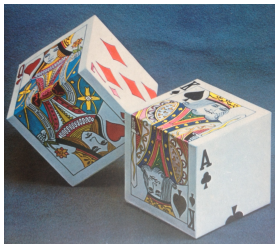


# Fundamentos Estatísticos para Ciência dos Dados

## Ajuste de distribuições

Renato Martins Assunção

DCC, UFMG - 2015



# Teste qui-quadrado

- Compara teórico e observado.
- Testa se os dados de uma amostra  $Y_1, Y_2, \dots, Y_n$  seguem uma certa distribuição de probabilidade (ou modelo teórico).
- O modelo teórico pode ser qualquer distribuição de probabilidade, contínua ou discreta.
- Isto é, observamos amostra  $Y_1, Y_2, \dots, Y_n$ .
- Supomos que estes dados são i.i.d. e com distribuição teórica  $F(y)$ : este é o modelo.
- $F(x)$  poderia ser uma  $\text{Bin}(20, 0.1)$ , ou uma  $\text{Poisson}(5)$ , ou uma  $\text{exp}(10)$ , ou uma  $N(0, 1)$ .

# Teste qui-quadrado

- Amostra  $Y_1, Y_2, \dots, Y_n$  são v.a.'s iid de um modelo teórico  $F(y)$ ?
- Como verificar se isto procede?
- O teste qui-quadrado é uma maneira de fazer isto.
- Ele possui duas vantagens: pode ser usado com distribuições contínuas OU discretas; e ele sabe como lidar com quantidades estimadas a partir dos dados (mais sobre isto + tarde).

# Teste qui-quadrado: PASSO 1

- Particione o conjunto de valores possíveis de  $Y$  em  $N$  categorias (ou intervalos). Por exemplo:
- Se o modelo teórico é uma  $\text{Bin}(20, 0.1)$ , podemos criar 5 categorias de valores possíveis:  $Y = 0$ ,  $Y = 1$ ,  $Y = 2$ ,  $Y = 3$  e  $Y \geq 4$ .
- Modelo é uma  $\text{Poisson}(5)$ , com 12 categorias:  $Y = 0$ ,  $Y = 1, \dots, Y = 10$  e  $Y \geq 11$ .
- Modelo é uma  $\text{exp}(10)$ , com 5 categorias-intervalos:  $[0, 0.05)$ ,  $[0.05, 0.1)$ ,  $[0.1, 0.2)$ ,  $[0.2, 0.4)$ ,  $[0.4, \infty)$
- Modelo é  $N(0, 1)$ , com 4 categorias-intervalos:  $(-\infty, -2)$ ,  $[-2, -1)$ ,  $[-1, 0)$ ,  $[0, 1)$ ,  $[1, 2)$ , e  $(2, \infty)$ .
- Em princípio os intervalos-categorias são arbitrários mas na prática escolhemos de forma que não tenham nem probabilidades muito altas nem muito altas.

## Teste qui-quadrado: PASSO 2

- Para cada  $k$ -ésimo intervalo-categoria, calcule o número  $N_k$  de elementos da amostra  $Y_1, Y_2, \dots, Y_n$  que caem no intervalo  $k$  (a frequência observada)
- Calcule também o número esperado  $E_k$  de observações que deveria cair no intervalo  $k$ .
- Isto é, calcule  $E_k = n \times P(Y \in \text{Intervalo } k)$ .
- Por exemplo, se o modelo teórico é uma  $\text{Bin}(20, 0.1)$ , se temos amostra de tamanho  $n = 53$  e se a categoria é  $Y = 0$ :
- Então o número esperado é
$$E = 53 * \mathbb{P}(Y = 0) = 53 * (1 - 0.1)^{20} = 6.44.$$
- Se observamos 53 repetições de uma  $\text{Bin}(20, 0.1)$  esperamos que 6.44 delas sejam iguais a zero.

