Variância, DP e Desigualdades

Renato Martins Assunção

DCC - UFMG

2016

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

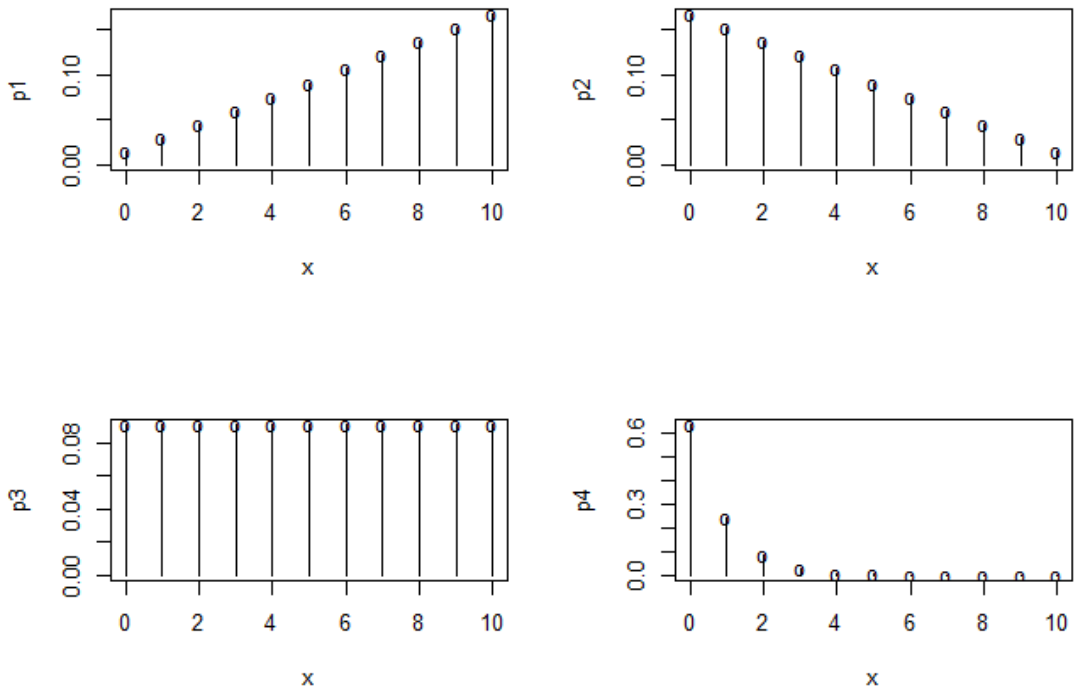

Figura: Sem fazer nenhuma conta, identifique as distribuições com as seguintes esperanças: $5, 6.67, 0.53, 3.33$

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

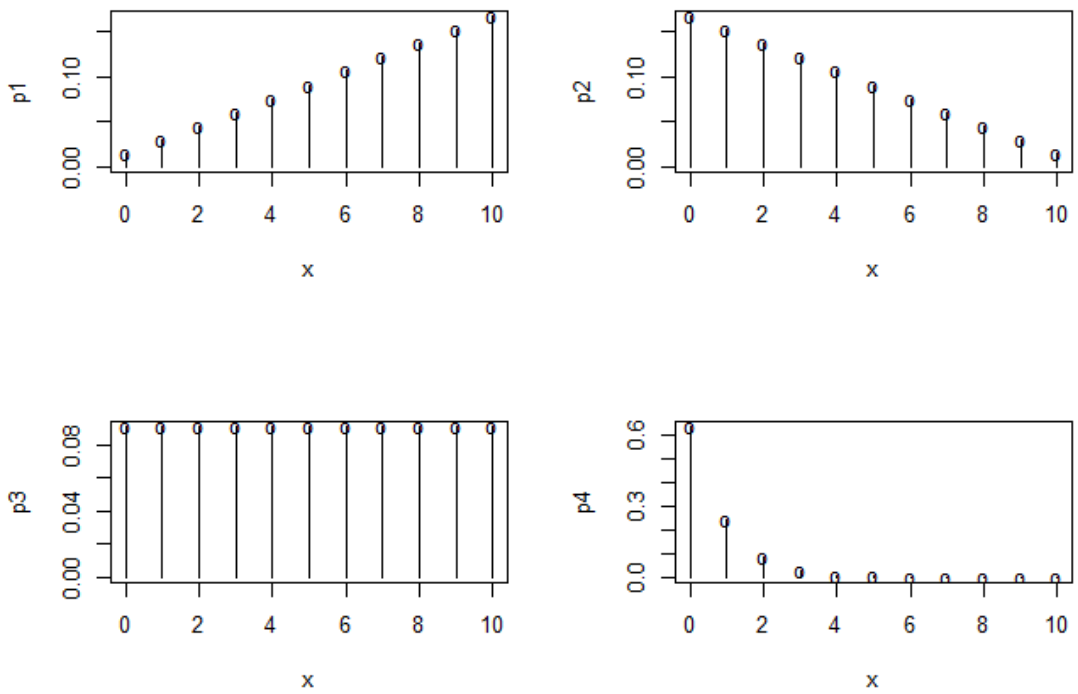

Figura: $p_1 = 6.67$, $p_2 = 3.33$, $p_3 = 5$, $p_4 = 0.53$.

$\mathbb{E}(Y)$ 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

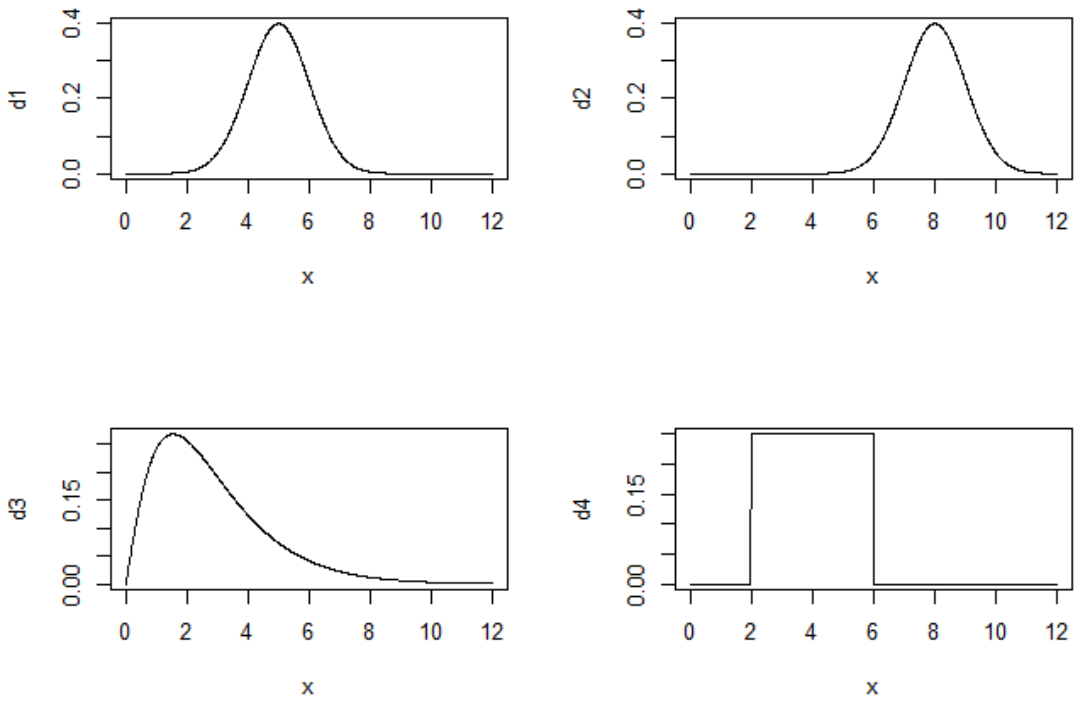

Figura: Sem fazer nenhuma conta, identifique as distribuições com as seguintes esperanças: $1.8, 8, 5, 4$

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, $\mathbb{E}(X) = c\mathbb{P}(X = c) = c \times 1 = c$

Propriedades da esperança

- Se $X_1, X_2, \ldots, X_n$ são v.a.'s INDEPENDENTES então

$$\mathbb{E}(X_1 X_2 \ldots X_n) = \mathbb{E}(X_1)\ \mathbb{E}(X_2)\ \ldots\ \mathbb{E}(X_n)$$

- Em particular, se as duas v.a.'s X e Y são independentes:

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

$\mathbb{E}(XY) = \mathbb{E}(X)\ \mathbb{E}(Y)$, se independentes

- Prova no caso particular de duas v.a.'s discretas:
- A v.a. XY possui os valores possíveis $x_i y_j$ onde x_i e y_j varrem todas as possibilidades para X e Y.
- Independência implica que

$$\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i)\ \mathbb{P}(Y = y_j)$$

pois os eventos $[X = x_i]$ e $[Y = y_j]$ são independentes.

