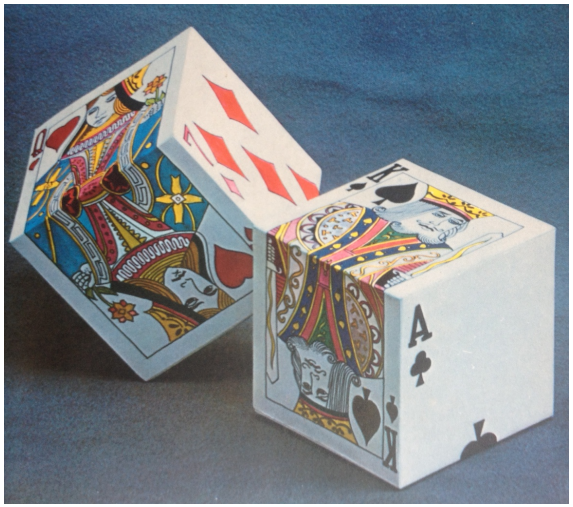
Fundamentos Estatísticos para Ciência dos Dados
Ajuste de distribuições

Renato Martins Assunção

DCC, UFMG - 2015



Plano do tópico

- Vamos aprender a verificar se uma amostra segue uma certa distribuição ou modelo probabilístico.
- Vamos aprender o teste de Kolmogorove o test qui-quadrado.
- Teste de Kolmogorov é aplicado quando o modelo é de uma v.a. contínua.
- Teste qui-quadrado pode sr aplicado a distibuições contínuas ou discretas mas exige uma cateorização arbitrária.

Teste qui-quadrado

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

Teste qui-quadrado

- Amostra $Y_1, Y_2, \ldots, Y_n$ 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 conínuas 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 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 Poisson(5), com 12 categorias: $Y = 0$, $Y = 1, \ldots, Y = 10$ e $Y \geq 11$.
- Modelo é uma 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 $Y_1, Y_2, \ldots, Y_n$ que caem no intervalo k (a freqüência 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 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 Bin$(20, 0.1)$ esperamos que 6.44 delas sejam iguais a zero.

Outro exemplo

- Modelo teórico é uma 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 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 freqüências observadas N_k e as freqüências 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 freqüências observadas N_k e as freqüências 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: Poisson(λ) com $\lambda = 0.9323$
- Particione o conjunto de valores possíveis em intervalos:
- $Y = 0$, $Y = 1, \ldots, Y = 5$, 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
$\lvert N_k - E_k \rvert$	11.5	-10.6	-0.9
N_k^*	1020	1001	1006
E_k^*	1008.5	1011.6	1006.9
$\lvert N_k^* - E_k^* \rvert$	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 distrbuição particular sob estudo (binmomial, 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:
$$X^2 = \sum_k \frac{(N_k - E_k)^2}{E_k}$$
- No caso das bombas em Londres
$$X^2 = \frac{(2.26)^2}{226.74} + \frac{(-0.39)^2}{211.39} + \ldots + \frac{(-0.50)^2}{1.5} = 1.13$$
- Como saber se X^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 Poisson(λ) 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 $X = 0$, $X = 1, \ldots, X = 4$, e $X \geq 5$ em 576 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 $X_1, \ldots, X_{576}$ iid Poisson($\lambda = 0.9323$)
  - Conte o número N_k de X_i's iguais a $0, 1, \ldots, 4, \geq 5$
  - Faça $Qui[i] \leftarrow X^2 = \sum_k (N_k - E_k)^2 / E_k$
- 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

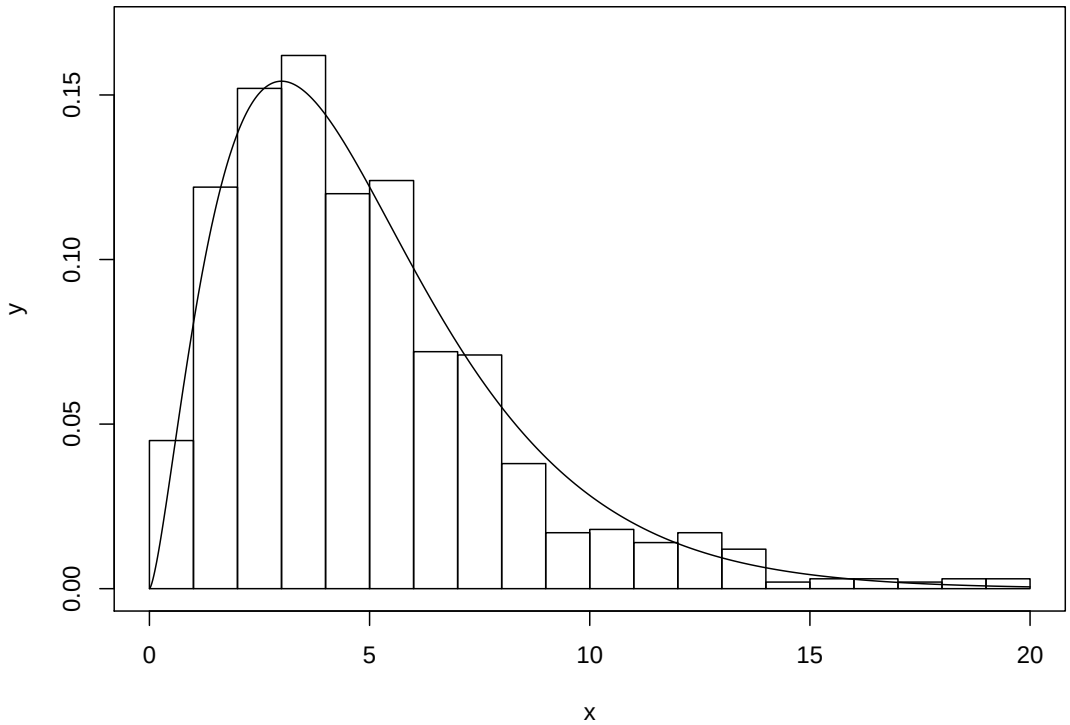


Figura: Histograma de 1000 simulações e densidade de Qui-quadrado com 5 df superimposta

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 λ's maiores, $n = 100$ pode ser suficiente. O fato que comamostras não muito grandes já podmeos 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 λ 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.
- $k =$ 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.