## Outro exemplo

- Modelo teórico é uma  $\text{Poisson}(2)$ .
- Temos amostra de tamanho  $n = 97$ .
- A categoria é  $Y \geq 4$ .
- Então o número esperado nesta categoria é

$$E = 97 * \mathbb{P}(Y \geq 4) = 97 * \sum_{j=4}^{\infty} \frac{2^j \exp(-2)}{j!} = 13.86$$

- Se observamos 97 repetições indep de uma  $\text{Poisson}(2)$ , esperamos que 13.86 delas sejam maiores ou iguais a 4.

## Mais um exemplo

- Modelo teórico é uma  $\exp(10)$ .
- Temos amostra de tamanho  $n = 147$ .
- O intervalo-categoria é  $X \in [0.2, 0.4)$ .
- Então o número esperado de observações neste intervalo é

$$E = 147 * \int_{0.2}^{0.4} 10 \exp(-10x) dx = 17.20$$

- Repete-se o cálculo nos demais intervalos.

## Teste qui-quadrado: PASSO 3

- $E_k$  é o valor esperado **caso o modelo teórico seja verdadeiro**.
- Compare as frequências observadas  $N_k$  e as frequências esperadas  $E_k$ .
- Caso  $E_k$  e  $N_k$  sejam muito diferentes, isto é uma evidência de que o modelo teórico não é próximo da realidade.
- Caso  $E_k$  e  $N_k$  sejam parecidos, isto é evidência de que o modelo gera valores parecidos com os observados.



## Teste qui-quadrado: PASSO 3

- Isto quer dizer que os dados observados REALMENTE sigam o modelo teórico? NÃO.
- TRÊS razões para o NÃO:
  - 1 Suponha que temos uma única amostra e dois (ou mais) modelos diferentes: os valores teóricos dos modelos podem estar bem próximos dos valores observados e não termos nenhum deles claramente melhor que o outro.
  - 2 ESTE ASPECTO do modelo (as contagens nos intervalos) é próximo da realidade. Outros aspectos do modelo, quando comparados com a realidade, podem mostrar que o modelo não é adequado. Por exemplo, uma análise de resíduos num modelo de regressão linear pode mostrar alguns problemas que não são aparentes na comparação entre  $E_k$  e  $N_k$ .
  - 3 Finalmente, ninguém acredita que a realidade siga fielmente uma fórmula matemática perfeita. Precisamos apenas que a fórmula seja uma BOA APROXIMAÇÃO para a realidade.

## Teste qui-quadrado: AINDA O PASSO 3

- Como então comparar as frequências observadas  $N_k$  e as frequências esperadas  $E_k$ ?
- Podemos ter uma boa aproximação numa categoria mas uma péssima aproximação em outra categoria. Assim, precisamos de um resumo, uma idéia global de como é a aproximação em geral, considerando todas as categorias.
- A medida-resumo é uma espécie de “média” das diferenças  $|N_k - E_k|$ .
- Note a presença do valor absoluto  $|N_k - E_k|$  ao invés das diferenças  $N_k - E_k$ .
- Se a medida-resumo for pequena, então  $N_k \approx E_k$  e adotamos o modelo teórico.
- Se a medida-resumo for grande, vamos precisar adotar outro modelo teórico para os dados.

# Medida-resumo de comparação

- Bombas em Londres: 576 quadrados com a contagem em cada um deles. Modelo:  $\text{Poisson}(\lambda)$  com  $\lambda = 0.9323$
- Particione o conjunto de valores possíveis em intervalos:
- $Y = 0, Y = 1, \dots, Y = 5$ , e  $Y \geq 6$ .
- Calcular  $N_k$ ,  $E_k$  e a diferença  $N_k - E_k$  para cada intervalo.