- Então

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

X e Y: Qual possui maior variância?

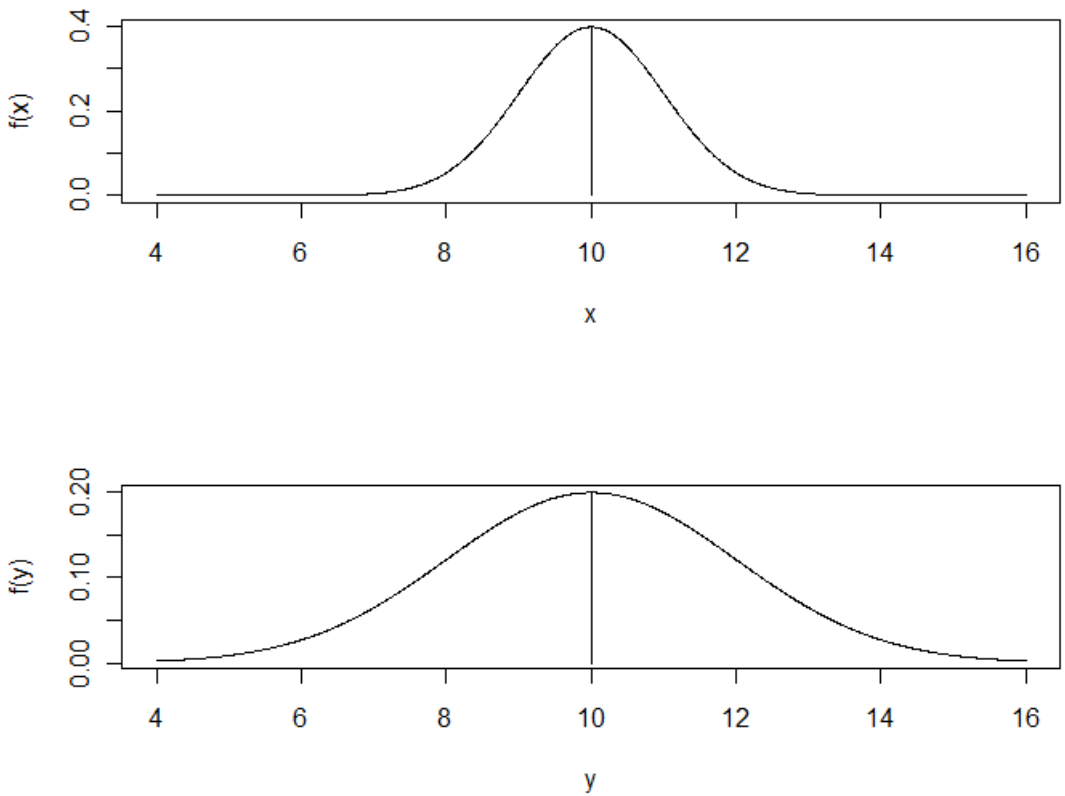

Figura: Densidades de X e Y com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 10$.

Gerando as amostras

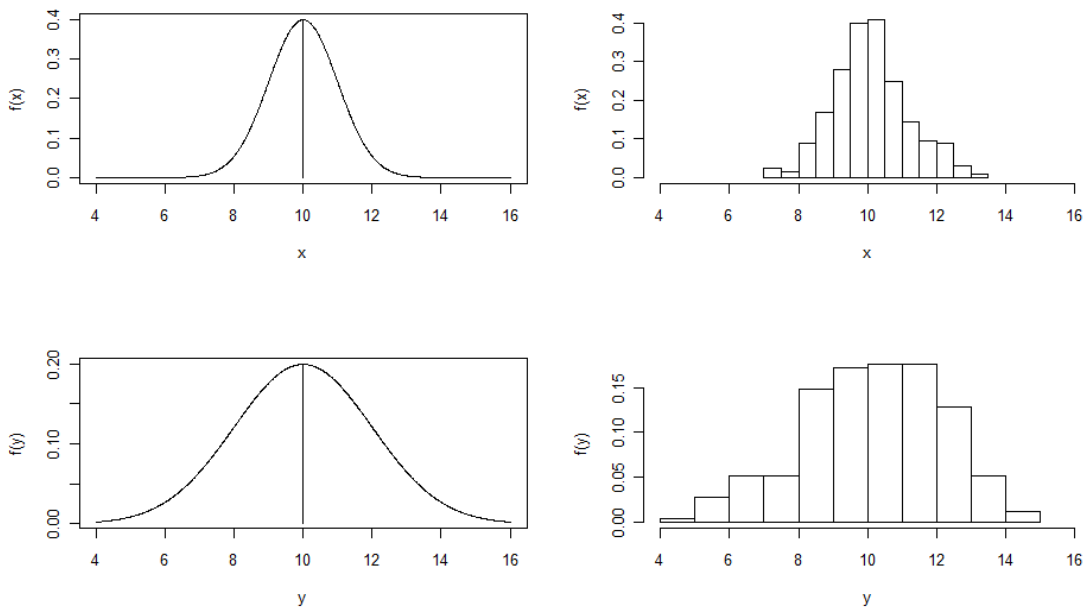

Figura: Histogramas de amostras e densidades de X e Y com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 10$. Qual das amostras varia mais em torno do seu valor médio?

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

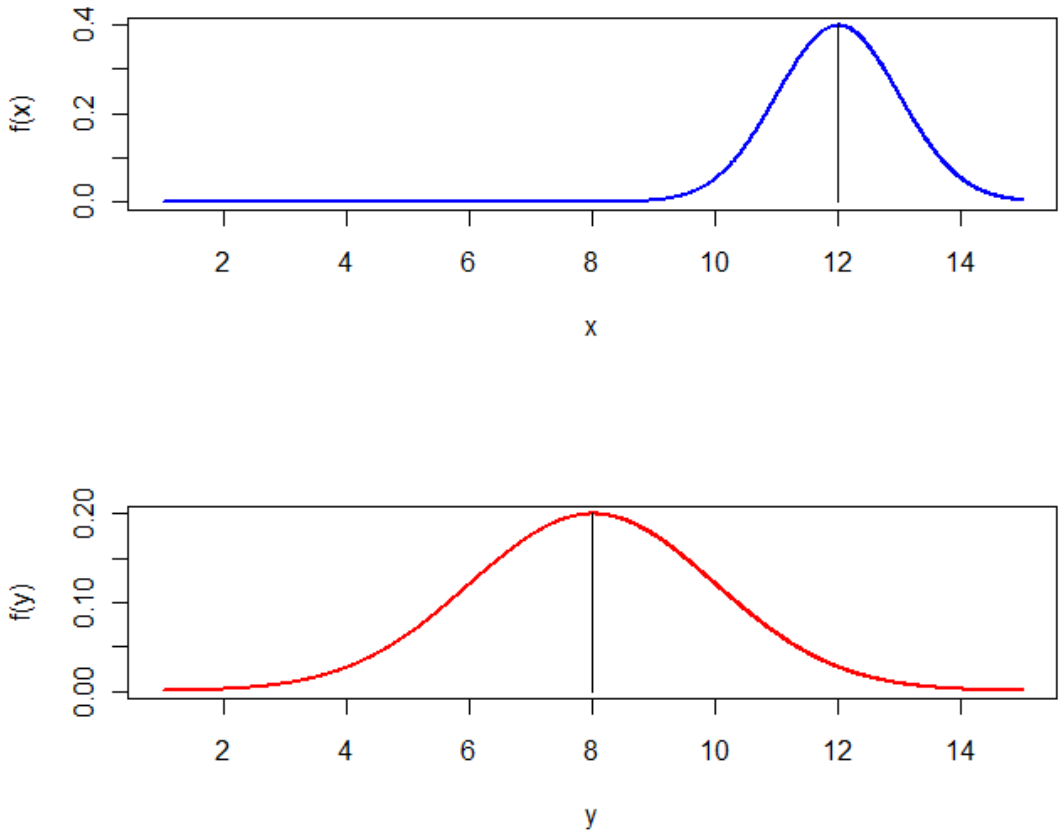

Figura: Densidades de X e Y com diferentes valores esperados: $\mathbb{E}(X) \neq \mathbb{E}(Y)$.
Qual das amostras varia mais em torno do seu valor médio?