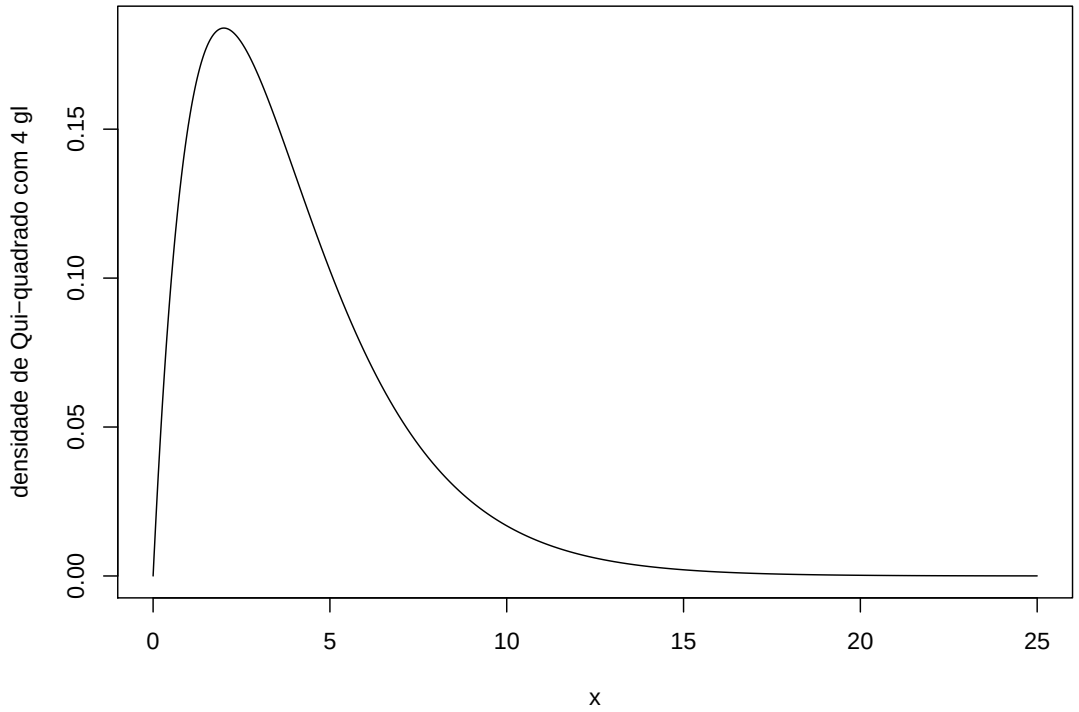


Figura: Densidade de uma distribuição qui-quadrado com 4 g.l.

Distribuição Qui-quadrado

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

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

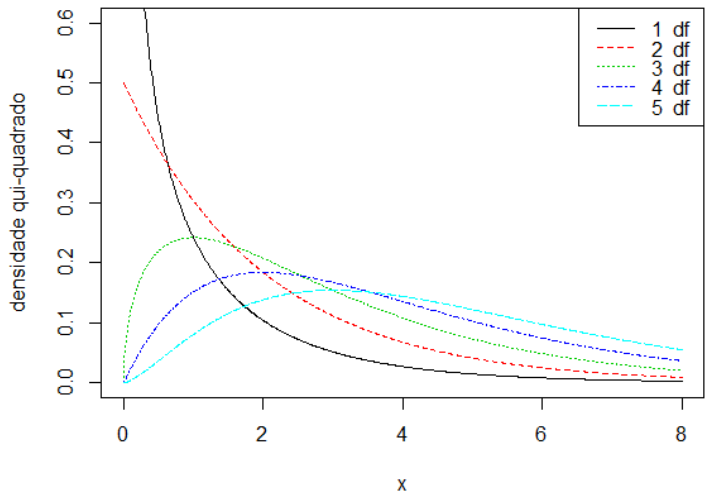
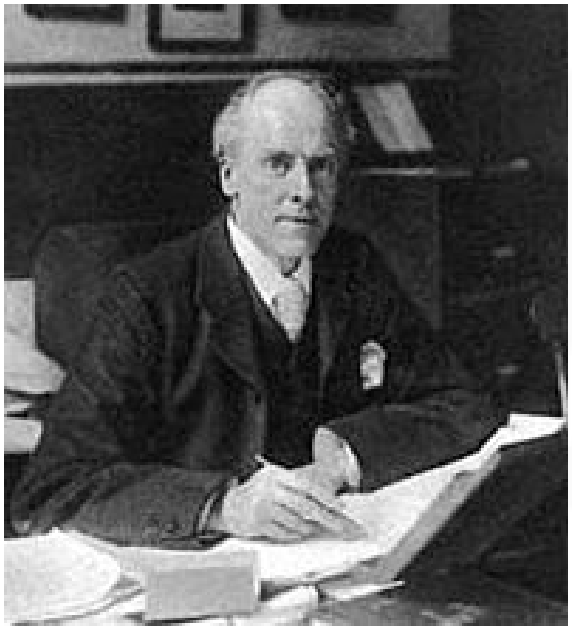


Figura: Densidades da distribuição qui-quadrado com $k = 1, 2, 3, 4, 5$

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 bomabas 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

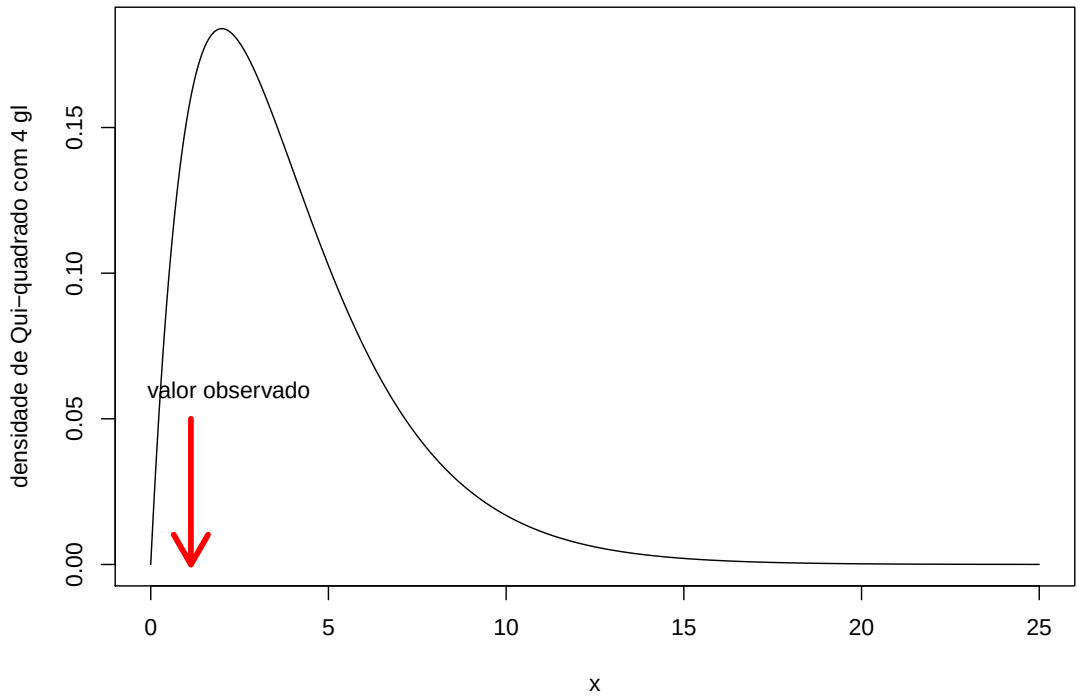


Figura: 1.13 é o valor observado de X^2 no caso das bombas em Londres. Gráfico da densidade de uma qui-quadrado com 4 g.l.

