Inferência para CS
Modelos univariados contínuos

Renato Martins Assunção

DCC - UFMG

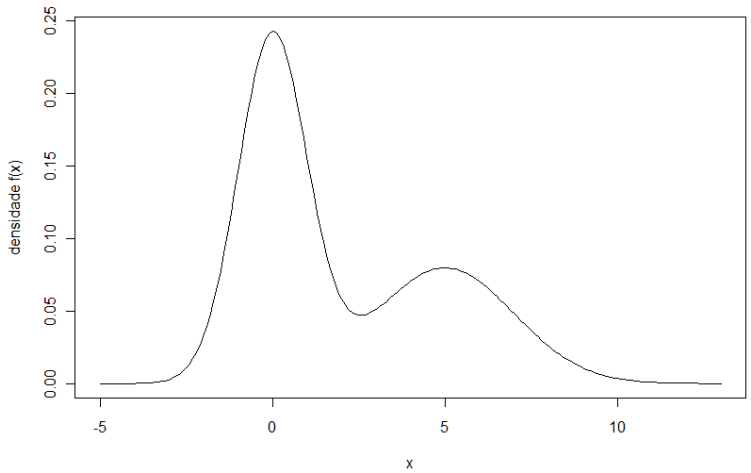
2014

V.A. Contínua

- Composta de um intervalo e uma função densidade.
- Um intervalo de valores reais que são os valores possíveis.
- Uma FUNÇÃO densidade de probabilidade definida neste intervalo.
- Exemplos:
  - $X \in [0, 1]$ com $f(x) = 1$ (distribuição uniforme).
  - $X \in (0, \infty)$ com $f(x) = \exp(-x)$ para $x \in (0, \infty)$.
  - $X \in \mathbb{R}$ com $f(x) = 1/\sqrt{2\pi}\exp(-x^2/2)$.
- A única restrição: $f(x) \geq 0$ para todo x e sua integral deve ser $= 1$.

Probabilidades estão associadas com áreas

- No caso contínuo, probabilidades estão associadas com áreas sob a função densidade.
- $$\mathbb{P}(X \in (a, b)) = \int_a^b f(x)dx$$
- Olhando o gráfico de $f(x)$ sabemos quais as faixas de valores mais prováveis.

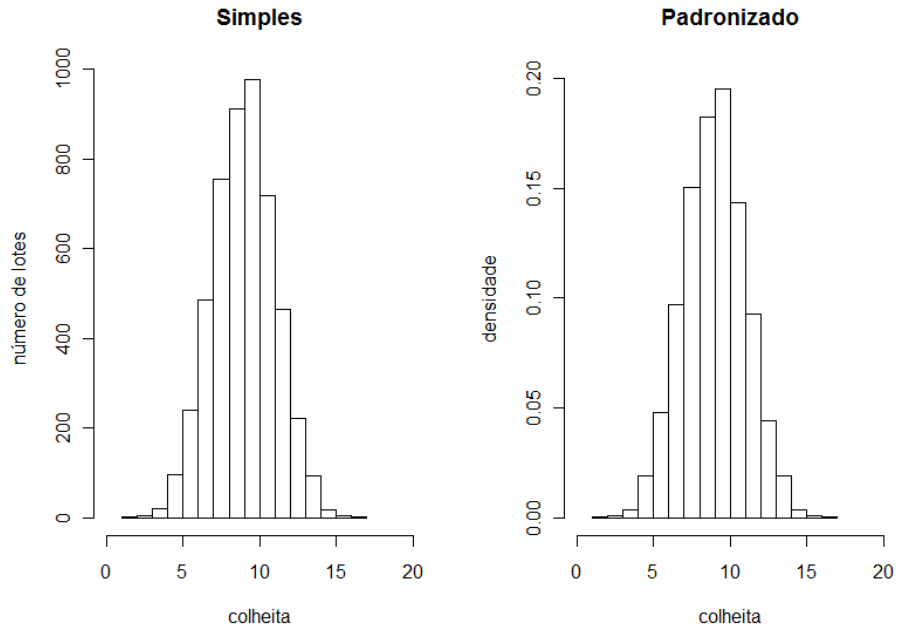


Modelos para dados contínuos

- v.a. Y contínua.
- Imagine uma amostra de 5000 lotes que constituem uma fazenda e onde se cultiva somente soja.
- Seja y_i a colheita do lote i.
- É muito pouco prático e um tanto sem sentido trabalharmos com uma distribuição discreta para uma situação como essa.
- É mais útil assumirmos que as colheitas dos lotes são os resultados de 5000 realizações de uma certa variável aleatória *contínua* que possua uma forma simples e já conhecida.
- Qual a densidade desta Y?
- Para saber isto, faça um histograma (com área total $= 1$).

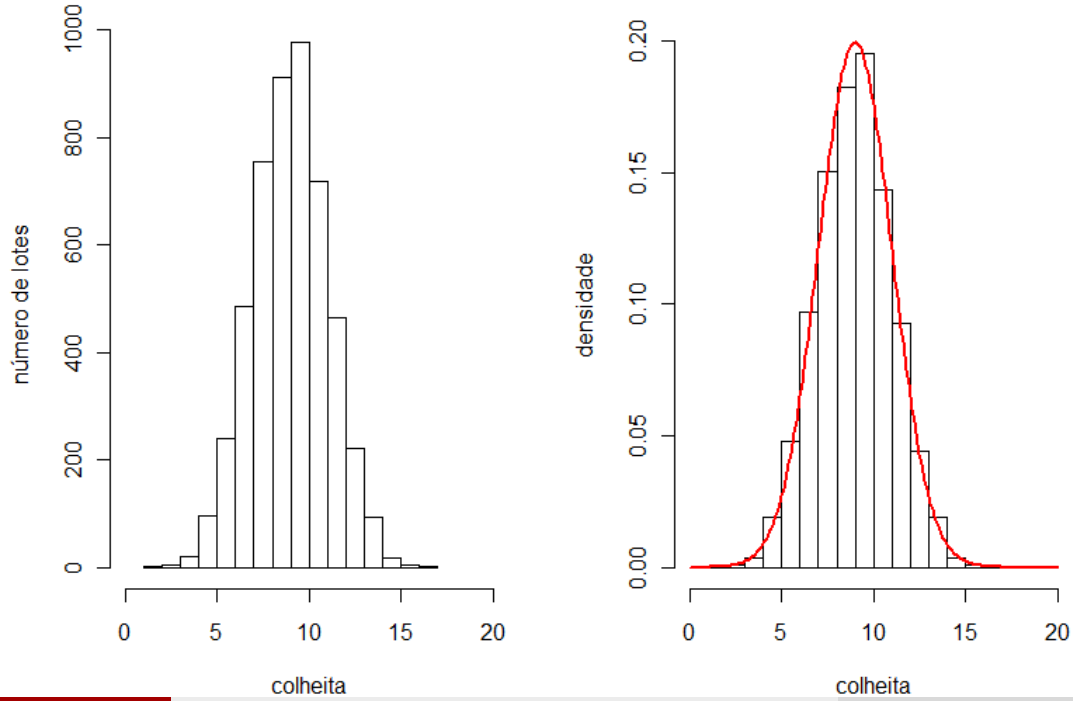
Histograma

- Quebre o eixo horizontal em pequenos intervalos de comprimento Δ.
- Em cada pequeno intervalo i, conte o número n_i de elementos em sua amostra que caíram no intervalo.
- Levente uma barra cuja altura seja igual a esta contagem (esquerda)
- Histograma padronizado tem área total $= 1$.
- Para isto: levante uma barra com altura $= n_i/(n\Delta)$.



