# 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 $f(x) = 1/\sqrt{\pi}\exp(-x^2/2)$.
- A única restrição: $f(x) \geq 0$ para todo $x$ e sua integral deve ser $= 1$.

# Probabilidades estão associadas com áareas

- 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.

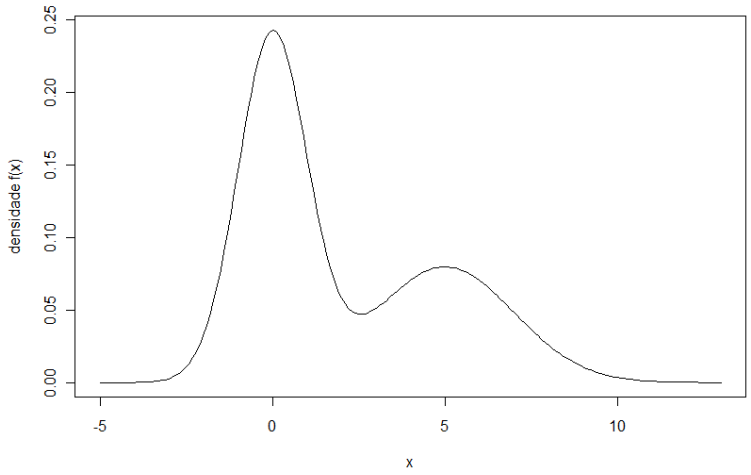


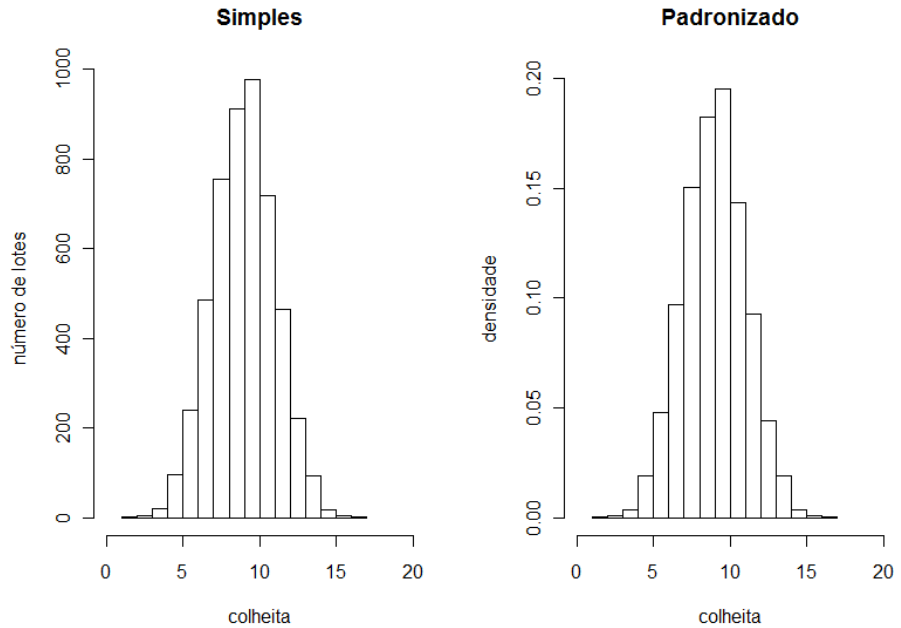

Figura: Densidade de probabilidade $f(x)$

# 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 $= 1$).

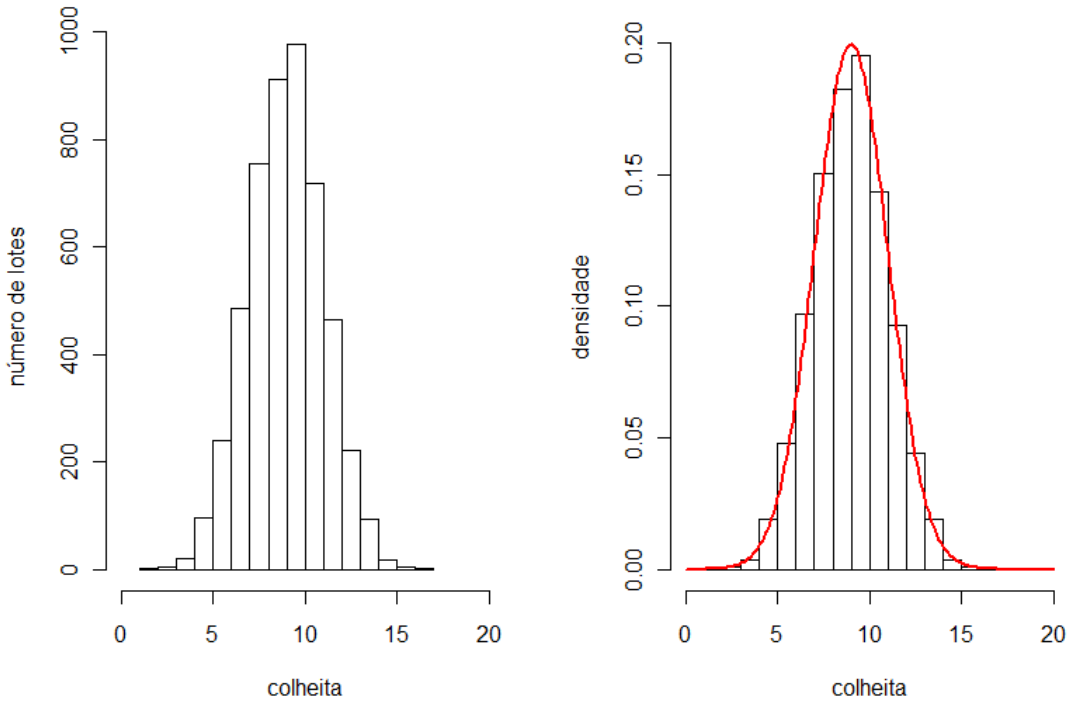
# Histograma

- Quebre o eixo horizontal em pequenos intervalos de comprimento $\Delta$.
- 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.

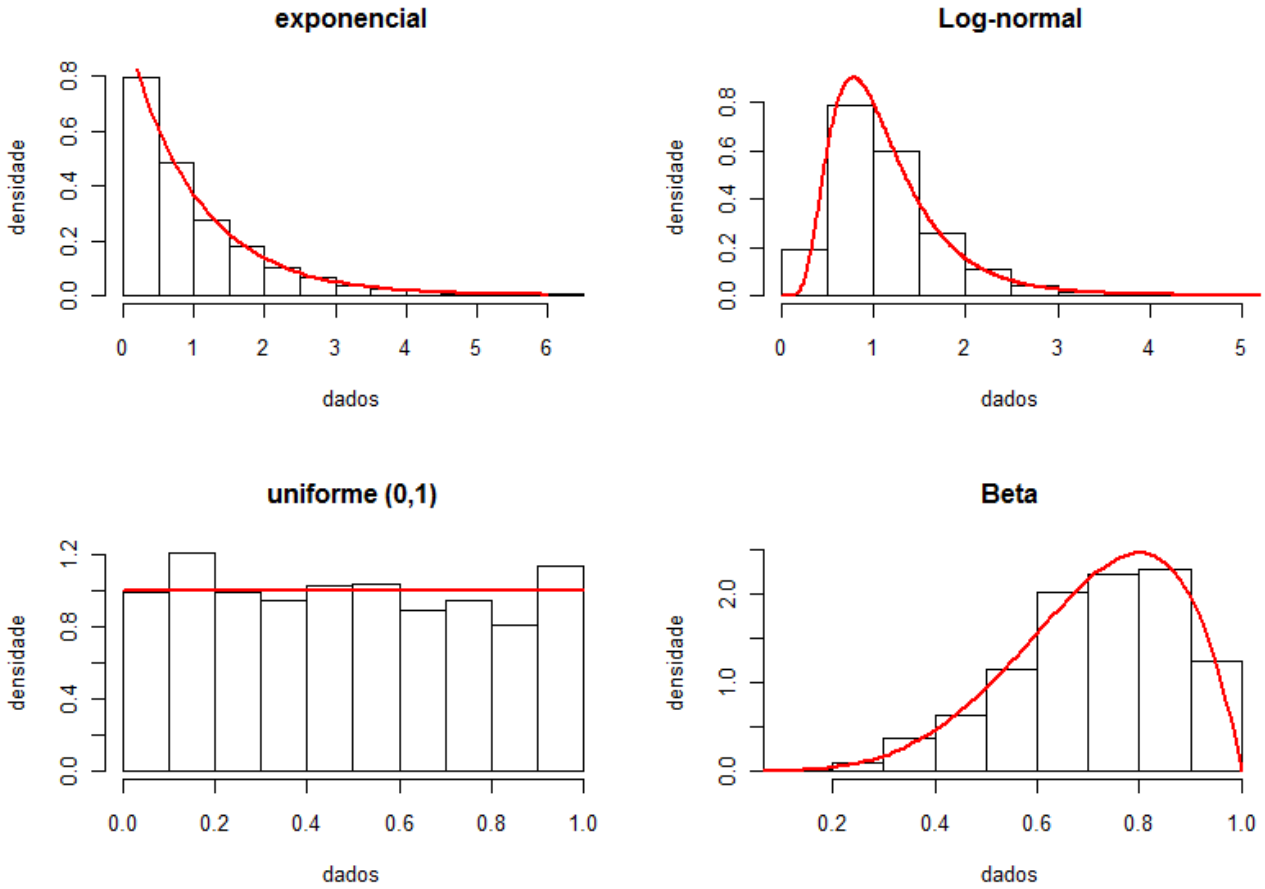



Figura: Pla

# 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 $\delta$.
- 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 $(y_0 - \delta/2, y_0 + \delta/2)$:

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

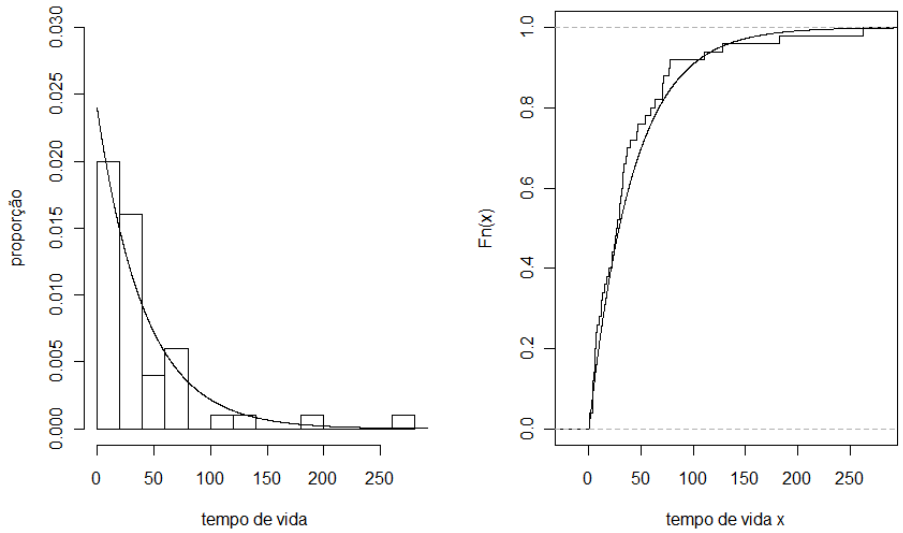
- Igualando as duas aproximações e dividindo por $\delta$ dos dois lados, temos

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

- 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).