As densidades e as amostras

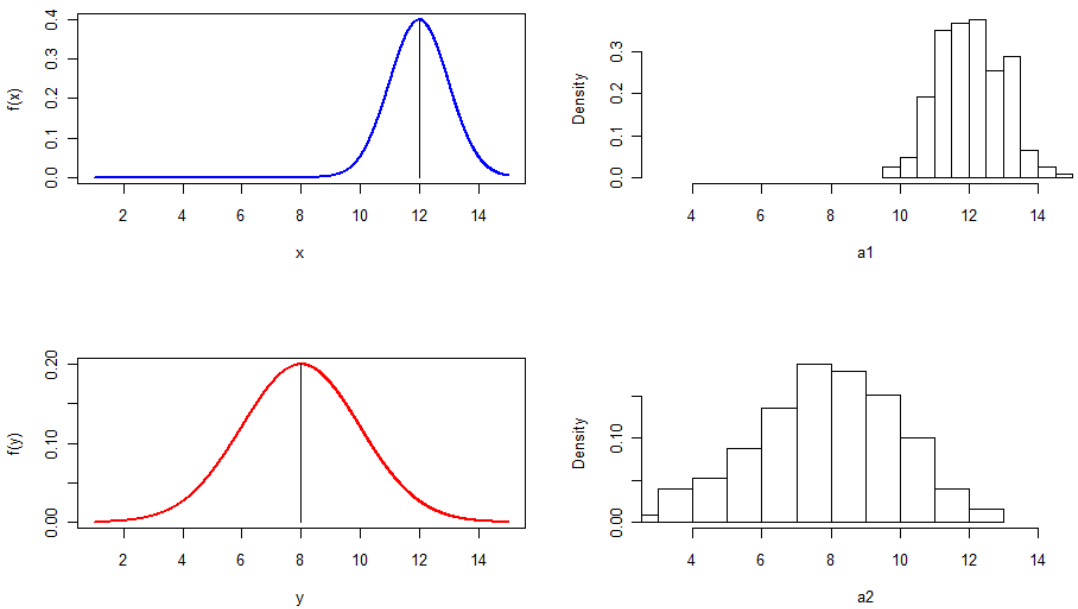

Figura: Histogramas de amostras e densidades de X e Y com diferentes valores esperados: $\mathbb{E}(X) \neq \mathbb{E}(Y)$. Qual das amostras varia mais em torno do seu valor médio?

Não precisa ter densidade simétrica

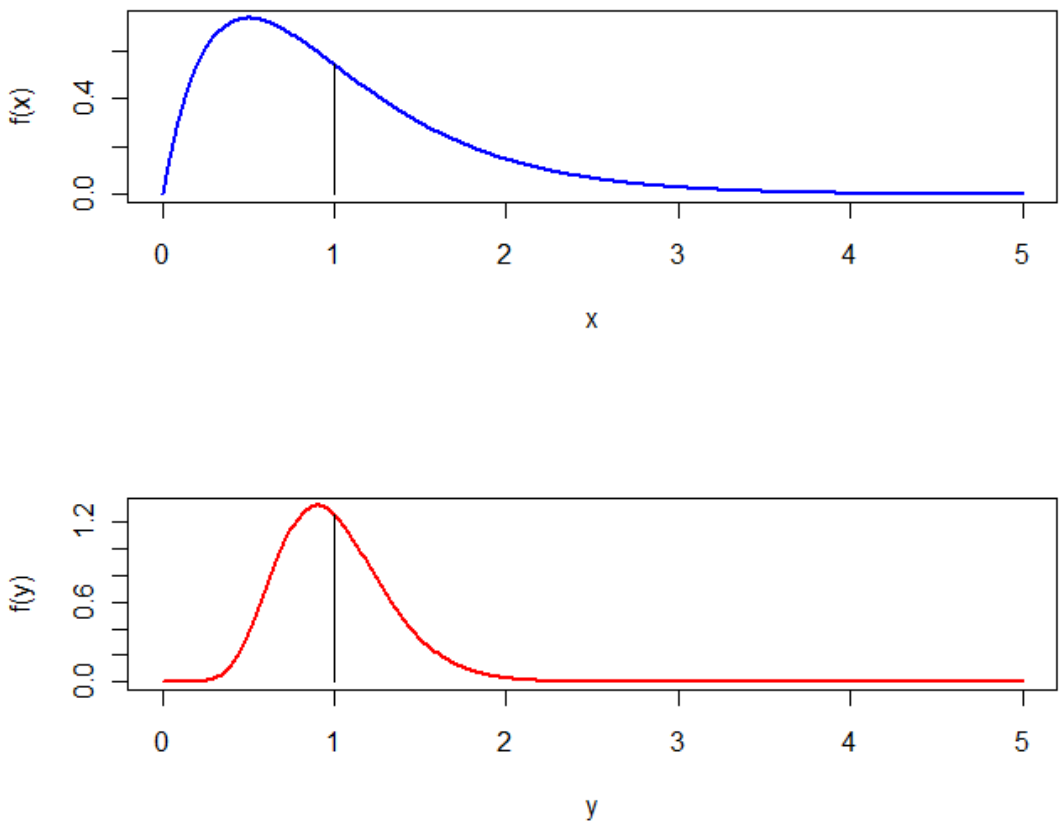

Figura: Densidades ASSIMÉTRICAS de X e Y mas mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 1$.

Com as amostras de cada distribuição

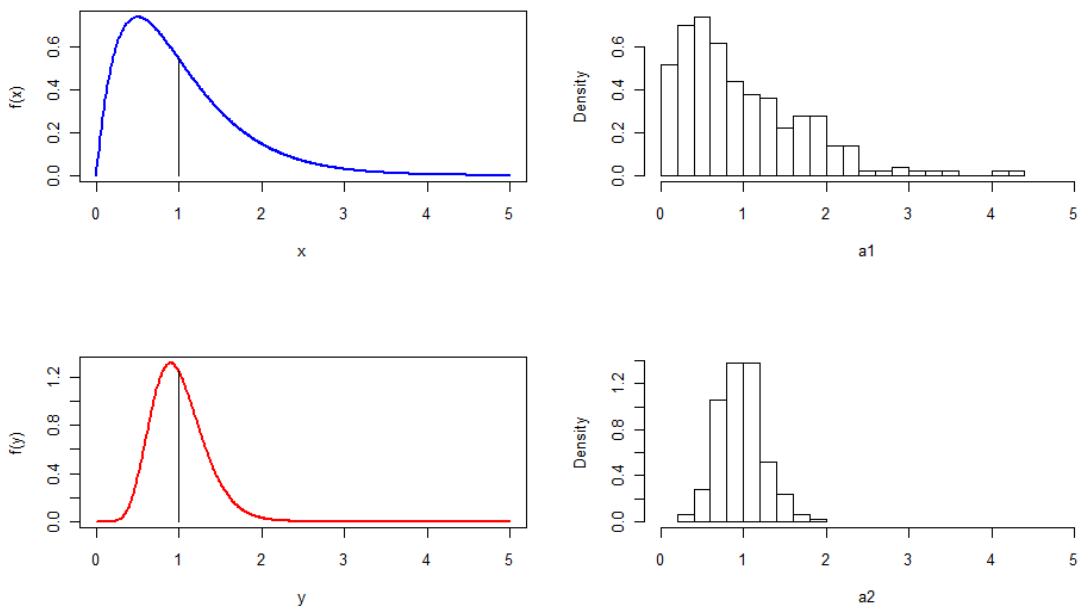

Figura: Histogramas de amostras e densidades ASSIMÉTRICAS de X e Y com mesmo valor esperado: $\mathbb{E}(X) = \mathbb{E}(Y) = 1$. Qual das amostras varia mais em torno do seu valor médio?