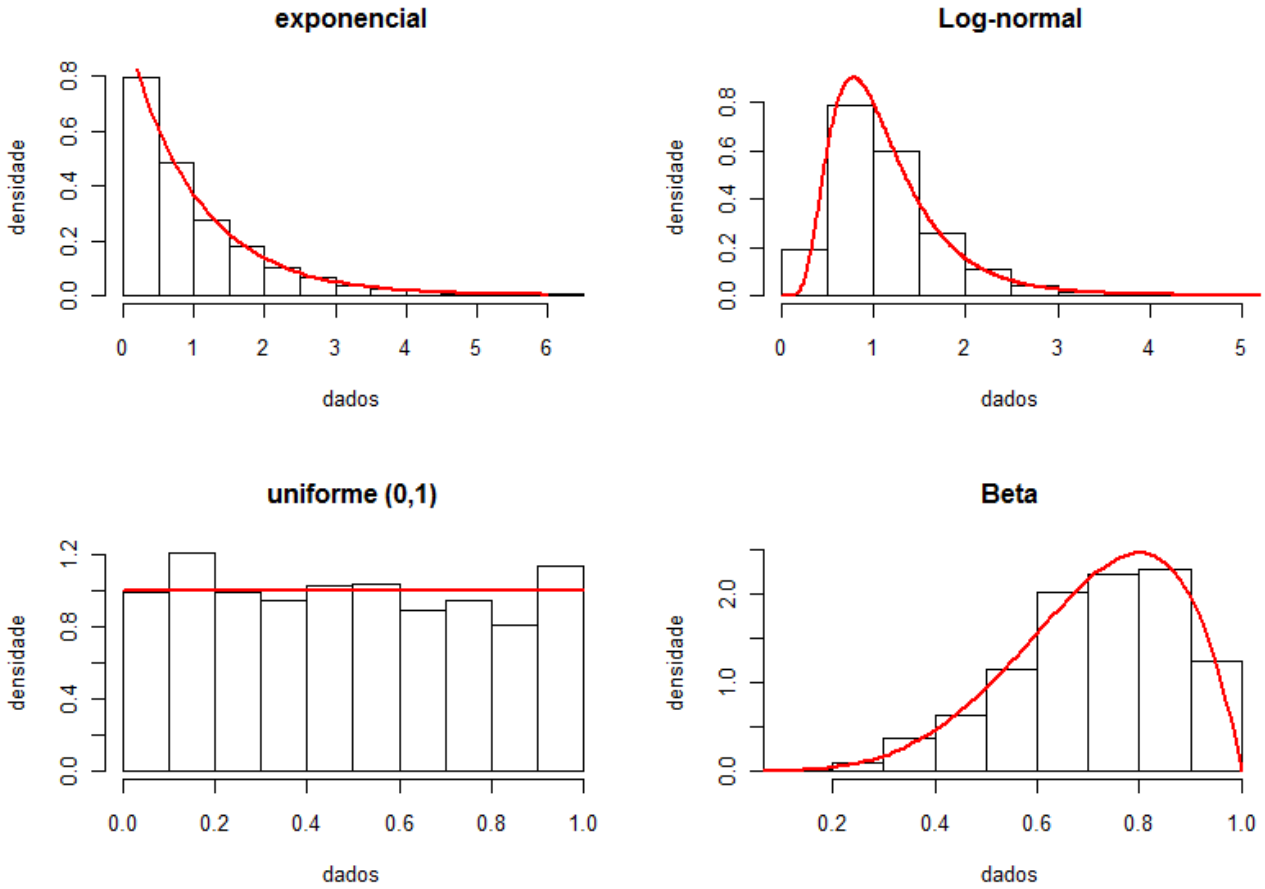
Modelos para dados contínuos

- No histograma padronizado, sobreponha uma densidade candidata.
- O histograma se parece com uma certa densidade gaussiana (ou normal, $N(9, 4)$).
- Então a distribuição real será *aproximada* por esta distribuição normal (veremos como escolher uma distribuição candidata mais tarde).



Mais exemplos para dados contínuos

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



Justificativa

- $f^*(y) =$ densidade verdadeira que gerou os dados.
- $f(y)$ modelo retirado da nosso catálogo de distribuições conhecidas.
- Se o histograma da amostra é bem aproximado por $f(y)$ então acreditamos $f(y) \approx f^*(y)$. Por quê?
- Seja $(y_0 - \delta/2, y_0 + \delta/2)$ um pequeno intervalo do histograma centrado em y_0 e de (pequeno) comprimento δ.
- Aproximando a área debaixo da curva por um retângulo:

$$\begin{aligned} P(Y \in (y_0 - \delta/2, y_0 + \delta/2) &= \int_{y_0-\delta/2}^{y_0+\delta/2} f^*(y)\, dy \\ &\approx f^*(y_0)\delta \end{aligned}$$

Justificativa

- A probabilidade também pode ser aproximada pela fração de elementos da amostra que caíram no intervalo $(y_0 - \delta/2, y_0 + \delta/2)$:

$$\frac{\#\{Y_i's \in (y_0 - \delta/2, y_0 + \delta/2)\}}{n} \approx P(Y \in (y_0 - \delta/2, y_0 + \delta/2)$$

- Igualando as duas aproximações e dividindo por δ dos dois lados, temos

$$\frac{\#\{Y_i's \in (y_0 - \delta/2, y_0 + \delta/2)\}}{n\delta} \approx f^*(y_0)$$

- O lado esquerdo é a altura do histograma no ponto y_0. O lado direito é a altura da curva densidade no memso ponto y_0.
- Assim, as alturas do histograma nos pontos centrais são $\approx$ iguais à densidade DESCONHECIDA.
- Olhar o histograma é olhar a densidade desconhecida (aproximadamente).

Esperança e Variância

- Suponha que você VAI SIMULAR uma distribuição $F(y)$.
- Isto é, vamos gerar números pseudo-aleatórios com distribuição $F(y)$.
- Como RESUMIR grosseiramente esta longa lista de números ANTES MESMO DE GERÁ-LOS?
- O valor TEÓRICO em torno do qual eles vão variar: a esperança $\mathbb{E}(Y)$.
- As vezes, $Y > \mathbb{E}(Y)$; as vezes, $Y < \mathbb{E}(Y)$. Podemos esperar os valores gerados de oscilando Y em torno de $\mathbb{E}(Y)$.
- Em torno, quanto?? DP $=$ desvio-padrão.
- DP é o valor TEÓRICO que mede o quanto os valores oscilam em torno de $\mathbb{E}(Y)$: $\sigma = \sqrt{\text{Var}(Y)}$.