$k$	0	1	2	3	4	5 e acima
$N_k$	229	211	93	35	7	1
$E_k$	226.74	211.39	98.54	30.62	7.14	1.5
$N_k - E_k$	2.26	-0.39	-5.54	4.38	-0.14	-0.50

- onde  $E_k = 576 \times \mathbb{P}(Y = k) = 576 \frac{0.9323^k}{k!} e^{-0.9323}$ .
- Talvez uma medida-resumo seja a média das diferenças (em valor absoluto):

$$\frac{1}{6} \sum_{k=0}^5 |N_k - E_k|$$

- Esta não é uma boa idéia...Vamos ver porque.

## Diferenças absolutas ou relativas?

- Imagine três categorias com as seguintes diferenças  $|N_k - E_k|$ : 11.5, 10.6 e 0.9.
- Estas diferenças são grandes ou pequenas?
- Depende... do quê? Do valor esperado nessas categorias.
- Considere duas possíveis situações com apenas três categorias

$k$	0	1	2
$N_k$	20	1	6
$E_k$	8.5	11.6	6.9
$ N_k - E_k $	11.5	-10.6	-0.9
$N_k^*$	1020	1001	1006
$E_k^*$	1008.5	1011.6	1006.9
$ N_k^* - E_k^* $	11.5	-10.6	-0.9

# Diferenças absolutas ou relativas?

- Repetindo a tabela:

$k$	0	1	2
$N_k$	20	1	6
$E_k$	8.5	11.6	6.9
$ N_k - E_k $	11.5	-10.6	-0.9
$N_k^*$	1020	1001	1006
$E_k^*$	1008.5	1011.6	1006.9
$ N_k^* - E_k^* $	11.5	-10.6	-0.9

- As diferenças são *idênticas* mas, RELATIVAMENTE AO QUE ESPERAMOS CONTAR EM CADA CATEGORIA, as diferenças são muito menores na segunda situação.
- Quando esperamos contar 11.6 numa categoria e observamos apenas 1, erramos por 10.6 e este erro parece grande.
- Mas quando esperamos 1011.6 e observamos 1001 o erro parece pequeno mesmo que a diferença seja a mesma de antes.
- Parece razoável considerar as diferenças  $|N_k - E_k|$  maiores (em ALGUM sentido) do que as diferenças  $|N_k^* - E_k^*|$ .

## Medida-resumo de comparação

- Talvez uma medida-resumo mais apropriada seja então a média das diferenças RELATIVAS ao esperado em cada categoria.
- Isto é, com  $N$  categorias ao todo, um candidato a medida-resumo seria:

$$\frac{1}{N} \sum_k \frac{|N_k - E_k|}{E_k}$$

- Pearson estudou esta medida e achou que, embora intuitiva e simples, ela não era matematicamente manejável.
- A razão é que a distribuição dessa media-resumo dependia de aspectos específicos do problema sendo analisado. Dependia do tamanho da amostra, da distribuição particular sob estudo (binomial, Poisson, exponencial, etc).
- Num toque de gênio ele propôs uma medida-resumo diferente.

# Medida-resumo de Pearson

- Calcule  $N_k$ ,  $E_k$  e a diferença  $N_k - E_k$  para cada valor possível.
- Ao invés de calcular

$$\frac{1}{N} \sum_k \frac{|N_k - E_k|}{E_k}$$

- calcule a medida-resumo de Pearson:

$$\chi^2 = \sum_k \frac{(N_k - E_k)^2}{E_k}$$

- No caso das bombas em Londres

$$\chi^2 = \frac{(2.26)^2}{226.74} + \frac{(-0.39)^2}{211.39} + \dots + \frac{(-0.50)^2}{1.5} = 1.13$$

- Como saber se  $\chi^2$  é grande ou pequeno? A resposta precisou do gênio de Pearson e, ao mesmo tempo, ela justifica por que usamos a diferença ao QUADRADO na medida-resumo.