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

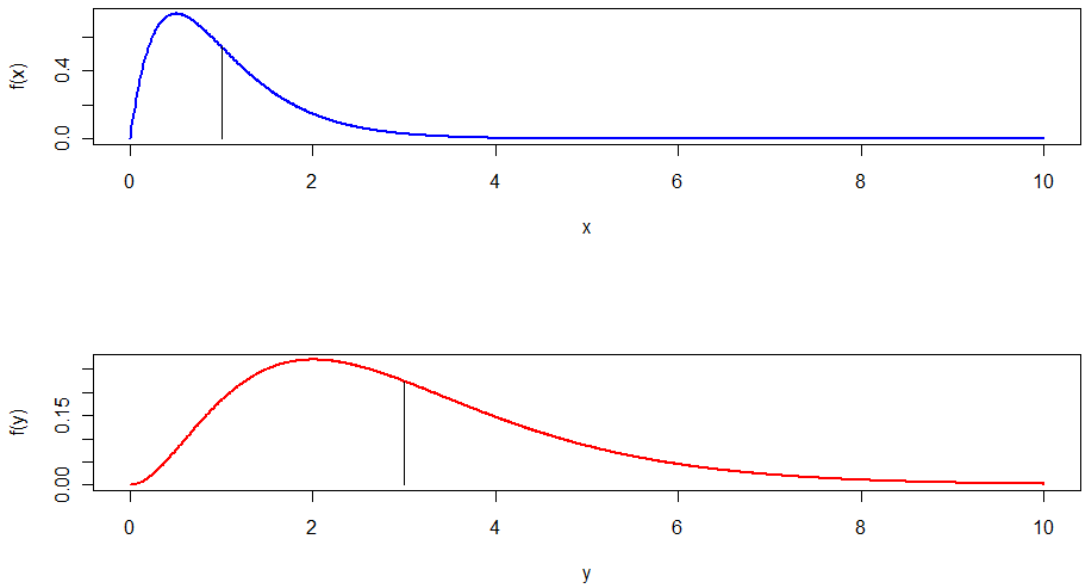

Figura: Densidades de X e Y com $1 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3$. Qual das amostras varia mais em torno do seu valor médio?

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

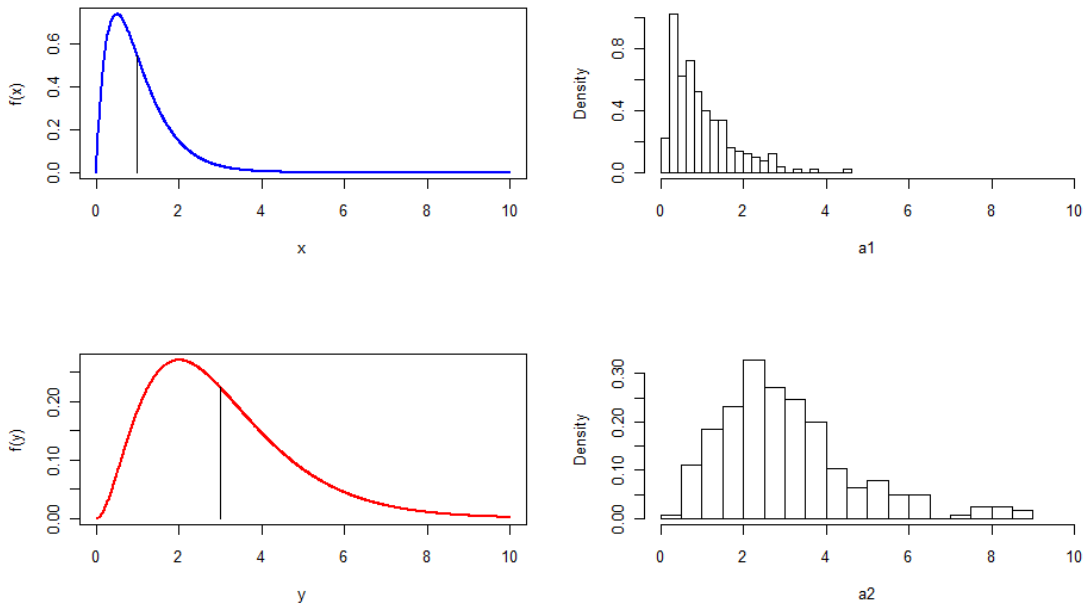

Figura: Histogramas e densidades de X e Y com $1 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3$. Qual das amostras varia mais em torno do seu valor médio?

Pode ser discreta

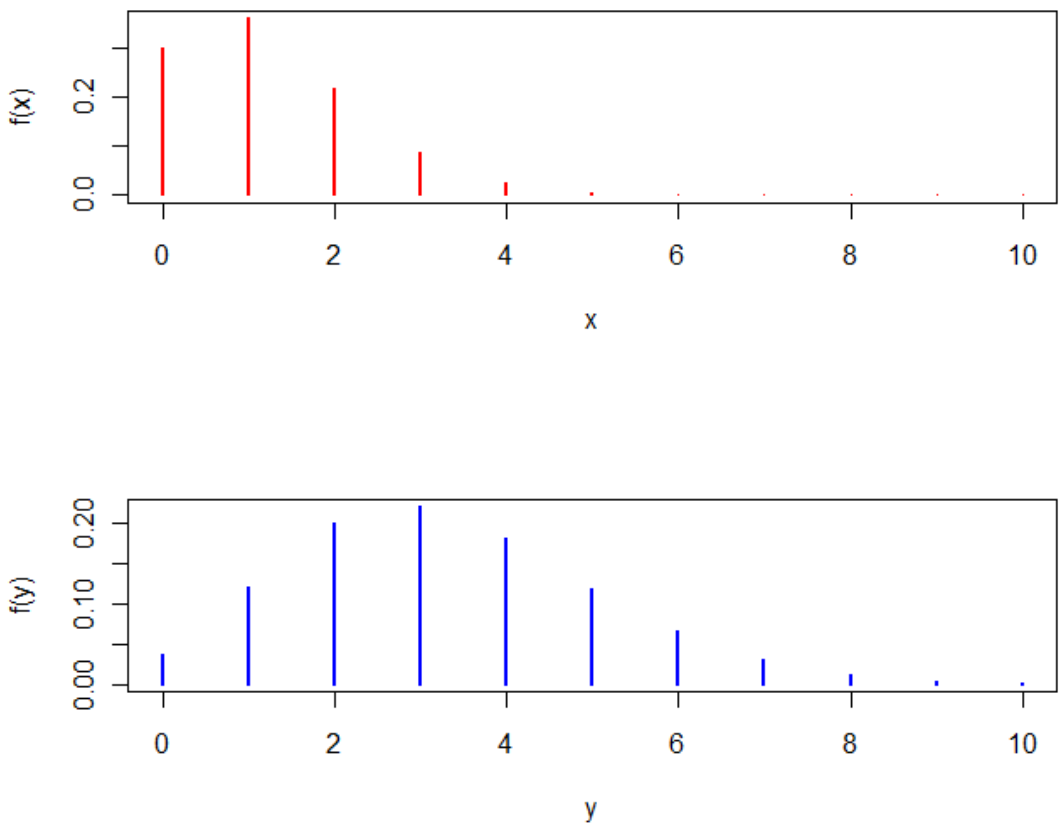

Figura: Funções de probabilidade de duas Poisson com $1.2 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3.3$. Qual das amostras varia mais em torno do seu valor

