



Introdução à análise de dados em FAE

Estatística básica - 1

PROFESSORES:

SANDRO FONSECA DE SOUZA

SHEILA MARA DA SILVA

ELIZA MELO DA COSTA

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

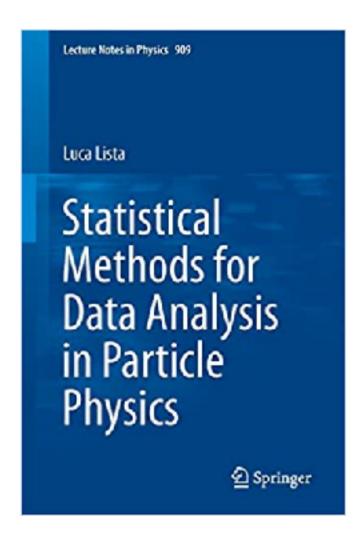
Livros de referência

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

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

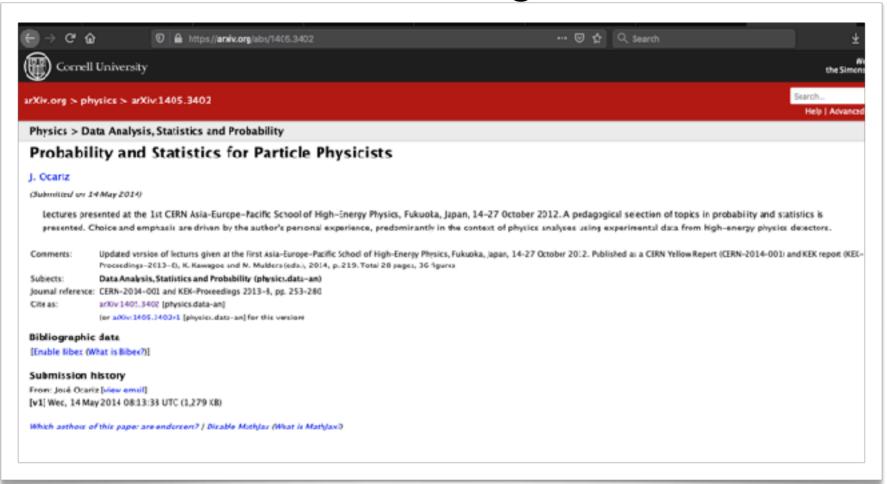
V. Oguri, et. al., Estimativas e erros em experimentos de Física, 2013

Bibliografia Sugerida





Leitura Sugerida



Tópicos

- 1) Introdução
- 2) χ^2
- 3) Estatística Frequentista e Bayesiana (na próxima aula)

Introdução

O que é estatística?

Probabilidade e estatística

Por que incertezas?

Incertezas sistematicas e estatísticas

Combinação de incertezas

Combinando dados de diferentes experimentos

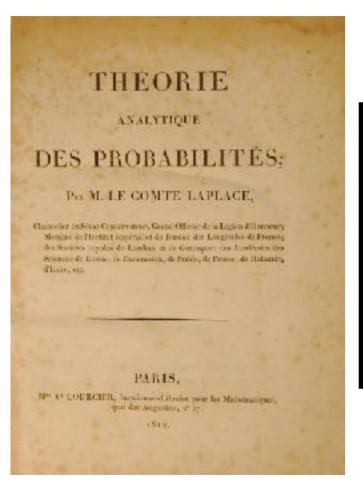
Distribuições: Binomial, Poisson e Gaussiana

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.

Definição de Probabilidade



A teoria dos jogos de azar consiste em reduzir todos os eventos do mesmo género a um certo número de casos igualmente possíveis



Pierre Simon Laplace

Exemplo: Vamos jogar dados

Probabilidade

Estatística

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

Determinação de parâmetros

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

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

Ajuste de dados

P(#par) = 2/3?

Teoria → Dados

Teste de hipóteses

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?

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 dáse 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

$$z = x - y$$

$$\delta_z = \delta_x - \delta_y[1]$$

$$\sigma_z^2 = \sigma_x^2 - \sigma_y^2[2]$$

1.[1] é para casos específicos ______

Também poderia ser

ou até mesmo

1.
$$\sigma_z^2=\overline{\delta_x^2}+\overline{\delta_y^2}-2\overline{\delta_x\delta_y}$$

$$=\sigma_x^2-\sigma_y^2$$

Combinação de incertezas

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

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

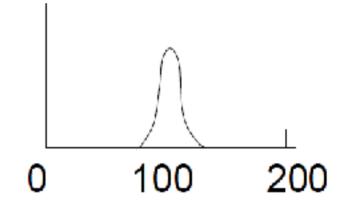
4. Vamos jogar moeda

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

Depois de 100 jogadas,

[1]
$$100 \pm 100$$
 ou [2] 100 ± 10 ?