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. X^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 X^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 $Z_1, Z_2, \ldots, Z_k$ são iid $N(0,1)$ então $Y = Z_1^2 + \ldots + Z_k^2 \sim \chi^2_k$, uma qui-quadrado com k g.l.
- Temos

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

- Porque X^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

$$\frac{N_k - E_k}{E_k} \approx N(0,1) \longmapsto \frac{(N_k - E_k)^2}{E_k} \approx N^2(0,1) = \chi^2_1$$

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

Ajuste clássico de Poisson: coices de cavalo

- Von Botkiewicz: 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	Freqüência N_k
0	109
1	65
2	22
3	3
≥ 4	1
Total	200=20 (anos) x 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
≥ 4	1	0.7

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

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

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

Alguns exemplos clássicos: coices de cavalo

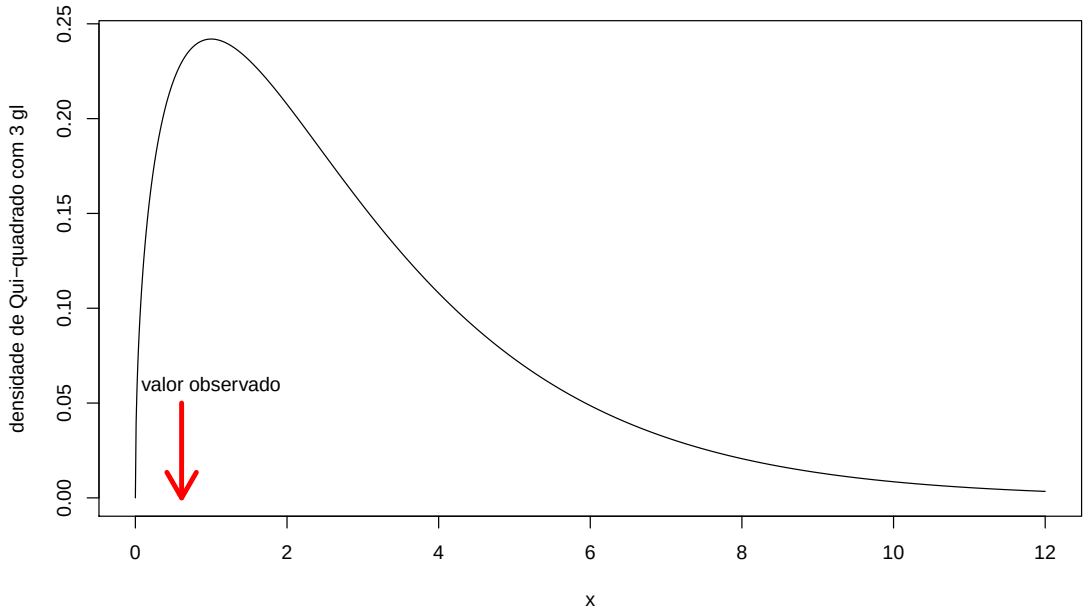
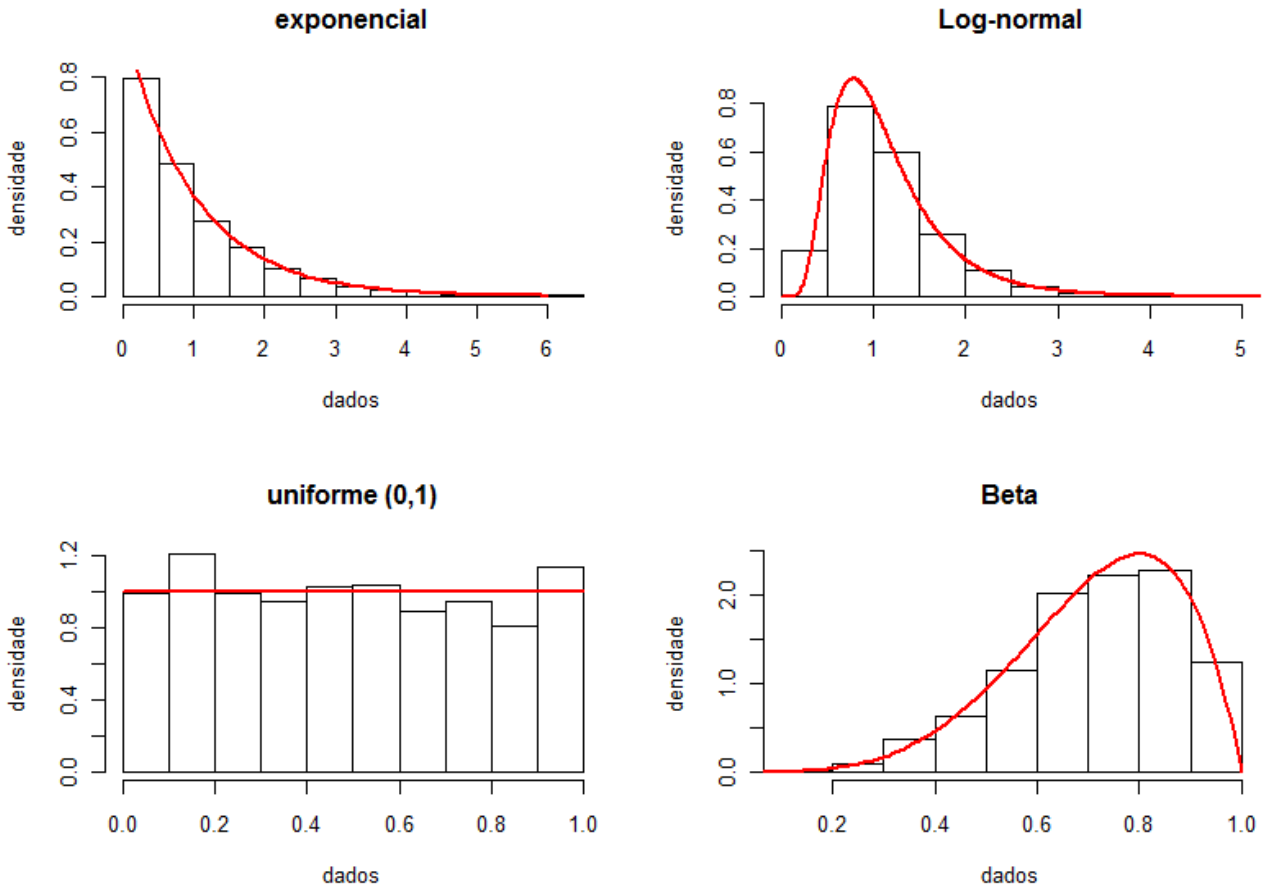


Figura: Valor observado $X^2 = 0.61$ e densidade de qui-quadrado com 3 g.l.

