Probabilidade

Fernando B. Sabino da Silva

Contents

1 Probabilidade de eventos

1.1 O conceito de probabilidade

- Experimento: Medir o tempo de espera em uma fila. Se o tempo exceder 2 minutos marque 1 e 0, caso contrário.

- O experimento é realizado n vezes com resultados $y_1, y_2, \ldots, y_n$. Há uma **variação aleatória** no resultado, i.e. às vezes ocorre 1 e em outras vezes ocorre 0.
- **Probabilidade empírica** de exceder 2 minutos:

$$p_n = \frac{\sum_{i=1}^{n} y_i}{n}.$$

- **Probabilidade teórica** de exceder 2 minutos:

$$p = \lim_{n \to \infty} p_n.$$

.

- Tentamos inferir o verdadeiro valor de p com base em uma amostra. Por exemplo, "$p > 0.1$?" ("mais de \$1\$0% dos clientes experimentaram um tempo de espera superior a 2 minutos?").
- A (inferência) estatística está preocupada com tais questões que são úteis para a tomada de decisões. Em geral, só temos acesso a uma amostra finita.

1.2 Experimento real

- Em um determinado mês de 2017, um grupo de estudantes respondeu a pergunta de quanto tempo eles precisaram esperar na fila em uma determinada cantina (em minutos):

```
y_cantina <- c(2, 5, 1, 6, 1, 1, 1, 1, 3, 4, 1, 2, 1, 2, 2, 2, 4, 2, 2, 5, 20, 2, 1, 1, 1, 1)
x_cantina <- ifelse(y_cantina > 2, 1, 0)
x_cantina
```

```
##  [1] 0 1 0 1 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0
```

- Probabilidade empírica de esperar mais de 2 minutos:

```
p_cantina <- sum(x_cantina) / length(x_cantina)
p_cantina
```

```
## [1] 0.2692308
```

- Questão: A probabilidade na população é $p > 1/3$?
- Nota: Um estudante disse que esperou por 20 minutos. Dado os outros resultados, podemos duvidar da veracidade e ignorar essa observação).