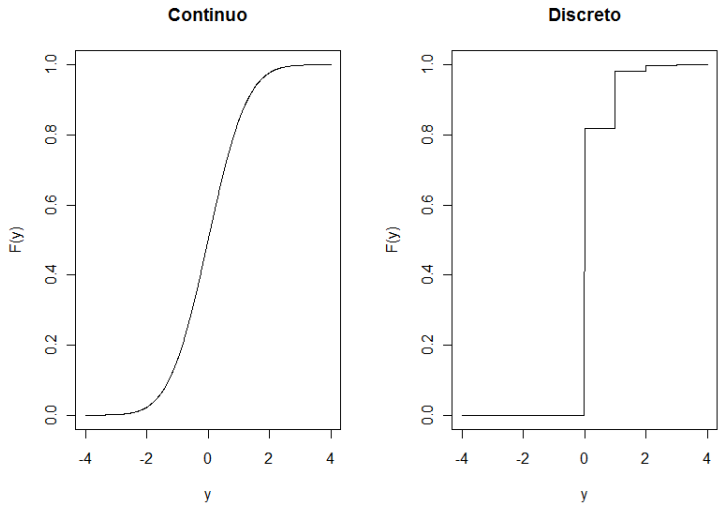
# Dificuldades...

- 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

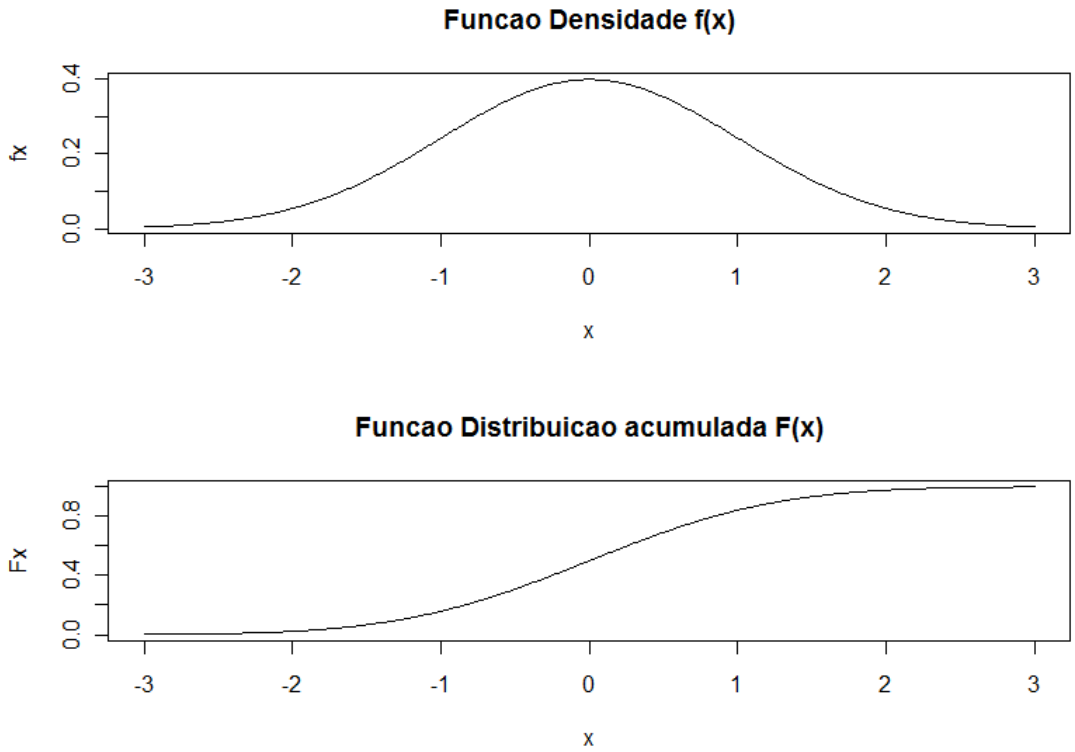



Figura: Densidade $f(x)$ e Função Distribuição Acumulada $F(y)$

# Caso discreto

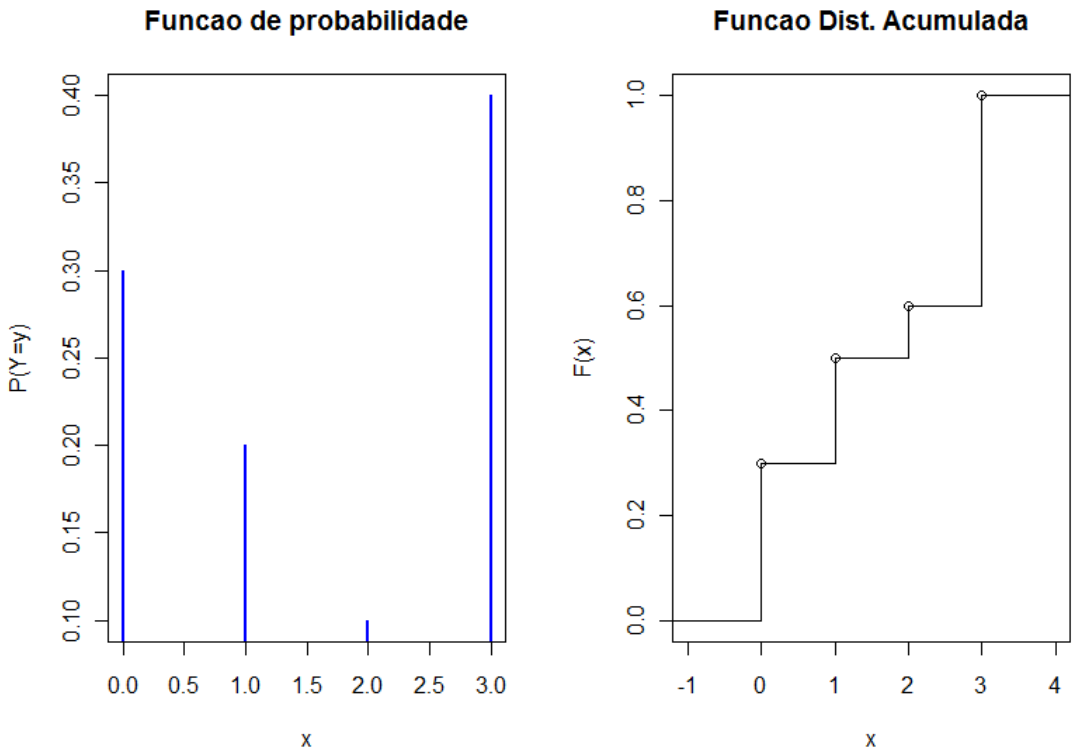



Figura: Função de probabilidade $\mathbb{P}(Y = y)$ e e função distribuição acumulada $F(y)$. $Y$ tem quatro valores possíveis, 0, 1, 2, 3, com probabilidades iguais a 0.3, 0.2, 0.1 e 0.4, respectivamente

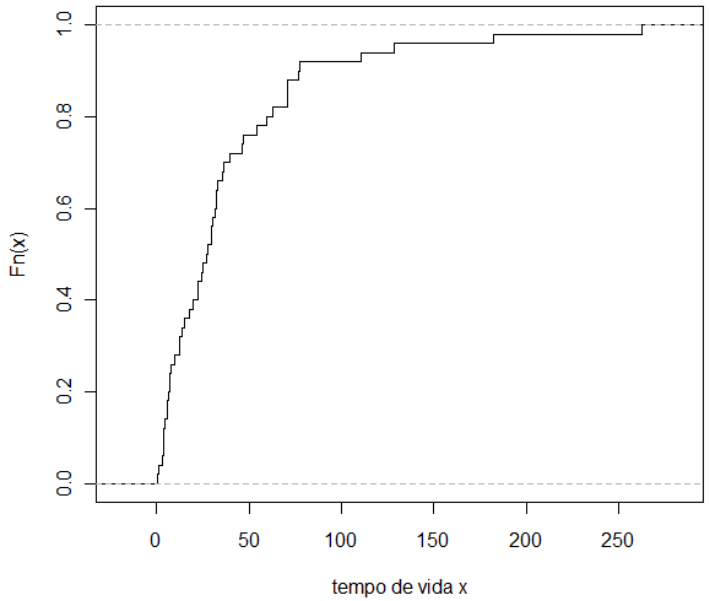
# Importância de $F(y)$

- 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

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



# 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 $D_n$.

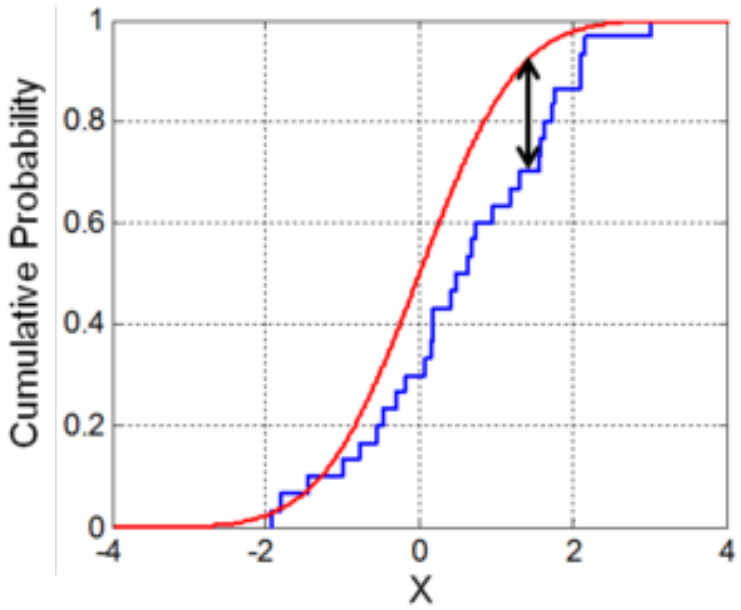



Figura: Empírica $\hat{F}_n(y)$ e a teórica $F(y)$.

