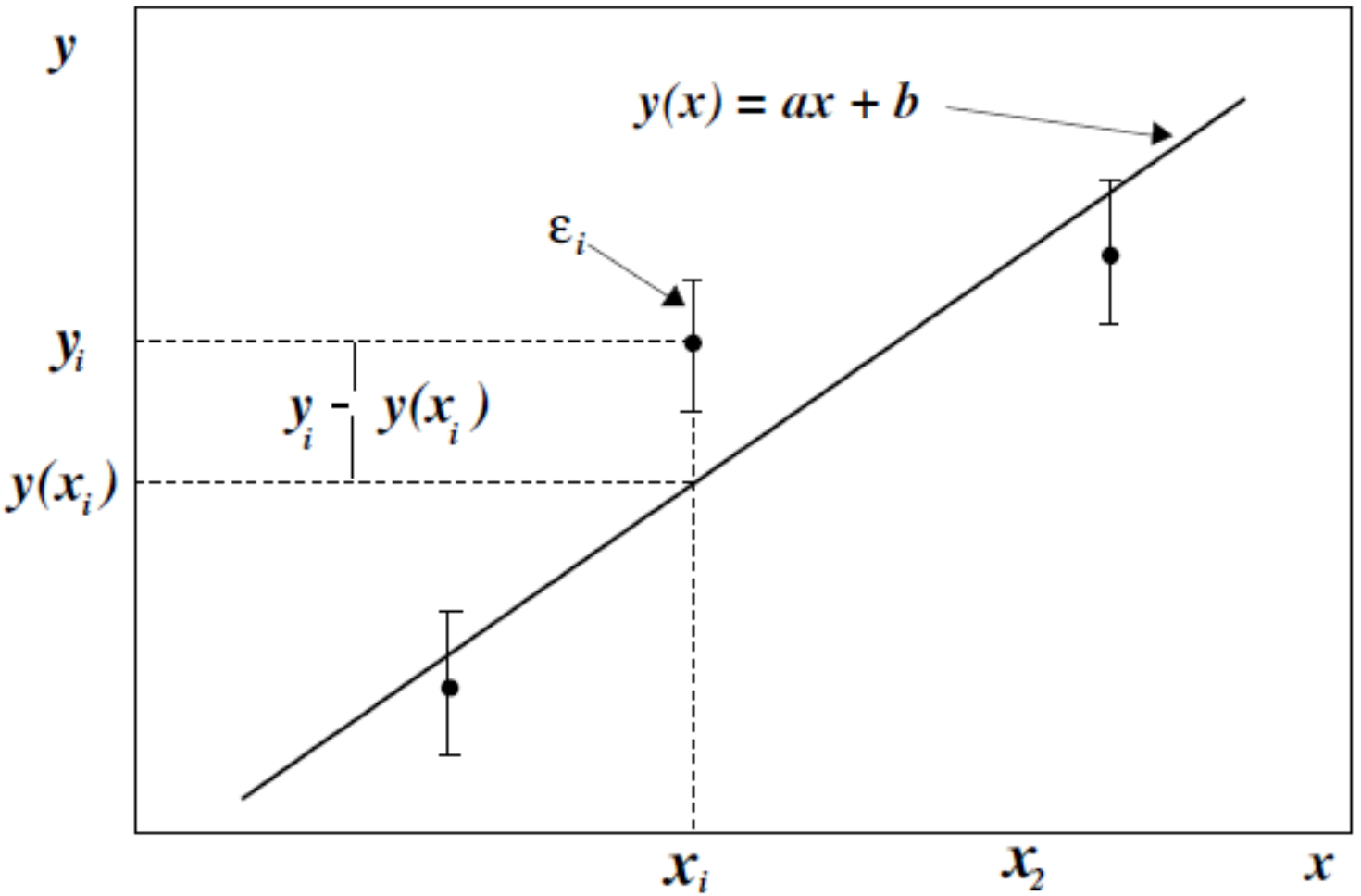
Para encontrar o melhor ajuste, é preciso minimizar os desvios entre o valor observado e o predito

$$\varepsilon_i = Y_i^{obs} - [ax_i + b]$$

Exercício: Minimize a soma dos quadrados dos desvios e encontre as expressões para os parâmetros a e b