$\mathbb{E}(Y)$ no caso discreto

- Caso discreto com valores possíveis $\{x_1, x_2, \ldots\}$: Então $\mathbb{E}(Y) = \sum_{x_i} x_i \mathbb{P}(Y = x_i)$
- É uma soma ponderada dos valores possíveis da v.a. Y.
- Os pesos são as probabilidades de cada valor.
- Os pesos são ≥ 0 e somam 1.
- $\mathbb{E}(Y)$ geralmente NÃO É um dos valores possíveis $\{x_1, x_2, \ldots\}$.
- É um valor TEÓRICO, não precisa de dados esatísticos para ser calculado.

Identifique $\mathbb{E}(Y)$ em cada caso

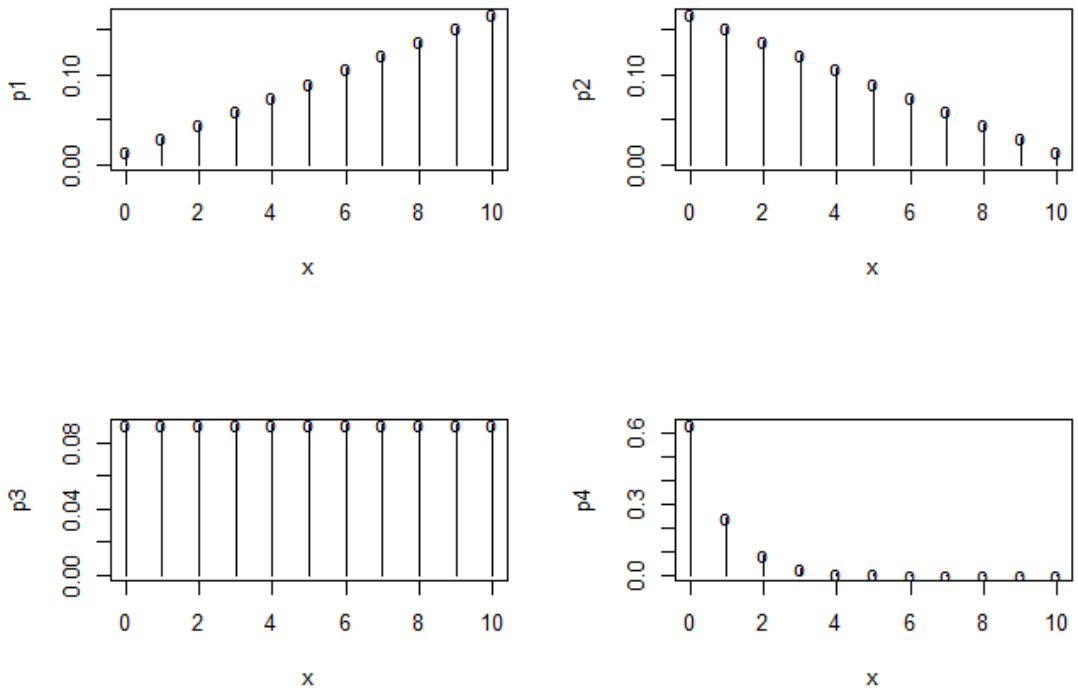


Figura: Sem fazer nenhuma conta, identifique as distribuições com as seguintes esperanças: $5, 6.67, 0.53, 3.33$

$\mathbb{E}(Y)$: resposta

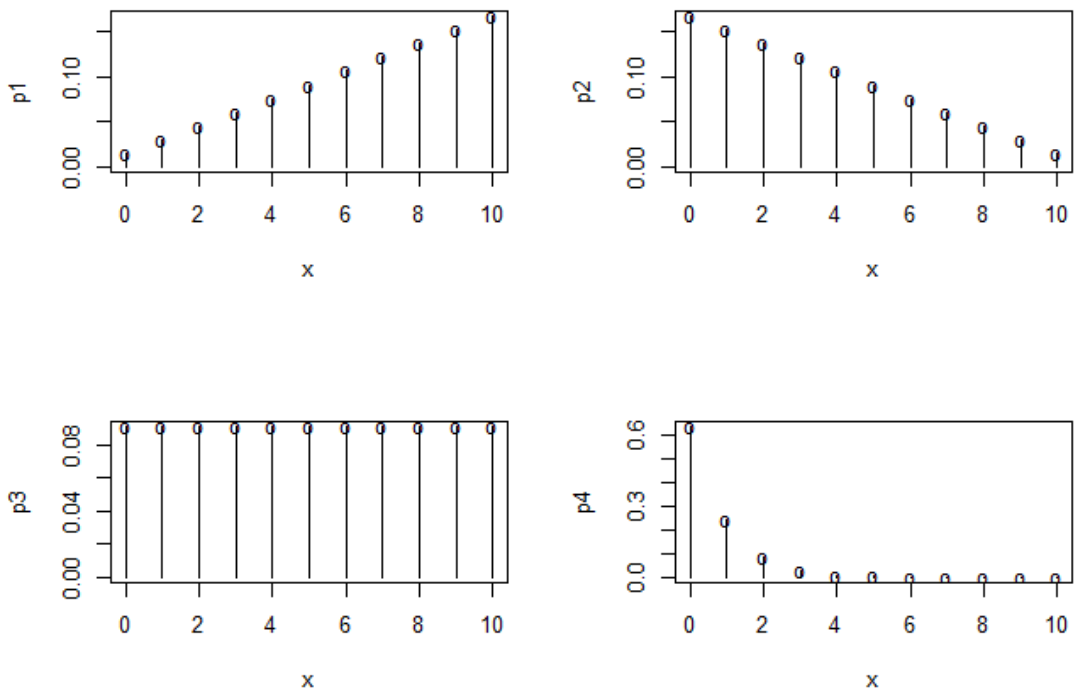


Figura: $p_1 = 6.67$, $p_2 = 3.33$, $p_3 = 5$, $p_4 = 0.53$.

$\mathbb{E}(Y)$ no caso contínuo

- Caso contínuo: $\mathbb{E}(Y) = \int_{-\infty}^{\infty} yf(y)dy$
- Podemos raciocionar intuitivamente EXATAMENTE como no caso discreto.
- Quebrar todo eixo real em pequenos bins de comprimento Δ e centrados em $\ldots, y_{-2}, y_{-1}, y_0, y_1, y_2, \ldots$.
- Então, em cada pequeno bin, aproxime a integral:

$$\int_{\text{bin}_i} yf(y)dy \approx y_i f(y_i)\Delta$$

- Portanto, $\mathbb{E}(Y) = \int_{-\infty}^{\infty} yf(y)dy$ é igual a

$$\int_{-\infty}^{\infty} yf(y)dy = \sum_{i=-\infty}^{\infty} \int_{\text{bin}_i} yf(y)dy \approx \sum_{i=-\infty}^{\infty} y_i f(y_i)\Delta \approx \sum_{i=-\infty}^{\infty} y_i \mathbb{P}(Y \in \text{bin}_i)$$

Desenhar

Assim, caso contínuo (esperança como integral) é a versão contínua do caso discreto.
Desenhar no quadro.