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



Compare com a idade do universo ~1018 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:
$$u = f(x, y)$$

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

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
$$u=x/y$$

$$u=x/y$$

$$u=x/y$$
 16

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

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

Veja exemplo no backup

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 \pm 30 \text{ k}
```

Média =
$$99 \pm 5 k$$

Total =
$$198 \pm 10 \text{ k}$$

Concepção teóricas adicionais (inquestionáveis) melhoram a precisão da resposta

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 ?

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

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)

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

Exemplo: N = 3 p = 0,6 1 - p = 0,4
$$P = \frac{3!}{(3-s)!s!}0, 6^s(0,4)^{N-s}$$

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

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

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

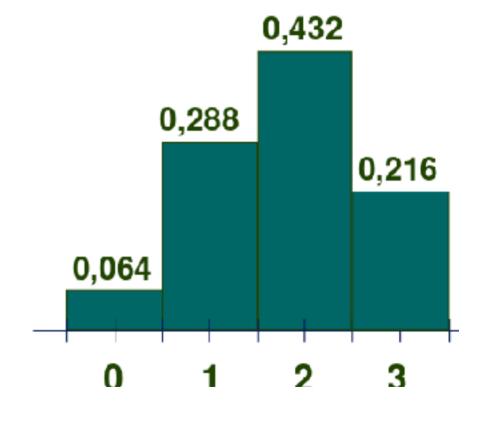
$$P(S=3) = \frac{3!}{(3-3)!2!}0, 6^3(1-0,6)^{3-3} = 1.0, 6^3.0, 4^0 = 0,216$$

Exemplo:
$$N = 3$$

$$p = 0.6$$

$$p = 0.6$$
 $1 - p = 0.4$

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



Estatística: Estime p e σ_{D} tendo s (e N)?

$$p = s/N$$

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

• $\mu = N p$, $\sigma_p^2 = N p$

Casos limite:



 $\rightarrow \infty$

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.

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

$$n \mapsto \infty$$

$$p \mapsto 0$$

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

Limite da Binomial

$$n \mapsto \infty$$

$$p \mapsto 0$$

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

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

$$(x=0, 1, 2, ...)$$

Distribuição de Poisson

As probabilidade de uma distribuição de Poisson:

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

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

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

Binomial

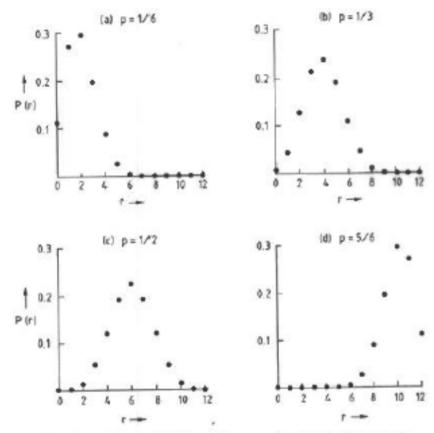