Com as amostras de X e Y

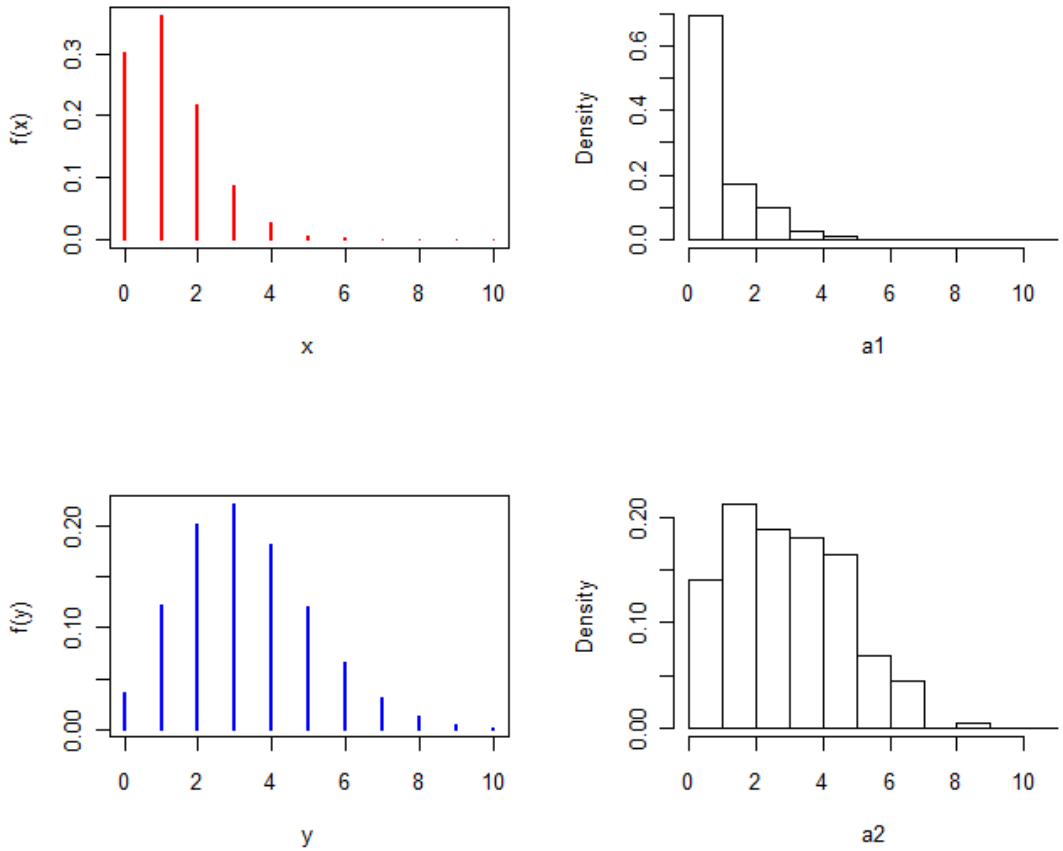

Figura: Histogramas e funções de probabilidade de duas Poisson com $1.2 = \mathbb{E}(X) \neq \mathbb{E}(Y) = 3.3$. Qual das amostras varia mais em torno do seu valor

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 μ
- As vezes, só um pouco maiores ou menores que μ.
- As vezes, MUITO maiores ou MUITO menores que μ.
- 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.?
- Pela sua densidade de probabilidade...
- Mas isto é muita coisa (uma lista de números possíveis e as 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 ALEATÓRIO em torno de μ.

Do desvio absoluto para o desvio quadrático: variância

- Queremos $E(\ |X-\mu|\ )$ para representar a variabilidade da v.a. X em torno de μ.
- Mas cálculos com valor absoluto são MUITO difíceis.
- Em particular, a função $f(x) = |x|$ tem seu mínimo num ponto sem derivada.
- Isto é, seu mínimo não pode ser obtido derivando-se $f(x)$ e igualando 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 para voltar à escala original.

Variância e DP

- Este é o desvio-padrão:

$$DP = \sigma = \sqrt{\sigma^2} = \sqrt{E(\ |X-\mu|^2\ )}$$

.

- $\sigma = \sqrt{E(\ |X-\mu|^2\ )} \neq E(\ |X-\mu|\ )$ mas eles costumam não ser muito diferentes.
- Assim, a interpretação do DP σ como sendo o tamanho esperado do desvio é aprox correto.
- Razão do nome: DP é o padrão para medir desvios (em torno do valor esperado).
- DP: Calcule a variância $\sigma^2 = \mathbb{E}\left(\ (Y-\mu)^2\ \right)$ e depois tire a sua raiz quadrada (obtendo então o desvio-padrão DP σ).

Identifique o DP em cada caso

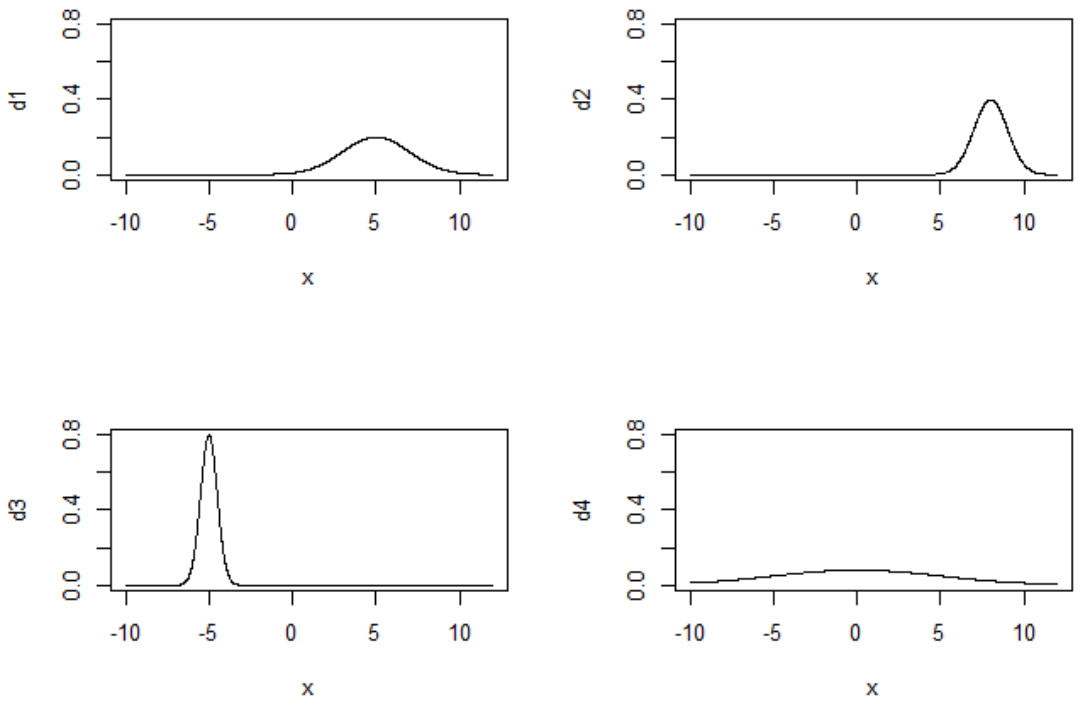

Figura: Sem fazer nenhuma conta, identifique as distribuições com os seguintes DPs: $5, 1/2, 2, 1$

Identifique o DP em cada caso

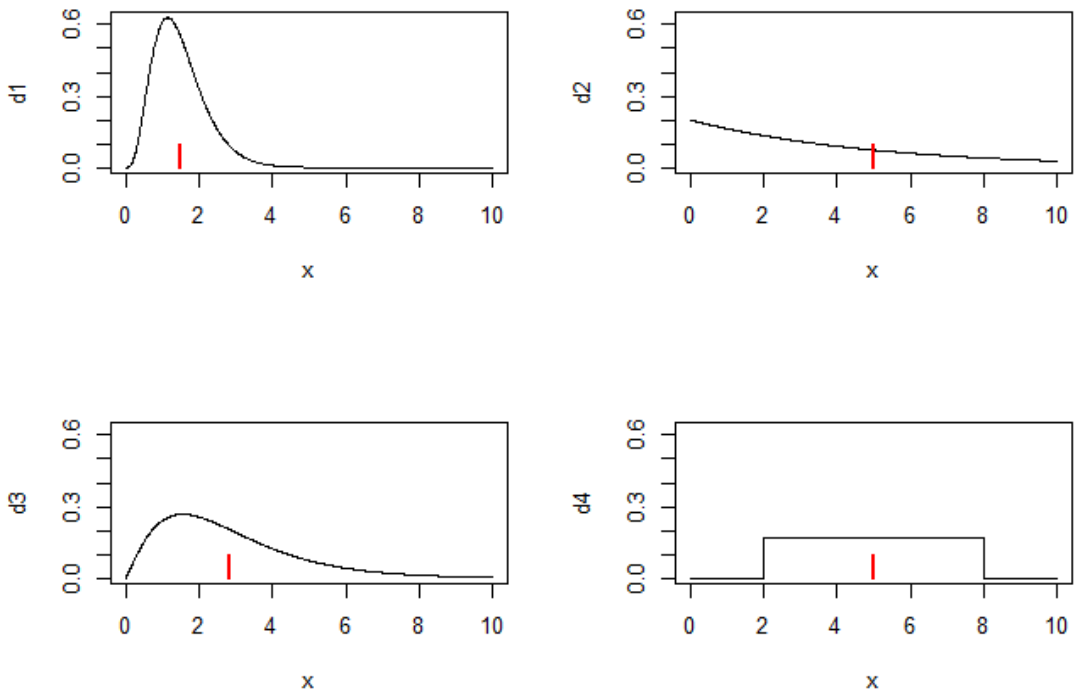

Figura: Sem fazer nenhuma conta, identifique as distribuições com os seguintes DPs: 3, 0.70, 0.58, 1.9