Identifique $\mathbb{E}(Y)$ em cada caso

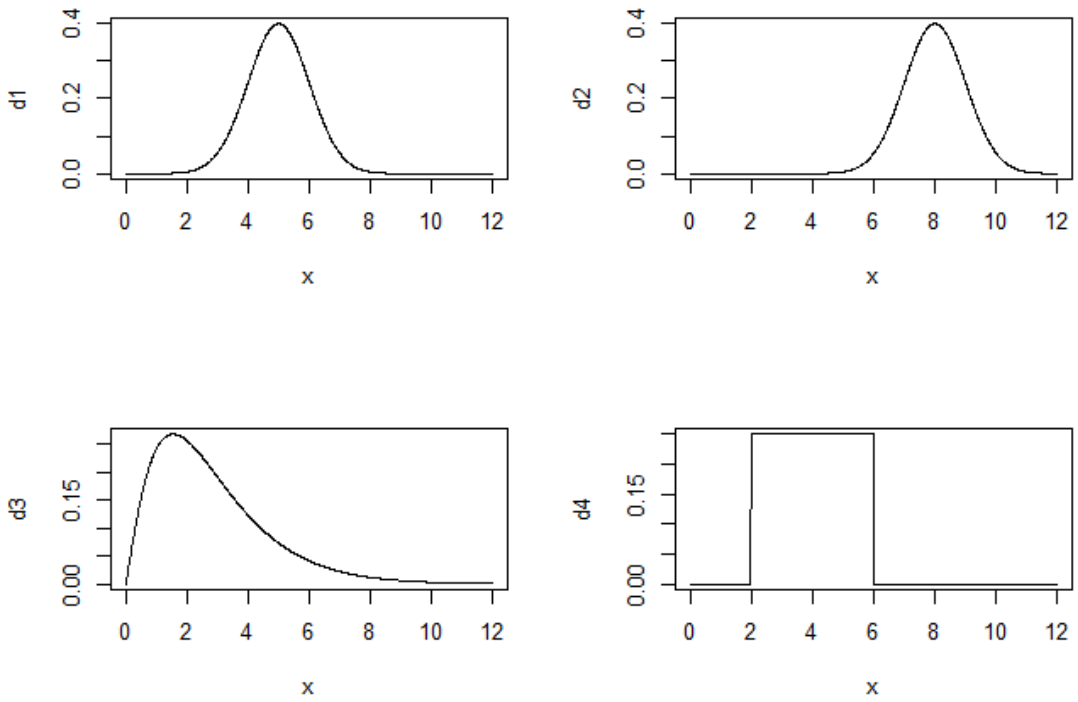


Figura: Sem fazer nenhuma conta, identifique as distribuições com as seguintes esperanças: $1.8, 8, 5, 4$

Propriedades da esperança: linearidade

- Seja Y uma v.a. e crie uma nova v.a. $Y = a + bX$ onde a e b são constantes.
- Por exemplo, suponha que medidmos a temperatura aleatória C de certo ambiente em graus Celsius.
- Suponha que o valor esperado de C seja $\mathbb{E}(C) = 28$ graus.
- Seja F a variável aleatória que mede a mesma temperatura em graus Fahrenheit.
- É claro que C e F estão relacionadas. Temos $F = 32 + (9/5)C$.
- Isto é, temos $a = 32$ e $b = 9/5$.
- $\mathbb{E}(F) = \mathbb{E}(a + bC)$ e $\mathbb{E}(C)$ estão relacionadas:
- A esperança da v.a. F pode ser obtida diretamente a partir daquela de C:

$$\mathbb{E}(F) = \mathbb{E}(32 + (9/5)C) = 32 + (9/5)\mathbb{E}(C) = 32 + (9/5) \times 28$$

Propriedades da esperança: linearidade

- Caso geral, $Y = a + bX$ onde a e b são constantes.
- Então $\mathbb{E}(X)$ e $\mathbb{E}(Y)$ estão relacionadas;

$$\mathbb{E}(X) = \mathbb{E}(a + bY) = a + b\mathbb{E}(Y)$$

$\mathbb{E}(a + bY) = a + b\mathbb{E}(Y)$

- Prova apenas num caso específico com v.a.'s discretas:
- Considere a v.a. X com os valores possíveis $x_1, x_2, x_3, \ldots$ onde
- Consider a NOVA v.a. $Y = 2 + 3X$ que tem os valores possíveis $y_1, y_2, y_3, \ldots$ onde $y_i = 2 + 3x_i$.
- Além disso, temos

$$\mathbb{P}(Y = y_i) = \mathbb{P}(Y = 2 + 3x_i) = \mathbb{P}(X = x_i)$$

pois $[Y = y_i]$ se, e somente se, $[X = x_i]$ onde $x_i = (y_i - 2)/3$ ou $y_i = 2 + 3x_i$.

- Por exemplo, $\mathbb{P}(Y = 8) = \mathbb{P}(Y = 2 + 3 \times 2) = \mathbb{P}(X = 2)$
- Assim, podemos calcular a esperança de $Y = 2 + 3X$:
-

$$\mathbb{E}(Y) = \sum_i y_i \mathbb{P}(Y = y_i) = \sum_i (2+3x_i)\mathbb{P}(X = x_i) = 2\sum_i \mathbb{P}(X = x_i) + 3\sum_i x_i \mathbb{P}(X = x_i) = 2\times$$

Propriedades da esperança

- Uma escolha muito especial para estas constantes é a seguinte:

$$a = -\mathbb{E}(X) = -\mu \text{ e } b = 1$$

- Neste caso, temos $Y = a + bX = X - \mu$ onde $\mathbb{E}(X) = \mu$.
- Isto é, estamos olhando para a v.a. $Y = X - \mathbb{E}(X)$, a v.a. X menos seu próprio valor esperado.
- Pela propriedade, temos

$$\mathbb{E}(Y) = \mathbb{E}(X - \mu) = \mathbb{E}(X) - \mu = \mu - \mu = 0$$

- Dizemos que a v.a. Y é a v.a. centrada (em sua esperança).

Propriedades da esperança: linearidade

- Se $X_1, X_2, \ldots, X_n$ são v.a.'s e $a_0, a_1, a_2, \ldots, a_n$ são constantes então

$$\mathbb{E}(a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_0 + a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \ldots + a_n\mathbb{E}(X_n)$$

- Em particular:

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

- Prova do caso particular de duas v.a.'s discretas.
- A v.a. X possui os valores possíveis $x_1, x_2, \ldots$
- A v.a. Y possui os valores possíveis $y_1, y_2, \ldots$
- A v.a. $X+Y$ possui os valores possíveis $x_i + y_j$ onde x_i e y_j varrem todas as possibilidades para X e Y.
- Assim, temos

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_i \sum_j (x_i + y_j)\mathbb{P}(X = x_i, Y = y_j) \\
&= \sum_i \sum_j x_i \mathbb{P}(X = x_i, Y = y_j) + \sum_j \sum_i y_j \mathbb{P}(X = x_i, Y = y_j) \\
&= \sum_i x_i \sum_j \mathbb{P}(X = x_i, Y = y_j) + \sum_j y_j \sum_i \mathbb{P}(X = x_i, Y = y_j)
\end{aligned}$$

- Vamos obter as somas destas probabs.