# A distribuição de $X^2$

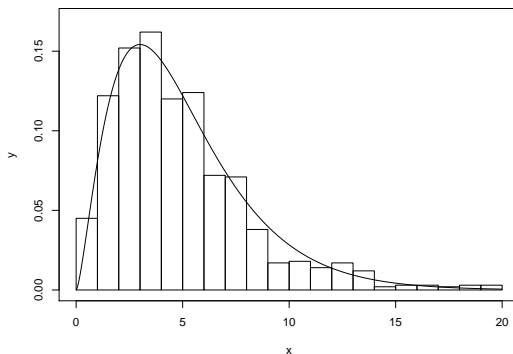
- Considerando o nosso problema das bombas em Londres como ilustração, vamos entender o que Pearson se perguntou.
- SUPONHA QUE O MODELO TEÓRICO  $\text{Poisson}(\lambda)$  SEJA VERDADEIRO.
- Mesmo neste caso,  $X^2$  nunca será exatamente a zero.
- Dependendo da amostra, ele pode ser pequenino ou um pouco maior.
- Não deve ser muito grande pois o modelo é verdadeiro e portanto  $N_k$  deve ser próximo de  $E_k$ .
- Qual a variação natural de  $X^2$  quando o modelo teórico é verdadeiro?
- Vamos responder isto com um experimento no R



# A distribuição por simulação

- Execute o seguinte algoritmo em R:
- Crie um vetor  $E$  de dimensão 6 com as contagens esperadas de  $X = 0, X = 1, \dots, X = 4$ , e  $X \geq 5$  em  $576 \text{ Poisson}(0.9323)$ .
- Isto é,  $E = c(226.74, 211.39, 98.54, 30.62, 7.14, 1.5)$
- Crie um vetor Qui com 1000 posições.
- for(i in 1:1000) faça
  - Gere  $X_1, \dots, X_{576}$  iid  $\text{Poisson}(\lambda = 0.9323)$
  - Conte o número  $N_k$  de  $X_i$ 's iguais a  $0, 1, \dots, 4, \geq 5$
  - Faça  $\text{Qui}[i] \leftarrow X^2 = \sum_k (N_k - E_k)^2 / E_k$
- Faça um histograma dos 1000 valores gerados do vetor Qui.
- Este histograma mostra a variabilidade que se pode esperar de  $X^2$  *quando o modelo é verdadeiro.*

# A distribuição por simulação



**Figura:** Histograma de 1000 simulações e densidade de Qui-quadrado com 5 df superimposta

# Distribuição Qui-quadrado

- Pearson descobriu que a distribuição de  $X^2$  era (aproximadamente) a MESMA qualquer que fosse a distribuição do modelo (Poisson, normal, gama, ou QUALQUER OUTRO MODELO PARA OS DADOS).
- Era uma distribuição universal.
- Ele conseguiu uma fita métrica para medir desvios dos dados observados em relação a QUALQUER modelo teórico.
- É um resultado incrível:
- QUALQUER QUE SEJA a DISTRIBUIÇÃO dos dados da amostra, a distribuição de  $X^2$  é uma só: uma qui-quadrado.
- Não precisamos fazer nenhuma simulação Monte Carlo para encontrar quais os valores razoáveis para  $X^2$  quando o modelo teórico for correto.

## Um pouco mais de rigor

- A distribuição de  $X^2$  não é EXATAMENTE igual a uma distribuição qui-quadrado (com a densidade que acabei de mostrar no gráfico).
- Ela é APROXIMADAMENTE igual a uma qui-quadrado quando o tamanho da amostra é grande.
- O que é uma amostra grande?
  - No caso Poisson, com  $\lambda \approx 1$ , basta ter  $n > 200$ . Com  $\lambda$ 's maiores,  $n = 100$  pode ser suficiente. O fato que com amostras não muito grandes já podemos usar a aproximação.
- Precisamos do número dos graus de liberdade  $k$ . Ele é igual ao número de categorias menos 1 e menos  $p$ , onde  $p$  é o número de parâmetros estimados.
- No caso das bombas de Londres, número de categorias é 6 e tivemos de estimar  $\lambda$  com os dados (obtivemos  $\hat{\lambda} = 0.9323$ ).
- Então, graus de liberdade  $k = 6 - 1 - 1$ .