# $\hat{F}_n(y)$ é aleatória

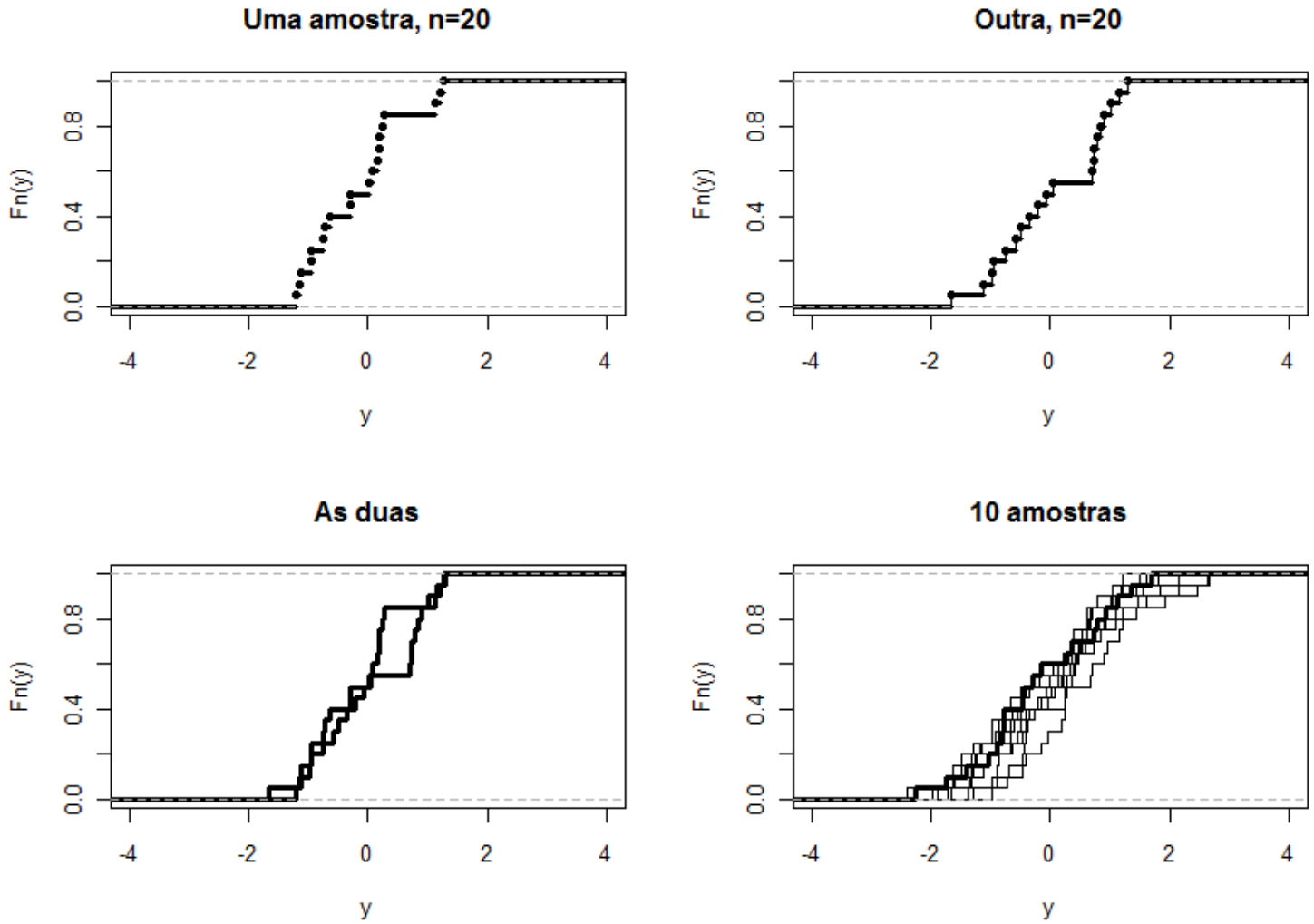



Figura: Caráter aleatório de $\hat{F}_n(y)$

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

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

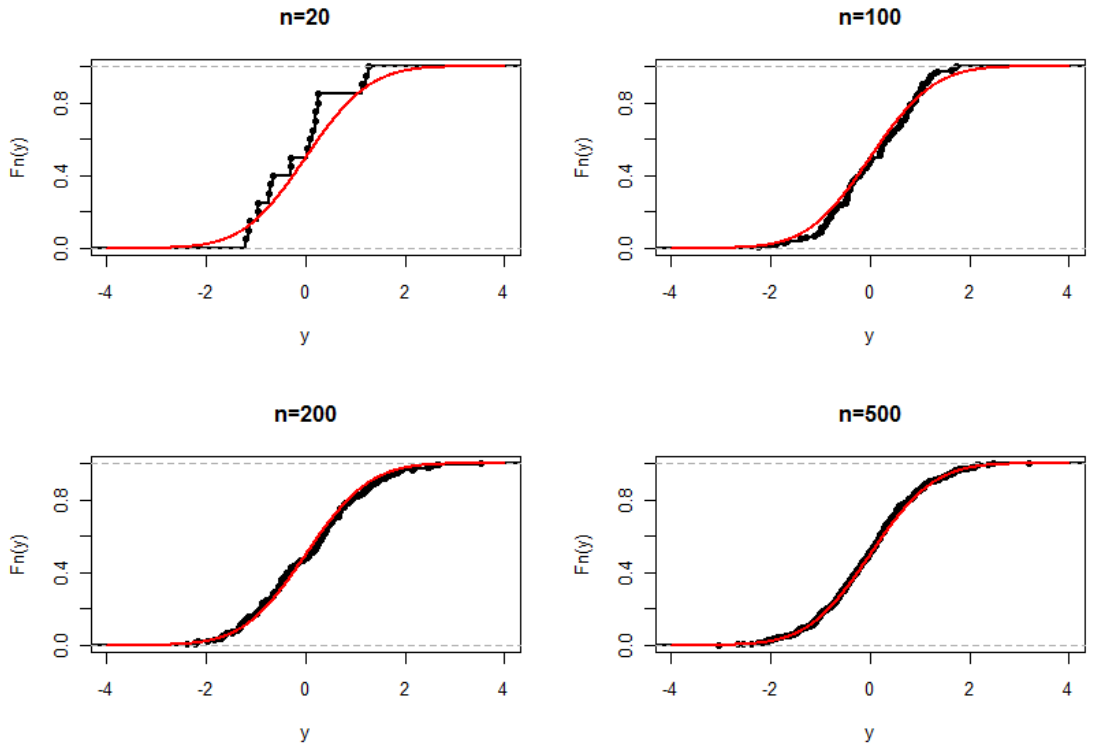



Figura: $D_n \to 0$ se o modelo é correto

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

- 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 $> 0$.

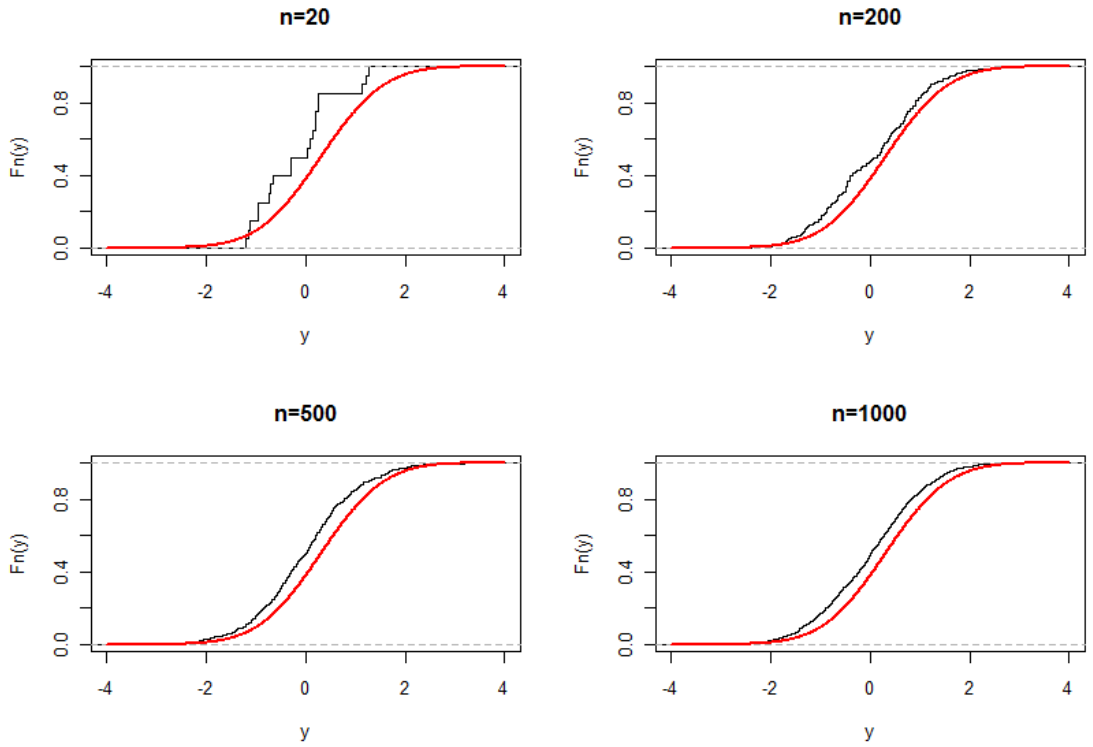



Figura: $D_n \to 0$ se o modelo é correto

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

- 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).
  - $\sqrt{n}D_n \not\to 0$ 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).
- Densidade de $K$ é dada por $f(x) = 8x \sum_{k=1}^{\infty} (-1)^{k+1} k^2 e^{-2k^2x^2}$.
- 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 $\sqrt{n}D_n \to \infty$.