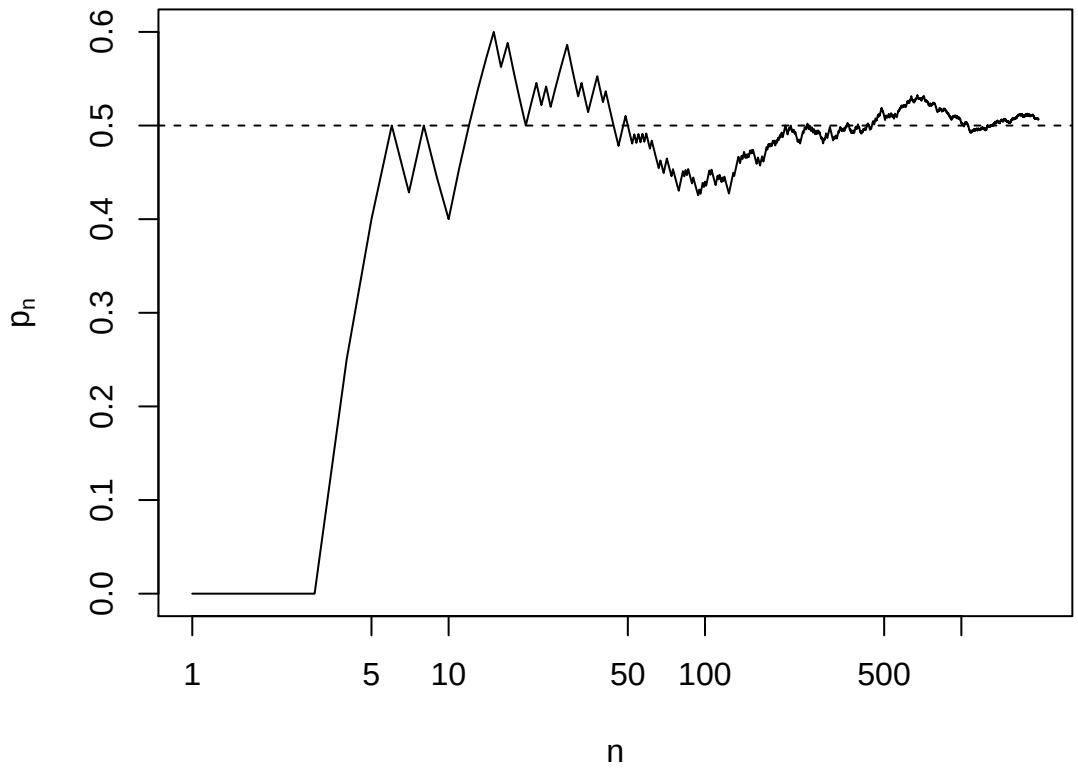
1.3 Outro experimento

- John Kerrich, um matemático sul-africano, estava visitando Copenhague quando a segunda Guerra Mundial eclodiu. Dois dias antes ele estava programado para viajar para a Inglaterra quando os alemães invadiram a Dinamarca. Kerrich passou o resto da guerra em um acampamento e para passar o tempo realizou uma série de experimentos. Em um destes, ele jogou uma moeda 10,000 vezes. Os resultados são mostrados a seguir.
- Abaixo, `x` é um vetor com os primeiros 2000 resultados do experimento de John Kerrich. (0 = coroa, 1 = cara):

```
head(x, 10)
```

```
##  [1] 0 0 0 1 1 1 0 1 0 0
```

- Gráfico da probabilidade empírica p_n de sair cara contra o número de lançamentos n:



(O eixo horizontal está na escala log).

1.4 Definições

- **Espaço Amostral**: Todos os possíveis resultados de um experimento.
- **Evento**: Qualquer subconjunto do espaço amostral.

Nós conduzimos o experimento n vezes. Seja $\#(A)$ o número de vezes que observamos o evento A.

- **Probabilidade empírica** do evento A:

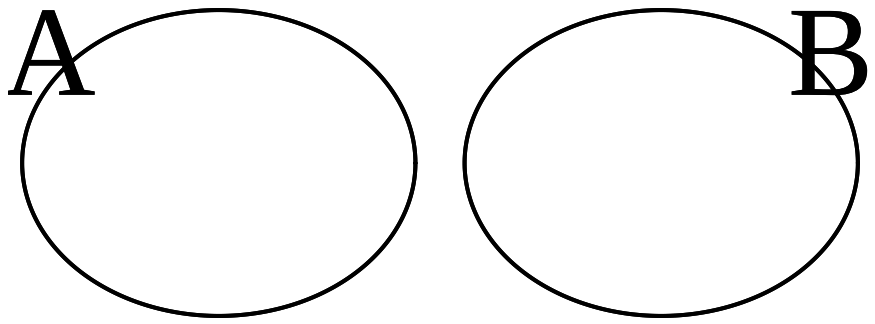
$$p_n(A) = \frac{\#(A)}{n}.$$

- **Probabilidade teórica** do evento A:

$$P(A) = \lim_{n\to\infty} p_n(A)$$

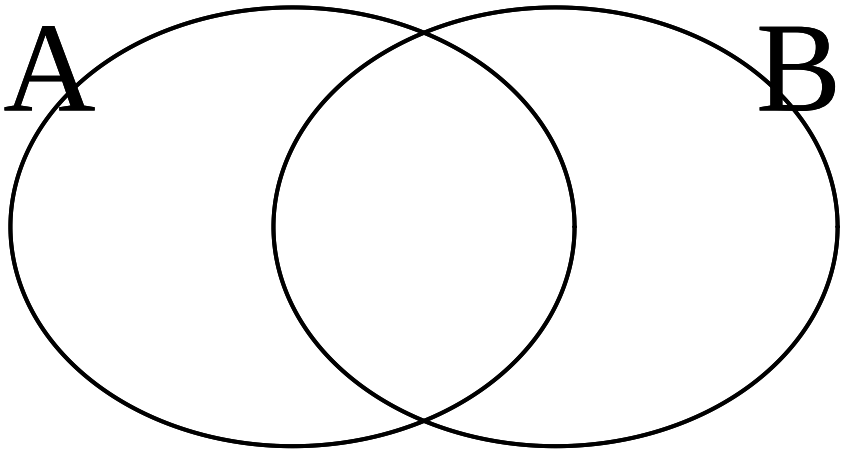
1.5 Probabilidades teóricas de dois eventos

- Se dois eventos A e B são **disjuntos** (sem intersecção) então
  - $\#(A \text{ e } B) = 0$ implica que $P(A \cap B) = 0$.
  - $\#(A \text{ ou } B) = \#(A) + \#(B)$ implica que $P(A \cup B) = P(A) + P(B)$.