# Como usar este resultado de Pearson?

- Quando o modelo teórico é verdadeiro, o valor de  $X^2$  segue (aproximadamente) uma distribuição qui-quadrado com  $k$  graus de liberdade.
- $k = \text{no. de categorias} - 1 - p$  onde  $p$  é o número de parâmetros estimados com os dados.
- No caso das bombas em Londres, temos  $k = 6 - 1 - 1 = 4$  g.l.
- Quais são os valores típicos de uma qui-quadrado com 4 g.l.?
- E quais são os valores não-típicos, os valores que dificilmente viriam de uma qui-quadrado com 4 g.l.?

## A densidade de uma qui-quadrado com 4 g.l.

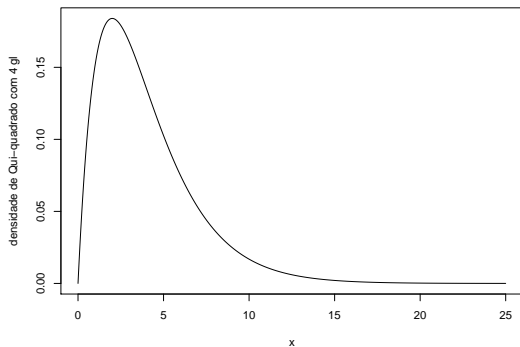


Figura: Densidade de uma distribuição qui-quadrado com 4 g.l.

# Distribuição Qui-quadrado

- Qui-quadrado com  $k$  graus de liberdade (é uma  $\text{Gama}(k/2, 1/2)$ ).
- Densidade

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

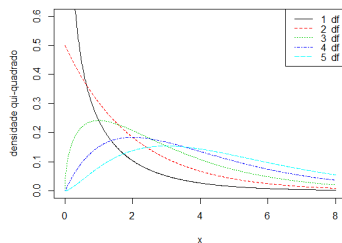


Figura: Densidades da distribuição qui-quadrado com  $k = 1, 2, 3, 4, 5$

# Karl Pearson

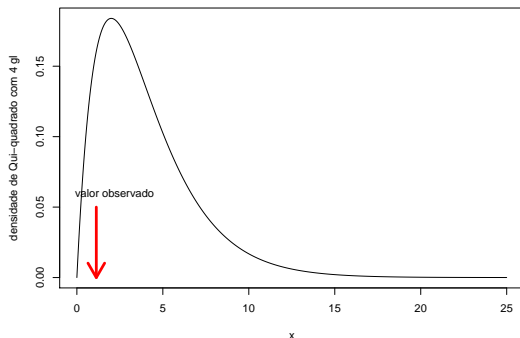




# Como usar o teste qui-quadrado?

- Assim, os valores típicos são aqueles entre 0 e 10.
- Os valores entre 10 e 15 são raros, tem probabilidade  $\mathbb{P}(X^2 \in (10, 15)) = 0.036$ .
- Os valores acima de 15 são possíveis mas ALTAMENTE improváveis. Eles ocorrem com probabilidade  $\mathbb{P}(X^2 > 15) = 0.005$ .
- 
- Calcule o valor realizado de  $X^2$  usando os dados da amostra.
- Este valor realizado é um número real positivo.
- Por exemplo, no caso das bombas em Londres, tivemos  $X^2 = 1.13$  com  $k = 4$  g.l.
- O valor 1.13 é um valor típico de uma qui-quadrado com 4 g.l.?

# Valor observado e densidade



**Figura:** 1.13 é o valor observado de  $X^2$  no caso das bombas em Londres. Gráfico da densidade de uma qui-quadrado com 4 g.l.