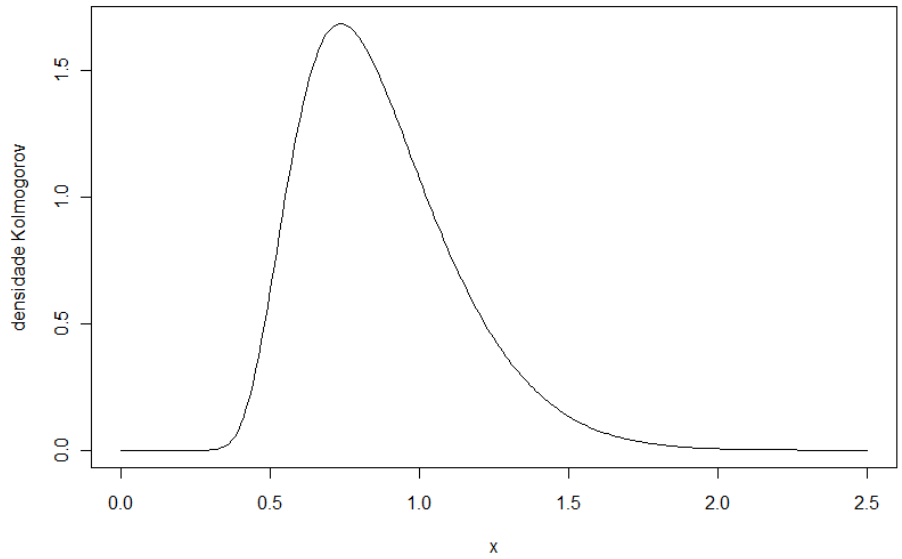



Figura: Densidade de $K \approx \sqrt{n}D_n$

# 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%.

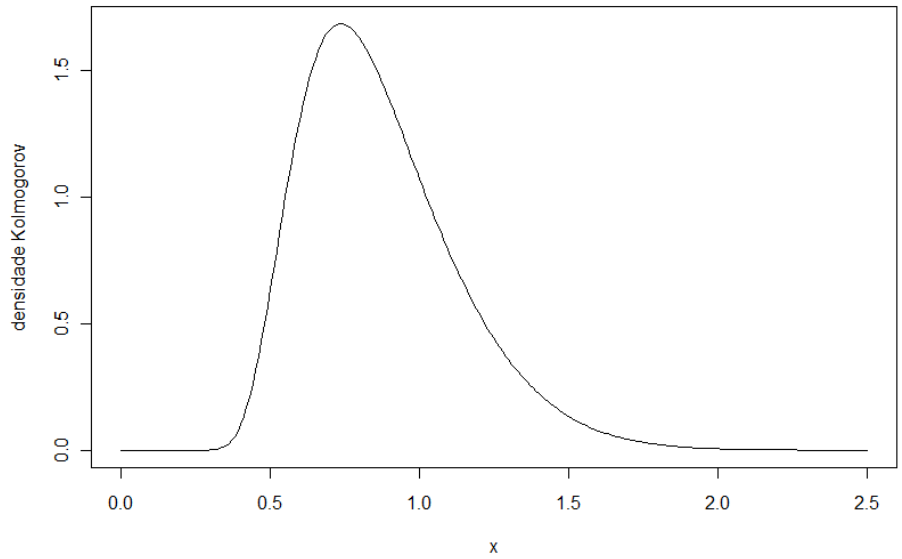
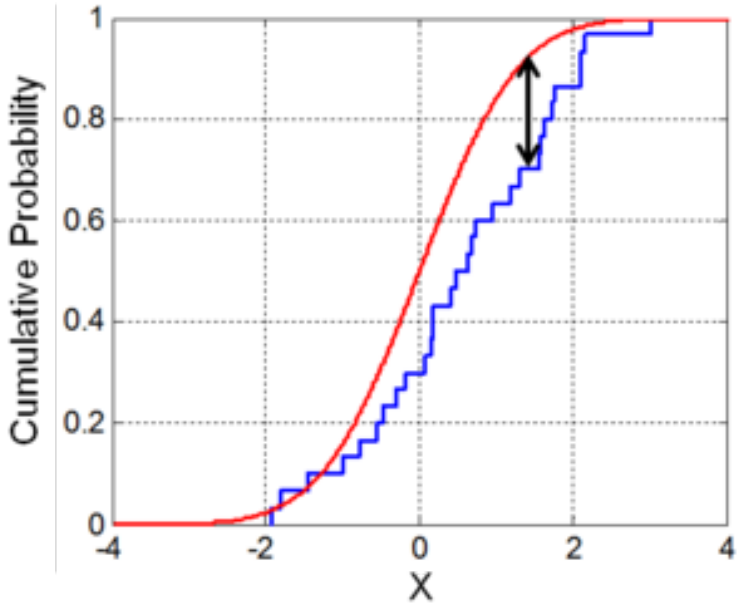



Figura: Densidade de $K \approx \sqrt{n}D_n$

# 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 $F(y)$.



# 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; 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 $\mu$ 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 $Y_1, Y_2, \ldots, Y_n$ 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 $F(y)$ (por EMV, + tarde), se for necessário
- Implementação muito fácil.
- Precisa especificar os intervalos ou classes onde as contagens vão ser feitas.
- Qual o efeito desta escolha? Quanto mais melhor, mas muitas classes podem levar a contagens 0 ou próximas de zero.
- Devemos escolher classes de forma que o número esperado em cada uma delas seja, de preferência, pelo menos 5. Classes com menos de 1 devem ser evitadas.

# 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 $\mathbb{E}(Y)$.
- Quão em torno?? DP = desvio-padrão.
- O valor TEÓRICO que mede o quanto os valores oscilam em torno de $\mathbb{E}(Y)$. Isto é, o desvio-padrão DP. $\sigma = \sqrt{\text{Var}(Y)}$.

# $\mathbb{E}(Y)$ 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 $\geq 0$ e somam 1.
- $\mathbb{E}(Y)$ geralmente NÃO É um dos valores possíveis $\{x_1, x_2, \ldots\}$.

# $\mathbb{E}(Y)$ no caso contínuo

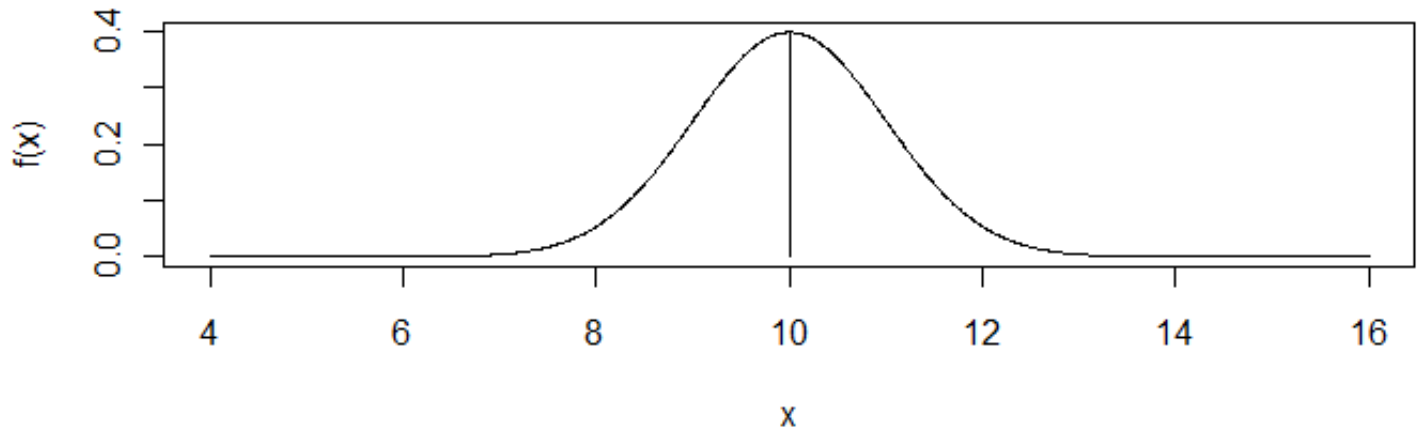
- Caso contínuo: $\mathbb{E}(Y) = \int_{-\infty}^{\infty} yf(y)dy$
- Podemos raciocionar intuitivamente EXATAMENTE como no caso discreto.
- Quebrar todo eixo real em pequenos bins de comprimento $\Delta$ e centrados em $\ldots, y_{-2}, y_{-1}, y_0, y_1, y_2, \ldots$.
- Então, em cada pequeno bin, aproxime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

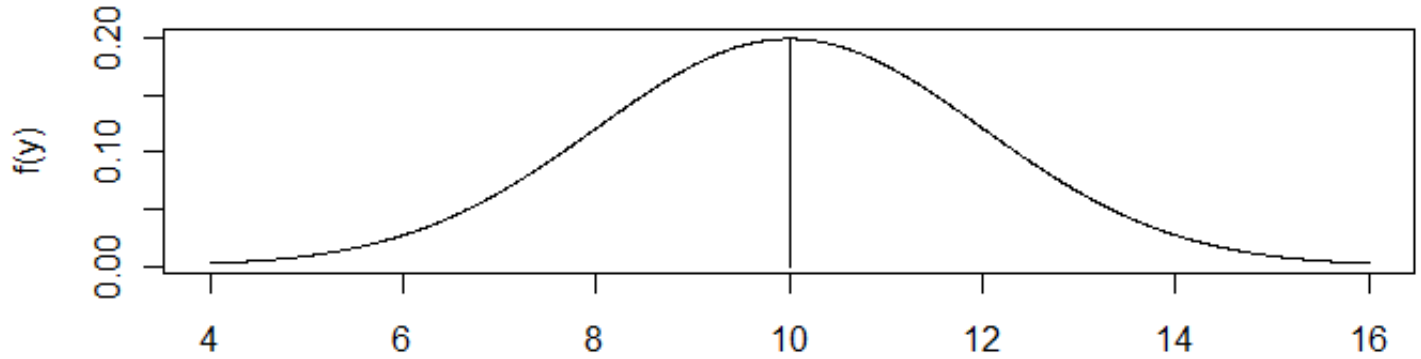
$$\int_{\text{bin}_i} yf(y)dy \approx y_i f(y_i)\Delta$$