Ajuste de funções



Ajuste de funções

- No caso anterior assumimos que as incertezas nas medidas de y e x são constantes. Em geral devemos considerar o erro em cada medida (σ_i):

$$S(a,b) = \sum_{i=1}^{N} \left(\frac{y_i - y(x_i)}{\sigma_i} \right)^2 = \sum_{i=1}^{N} \left[\frac{y_i - (ax_i + b)}{\sigma_i} \right]^2$$

Erro efetivo em cada medida

Ajuste de funções

- Podemos mostrar (Exercício - Ver Apêndice F do livro texto) que as estimativas dos parâmetros e suas incertezas são dadas por:

$$a = r\frac{\sigma_y}{\sigma_x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$b = \bar{y} - a\bar{x}$$

$$\sigma_a = \frac{1}{\sigma_x}\frac{\epsilon_y}{\sqrt{N}}$$

$$\sigma_b = \sigma_a\sqrt{\overline{x^2}}$$

$$\epsilon_y = \sqrt{\sum_{i=1}^{N}\frac{[y_i - (ax_i + b)]^2}{N-2}} = \sigma_y\sqrt{\frac{N}{N-2}(1-r^2)}$$

Ajuste de funções

- Plote os dados
- Determine os parâmetros com seus erros

  a e b, por exemplo.
- Veja se o χ^2 é bom

O teste do χ^2 é um teste, não paramétrico, de hipótese para a qualidade de um ajuste, associado à frequência de observação ou às próprias medidas de uma grandeza. Avaliar erros aleatórios.

Ajuste de funções

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i^{obs} - y_i^{esp}}{\sigma_i} \right)^2$$

Karl Pearson

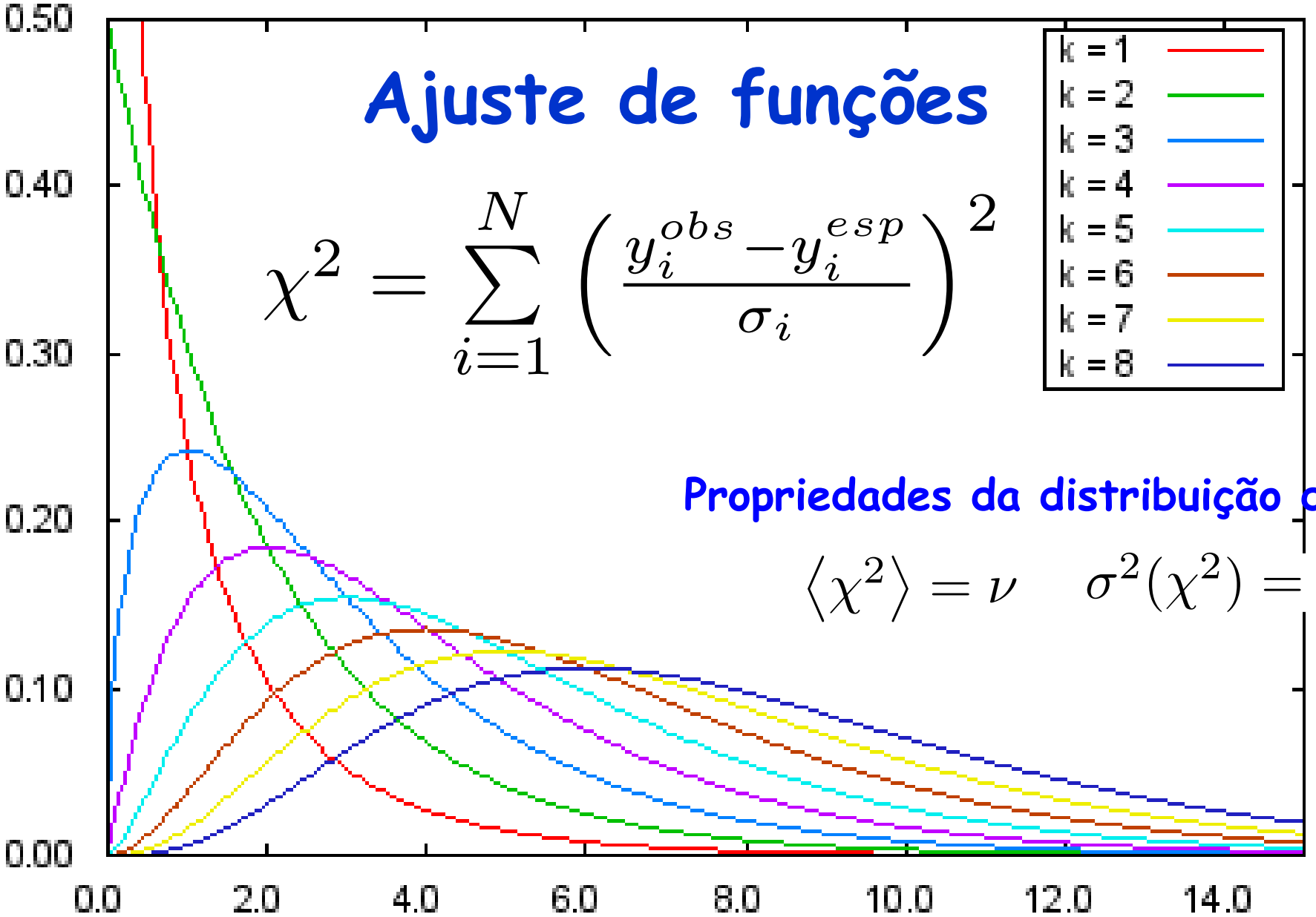
- Usualmente, y_i^{esp} dependem de p parâmetros (obtidos dos dados)
- Assim, na expressão de χ^2, apenas $\nu = N - p$ são termos independentes, número de graus de liberdade da distribuição