Identifique o DP em cada caso

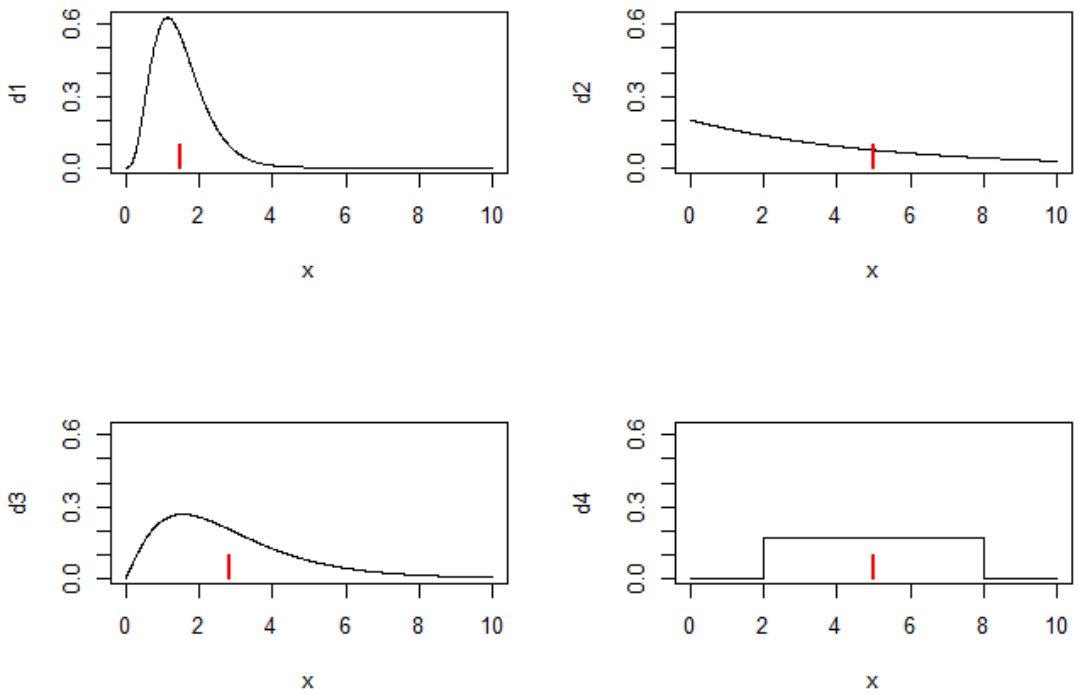

Figura: Ordem correta: linha de cima: 0.70, 3; linha de baixo: 1.9, 0.58

Calculando a variância e o DP

- $DP = \sqrt{\mathbb{V}(Y)}$ onde a variância é

$$\mathbb{V}(X) = E(\ (X-\mu)^2\ )$$

.

- Caso Discreto: como calcular?
- Lembre-se $\mathbb{E}(g(X)) = \sum_i g(x_i)\mathbb{P}(X = x_i)$
- Tome $g(X) = (X-\mu)^2$. Então

$$\mathbb{V}(X) = E(\ (X-\mu)^2\ ) = \sum_i (x_i-\mu)^2 \mathbb{P}(X = x_i)$$

.

Calculando a variância e o DP

- Por exemplo:

x_i	$\mathbb{P}(X = x_i)$
0	0.1
1	0.7
2	0.2

- Então $\mu = \mathbb{E}(X) = 0 * 0.1 + 1 * 0.7 + 2 * 0.2 = 1.1$
- Variância:

$$\begin{aligned} \mathbb{V}(X) &= E\left((X - 1.1)^2\right) \\ &= (0 - 1.1)^2 * 0.1 + (1 - 1.1)^2 * 0.7 + (2 - 1.1)^2 * 0.2 \\ &= 0.29 \end{aligned}$$

Calculando a variância e o DP: contínuo

- $DP = \sqrt{\mathbb{V}(Y)}$ onde a variância é

$$\mathbb{V}(X) = \mathbb{E}(\ (X-\mu)^2\ )$$

.

- Caso contínuo:

$$\mathbb{V}(X) = \mathbb{E}(\ (X-\mu)^2\ ) = \int (x-\mu)^2 f(x) \mathrm{d}x$$

.

Calculando a variância e o DP

- Por exemplo, $X \sim \exp(3)$ o que implica na densidade

$$f(x) = \begin{cases} 0, & \text{se } x < 0 \\ 3\exp(-3x), & \text{se } x \geq 0 \end{cases}$$

- Então

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x = \int_{0}^{\infty} x 3\exp(-3x) \mathrm{d}x = \frac{1}{3}$$

e

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}\left((X - 1/3)^2\right) \\ &= \int_{-\infty}^{\infty} (x - 1/3)^2 f(x) \mathrm{d}x \\ &= \int_{-\infty}^{\infty} (x - 1/3)^2 3\exp(-3x) \mathrm{d}x \\ &= 1/3^2 \end{aligned}$$

Outra fórmula

- Pode-se mostrar que

$$\mathbb{V}(X) = \mathbb{E}\left((X-\mu)^2\right) = \mathbb{E}\left(X^2\right) - (\mu)^2 = \mathbb{E}\left(X^2\right) - (\mathbb{E}(X))^2$$

- A prova é muito simples:

$$\begin{aligned}
\mathbb{V}(X) &= \mathbb{E}\left((X-\mu)^2\right) \\
&= \mathbb{E}\left(X^2 - 2X\mu + \mu^2\right) \\
&= \mathbb{E}\left(X^2\right) - \mathbb{E}\left(2X\mu\right) + \mathbb{E}\left(\mu^2\right) \\
&= \mathbb{E}\left(X^2\right) - 2\mu\mathbb{E}\left(X\right) + \mu^2 \\
&= \mathbb{E}\left(X^2\right) - 2\mu\mu + \mu^2 \\
&= \mathbb{E}\left(X^2\right) - \mu^2
\end{aligned}$$

$\mathbb{V}(Y)$ de Bernoulli

- Variância:
$$\mathbb{V}(Y) = \mathbb{E}(Y-\mu)^2 = \mathbb{E}(Y^2) - \mu^2$$
- Bernoulli:
  - $Y = 0$ com probabilidade θ e $Y = 1$ com probabilidade $1-\theta$.
  - Então $\mu = \mathbb{E}(Y) = 1 * \theta + 0 * (1-\theta) = \theta$
  - $\mathbb{E}(Y^2) = 1^2 * \theta + 0^2 * (1-\theta) = \theta$
  - Assim, $\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mu^2 = \theta - \theta^2 = \theta(1-\theta)$

$\mathbb{V}(Y)$ de Binomial

- Binomial: requer alguma manipulação de fórmulas matemáticas
- Vamos obter primeiro $\mathbb{E}(Y) = \mu$.
- $Y \sim$ Bin(n, θ).
- Valores possíveis são $0, 1, 2, \ldots, n$.
- Probabilidades associadas:
- Então

$$\mathbb{P}(Y = k) = \frac{n!}{k!}(n-k)!\theta^k(1-\theta)^{n-k}$$

- Queremos

$$\mu = \mathbb{E}(Y) = \sum_{k=0}^{n} k\mathbb{P}(Y = k) = \sum_{k=0}^{n} k\frac{n!}{k!}(n-k)!\theta^k(1-\theta)^{n-k}$$

$\mathbb{E}(Y)$ de Binomial

- Temos

$$\begin{aligned}
\mu &= \mathbb{E}(Y) = \sum_{k=0}^{n} k\mathbb{P}(Y=k) = \sum_{k=0}^{n} k\frac{n!}{k!}(n-k)!\theta^k(1-\theta)^{n-k} \\
&= \sum_{k=1}^{n} k\frac{n!}{k!}(n-k)!\theta^k(1-\theta)^{n-k} \text{ eliminando } k=0 \text{ da soma} \\
&= n\theta\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!}(n-k)!\theta^{k-1}(1-\theta)^{n-k} \text{ com } n\theta \text{ em evidência} \\
&= n\theta\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!}((n-1)-(k-1))!\theta^{k-1}(1-\theta)^{(n-1)-(k-1)} \text{ manipulando...} \\
&= n\theta\sum_{j=0}^{n-1}\frac{(n-1)!}{j!}((n-1)-j)!\theta^{j}(1-\theta)^{(n-1)-j} \text{ mudança de variáveis} \\
&= n\theta \times 1
\end{aligned}$$

- A última soma acima é 1 pois a parcela j é a probabilidade $\mathbb{P}(Z=j)$ onde $Z \sim$ Bin$(n-1,\theta)$ e estamos somando todas estas probabilidades.