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

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

# $X$ e $Y$: Qual possui maior variância?



# Gerando as amostras

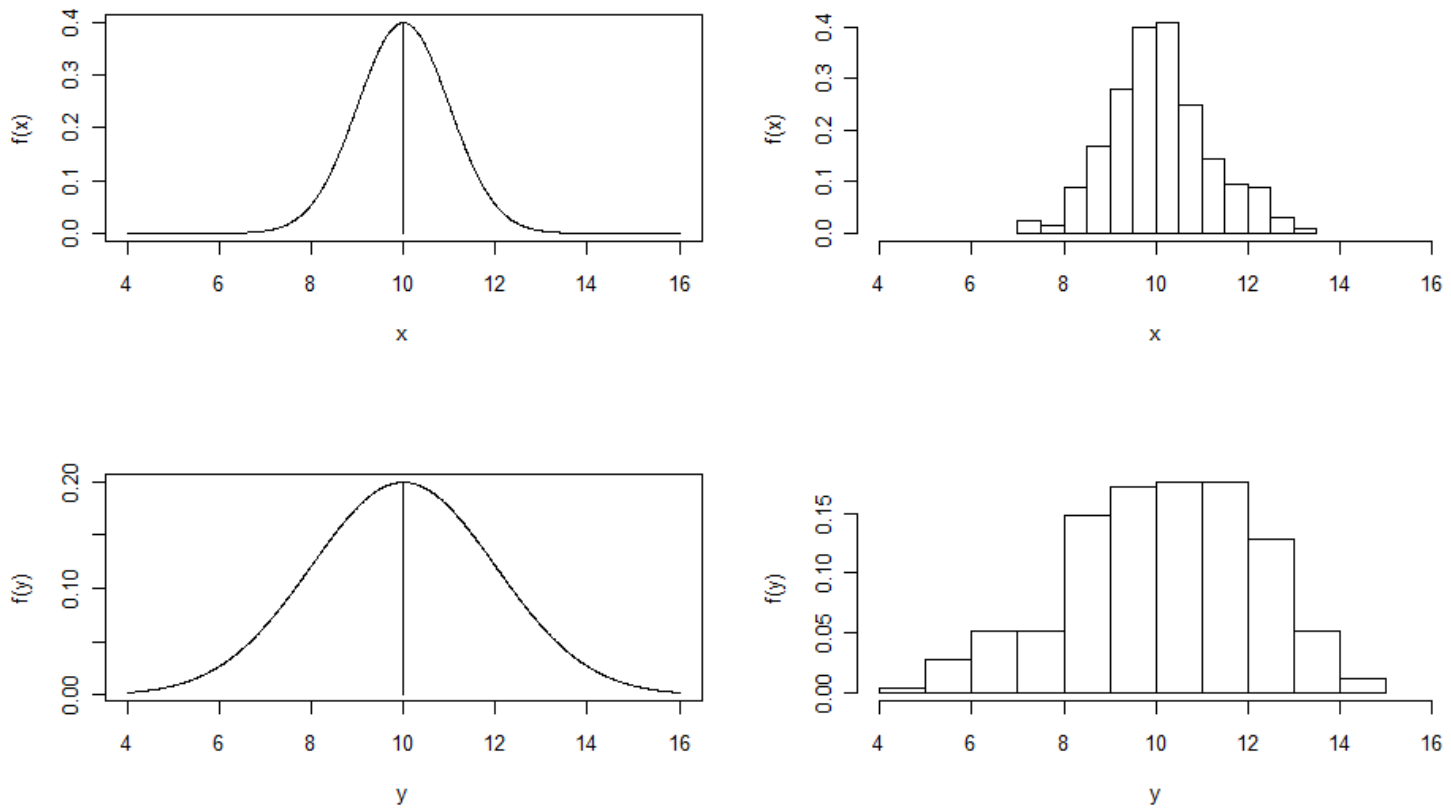
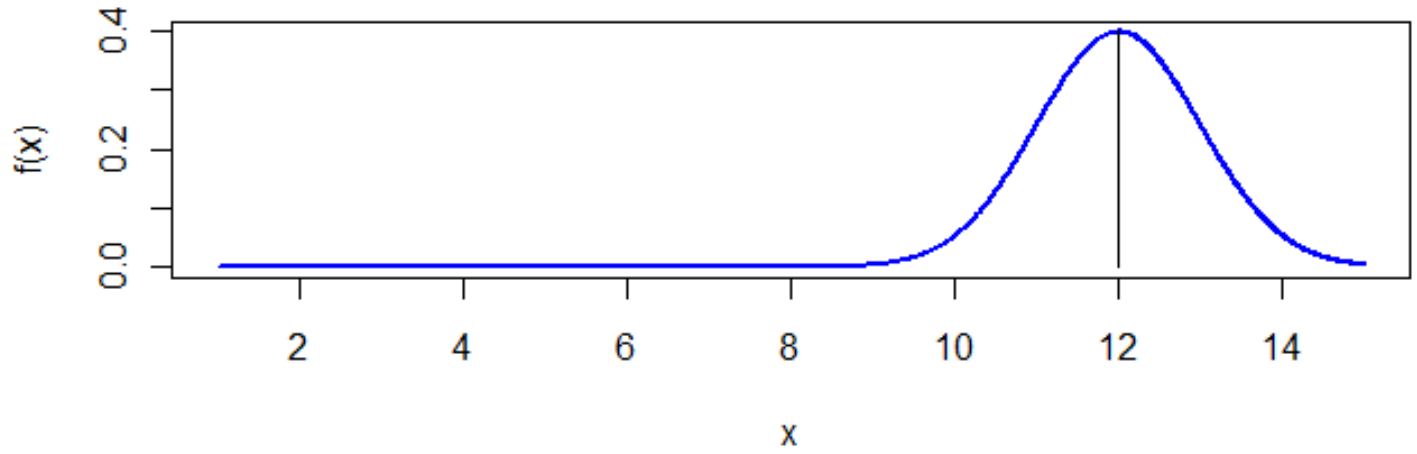



Figura: Histogramas de amostras e densidades de $X$ e $Y$ com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 10$. Qual das amostras varia mais em torno do seu valor médio?

# Não precisa ter $\mathbb{E}(X) = \mathbb{E}(Y)$




# As densidades e as amostras

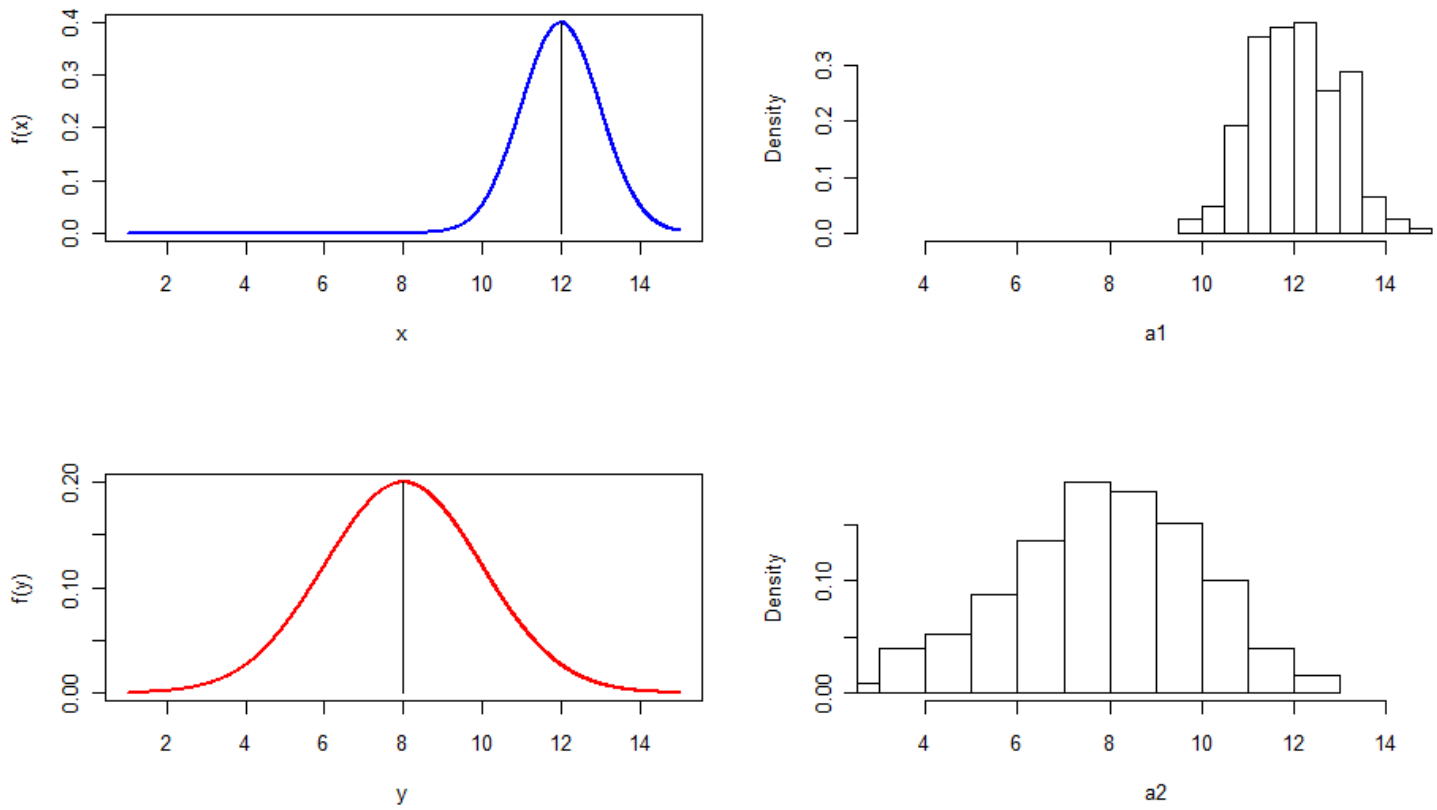
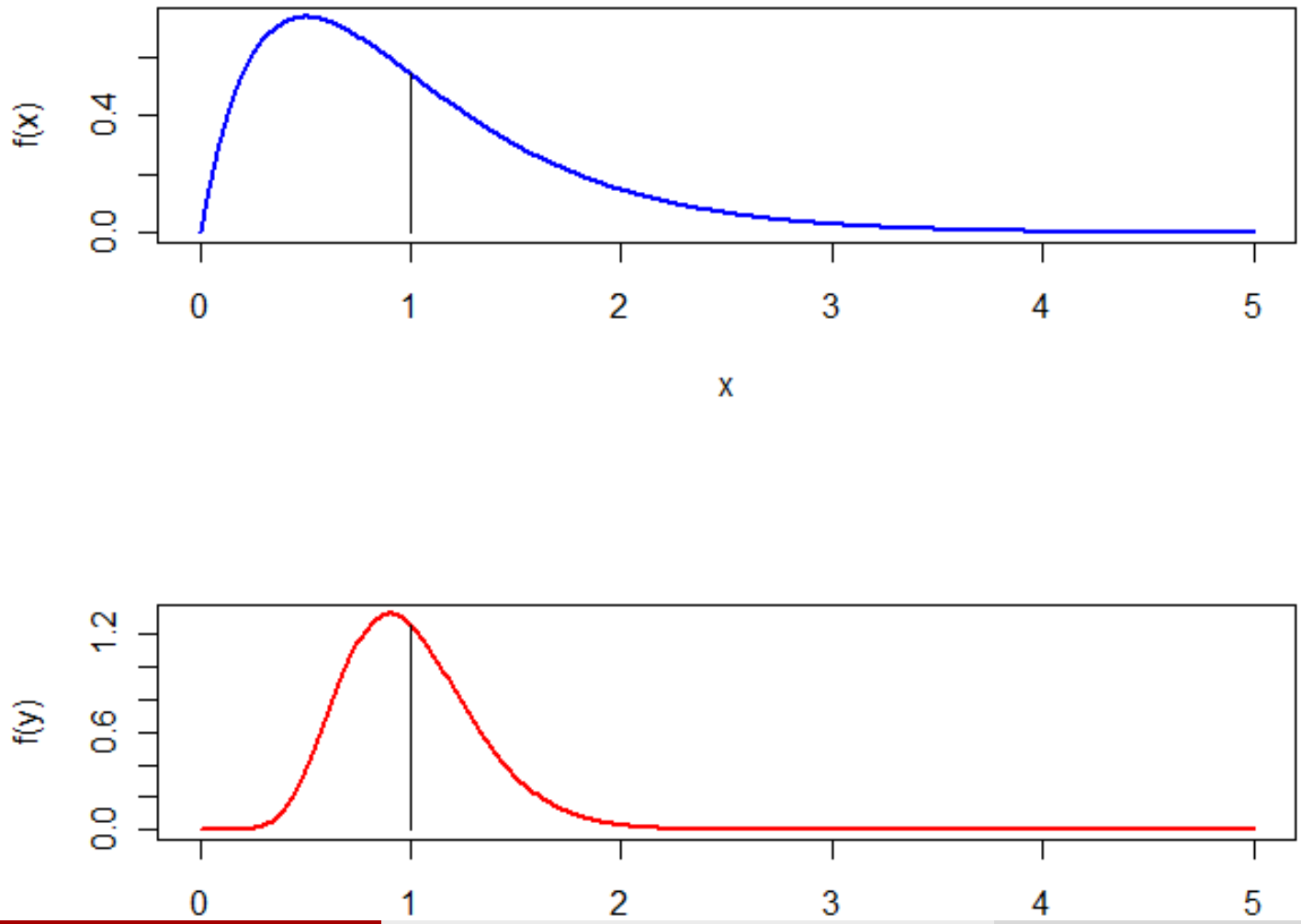