Distribuição de χ^2

- Grau de liberdade
  - Consideremos que 10 estudantes obtiveram em um teste média 8,0. Assim, a soma das 10 notas deve ser 80 (restrição). Portanto, neste caso, temos um grau de liberdade de 10 – 1 = 9, pois as nove primeiras notas podem ser escolhidas aleatoriamente, contudo a 10ª nota deve ser igual a [80 – (soma das 9 primeiras)].

Ajuste de funções

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i^{obs} - y_i^{esp}}{\sigma_i} \right)^2$$



Propriedades da distribuição do χ^2

$$\langle \chi^2 \rangle = \nu \qquad \sigma^2(\chi^2) = 2\nu$$

Ajuste de funções

- Aceita-se a validade da hipótese de que uma função seja adequada para a determinação de valores esperados, quando: $\frac{\chi^2}{\nu} \sim 1$
- No caso de um ajuste linear (v = N - 2), $S_{min} = \chi^2$

$$\frac{\chi^2}{\nu} = \frac{1}{N-2}\frac{\sigma_y^2}{\sigma^2}\left(1 - r^2\right) \sim 1$$

O teste do χ^2 permite uma análise sobre a subestimação ou sobrestimação dos erros nos N pares de medidas das grandezas envolvidas.

Tabela do x^2

Degrees of freedom (df)	χ^2 value[18]										
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	4.87	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

Frequentista e Bayesiana

- A diferença básica
  - Bayesiana: Probabilidade (parâmetros, a partir dos dados)
    - Grau de liberdade, aplica-se a um único evento ou constante física
  - Frequentista: Probabilidade (dados, a partir dos parâmetros)
    - Frequências ($n \to \infty$), não aplica-se a um único evento ou constante física

Frequentista e Bayesiana

- Bayesiana:
  - "Bayesians abordar a questão em que todos estão interessados, usando suposições que ninguém acredita"
- Frequentista:
  - “Frequentistas usam a lógica de forma impecável para lidar com um problema que não interessa a ninguém”