$\mathbb{V}(Y)$ de Binomial

- Calculando $\mathbb{E}(Y^2)$ no caso binomial:

$$\begin{aligned}\mathbb{E}(Y^2) &= \sum_{k=0}^{n} k^2 \mathbb{P}(Y=k) = \sum_{k=0}^{n} k^2 \frac{n!}{k!}(n-k)!\theta^k(1-\theta)^{n-k} \\ &= \sum_{k=1}^{n} k^2 \frac{n!}{k!}(n-k)!\theta^k(1-\theta)^{n-k} \text{ eliminando } k=0 \text{ da soma}\end{aligned}$$

- Uma série de manipulações algébricas leva ao resultado final

$$\mathbb{E}(Y^2) = n^2\theta^2 + n\theta(1-\theta)$$

- Para a prova completa, ver `https://proofwiki.org/wiki/Variance_of_Binomial_Distribution`
- Assim,

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = n\theta(1-\theta)$$

$\mathbb{V}(Y)$ de Uniforme

- Calculando $\mathbb{E}(Y)$ no caso Uniforme $U(0,1)$:

$$\begin{aligned}
\mathbb{E}(Y) &= \int_{\mathbb{R}} x f(x) dx \\
&= \int_{-\infty}^{0} x \times 0 \;\; dx + \int_{0}^{1} x \;\; 1 \;\; dx + \int_{0}^{\infty} x \times 0 \;\; dx \\
&= \frac{x^2}{2} \Big|_0^1 \\
&= \frac{1}{2}
\end{aligned}$$

$\mathbb{V}(Y)$ de Uniforme

- Calculando $\mathbb{E}(Y^2)$ no caso Uniforme $U(0,1)$:

$$\begin{aligned}
\mathbb{E}(Y^2) &= \int_{\mathbb{R}} x^2 f(x) dx \\
&= \int_{-\infty}^{0} x^2 \times 0 \;\; dx + \int_{0}^{1} x^2 \;\; 1 \;\; dx + \int_{0}^{\infty} x^2 \times 0 \;\; dx \\
&= \frac{x^3}{3} \Big|_0^1 \\
&= \frac{1}{3}
\end{aligned}$$

- Assim,

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = \frac{1}{3} - \frac{1}{2^2} = \frac{1}{12}$$

$\mathbb{V}(Y)$ de Uniforme

- Assim,

$$\mathbb{V}(Y) = \frac{1}{12}$$

e portanto

$$DP = \frac{1}{\sqrt{12}} \approx 0.289$$

- Note que, como $\mathbb{E}(Y) = 1/2$ e a distribuição é uniforme você poderia estar esperando que DP fosse exatamente $1/4 = 0.25$, a metade do intervalo $(0, 1/2)$.
- Isto não é estritamente verdade.
- A razão é que calculamos o valor esperado do desvio AO QUADRADO e somente depois tiramos a raiz quadrada.
- Isto não é exatamente o mesmo que tirar a média do desvio absoluto.

Propriedades da variância

- Seja X uma v.a. qualquer.
- Denote $\mu = \mathbb{E}(X)$ e $\sigma^2 = \mathbb{V}(X)$.
- Crie uma nova v.a. que seja uma transformação linear de X (como ao passar de Celsius para Fahrenheit).
- $Y = a + bX$ onde a e b são constantes.
- Então já sabemos que $\mathbb{E}(Y) = a + b\mathbb{E}(X)$
- Como $\mathbb{V}(Y)$ e $\mathbb{V}(X)$ se relacionam?

Propriedades da variância

- Seja X uma v.a. com $\mu_x = \mathbb{E}(X)$ e $\sigma_x^2 = \mathbb{V}(X)$.
- Se $Y = a + bX$ então $\mu_y = \mathbb{E}(Y) = a + b\mu_x$ e

$$\sigma_y^2 = \mathbb{V}(Y) = \mathbb{V}(a + bX) = b^2\mathbb{V}(X) = b^2\sigma_x^2$$

- Em termos do DP das v.a.'s:

$$DP_y = |b|\ DP_x$$

- Curioso: a não tem efeito na variância (ou no DP), apenas no valor esperado.

$\mathbb{V}(a+X)$

- Se $Y = a + bX$ então $\mathbb{V}(Y) = b^2\mathbb{V}(X)$
- Por quê isto é verdade intuitivamente? Primeiro, vamos ver que *a não* deve ter efeito.
- Suponha que $Y = 2 + X$. Deslocamos todos os valores de X por duas unidades.
- Deslocamos também seu valor esperado de 2: $\mathbb{E}(Y) = 2 + \mathbb{E}(X)$
- Assim os *desvios* de $Y = 2 + X$ em torno de *sua* média $\mathbb{E}(Y) = 2 + \mathbb{E}(X)$ ficarão inalterados pois tanto X quanto $\mathbb{E}(X)$ são deslocados de 2:

$$Y - \mathbb{E}(Y) = (2 + X) - (2 + \mathbb{E}(X)) = X - \mathbb{E}(X)$$

- Assim, a não tem efeito na variância (ou no DP) de $Y = a + X$, apenas no seu valor esperado.

$\mathbb{V}(bX)$

- Se $Y = bX$ então $\mathbb{V}(Y) = b^2\mathbb{V}(X)$ e portanto

$$DP_y = |b|\ DP_x$$

- Qual a intuição disso?
- Se todos os valores de X são multiplicados por $b = 2$ (digamos) então sua média também é multiplicada por 2.
- Isto é,

$$\mu_y = \mathbb{E}(Y) = \mathbb{E}(2X) = 2\mathbb{E}(X) = 2\mu_x$$

- Mas então os desvios de Y em relaạo a sua média também ficarão multiplicados por 2:

$$Y - \mu_y = 2X - 2\mu_x = 2(X - \mu_x)$$

- Como olhamos para o desvio ao quadrado, teremos 2^2 multiplicando $\mathbb{V}(X)$:

$$\mathbb{V}(Y) = \mathbb{E}((Y - \mu_y)^2 = \mathbb{E}\left(2(X - \mu_x)^2\right) = 2^2\mathbb{E}\left((X - \mu_x)^2\right)$$

Propriedades da variância

- Se X e Y são v.a.'s independentes, temos:

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

- Esta propriedade é quase que a justificativa (sutil) para adotarmos a variância como medida por excelência de variabilidade em probabilidade.
- Lembre-se que a esperança do desvio absoluto, $\mathbb{E}(|Y-\mu|)$, é muito mais intuitiva que $DP = \sqrt{\mathbb{V}(Y)} = \sqrt{\mathbb{E}(|Y-\mu|^2)}$.
- A raz ao para adotarmos a noção muito menos intuitiva da variância $\mathbb{V}(Y)$ é esta propriedade: a variância de uma soma é decomposta na soma das variâncias quando as v.a.'s são independentes.
- Veremos como este mesmo resultado vai ficar muito mais elaborado nas técnicas de análise de prediç ao.
- Ela vai permitir, por exemplo,decompor a varibilidade de um sinal (acústico ou elétrico) formado pela soma de vários sinais independentes na soma de seus componentes. Estudando estes

$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y)$, se independentes

- Se X e Y são v.a.'s independentes, temos:

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

- Prova no caso particular de duas v.a.'s discretas: Como

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

- Além disso, como X e Y são independentes, então $X - \mu_x$ e $Y - \mu_y$ também são independentes e portanto

$$\mathbb{E}\left((X-\mu_x)(Y-\mu_y)\right) = \mathbb{E}(X-\mu_x)\ \mathbb{E}(Y-\mu_x) = 0 = 0$$

E se não forem independentes?

- Se X e Y não são v.a.'s independentes, teremos também uma fórmula.
- Esta fórmula vai depender de ρ, o grau de não-independência (ou CORRELAÇÃO) entre X e Y.
- Teremos:

$$\mathbb{V}(X+Y) = \mathbb{V}(X) + \mathbb{V}(Y) + 2\rho\sqrt{\mathbb{V}(X) + \mathbb{V}(Y)}$$

- Voltaremos a esta fórmula mais tarde.
- Ela vai se tornar mais clara quando estudarmos ρ.