Fig. A3.1 The probabilities P(r), according to the binomial distribution, for r successes out of 12 independent trials, when the probability p of success in an individual trial is as specified in the diagram. As the expected number of successes is 12p, the peak of the distribution moves to the right 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 $p=\frac{\pi}{6}$, 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.

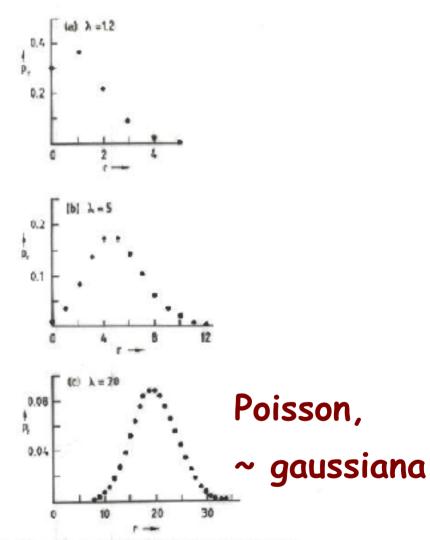
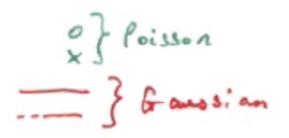
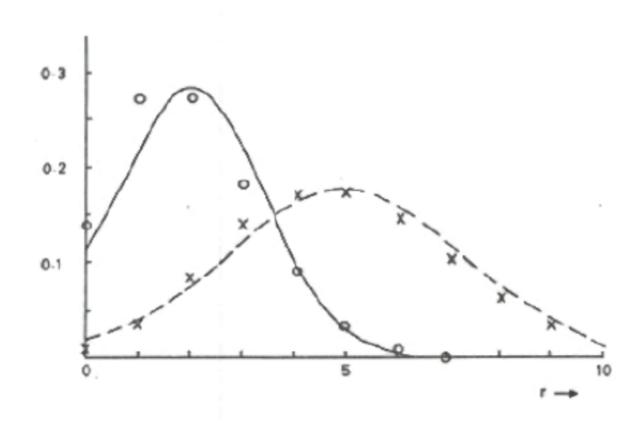
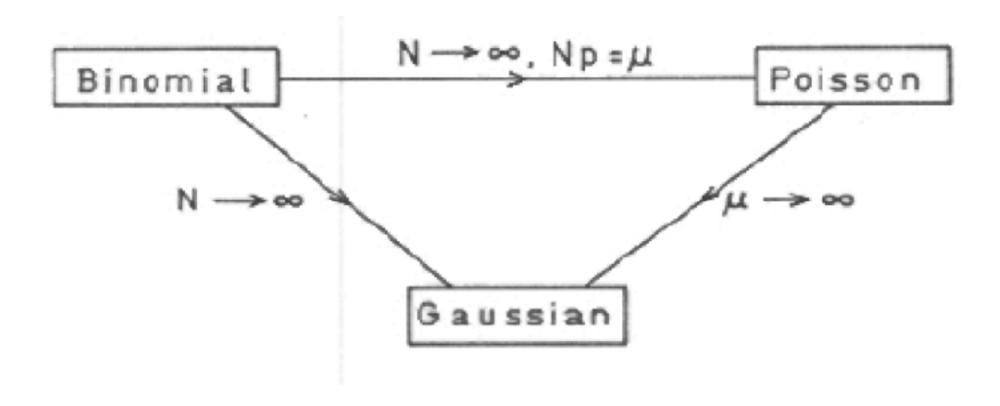


Fig. A4.1 Poisson distributions for different values of the parameter λ . (a) $\lambda = 1.2$; (b) $\lambda = 5.0$; (c) $\lambda = 20.0$. F_r is the probability of observing τ events. (Note the different scales on the three figures.) For each value of λ , the mean of the distribution is at λ , and the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

Relevante para o melhor acordo do ajuste







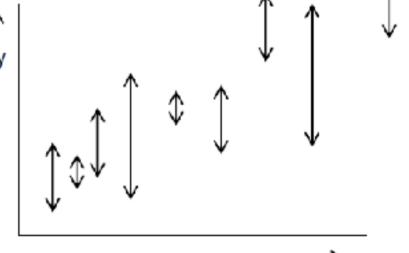
Exemplos práticos do limite das distribuições

- Decaimento radioativo -> Distribuição de Poisson
- Decaimento de partícula (ns) + background associado (nb) a contagem total (n) : n=ns+nb obedece a distribuição de Poisson

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

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

Teoria: y = ax + b



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

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

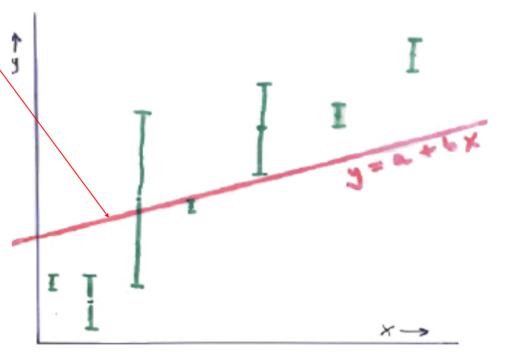
Esse método não é único e pode ser utilizado com outras funções!

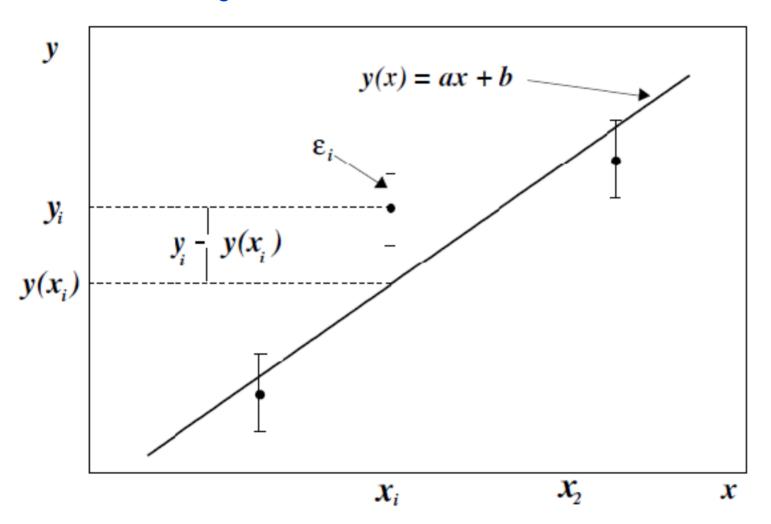
Esse é o melhor ajuste possível?

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

$$\varepsilon_i = Y_i^{\text{obs}} - [\alpha x_i + b]$$

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





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

$$S\left(a,b
ight) = \sum_{i=1}^{N} \left(rac{y_i - y\left(x_i
ight)}{\sigma_{\!i}}
ight)^2 = \sum_{i=1}^{N} \left[rac{y_i - (ax_i + b)}{\sigma_i}
ight]^2$$
 Erro efetivo em cada medida

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

$$a=rrac{\sigma_{oldsymbol{y}}}{\sigma_{oldsymbol{x}}}=rac{\sigma_{oldsymbol{x}oldsymbol{y}}}{\sigma_{oldsymbol{x}}^2}$$