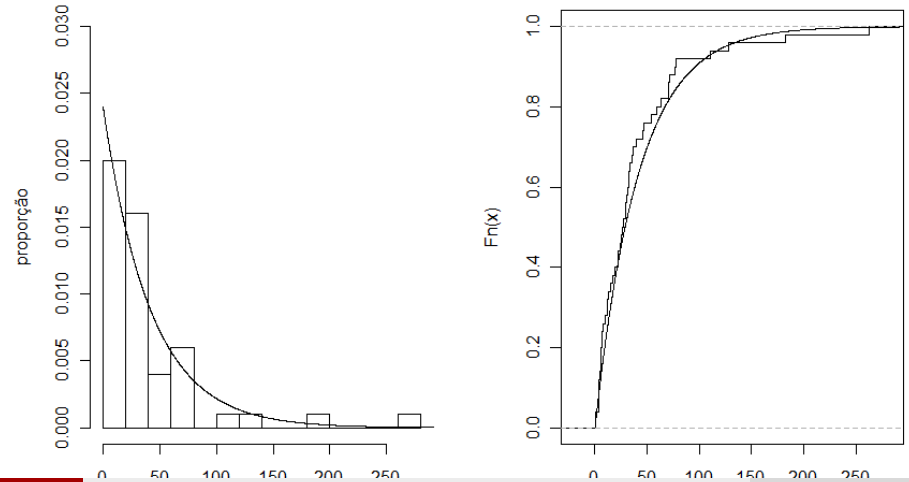
Histogramas e Densidades de modelos contínuos

- Amostras de tamanho $n = 1000$ geradas de 4 distribuições, seu histograma padronizado e a densidade correspondente sobreposta.



Dificuldades...

- Olhar o histograma pode não ser suficiente.
- Abaixo, um histograma de uma amostra de uma v.a. contínua com uma densidade candidata sobreposta.
- Não parece óbvio que a densidade ajusta-se perfeitamente ao histograma.
- Até onde podmeos tolerar um desajuste?
- O segundo gráfico é da função de distribuição acumulada de probabilidade, menos intuitiva mas mais útil.



$\mathbb{F}(y)$ - Caso discreto

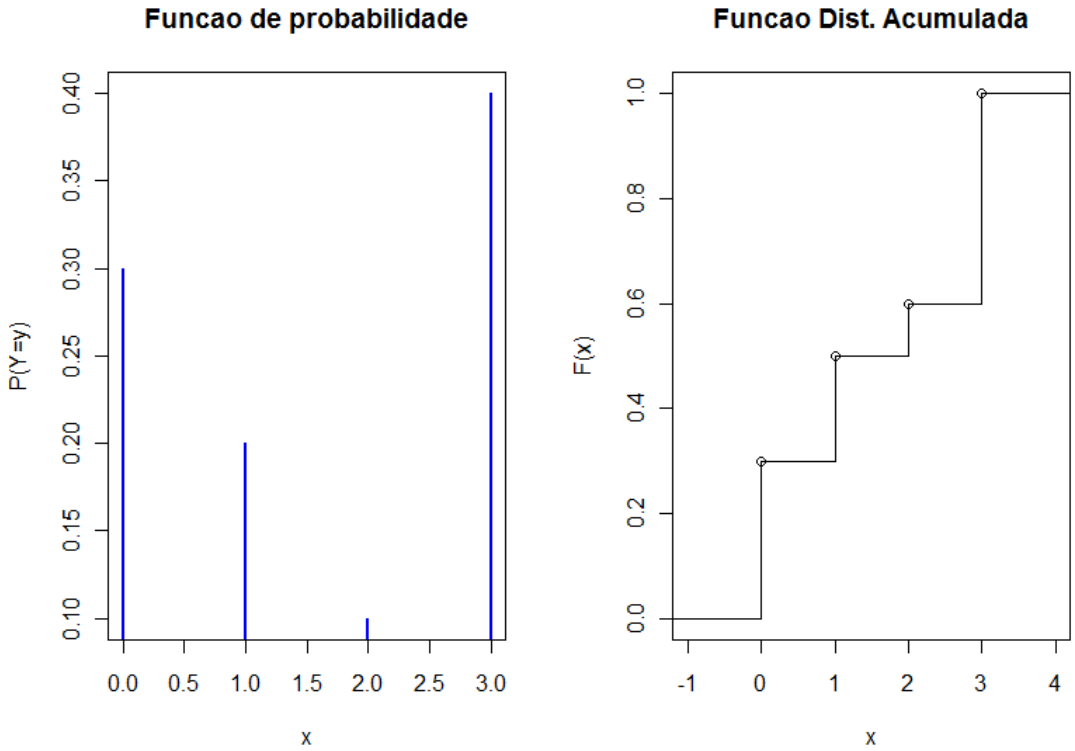


Figura: Função de probabilidade $\mathbb{P}(Y = y)$ e e função distribuição acumulada $F(y)$. Y tem quatro valores possíveis, 0, 1, 2, 3, com probabilidades iguais a 0.3, 0.2, 0.1 e 0.4, respectivamente