$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$

- O evento $[X = x_i]$ é a união dos eventos *disjuntos* $[X = x_i, Y = y_1]$, $[X = x_i, Y = y_2], \ldots, [X = x_i, Y = y_m]$:

$$[X = x_i] = [X = x_i, Y = y_1] \bigcup [X = x_i, Y = y_2] \bigcup \ldots \bigcup [X = x_i, Y = y$$

- A probab da união de eventos DISJUNTOS é a soma das probabs:

$$\mathbb{P}(X = x_i) = \mathbb{P}(X = x_i, Y = y_1) + \mathbb{P}(X = x_i, Y = y_2) + \ldots + \mathbb{P}(X = x_i, Y = y_n)$$

Propriedades da esperança: linearidade

- Assim, temos

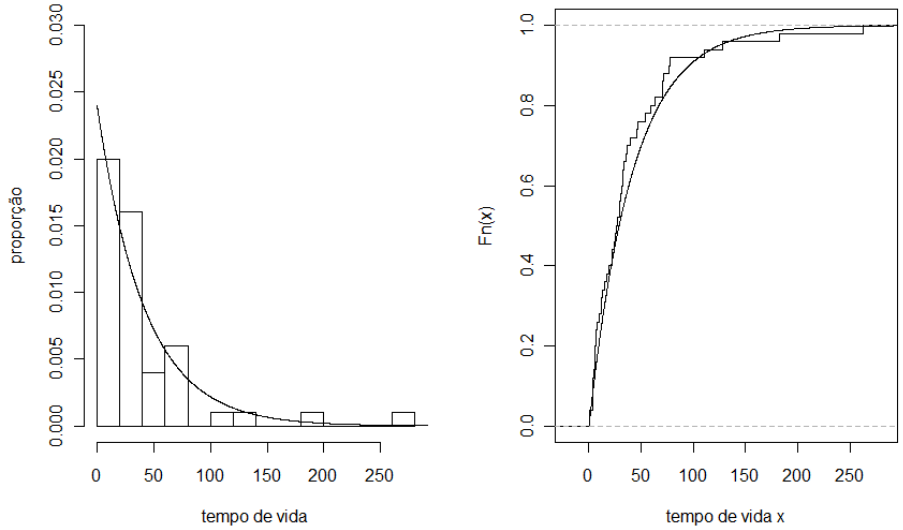
$$\begin{aligned}
\mathbb{E}(X+Y) &= \dots \\
&= \sum_i x_i \sum_j \mathbb{P}(X=x_i, Y=y_j) + \sum_j y_j \sum_i \mathbb{P}(X=x_i, Y=y_j) \\
&= \sum_i x_i \mathbb{P}(X=x_i) + \sum_j y_j \mathbb{P}(Y=y_j) \\
&= \mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned}$$

Propriedades da esperança

- Suponha que a v.a. X seja um valor constante.
- Isto é, para todo resultado ω do experimento a v.a. assume o valor $X(\omega) = c$.
- Um resultado particular óbvio mas muito útil é que, para esta variável que é sempre igual a c, o valor que podemos esperar para ela é ... c.
- A prova é simples: X é discreta com um único valor possível, c.
- Portanto, $\mathbb{E}(X) = c\mathbb{P}(X = c) = c \times 1 = c$

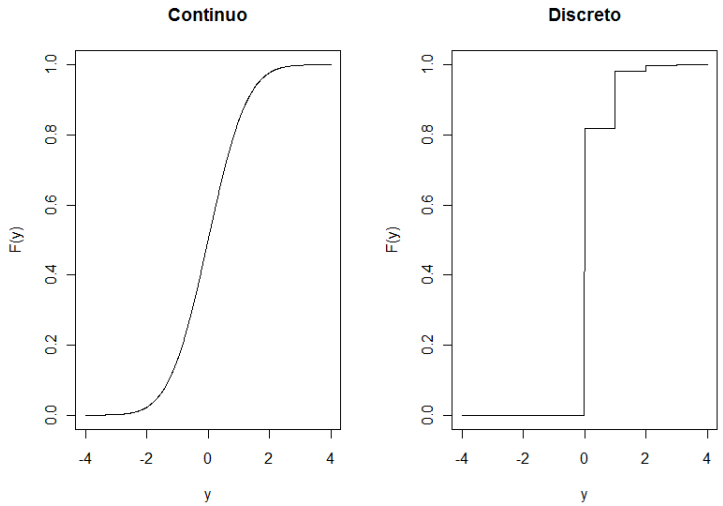
Dificuldades...

- O caso contínuo pode não ser tão simples: pode ser que o histograma não seja suficiente.
- Abaixo, um histograma de uma amostra de uma v.a. contínua com uma densidade candidata sobreposta.
- Como decidir? Qui-quadrado é uma opção mas precisa criar as categorias.
- O segundo gráfico é uma função menos intuitiva mas mais útil.



Função Distribuição Acumulada

- A função distribuição acumulada é uma função *matemática* que mostra como as probabilidades vão se acumulando no eixo real.
- Temos sempre $F : \mathbb{R} \to [0, 1]$
- 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 $F(y) = \sum_{x_i \leq y} \mathbb{P}(Y = x_i)$