$$b=ar{y}-aar{x}$$

$$\sigma_{m{a}} = rac{1}{\sigma_{m{x}}} rac{\epsilon_{m{y}}}{\sqrt{N}}$$

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

$$\epsilon_{m{y}} = \sqrt{\sum_{i=1}^{N} rac{\left[y_i - (ax_i + b)
ight]^2}{N-2}} = \sigma_{m{y}} \sqrt{rac{N}{N-2} \left(1 - r^2
ight)}$$

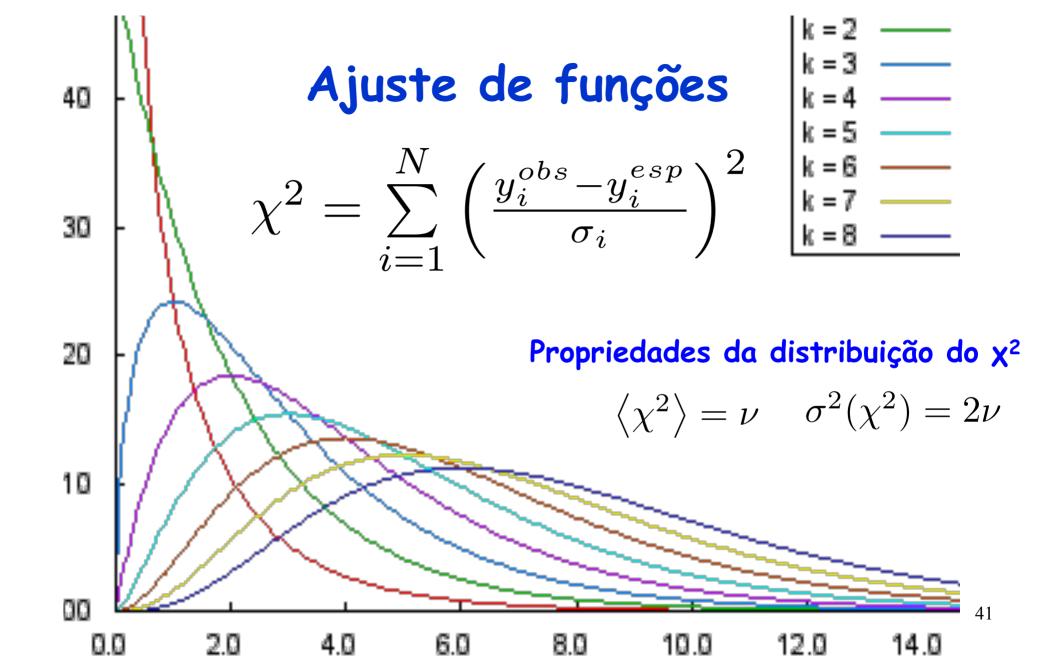
- Plote os dados
- Determine os parâmetros com seus erros a e b, por exemplo.
- Veja se o x² é bom
- O teste do χ^2 é um teste, não paramétrico, de hipótese para a qualidade de um ajuste, associado à frequência de observação ou às próprias medidas de uma grandeza. Avaliar erros aleatórios.

$$\chi^2 = \sum_{i=1}^{N} \left(rac{y_i^{obs} - y_i^{esp}}{\sigma_i}
ight)_{ ext{Karl Pearson}}^2$$

- Usualmente, y_i^{esp} dependem de p parâmetros (obtidos dos dados)
- Assim, na expressão de x², apenas v = N p são termos independentes, número de graus de liberdade da distribuição

Distribuição de x²

- · 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 10^a nota deve ser igual a [80 (soma das 9 primeiras)].



- 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 aiuste linear (v = N 2). S $_{\rm nin}$ = x² $\frac{\chi^2}{\nu}=\frac{1}{N-2}\frac{\sigma_y^2}{\sigma^2}(1-r^2)\sim 1$

O teste do x² permite uma análise sobre a subestimação ou sobrestimação dos erros nos N pares de medidas das grandezas envolvidas.

42

Tabela do x²

Degrees of freedom (df)	χ² 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

Tabela do x²

Degrees of freedom (df) χ^2 value ^[18]											
The <u>p-value</u> is the probability of observing a test statistic at least as extreme in a chi-square distribution.											10.83
Accordingly, since the <u>cumulative distribution function</u> (CDF) for the appropriate											13.82
degrees of freedom (df) gives the probability of having obtained a value less extreme than this point, subtracting the CDF value from 1 gives the p -value. A											16.27
low p -value, below the chosen significance level, indicates statistical										18.47	
significance, i.e., sufficient evidence to reject the null hypothesis. A significance level of 0.05 is often used as the cutoff between significant and non-significant											
results.											22.46