$\mathbb{F}(y)$ - Caso contínuo

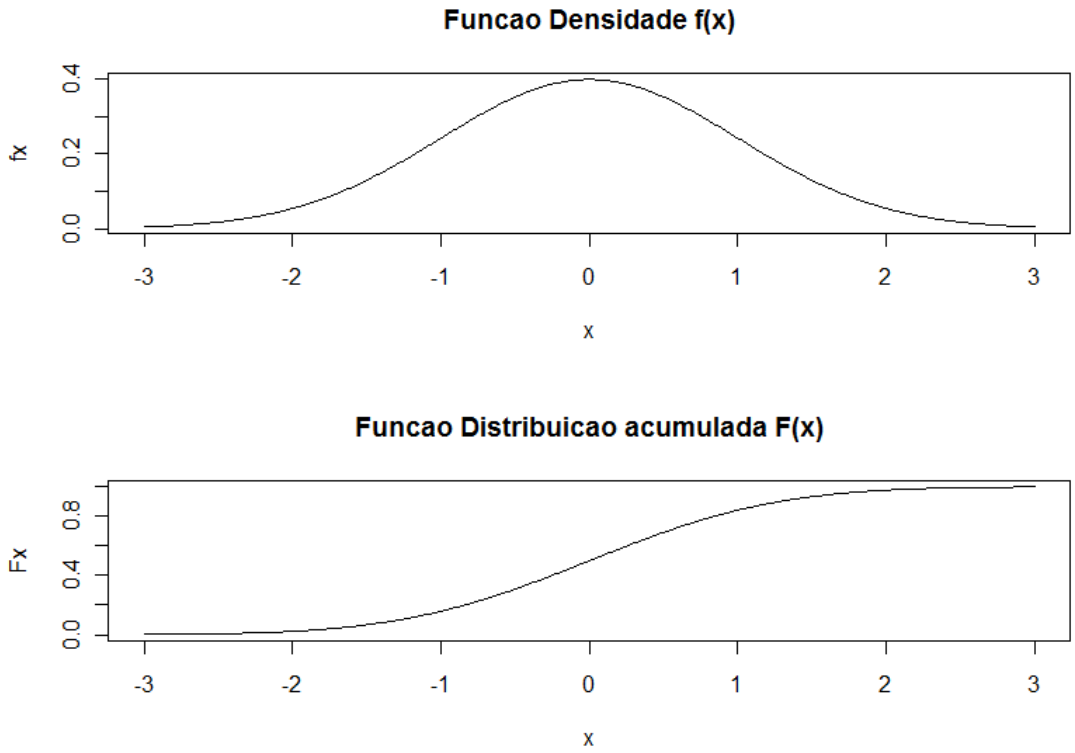


Figura: Densidade $f(x)$ e Função Distribuição Acumulada $F(y)$

Função Distribuição Acumulada

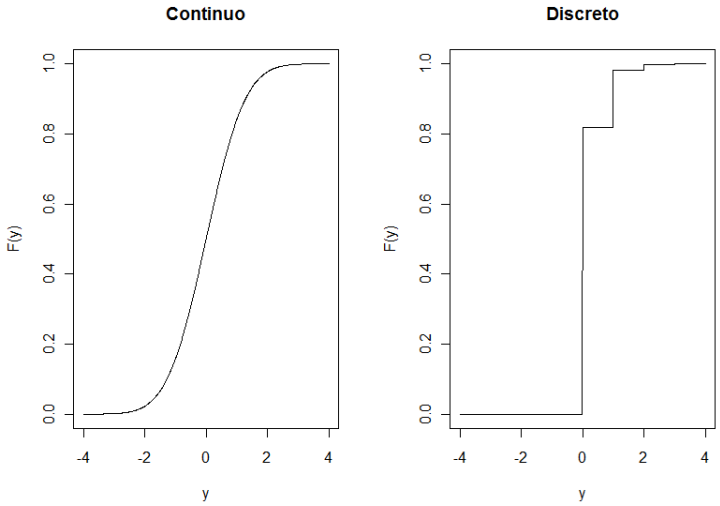


Figura: Função Distribuição Acumulada $F(y)$

Função Distribuição Acumulada

- Função *matemática* $\mathbb{F}$ que mostra como as probabilidades de uma v.a. vão se acumulando no eixo real.
- $\mathbb{F} : \mathbb{R} \to [0, 1]$
- $\mathbb{F}$ é não-decrescente.
- $\mathbb{F}(x) \to 1$ quando $x \to \infty$
- $\mathbb{F}(x) \to 0$ quando $x \to -\infty$
- Se Y é uma v.a. qualquer e y é um ponto da reta real então $F(y) = \mathbb{P}(Y \leq y)$:
  - Caso contínuo: $F(y) = \int_{-\infty}^{y} f(x)dx$
  - Caso discreto com valores possíveis $\{x_1, x_2, \ldots\}$: Então $F(y) = \sum_{x_i \leq y} \mathbb{P}(Y = x_i)$

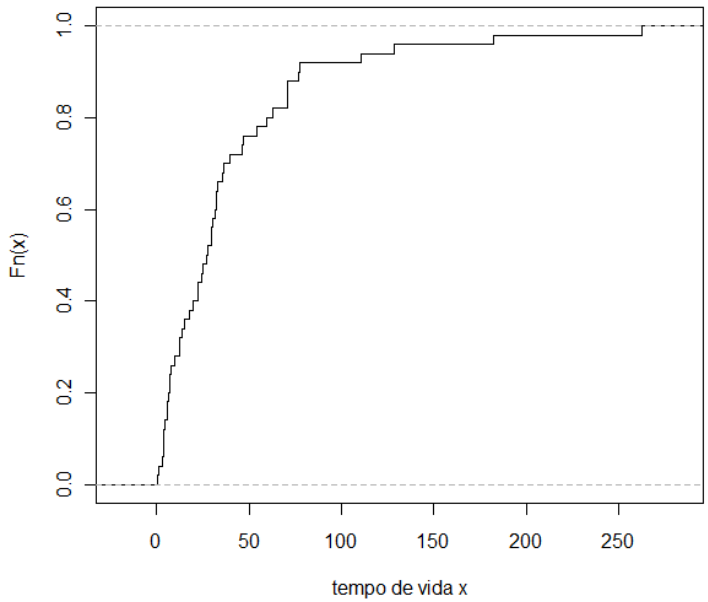
Importância de $F(y)$

- A função distribuição acumulada $F(y)$ é menos intuitiva que a densidade.
- Tem importância teórica:
  - é muito mais fácil provarmos teoremas com ela (existe sempre, tanto faz se a v.a. é discreta ou contínua)
  - tem seus limites entre $[0, 1]$,
  - é sempre crescente (não-decrescente),
  - serve para medir distâncias entre distribuições de probab, etc.
- Tem importância prática: alguns testes e técnicasde simulação Monte Carlo.
- Vamos ver uma delas agora: o teste de Kolmogorov.

Função Distribuição Acumulada EMPÍRICA

- **Definição:** Seja $y_1, y_2, \ldots, y_n$ um conjunto de números reais. A *função distribuição acumulada empírica* $\hat{F}_n(y)$ é uma função $\hat{F}_n : \mathbb{R} \to [0, 1]$ tal que, para qualquer $y \in \mathbb{R}$ temos

$$\hat{F}_n(y) = \frac{\#\{y_i \leq y\}}{n} = \text{Proporção dos } y_i \text{ que são} \leq y$$



Usando $\hat{F}_n(y)$ com distribuições contínuas

- Suponha que Y seja uma v.a. contínua.
- Adotamos um modelo para Y, tal como uma exponencial com parâmetro $\lambda = 0.024$.
- Calculamos a função acumulada teórica $F(y)$: não precisa dos dados para isto, um cálculo matemático-probabilístico.
- Com base na amostra de dados, E SOMENTE NELA, construímos a função distribuição acumulada empÃrica $\widehat{F}_n(y)$.
- Se tivermos $\widehat{F}_n(y) \approx F(y)$ para todo y concluímos que o modelo adotado ajusta-se bem aos dados.

Usando $\hat{F}_n(y)$ com distribuições contínuas

- Como saber se $\widehat{F}_n(y) \approx F(y)$?

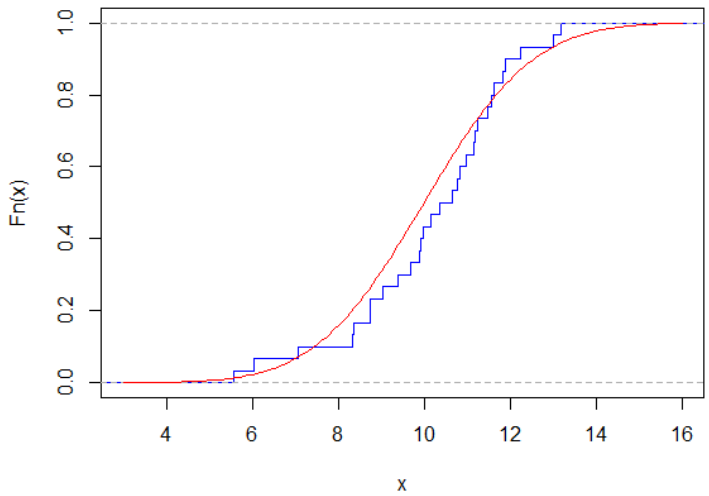


Figura: Empírica $\hat{F}_n(y)$ e a teórica $F(y)$.