Caso contínuo

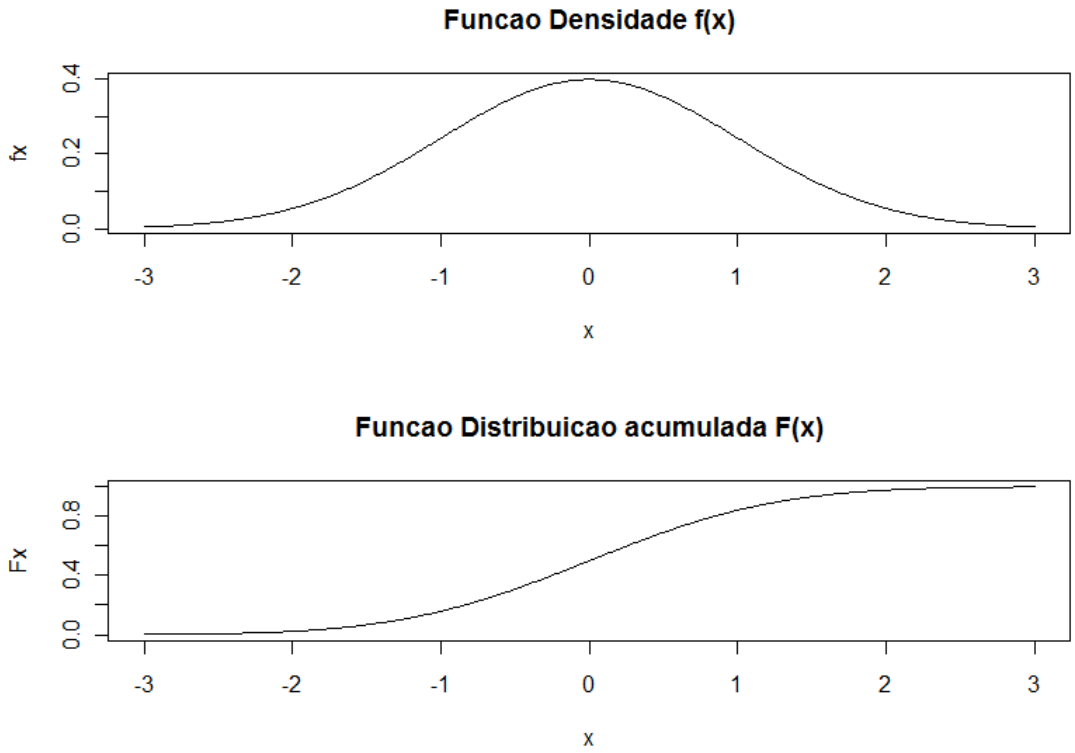


Figura: Densidade $f(x)$ e Função Distribuição Acumulada $F(y)$

Caso discreto

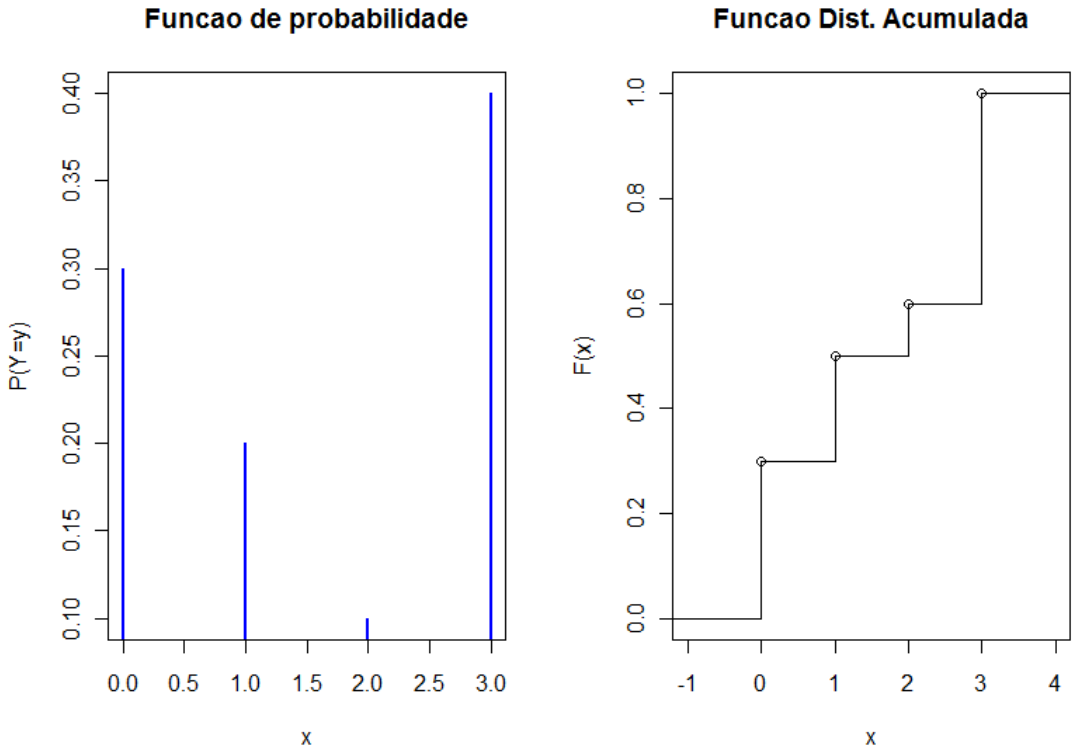


Figura: Função de probabilidade $\mathbb{P}(Y = y)$ e e função distribuição acumulada $F(y)$. Y tem quatro valores possíveis, 0, 1, 2, 3, com probabilidades iguais a 0.3, 0.2, 0.1 e 0.4, respectivamente

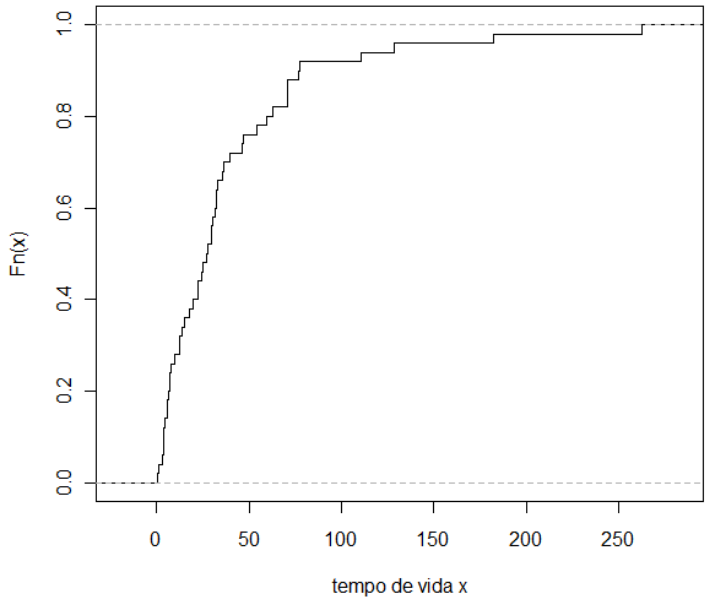
Importância de $F(y)$

- 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 técnicas.
- Vamos ver uma delas agora.

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 $F(y)$.
- Com base na amostra, E SOMENTE NELA, construímos a função distribuição acumulada empírica $\widehat{F}_n(y)$.
- Se tivermos $\widehat{F}_n(y) \approx F(y)$ para todo y concluimos que o modelo adotado ajusta-se bem aos dados.
- Como saber se $\widehat{F}_n(y) \approx F(y)$?

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.