## P-valor

- É óbvio que 1.13 é um valor típico de uma qui-quadrado com 4 g.l.
- Ele está bem no meio da faixa de variação razoável dos valores da v.a.  $\chi^2$ .
- Isto é sinal de que as diferenças entre as contagens observadas na amostra e as contagens esperadas pelo modelo são aquelas que se espera QUANDO O MODELO É VERDADEIRO.
- Uma forma de expressar quão discrepante é o valor observado de  $\chi^2$  é calcular a probabilidade de observar uma v.a. qui-quadrado com 4 g.l. MAIOR OU IGUAL a 1.13
- Esta probabilidade é chamada de **p-valor** e, no caso das bombas em Londres, ela é igual a 0.89.
- É a área da densidade da qui-quadrado com 4 g.l. que está acima do valor 1.13 observado na amostra.
- Um p-valor próximo de zero é sinal de que o modelo não se ajusta bem aos dados. Não foi este o caso aqui.

# Como provar o resultado de Pearson? Idéia da prova

- Se  $Z_1, Z_2, \dots, Z_k$  são iid  $N(0, 1)$  então  $Y = Z_1^2 + \dots + Z_k^2 \sim \chi_k^2$ , uma qui-quadrado com  $k$  g.l.

- Temos

$$\chi^2 = \sum_k \frac{(N_k - E_k)^2}{E_k}$$

- Porque  $\chi^2$  segue uma qui-quadrado?
- $N_k$  é a contagem dos elementos da amostra que caem na categoria  $k$
- $N_k \sim \text{Bin}(n, \theta)$  onde  $\theta = \mathbb{P}(X \in \text{categoria } k)$
- Se  $n$  é grande, pelo Teorema Central do Limite,

$$\frac{N_k - n\theta}{n\theta(1 - \theta)} = \frac{N_k - E_k}{E_k(1 - \theta)} \approx N(0, 1)$$

- Se  $(1 - \theta) \approx 1$  então

$$\frac{N_k - E_k}{E_k} \approx N(0, 1) \mapsto \frac{(N_k - E_k)^2}{E_k} \approx N^2(0, 1) = \chi_1^2$$

- Somando sobre as categorias, teremos uma qui-quadrado (estou omitindo vários detalhes e sutilezas).

# Ajuste clássico de Poisson: coices de cavalo

- Von Bortkiewicz: mortos por coices de cavalo em certas corporações do exército prussiano durante vinte anos (1875-1894)
- Em cada ano, durante 20 anos: número de mortos em cada uma das 10 corporações.

nº $k$ de mortes no ano	Frequência $N_k$
0	109
1	65
2	22
3	3
$\geq 4$	1
Total	200=20 (anos) $\times$ 10 (corporações)

# Ajuste clássico de Poisson: coices de cavalo

- Comparando observado e esperado:

$k$	$N_k$	$E_k$
0	109	108.7
1	65	66.3
2	22	20.2
3	3	4.1
$\geq 4$	1	0.7

- Muito próximos. Verificando pelo critério do teste qui-quadrado:

$$\chi^2 = \frac{(109 - 108.7)^2}{108.7} + \frac{(65 - 66.3)^2}{66.3} + \frac{(22 - 20.2)^2}{20.2} + \frac{(3 - 4.1)^2}{4.1} + \frac{(1 - 0.7)^2}{0.7} = 0.61$$

- Este valor realizado de  $\chi^2$  deve ser comparado com os valores prováveis da v.a.  $\chi^2$  caso o modelo seja verdadeiro (que segue uma qui-quadrado com  $k = 5 - 1 = 4$  g.l.).

# Alguns exemplos clássicos: coices de cavalo

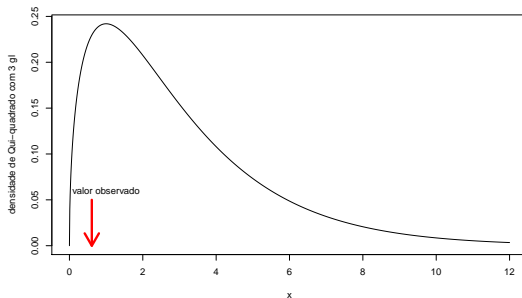


Figura: Valor observado  $X^2 = 0.61$  e densidade de qui-quadrado com 3 g.l.