Figura: Histogramas de amostras e densidades de $X$ e $Y$ com diferentes valores esperados: $\mathbb{E}(X) \neq \mathbb{E}(Y)$. Qual das amostras varia mais em torno do seu valor médio?

# Não precisa ter densidade simétrica



# Com as amostras de cada distribuição

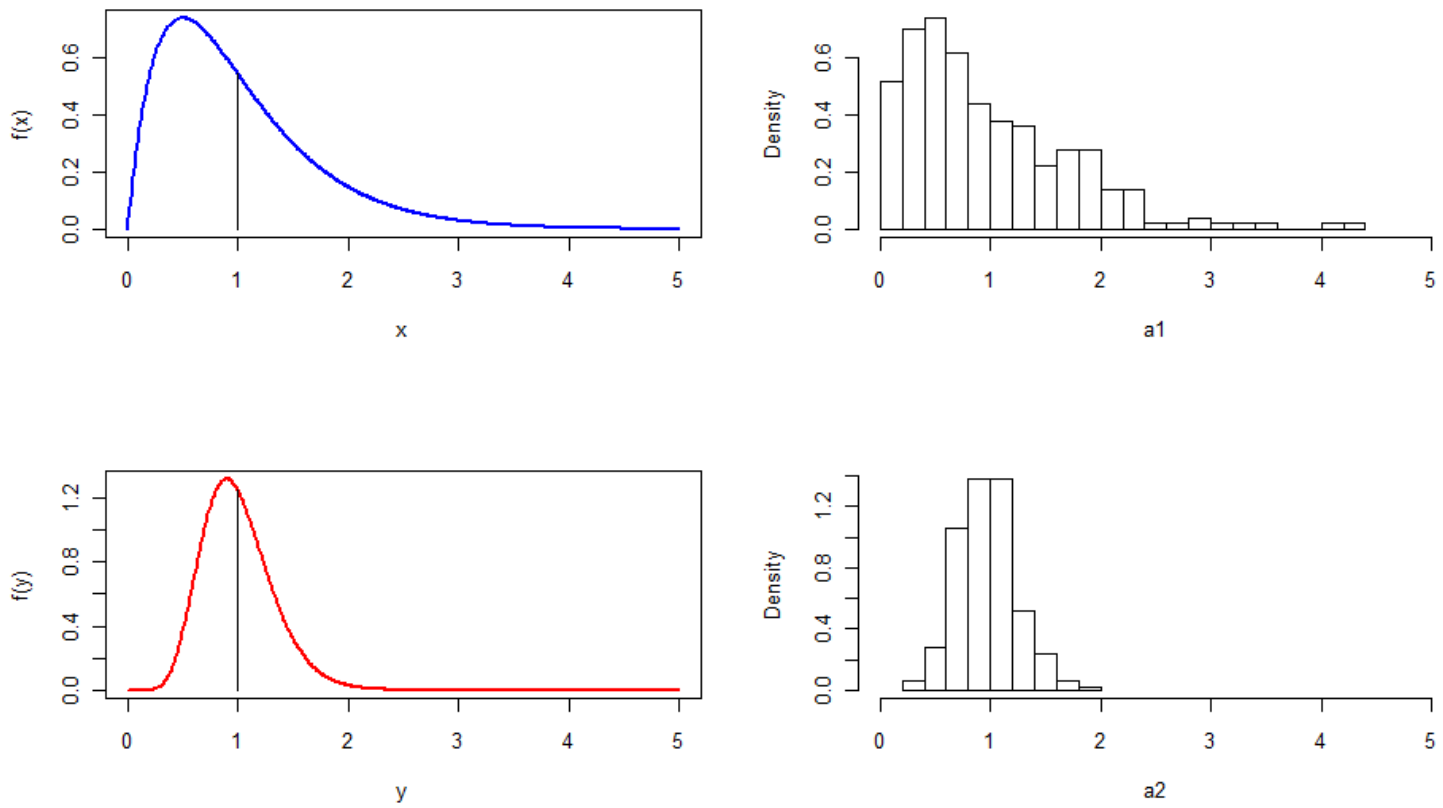



Figura: Histogramas de amostras e densidades ASSIMÉTRICAS de $X$ e $Y$ com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 1$. Qual das amostras varia mais em torno do seu valor médio?

# Assimétricas e com $\mathbb{E}(X) \neq \mathbb{E}(Y) = 1$

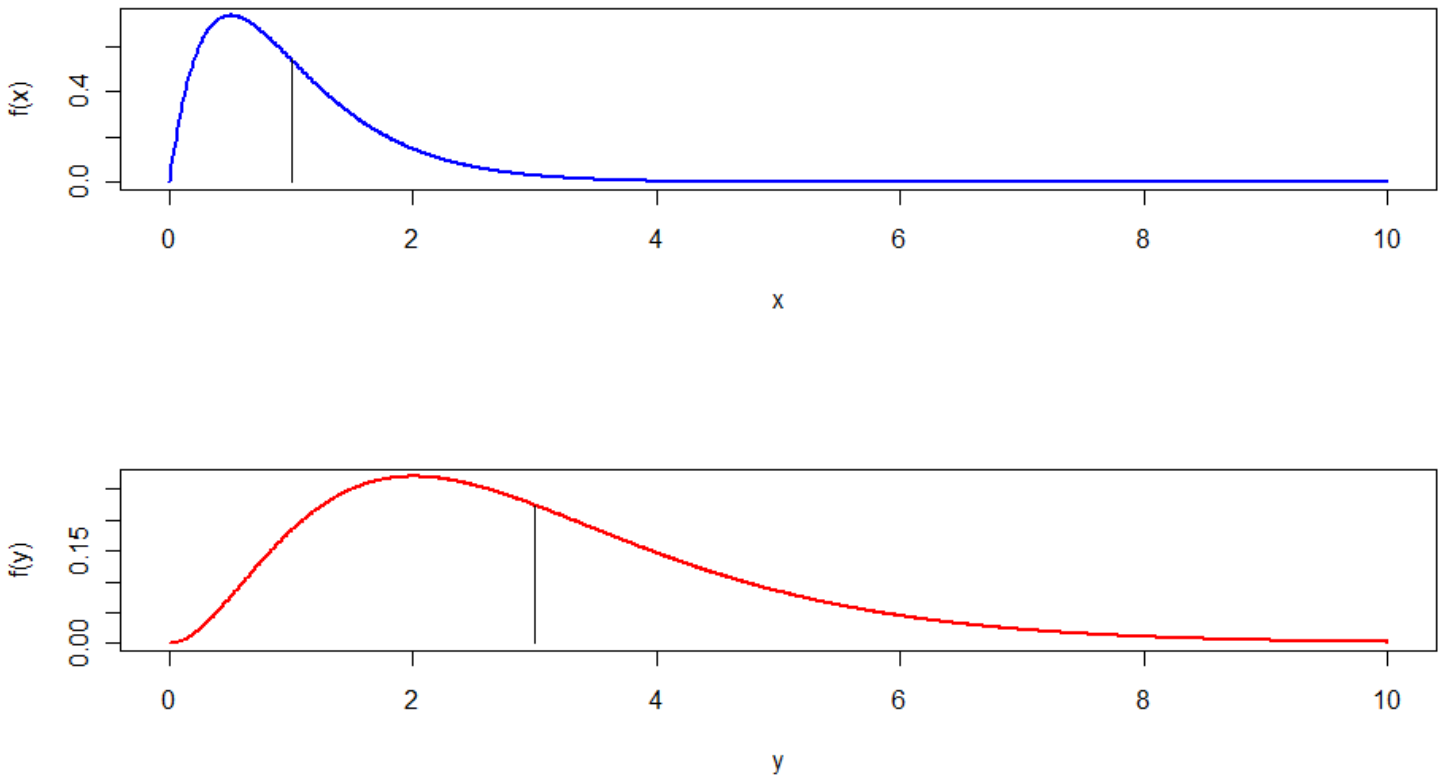



Figura: Densidades de $X$ e $Y$ com $1 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3$. Qual das amostras varia mais em torno do seu valor médio?