Teste de Kolmogorov

- Considere $D_n = \max_y |\hat{F}_n(y) - F(y)|$
- Se $D_n \approx 0$ então o modelo adotado ajusta-se bem aos dados.
- Como saber se $D_n \approx 0$? Kolmogorov estudou o comportamento de D_n.

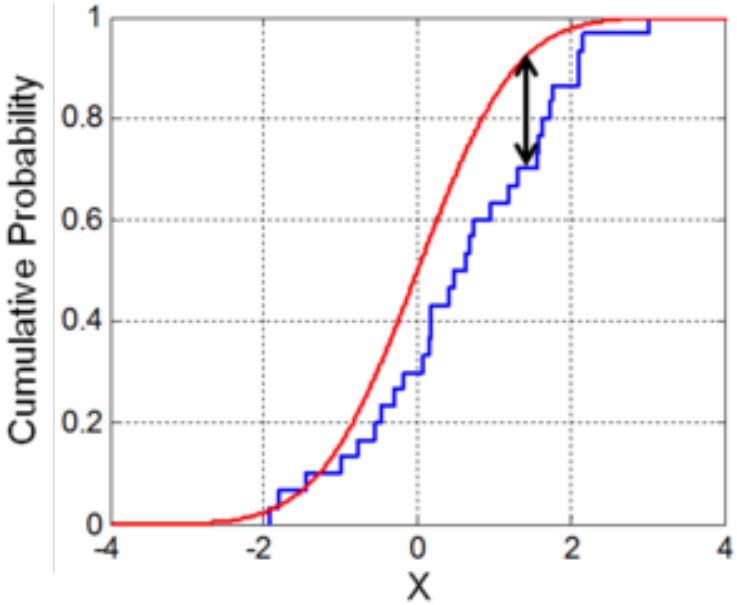
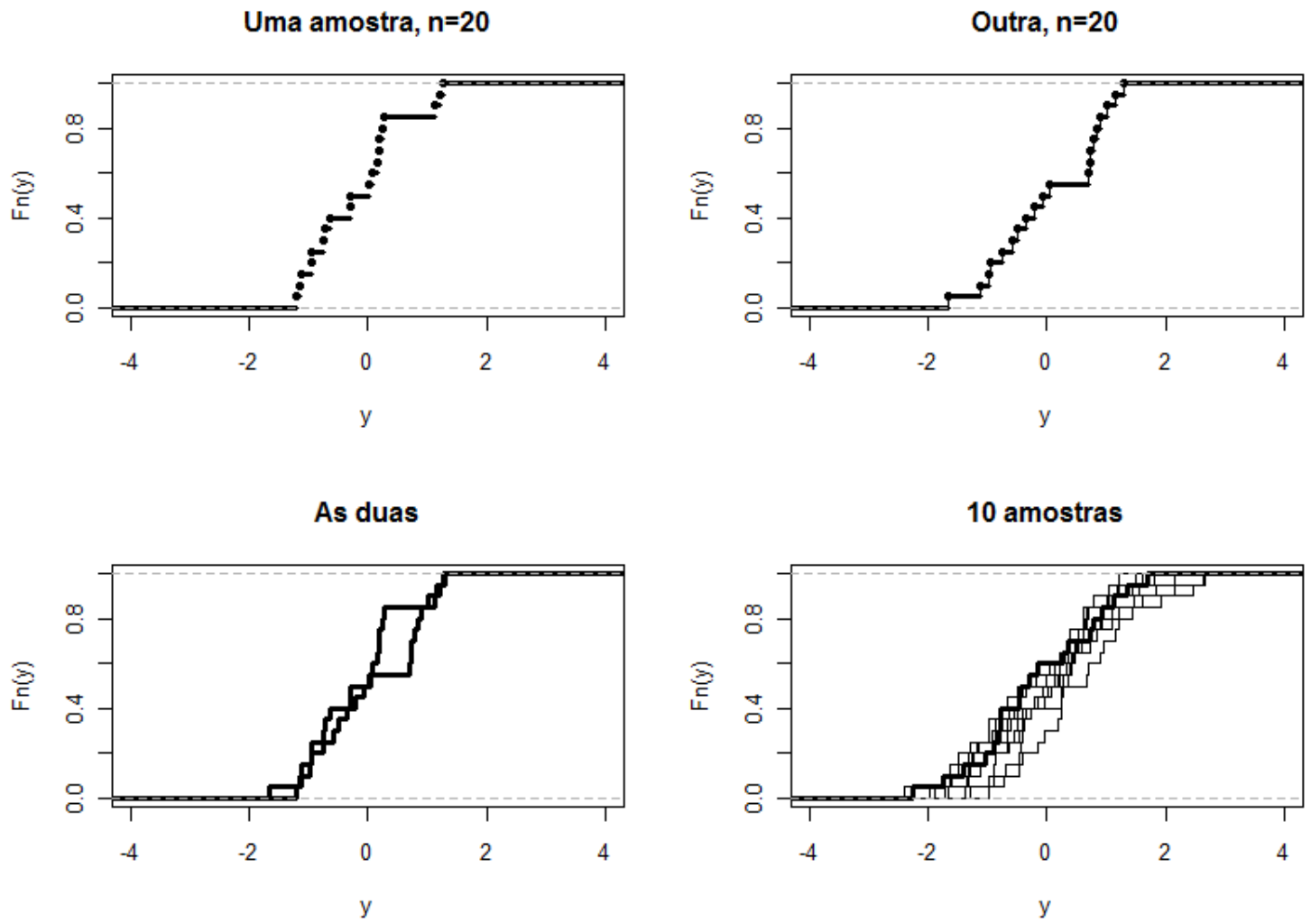


Figura: Empírica $\hat{F}_n(y)$ e a teórica $F(y)$.

$\hat{F}_n(y)$ é função aleatória!!!



$D_n = \max_y |\hat{F}_n(y) - F(y)|$

- Suponha que $F(y)$ é o verdadeiro modelo gerador dos dados (nos gráficos abaixo, usei uma $N(0,1)$).
- Pode-se mostrar que $D_n \to 0$ se $n \to \infty$ qualquer que seja $F(y)$.