Propriedades da variância

- Caso geral, com mais de duas v.a.'s independentes
- Se $X_1, X_2, \ldots, X_n$ são v.a.'s INDEPENDENTES então

$$\mathbb{V}(X_1 + X_2 + \ldots + X_n) = \mathbb{V}(X_1) + \ldots + \mathbb{V}(X_n)$$

Desigualdade de Tchebyshev

- O DP é o padrão universal para medir desvios (em torno do valor esperado).
- Desigualdade de Tchebyshev justifica esta última afirmação.
- 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.,

$$\mathbb{P}(|Y - \mu| > 2\sigma) \leq 1/2^2 = 0.25$$

Desigualdade de Tchebyshev

- Tchebyshev:
$$\mathbb{P}(|Y - \mu| > k\sigma) \leq 1/k^2$$
- Para $k = 4$, a probabilidade se reduz a 0.06:
$$\mathbb{P}(|Y - \mu| > 4\sigma) \leq 1/4^2 \approx 0.06$$

.

- A chance é apenas de 6% de que Y se desvie por mais que 4 DPs do seu valor esperado $\mathbb{E}(Y)$.
- Isto vale PARA TODA E QUALQUER v.a.
- o DP serve como uma métrica universal de desvios estatísticos: 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:

$$\begin{aligned}
\mathbb{P}(|Y - \mu| > 2\sigma) &\leq 1/2^2 = 0.25 \\
\mathbb{P}(|Y - \mu| > 4\sigma) &\leq 1/4^2 \approx 0.06 \\
\mathbb{P}(|Y - \mu| > 6\sigma) &\leq 1/6^2 \approx 0.03 \\
\mathbb{P}(|Y - \mu| > 10\sigma) &\leq 1/10^2 \approx 0.01 \\
\mathbb{P}(|Y - \mu| > 20\sigma) &\leq 1/20^2 \approx 0.003
\end{aligned}$$

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: Existe uma v.a. em que a desigualdade vira igualdade. Então não podemos ter nada mais apertado que Tchebyshev a não ser que joguemos fora a distribuição desta v.a. Esta distribuição é a seguinte v.a. discreta Y:

Outro comentário sobre Tchebyshev

- A v.a. em que Tchebyshev é uma igualdade:

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

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

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

- A desigualdade de Tchebyshev atinge a igualdade perfeita para esta distribuição. Então não podemos teruma cota MENOR que a fornecida para Tchebyshev se quisermos que ela atenda a TODAS as v.a.'s

Finalizando o segundo comentário

- A força da desigualdade de Tchebyshev é a sua generalidade: vale para TODA v.a.
- 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.

Outro comentário sobre Tchebyshev

- Por exemplo, se $Y \sim N(\mu, \sigma)$ e $k = 2$, então usando a densidade da normal podemos calcular a probabilidade

$$\mathbb{P}(|Y - \mu| \geq 2\sigma) = 0.04550026 \approx 0.05 = 1/20$$

- A desigualdade de Tchebyshev garante apenas que

$$\mathbb{P}(|Y - \mu| \geq 2\sigma) \leq 1/4 = 0.25$$

- Veja que a desigualdade de Tchebyshev está correta neste caso mas longe do valor exato da probabilidade no caso da normal.
- Tchebyshev garante que a chance de uma gaussiana X, bem como QUALQUER outra distribuição, afastar-se de sua esperança μ por mais de 2σ é menor que 0.25.
- O que Tchebyshev não consegue perceber é que, no caso particular da gaussiana, esta chance é muito menor, não passa de 5%.

Outro comentário sobre Tchebyshev

- A razão é que a cota de $1/4$ da Tchebyshev para um desvio de 2σ é válida para TODA v.a.
- A cota de Tchebyshev tem de valer para a normal-gaussiana mas também para TODA e qualquer outra distribuição.
- O preço dessa generalidade é terminar com uma cota que é muito exagerada em muitos casos particulares, como no caso da gaussiana.
- Apesar desse comentário, não se deixe enganar: Tchebyshev é um grande resultado.
- Ele vale SEMPRE e nada melhor pode ser obtido para TODAS as distribuições.

OPCIONAL: Prova da desigualdade de Tchebyshev

- A demonstraa̧ão não será cobrada nas provas e exercícios. Leia se houver interesse (espero que haja...)
- Tchebyshev é importante e sua prova é muito ilustrativa de como resultados gerais de probabilidade avançada são obtidos.
- Vamos assumir que X é uma v.a. contínua com densidade $f(x)$. No caso discreto, basta trocar a integral por somas.
- Seja A um conjunto qualquer da reta real $\mathbb{R}$.
- Então

$$\mathbb{P}(X \in A) = \int_A f(x)\mathrm{d}x \,,$$

pois a probabilidade de um conjunto A é a área debaixo da densidade $f(x)$ em A.

Prova da desigualdade de Tchebyshev

- Seja $A = \{x \in \mathbb{R} \text{ tais que } |x - \mu| > k\sigma\}$
- O evento $[|X - \mu| > k\sigma]$ é idêntico ao evento $[X \in A]$ pois estes dois subconjuntos de Ω são os mesmos:

$$\{\omega \in \Omega \text{ tais que } |X(\omega) - \mu| > k\sigma\}$$

é o mesmo que

$$\{\omega \in \Omega \text{ tais que } X(\omega) \in A\}$$

(veja a definição do conjunto $A \subset \mathbb{R}$ acima)

- Então

$$\mathbb{P}(|X - \mu| > k\sigma) = \mathbb{P}(X \in A) = \int_{\{x:|x-\mu|>k\sigma\}} f(x)\mathrm{d}x$$

Prova da desigualdade de Tchebyshev

- Por definição, temos

$$\mathbb{V}(X) = \sigma^2 = \mathbb{E}(X-\mu)^2 = \int_{\mathbb{R}} (x-\mu)^2 f(x) \mathrm{d}x$$

- A reta pode ser particionada em dois conjuntos disjuntos: $\mathbb{R} = A \cup A^c$.
- Então

$$\begin{aligned}
\sigma^2 &= \int_{A \cup A^c} (x-\mu)^2 f(x) \mathrm{d}x \\
&= \int_A (x-\mu)^2 f(x) \mathrm{d}x + \int_{A^c} (x-\mu)^2 f(x) \mathrm{d}x \\
&\geq \int_A (x-\mu)^2 f(x) \mathrm{d}x
\end{aligned}$$

pois o segundo termo é sempre positivo (ou zero).

Prova da desigualdade de Tchebyshev

- Resumindo:

$$\sigma^2 \geq \int_A (x-\mu)^2 f(x) \mathrm{d}x$$

onde $A = \{x : |x-\mu| > k\sigma\}$.

- Mas se $x \in A$ temos $(x-\mu)^2 > k^2\sigma^2$ e portanto

$$\begin{aligned}
\sigma^2 &\geq \int_A (x-\mu)^2 f(x) \mathrm{d}x \\
&\geq \int_A (k^2\sigma^2) f(x) \mathrm{d}x \\
&= (k^2\sigma^2) \int_A f(x) \mathrm{d}x \\
&= (k^2\sigma^2)\, \mathbb{P}(X \in A)
\end{aligned}$$

Prova da desigualdade de Tchebyshev

- Concluindo, encontramos que

$$\sigma^2 \geq (k^2\sigma^2)\mathbb{P}(X \in A)$$

- Ou seja

$$\mathbb{P}(X \in A) \leq \frac{1}{k^2}$$

- Mas

$$\mathbb{P}(X \in A) = \mathbb{P}\left(|X - \mu| \geq k\sigma\right)$$

- Assim, temos a desigualdade de Tchebyshev

$$\boxed{\mathbb{P}\left(|X - \mu| \geq k\sigma\right) \leq \frac{1}{k^2}}$$

Estimando μ e σ com dados

- O valor esperado $\mathbb{E}(Y) = \mu$ e o DP $\sigma = \sqrt{\mathbb{E}\left(Y - \mu\right)^2}$ são valores que não dependem dos dados.
- Por exemplo, suponha que Y possua distribuição 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 distribuição com densidade $f(y)$ e que NAO CONHECEMOS $f(y)$.
- Portanto, não podemos obter $\mathbb{E}(Y) = \mu$ e o DP σ.
- No entanto, podemos ESTIMAR $\mathbb{E}(Y) = \mu$ e o DP σ 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 (teorema que ainda será estudado),

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

  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 μ).
- 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 μ.

σ e o segundo momento

- DP σ 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.

σ 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.

V.A.s limitadas: desigualdade de Hoeffding

- Veremos no futuro...