# Assimétricas e com $\mathbb{E}(X) \neq \mathbb{E}(Y) = 1$

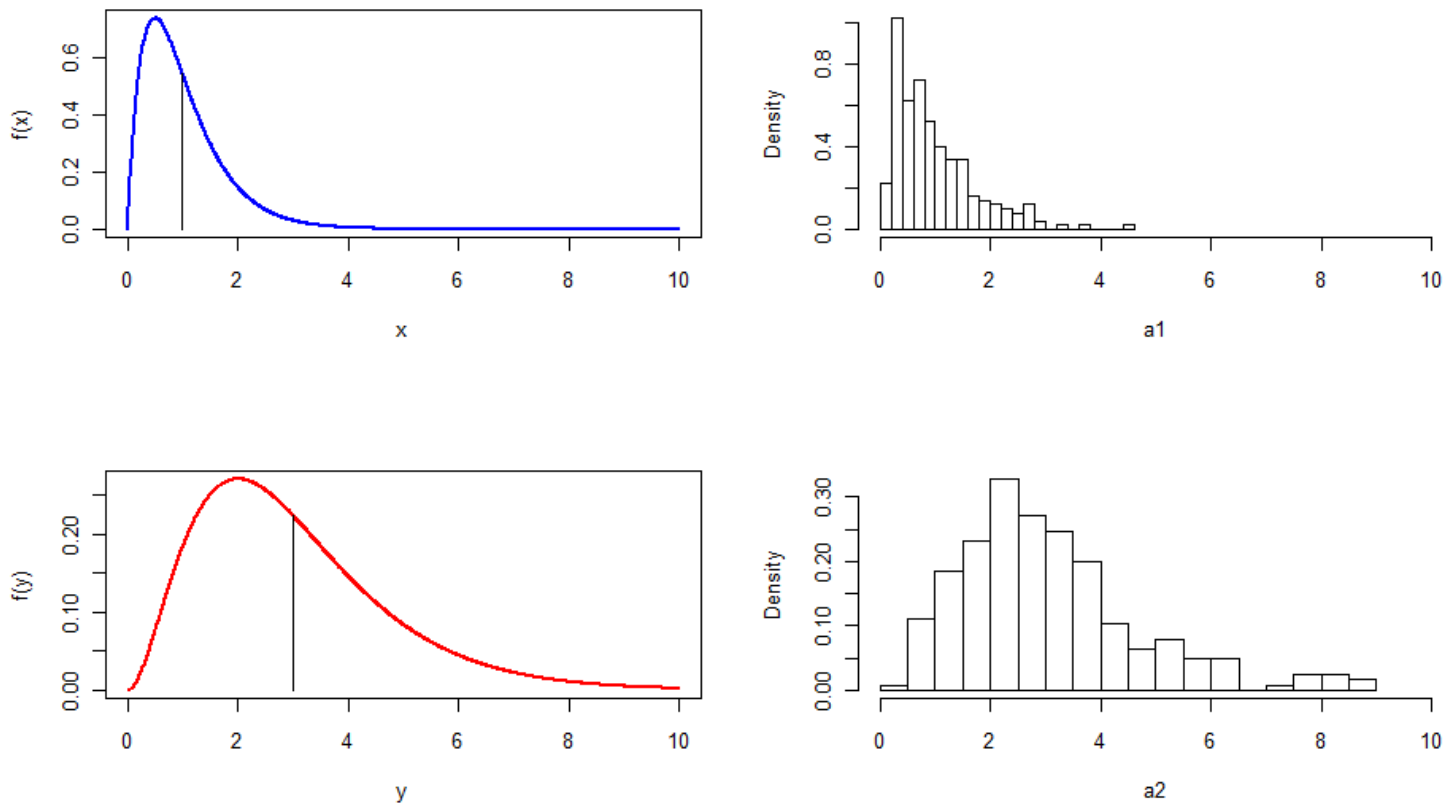
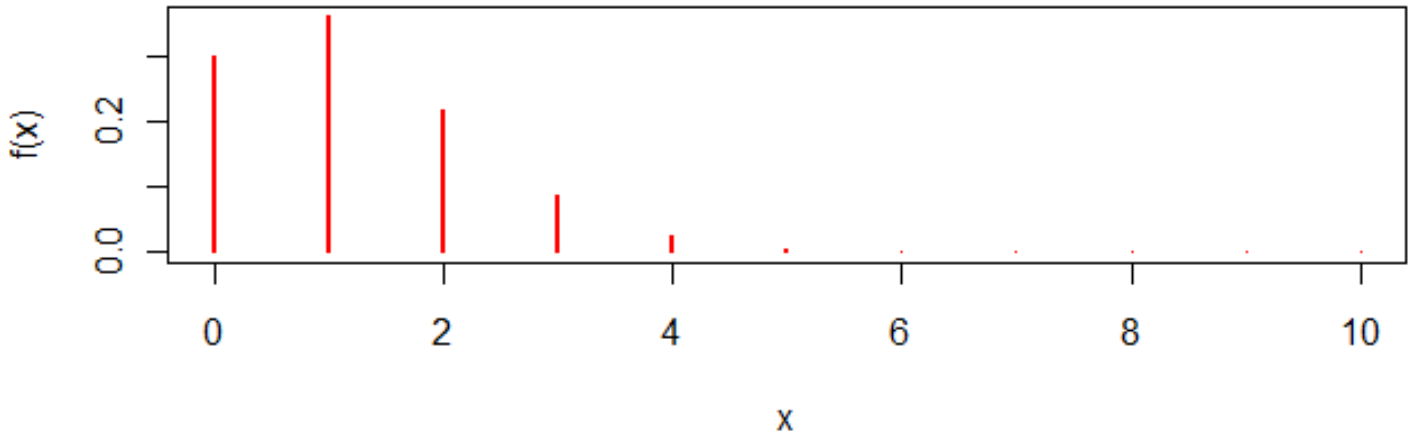
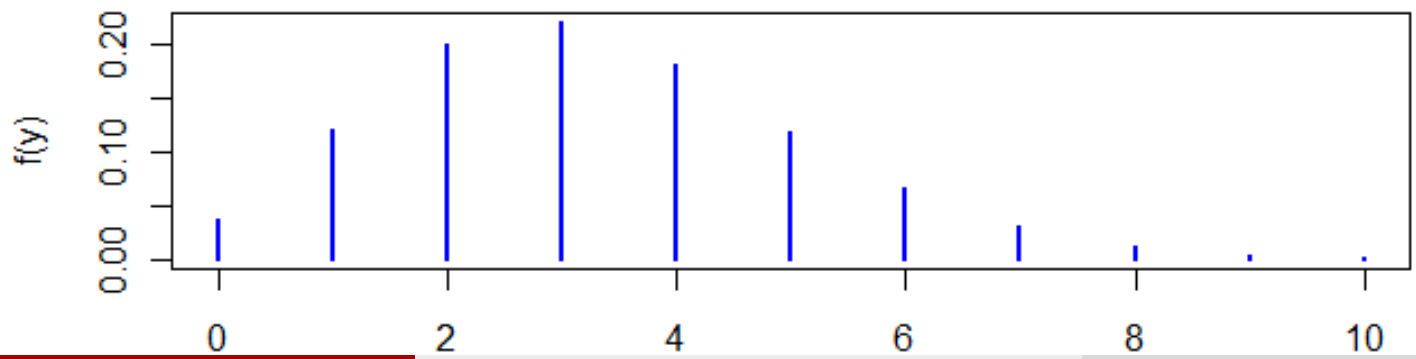


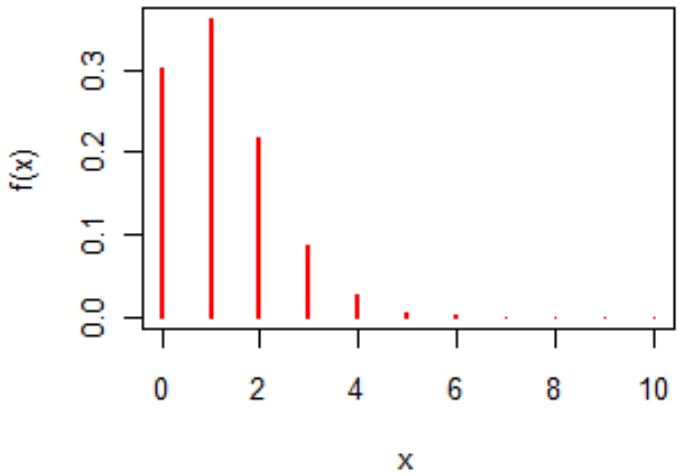

Figura: Histogramas e densidades de $X$ e $Y$ com $1 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3$. Qual das amostras varia mais em torno do seu valor médio?

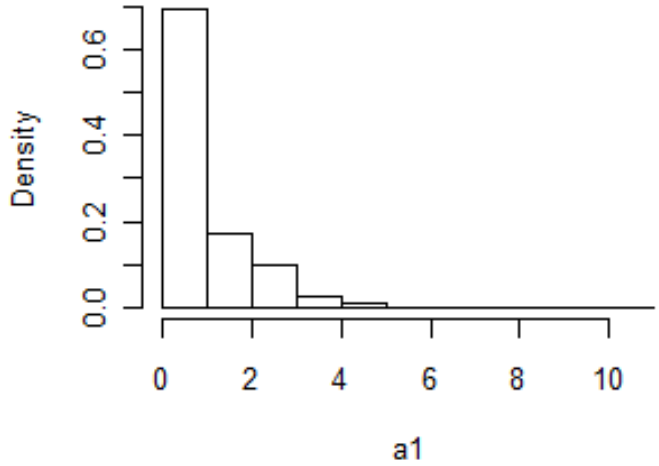
# Pode ser discreta

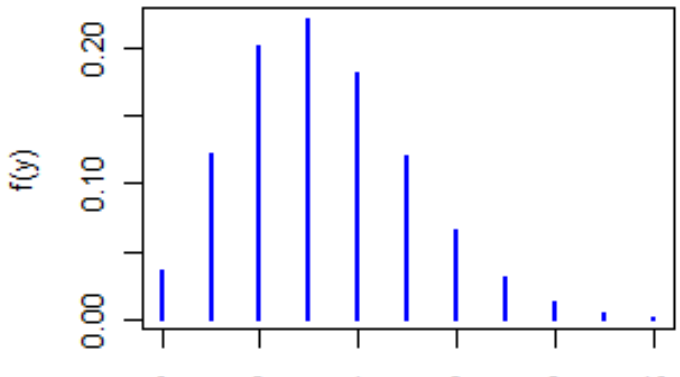


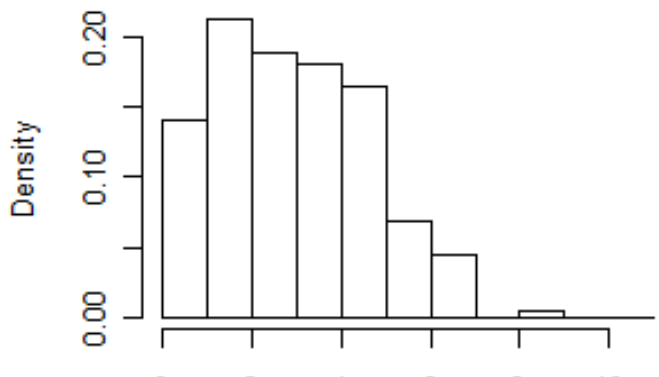



# Com as amostras de $X$ e $Y$



# OK, mas como definir Var($Y$)?

- Falta agora definir matematicamente esta noção intuitiva.
- Queremos medir o grau de variação da v.a. $X$ em torno de seu valor esperado $\mu = \mathbb{E}(Y)$.
- Podemos olhar para o DESVIO $Y - \mu$
- As vezes, $Y - \mu$ é positivo, as vezes é negativo.
- Queremos ter uma idéia do TAMANHO do desvio e não de seu sinal.
- Vamos olhar então para o desvio absoluto $|Y - \mu|$
- Mas $|Y - \mu|$ é uma variável aleatória!!!