The table below gives a number of <i>p</i> -values matching to for the first 10 degrees of freedom										24.32	
Of ITCCGOIT											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-→∞), 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"

backup slide

Combinação de resultados compatíveis

A partir de várias estimativas independentes $\{x_i\}$ do valor esperado de uma grandeza e respectivos erros padrão $\{\sigma_i\}$, o resultado *combinado* pode ser obtido da seguinte forma:

Estimativa padrão para o valor esperado:

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

Baseado nos slides do curso de Física Geral do Prof. Antonio Vilela-UERJ/DFNAE Erro padrão associado:

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

ou

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

Combinação de resultados compatíveis

A partir de várias estimativas independentes $\{x_i\}$ do valor esperado de uma grandeza e respectivos erros padrão $\{\sigma_i\}$, o resultado *combinado* pode ser obtido da seguinte forma:

Exemplo:

Estimativa I: $\bar{x}_1 \pm \sigma_{\bar{x}_1}$

Estimativa 2: $\bar{x}_2 \pm \sigma_{\bar{x}_2}$

$$\sigma_{\bar{x}} = \sigma = \frac{1}{\sqrt{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}}$$

$$\bar{x} = \sum_{i=1}^{N} \left(\frac{\sigma}{\sigma_i}\right)^2 x_i = \left(\frac{\sigma}{\sigma_1}\right)^2 x_1 + \left(\frac{\sigma}{\sigma_2}\right)^2 x_2$$

Baseado nos slides do curso de Física Geral do Prof. Antonio Vilela-UERJ/DFNAE Exercício (3.7.9): Dois experimentos (D0 e CDF) mediram a massa do quark top. As medições são dadas por:

$$m_t(D0) = (179,0 \pm 5,1) \text{ GeV/c}^2$$

 $m_t(CDF) = (176,1 \pm 6,6) \text{ GeV/c}^2$

Qual o resultado combinado dos dois experimentos para a massa do quark top?

i) Erro padrão da combinação de $m_t(D0)$ e $m_t(CDF)$:

$$\sigma$$
 = 4,03555 GeV/c²

$$\sigma_{\bar{x}} = \sigma = \frac{1}{\sqrt{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}}$$

ii) Estimativa padrão do valor esperado da combinação de $m_t(D0)$ e $m_t(CDF)$:

$$ar{x} = 177,916 ext{ GeV/c}^2$$
 $ar{x} = \sum_{i=1}^N \left(rac{\sigma}{\sigma_i}
ight)^2 x_i = \left(rac{\sigma}{\sigma_1}
ight)^2 x_1 + \left(rac{\sigma}{\sigma_2}
ight)^2 x_2$

Estimativa padrão para o resultado da medição:

$$m_t = (177.9 \pm 4.0) (GeV/c^2)$$