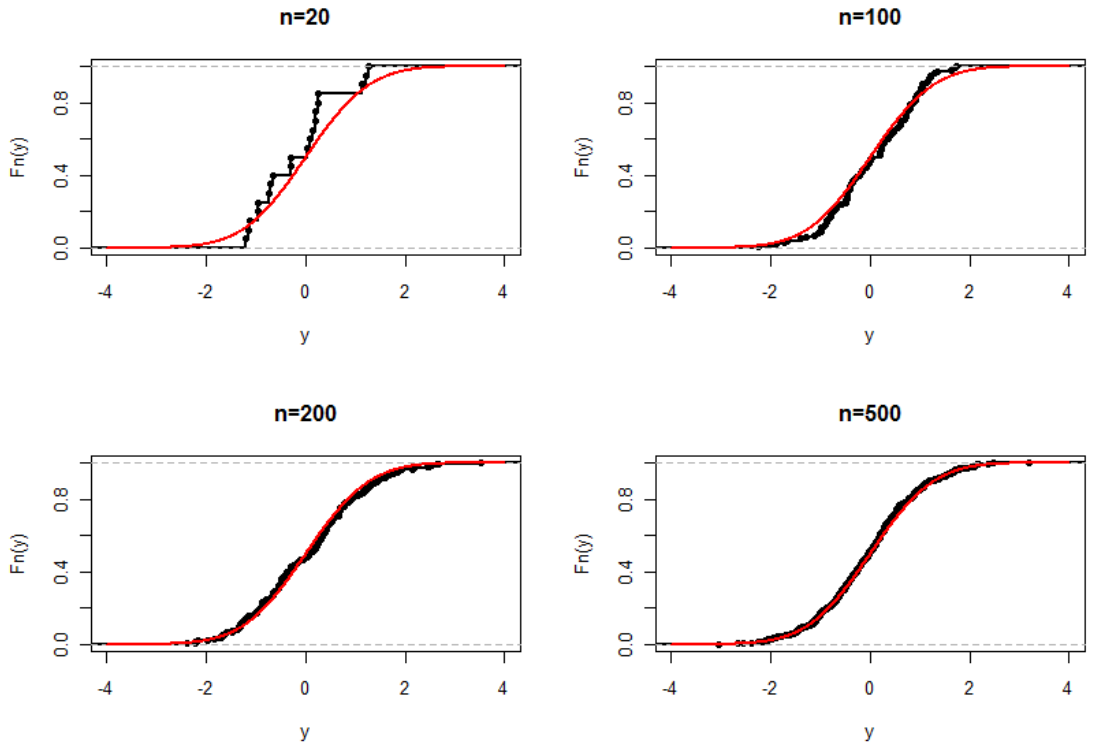
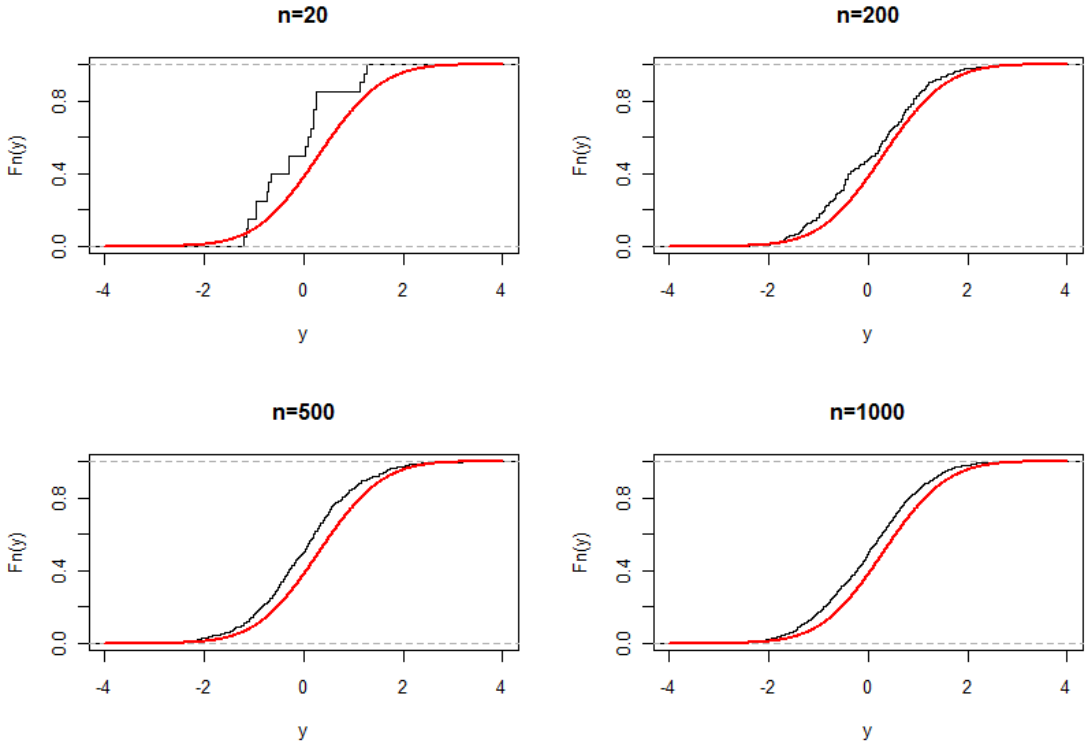


Figura: $D_n \to 0$ se o modelo é correto

$D_n = \max_y |\hat{F}_n(y) - F(y)|$

- Suponha que $F(y)$ NÃO é o modelo verdadeiro.
- Uso $F(y) \sim N(0,1)$ mas, NA VERDADE, dados são gerados de $N(0.3,1)$.
- Então pode-se mostrar que D_n converge para um valor > 0.



$D_n = \max_y |\hat{F}_n(y) - F(y)|$

- RESUMINDO:
- Suponha que $F(y)$ é o modelo verdadeiro. Então $D_n \to 0$ se $n \to \infty$.
- Se $F(y)$ não é o modelo verdadeiro, $D_n \to a > 0$.
- Mas continuamos com o problema: quão próximo de zero D_n tem de ser para aceitarmos o modelo teórico $F(y)$?
- $D_n = 0.01$ é pequeno?
- Com certeza, a resposta depende de n já que $D_n \to 0$ se $n \to \infty$.
- A resposta depende do modelo teórico $F(y)$ considerado?
- Por exemplo, o comportamento de D_n quando $F(y)$ for uma gaussiana é diferente do comportamento quando $F(y)$ for uma expoencial?

$D_n = O(1/\sqrt{n})$

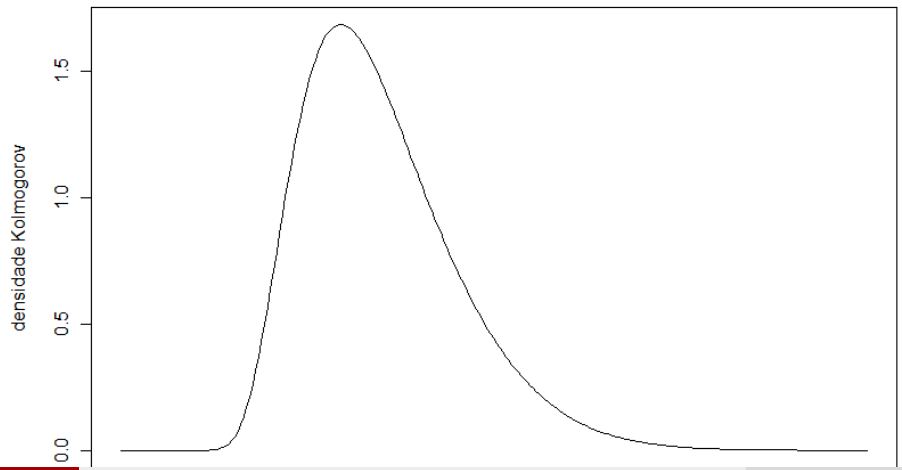
- Vimos que $D_n \to 0$ se $n \to \infty$.
- Com que rapidez ele decresce em direção a 0?
- Kolmogorov mostrou que:
  - $nD_n \to \infty$ (degenera).
  - $\log(n)D_n \to 0$ (degenera).
  - $\sqrt{n}D_n \not\to 0$ e também $\not\to \infty$.
  - $\sqrt{n}D_n$ fica (aleatoriamente) estabilizado.
  - Qualquer outra potência leva a resultados denegerados.
  - $n^{0.5+\epsilon}D_n \to \infty$.
  - $n^{0.5-\epsilon}D_n \to 0$.
- Mas e daí???