# Visão empírica do desvio $|Y - \mu|$

- Suponha que $Y$ seja uma v.a. qualquer (discreta ou continua) com $\mathbb{E}(Y) = \mu$
- Simule $Y$ várias vezes por Monte Carlo.
- Os valores aleatórios gerados sucessivamente vão variar em torno de $\mu$
- As vezes, só um pouco maiores ou menores que $\mu$.
- As vezes, MUITO maiores ou MUITO menores que $\mu$.
- Queremos ter uma ideia do tamanho do desvio $|Y - \mu|$.
- Mas como fazer isto se $|Y - \mu|$ é aleatório?

# Variação em torno de $\mathbb{E}(Y) = \mu$

- Como caracterizar uma v.a.?
- Sua densidade de probabilidade...
- Mas isto é muita coisa (uma lista de números possíveis e probabilidades associadas).
- Não existe uma forma de ter apenas um único número resumindo TODA a distribuição?
- SIM: o valor esperado do desvio absoluto: $E(|X - \mu|)$
- $E(\ |X - \mu|\ )$ é o valor esperado do desvio em torno de $\mu$.

# Do desvio absoluto para o desvio quadrático

- Queremos $E(\ |X-\mu|\ )$ para representar a variabilidade (em torno de $\mu$).
- Mas cálculos com valor absoluto são MUITO difíceis.
- Em particular, a função $f(x) = |x|$ possui mínimo num ponto sem derivada.
- Isto é, seu mínimo não pode ser obtido derivando-se $f(x)$ e igulando a zero
- Isto tem consequências de longo alcance em otimização.
- Saída: Calculamos a variância $\sigma^2 = E(\ |X-\mu|^2\ )$, que é mais fácil, e a seguir calculamos sua raiz quadrada (o desvio-padrão).
- OBS: $\sigma = \sqrt{E(\ |X-\mu|^2\ )} \neq E(\ |X-\mu|\ )$ mas eles costumam não ser muito diferentes.
- Assim, a interpretação do DP $\sigma$ como sendo o tamanho esperado do desvio é aprox correto.

# Desvio-padrão

- Padrão para medir desvios (em torno do valor esperado).
- Solução: Calcule a variância $\sigma^2 = \mathbb{E}\left( (Y-\mu)^2 \right)$ e depois tire a sua raiz quadrada (obtendo então desvio-padrão DP $\sigma$).
- O DP é o padrão universal para medir desvios (em torno do valor esperado).
- Desigualdade de Tchebyshev justifica este nome de desvio-padrão.

# Desigualdade de Tchebyshev

- Seja $Y$ uma v.a. QUALQUER com $\mathbb{E}(Y) = \mu$.
- Então $\mathbb{P}(|Y - \mu| > k\sigma) \leq 1/k^2$.
- Exemplo: se $k = 2$ então, para QUALQUER v.a., $\mathbb{P}(|Y - \mu| > 2\sigma) \leq 1/4$.
- Para $k = 4$, a probabilidade se reduz a 0.06.
- A chance é apenas de 6% de que $Y$ se desvie de seu valor esperado por mais que 4 DPs.
- Isto vale PARA TODA E QUALQUER v.a.
- o DP serve como uma métrica universal de desvios estatístico: desviar-se por mais de 4 DPs de sua média pode ser considerado um tanto raro.

# Um Comentário sobre Tchebyshev

- Tchebyshev: $\mathbb{P}(|Y - \mu| > k\sigma) \leq 1/k^2$
- Observe que a probabilidade decai com $1/k^2$.
- Nos primeiros inteiros temos uma queda rápida mas depois temos uma queda lenta:

| $k$ | 2 | 4 | 6 | 10 | 20 |
|---|---|---|---|---|---|
| $100\% \times \mathbb{P}$ | 25% | 6% | 3% | 1% | 0.3% |

# Outro comentário sobre Tchebyshev

- Tchebyshev: $\mathbb{P}(|Y-\mu| > k\sigma) \leq 1/k^2$
- A sua força é a generalidade: vale para toda e qualquer v.a.
- Mais ainda: ele é ótimo no seguinte sentido: não é possível obter cota mais apertada valendo para TODA v.a.
- Prova: Considere a seguinte v.a. discreta $Y$:

$$Y = \begin{cases} -1, & \text{com probab } \frac{1}{2k^2} \\ 0, & \text{com probab } 1 - \frac{1}{k^2} \\ 1, & \text{com probab } \frac{1}{2k^2} \end{cases}$$

- Ela tem $\mathbb{E}(Y) = 0$ e $DP = \sigma = 1/k$ e portanto

$$\mathbb{P}(|Y-\mu| \geq k\sigma) = \Pr(|Y| \geq 1) = \frac{1}{k^2}$$

e a desigualdade de Tchebyshev atinge a igualdade perfeita para esta distribuição.

# Finalizando o segundo comentário

- A fraqueza da desigualdade de Tchebyshev é ... a sua generalidade
- Para ser válida para toda e qualquer v.a. a desigualdade não é muito "apertada"
- Isto é, podemos obter cotas muito melhores para a chance de ter um desvio grande QUANDO CONSIDERAMOS APENAS UMA DISTRIBUIÇÃO ESPECÍIFICA.
- Por exemplo, se $Y \sim N(\mu, \sigma)$ e $k = 2$, então sabemos que

$$\mathbb{P}(|Y - \mu| \geq 2\sigma) \approx 1/20 = 0.05$$

- enquanto a desigualdade de Tchebyshev garante apenas que

$$\mathbb{P}(|Y - \mu| \geq 2\sigma) \leq 1/4$$

# Estimando $\mu$ e $\sigma$ com dados

- O valor esperado $\mathbb{E}(Y) = \mu$ e o DP $\sigma = \sqrt{\mathbb{E}\left(Y - \mu\right)^2}$ sao valores que nao dependem dos dados.
- Por exemplo, suponha que $Y$ possua distribuicao exponencial, cuja densidade é $f(y) = 3exp(-3y)$. Então $\mathbb{E}(Y) = \mu = 1/3$ e $\sigma = 1/3$.
- Suponha que temos uma amostra de dados retirados de certa distribuicao com densidade $f(y)$ e que NAO CONHECEMOS $f(y)$.
- Portanto, nao podmeos obter $\mathbb{E}(Y) = \mu$ e o DP $\sigma$.
- No entanto, podemos ESTIMAR $\mathbb{E}(Y) = \mu$ e o DP $\sigma$ com os dados.

# Momentos amostrais e teóricos

- A ideia é igualar os momentos teóricos com momentos amostrais correspondentes.
- Momentos:

| Momento Teórico | Momento Amostral |
|---|---|
| $\mathbb{E}(Y)$ | $m_1 = \overline{Y} = \sum_{i=1}^{n} Y_i/n$ |
| $\mathbb{E}(Y^2)$ | $m_2 = \sum_i Y_i^2/n$ |
| $\mathbb{E}(Y^3)$ | $m_3 = \sum_i Y_i^3/n$ |
| $\vdots$ | $\vdots$ |

- Veja que $m_1$ é a média dos dados na amostra.
- É comum escrevermos $m_1$ como $\overline{Y}$.

# Primeiro Momento $\overline{Y}$

- Pela lei dos grandes números,

$$\overline{Y} = \frac{1}{n}\sum_i Y_i \to \mathbb{E}(Y) = \mu$$

  quando $n \to \infty$.
- Assim, se o tamanho $n$ da amostra não for pequeno demais, podemos esperar a média amostral $\overline{Y} \approx \mathbb{E}(Y)$.
- Assim, podemos usar os dados para estimar o valor desconhecido (e teórico $\mu$).
- Note que, a não ser em casos de distribuições muito especiais (e bizarras), devemos ter $\mu \neq \overline{Y}$.
- Isto é, a média amostral $\overline{Y}$ não é igual à esperança $\mu$.

# $\sigma$ e o segundo momento

- DP $\sigma$ é a raiz quadrda da variância.
- Variância: $\sigma^2 = \mathbb{E}(Y - \mu)^2$.
- Temos

$$\sigma^2 = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2$$

- Como $\mathbb{E}(Y) \approx \overline{Y}$, podemos esperar então que

$$\sigma^2 \approx \mathbb{E}(Y^2) - \left(\overline{Y}\right)^2$$

- E o primeiro termo? Usamos a Lei dos Grandes Números de novo.

# $\sigma$ e o segundo momento

- Pela mesma lei dos grandes números, $m_k = \sum_i Y_i^k / n$ converge para $\mathbb{E}(Y^k)$.
- Assim, $m_2 \approx \mathbb{E}(Y^2)$
- Portanto,

$$\sigma^2 \approx \frac{1}{n} \sum_i Y_i^2 - \left(\overline{Y}\right)^2$$

- Podemos mostrar que

$$\frac{1}{n} \sum_i Y_i^2 - \left(\overline{Y}\right)^2 = \frac{1}{n} \sum_i \left(Y_i - \overline{Y}\right)^2$$

- Esta é a variância amostral.
- O DP pode ser estimado tomando sua raiz quadrada.