# Resumo: Kolmogorov versus Qui-quadrado

- Dados  $Y_1, Y_2, \dots, Y_n$  forma uma amostra i.i.d. de uma distribuição-modelo  $F(y)$ ?
- Duas opções: Kolmogorov e Qui-quadrado.
- Kolmogorov: modelo  $F(y)$  tem de ser contínuo; Teoria não vale se for discreta.
- Kolmogorov: Teste só é válido se não precisarmos estimar parâmetros de  $F(y)$ .
- Por exemplo,  $Y_1, Y_2, \dots, Y_n$  segue uma  $N(\mu, \sigma^2)$ ? Podemos usar Kolomogorov?
- Se  $\mu$  e  $\sigma^2$  forem especificados de antemão, antes de olhar os dados, OK, é válido.
- Se eles NÃO são especificados de antemão mas, ao contrário, precisam ser estimados a partir dos dados observados: a distribuição de  $\sqrt{n}D_n$  não é conhecida e não podemos usar Kolomogorov a não ser INFORMALMENTE.



## Resumo: Kolmogorov versus Qui-quadrado

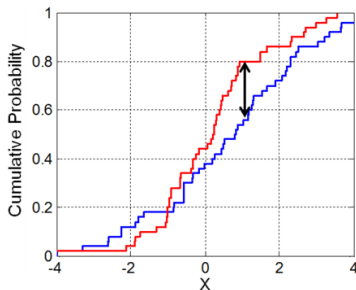
- Dados  $Y_1, Y_2, \dots, Y_n$  forma uma amostra i.i.d. de uma distribuição-modelo  $F(y)$ ?
- Qui-quadrado de Pearson: pode ser aplicado com qualquer modelo, contínuo ou discreto.
- Consegue incorporar o efeito de estimar parâmetros de  $F(y)$ , se for necessário.
- Implementação é muito fácil.
- Precisa especificar os intervalos ou classes onde as contagens vão ser feitas.
- Qual o efeito desta escolha? Quanto mais clases, melhor;
- Mas usar muitas classes pode levar a categorias com probabilidades próximas de zero.
- Devemos escolher classes de forma que o número esperado em cada uma delas seja, de preferência, pelo menos 5. Classes com contagens esperadas menores que 1 devem ser evitadas.

# Teste de hipótese de Kolmogorov

- O teste de Kolmogorov requer um modelo teórico de referência.
- Temos dados uma amostra:  $y_1, y_2, \dots, y_n$  com distribuição  $F(y)$  \*\*  
*desconhecida* \*\*.
- Existe um modelo teórico com distribuição  $F_0(y)$  (gaussiana, exponencial, Poisson, etc)
- Queremos testar a seguinte hipótese:
  - A distribuição desconhecida  $F(y)$  é igual à distribuição do modelo teórico  $F_0(y)$ ?
  - Isto é, posso acreditar na hipótese de que  $F(y) = F_0(y)$ ?
- O que é “testar”?
  - É verificar se os dados observados são compatíveis com a hipótese de que  $F(y) = F_0(y)$ .
- Rejeitamos a hipótese se julgarmos que os dados são incompatíveis com ela.
- Decisão baseada no p-valor da estatística  $\sqrt{n}D_n$

# Teste de Kolmogorov-Smirnov

- O teste de Kolmogorov-Smirnov testa se duas amostras possuem uma *mesma* distribuição de probabilidade
- Ele *não requer que a gente especifique qual seria esta distribuição comum*.
- O teste é baseado na distância entre as distribuições acumuladas empíricas das duas amostras.