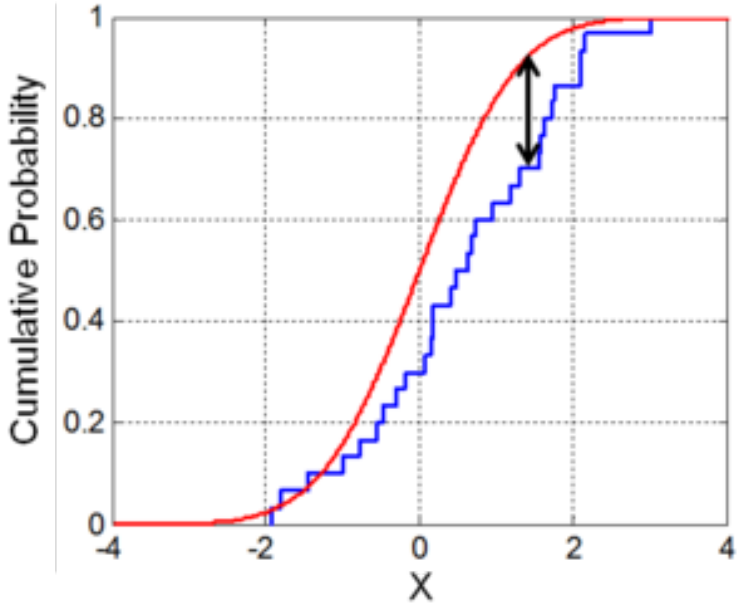
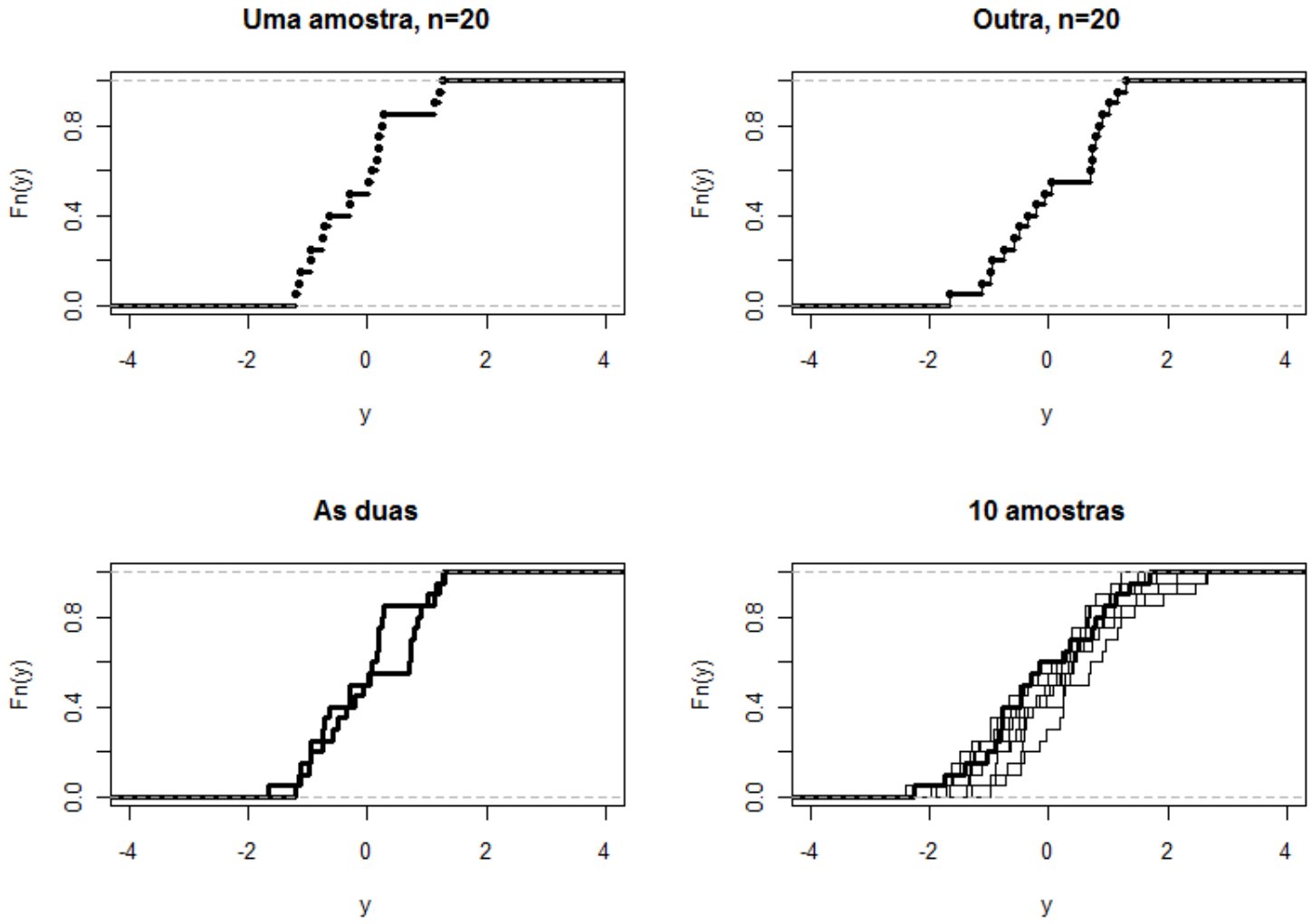


Figura: Empírica $\hat{F}_n(y)$ e a teórica $F(y)$.

$\hat{F}_n(y)$ é aleatória



- Figura: Caráter aleatório de $\hat{F}_n(y)$

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

- Suponha que $F(y)$ é o modelo verdadeiro (neste caso, uma $N(0,1)$).
- Então $D_n \to 0$ se $n \to \infty$.