Como saber se D_n é pequeno?

- Suponha que $F(y)$ é o modelo verdadeiro.
- Kolmogorov: $\sqrt{n}D_n \to K$ onde K é uma distribuição que NÃO DEPENDE de $F(y)$.
- Isto é, $\sqrt{n}D_n$ é aleatório mas sua distribuição é a mesma EM TODOS OS PROBLEMAS!!
- Sabemos como $\sqrt{n}D_n$ pode variar se o modelo for verdadeiro, qualquer que seja este modelo verdadeiro.
- Isto significa que temos uma métrica UNIVERSAL para medir distância entre $\hat{F}_n(y)$ e a distribuição verdadeira QUALQUER QUE SEJA esta distribuição verdadeira!!!

Densidade $\approx$ de $\sqrt{n}D_n$

- K é a distribuição de uma ponte browniana (assunto muito técnico).
- Comocuriosidade, a densidade de K é dada por $f(x) = 8x \sum_{k=1}^{\infty} (-1)^{k+1} k^2 e^{-2k^2x^2}$.
- Gráfico abaixo mostra esta densidade $f(x)$.
- Se calcularmos D_n usando o VERDADEIRO modelo $F(y)$ que gerou os dados então $\sqrt{n}D_n$ deve estar entre 0.4 e 1.5 com alta probabilidade.
- Se não usarmos o modelo verdadeiro, sabemos que $\sqrt{n}D_n \to \infty$.



Densidade $\approx$ de $\sqrt{n}D_n$

- Nunca teremos $\sqrt{n}D_n$ EXATAMENTE igual a zero.
- Se $\sqrt{n}D_n > 1.8$ teremos uma forte evidência de que o $F(y)$ escolhido não é o modelo gerador dos dados.
- Um ponto de corte menos extremo: se $F(y)$ é o modelo que gerou os dados, então a probab de $\sqrt{n}D_n > 1.36$ é apenas 5%.

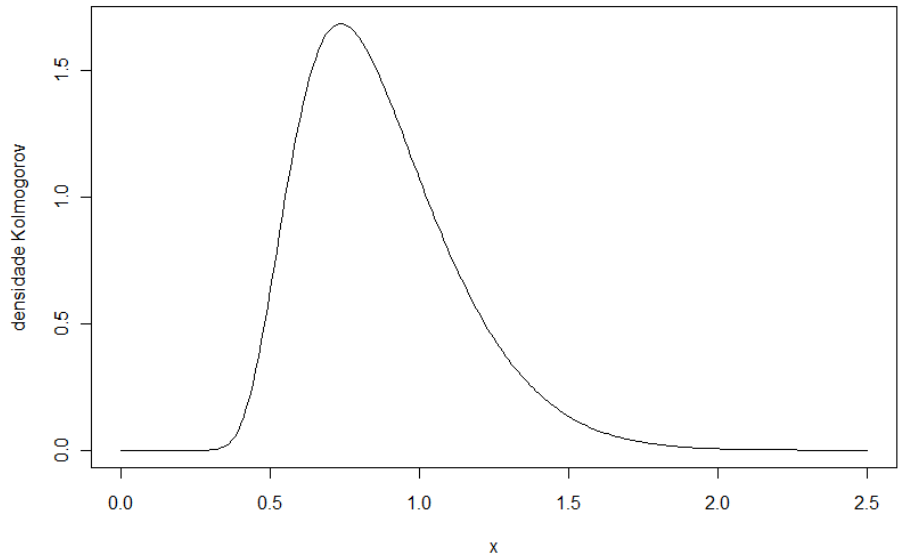
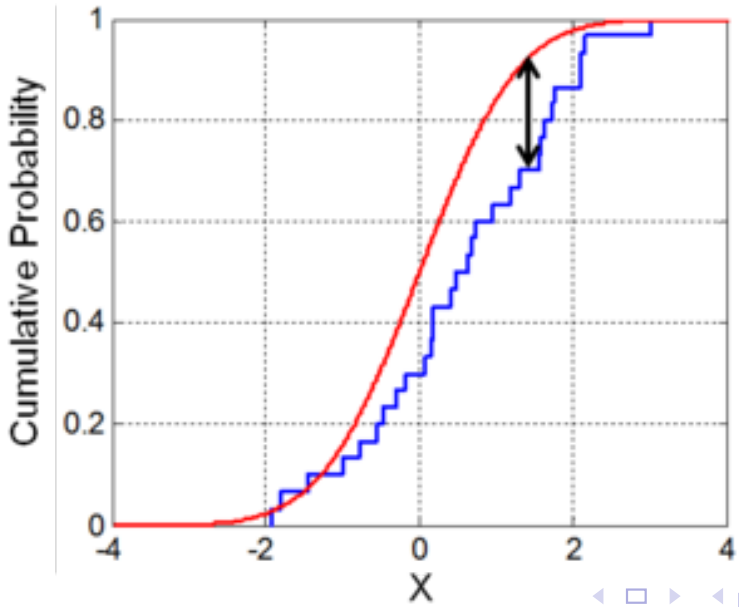


Figura: Densidade de $K \approx \sqrt{n}D_n$

Resumo da ópera

- Dados de uma amostra: $y_1, y_2, \ldots, y_n$.
- Eles foram gerados i.i.d. com a distribuição $F(y)$? (distribuição = hipótese = modelo)
- Calcule a distribuição acumulada empírica $\hat{F}_n(y)$.
- Calcule $D_n = \max_y |\hat{F}_n(y) - F(y)|$
- Se $\sqrt{n}D_n > 1.36$, rejeite $F(y)$ como modelo para os dados
- Se $\sqrt{n}D_n \leq 1.36$, siga em frente com o modelo $F(y)$.



Resumo: Kolmogorov versus Qui-quadrado

- Dados $Y_1, Y_2, \ldots, Y_n$ 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, $Y_1, Y_2, \ldots, Y_n$ segue uma $N(\mu, \sigma^2)$? Podemos usar Kolomogorov?
- Se μ e σ^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 $Y_1, Y_2, \ldots, Y_n$ 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 categoias 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.