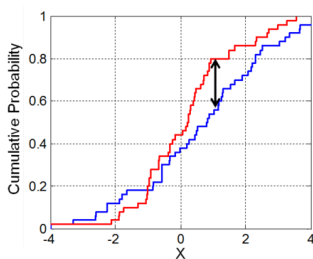


# Kolmogorov-Smirnov

- Temos dados de DUAS amostras:
  - Amostra 1:  $y_1, y_2, \dots, y_n$  com distribuição *desconhecida*  $F(y)$
  - Amostra 2:  $y_1^*, y_2^*, \dots, y_m^*$  com distribuição *desconhecida*  $G(y)$
- Amostras podem ter  $m \neq n$ : tamanhos diferentes.
- Exemplos:
  - pressão sistólica entre homens (amostra 1) e pressão sistólica entre mulheres (amostra 2)
  - pressão entre homens jovens (amostra 1) e entre homens idosos (amostra 2)
  - várias medições de  $y$  antes (amostra 1) e depois (amostra 2) de uma intervenção
  - medições de  $y$  no lugar 1 e no lugar 2
- Hipótese a ser testada:  $F(y) = G(y)$  *sem precisar dizer quem são*  $F(y)$  e  $G(y)$ .

# Kolmogorov-Smirnov

- Acumuladas empíricas das duas amostras:  $\hat{F}_n(y)$  e  $\hat{G}_m(y)$ .
- Se amostras são grandes devemos ter  $\hat{F}_n(y) \approx F(y)$  e  $\hat{G}_m(y) \approx G(y)$ .
- Portanto, se  $F(y) = G(y)$ , devemos ter  $\hat{F}_n(y) \approx \hat{G}_m(y)$ .
- Calcule  $D_{n,m} = \max_y |\hat{F}_n(y) - \hat{G}_m(y)|$



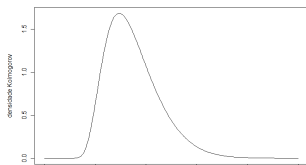
**Figura:** Distância entre as distribuições acumuladas empíricas  $\hat{F}_n(y)$  e  $G_n(y)$ .

# Kolmogorov-Smirnov

- Se  $F(y) = G(y) \Rightarrow \hat{F}_n(y) \approx \hat{G}_m(y)$ .
- Calcule  $D_{n,m} = \max_y |\hat{F}_n(y) - \hat{G}_m(y)|$
- If  $D_{n,m}$  is small, accept hypothesis  $F(y) = G(y)$ .
- If  $D_{n,m}$  is large, reject hypothesis  $F(y) = G(y)$ .
- How to establish a threshold for  $D_{n,m}$ ?
- When we should consider that  $D_{n,m}$  is large enough to reject the hypothesis.

# Kolmogorov-Smirnov

- K-S provaram que, se  $n$  e  $m$  são grandes,  $\sqrt{\frac{mn}{m+n}} D_{n,m}$  converge para uma distribuição que não depende da distribuição comum desconhecida.
- That is,  $\sqrt{\frac{mn}{m+n}} D_{n,m} \approx K$ , where  $K$  is the SAME distribution of the Kolmogorov test seen previously.
- The density of  $K$  is  $f(x) = 8x \sum_{k=1}^{\infty} (-1)^{k-1} k^2 x e^{-2k^2 x^2}$  and its graph is below.
- The p-value is the area under this density in the interval  $(t, \infty)$  where  $t = \sqrt{\frac{mn}{m+n}} D_{n,m}$  is calculated with the data



# p-value of Kolmogorov-Smirnov

- Suppose, for example, that two samples provided the value  $\sqrt{\frac{mn}{m+n}} D_{n,m} = 1.25$ .
- Put this 1.25 in the horizontal axis and look at the area under the Kolmogorov density in the interval  $(1.25, \infty)$ : area = 0.088.
- Interpretation of p-value: If the two samples were randomly sampled from identical distributions, what is the probability that the two cumulative frequency distributions would be as far apart as observed?
- What is the probability that  $\sqrt{\frac{mn}{m+n}} D_{n,m}$  would be equal to 1.25 or even larger than that?