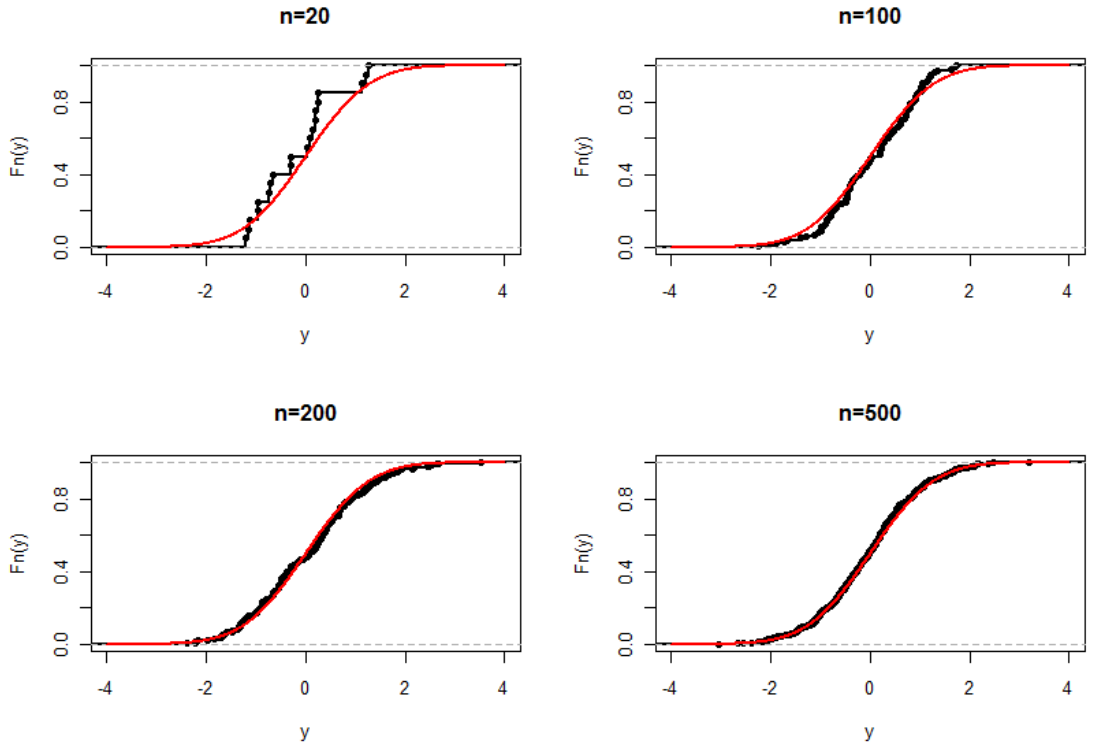
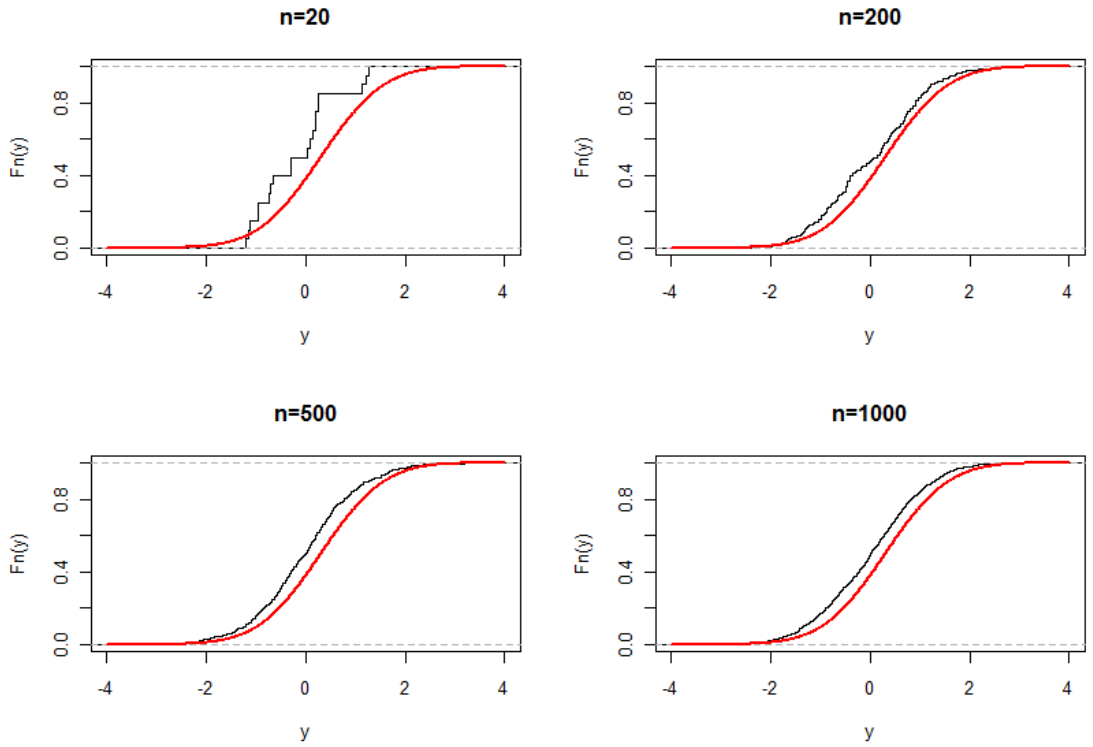


Figura: $D_n \to 0$ se o modelo é correto

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

- 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 D_n converge para um valor > 0.



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

- 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, depende de n já que $D_n \to 0$ se $n \to \infty$.
- A distância a zero para ser considerado próximo o suficiente depende do modelo $F(y)$?
- Por exemplo, o comportamento de D_n quando $F(y)$ for uma gaussiana é diferente do comportamento quando $F(y)$ for uma Pareto (power-law)?

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

- 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).
  - $\sqrt{n}D_n \nrightarrow 0$ e também $\nrightarrow \infty$.
  - $\sqrt{n}D_n$ 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).
- Densidade de K é dada por $f(x) = 8x\sum_{k=1}^{\infty}(-1)^{k+1}k^2e^{-2k^2x^2}$.
- 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 1.8.
- Se não usarmos o modelo verdadeiro, sabemos que $\sqrt{n}D_n \to \infty$.

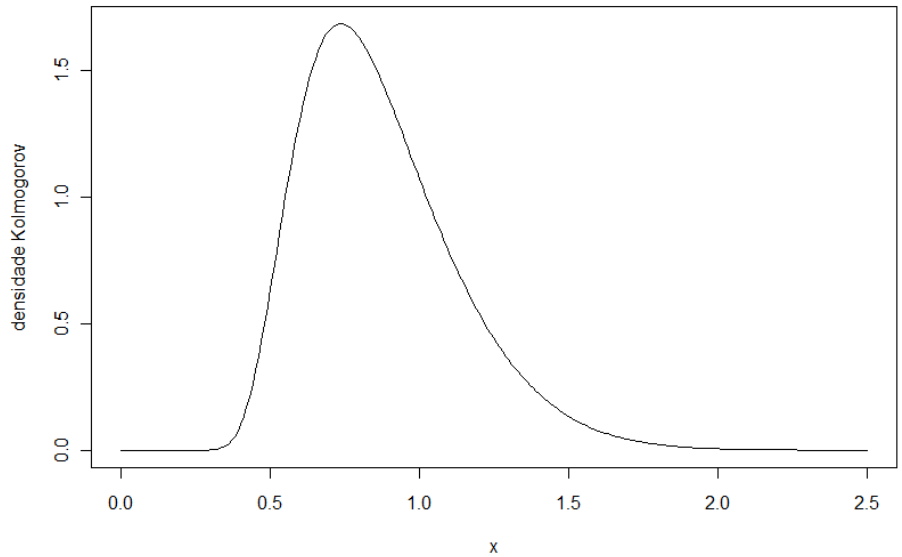


Figura: Densidade de $K \approx \sqrt{n}D_n$

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%.

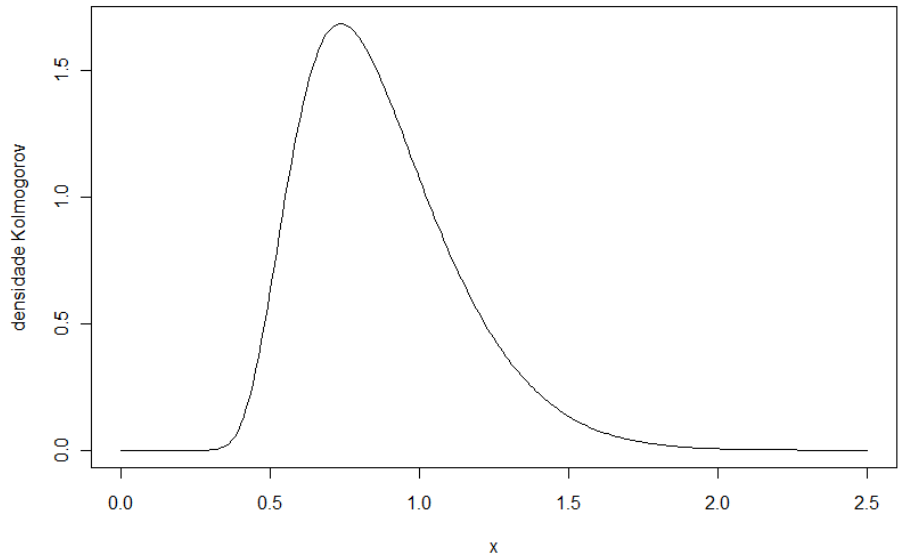


Figura: Densidade de $K \approx \sqrt{n}D_n$

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 $F(y)$.