- Se dois eventos A e B **não são disjuntos** então a fórmula mais geral é

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$



1.6 Probabilidade Condicional

- Digamos que consideramos dois eventos A e B. Então a **probabilidade condicional** de A dado (ou condicional ao) o evento B é escrito $P(A \mid B)$ e é definido por

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

- A probabilidade acima pode ser entendida do seguinte modo: "quão provável é o evento A se nós sabemos que B ocorreu".

1.6.1 Exemplo com os dados das revistas:

```
magAds <- read.delim("C:/Users/fsabino/Desktop/Codes/papers/Introductory_Stat_I/notebook/datasets_ads.t

# Criando dois novos fatores: 'words' e 'education':
magAds$words <- cut(magAds$WDS, breaks = c(31, 72, 146, 230), include.lowest = TRUE)
magAds$education <- factor(magAds$GROUP, levels = c(1, 2, 3), labels = c("high", "medium", "low"))

library(mosaic)
```

```
## Warning: package 'mosaic' was built under R version 3.4.3
## Warning: package 'dplyr' was built under R version 3.4.3
## Warning: package 'ggformula' was built under R version 3.4.3
## Warning: package 'ggplot2' was built under R version 3.4.3
## Warning: package 'mosaicData' was built under R version 3.4.3
## Warning: package 'Matrix' was built under R version 3.4.4
```

```
tab <- tally( ~ words + education, data = magAds)
tab
```

```
##            education
## words       high medium low
##   [31,72]      4      6   5
##   (72,146]     5      6   8
##   (146,230]    9      6   5
```

- O evento A={words=(146,230]} (o anúncio é um texto "difícil") tem probabilidade empírica

$$p_n(A) = \frac{9+6+5}{54} = \frac{20}{54} \approx 37\%.$$

- Digamos que só estamos interessados na probabilidade de um texto "difícil" (evento A) para revistas de educação superior (high education), i.e. condicionando no evento B = {education=high}. Então a probabilidade condicional empírica pode ser calculada a partir da tabela acima:

$$p_n(A \mid B) = \frac{9}{4+5+9} = \frac{9}{18} = 0.5 = 50\%.$$

- A probabilidade condicional de A dado B pode ser (teoricamente) expressada por

$$\begin{aligned} P(A \mid B) &= P(\text{words} = (146, 230] \mid \text{education} = \text{high}) \\ &= \frac{P(\text{words} = (146, 230] \cap \text{education} = \text{high})}{P(\text{education} = \text{high})}, \end{aligned}$$

que traduzindo para a probabilidade empírica (substituindo P por p_n) dará

$$
\begin{aligned}
p_n(A \mid B) &= \frac{p_n(\text{words} = (146, 230] \cap \text{education} = \text{high})}{p_n(\text{education} = \text{high})} \\
&= \frac{\frac{9}{54}}{\frac{4+5+9}{54}} \\
&= \frac{9}{4+5+9} \\
&= 50\%
\end{aligned}
$$

como calculado acima.

1.7 Probabilidade condicional e independência

- Se a informação sobre B não muda a probabilidade de A nós dizemos que A é **independente** de B e escrevemos
$$P(A \mid B) = P(A) \quad \Leftrightarrow \quad P(A \cap B) = P(A)P(B)$$
Quando isto acontecer nós dizemos que A e B são **eventos independentes**.
- Em geral, os eventos $A_1, A_2, ..., A_k$ são independentes se
$$P(A_1 \cap A_2 \cap ... \cap A_k) = P(A_1)P(A_2) \cdots P(A_k)$$
e se o produto de todas as combinações de ordem 2 a k-1 também são válidas, i.e., k eventos são ditos independentes se são **independentes k a k, (k-1) a (k-1),. . . , dois a dois**, isto é, para cada subconjunto S de $1, 2, \ldots k$.

1.7.1 Dados das revistas revisitados

- Lembre-se das probabilidades empíricas calculadas acima:
$$p_n(A) = 37\% \quad \text{e} \quad p_n(A \mid B) = 50\%.$$
- Isto indica (não podemos dizer com certeza, pois só temos a disposição uma amostra finita - no curso de estatística II vemos detalhes de como testar isso) que a probabilidade teórica

$$P(A) \neq P(A \mid B)$$

e, portanto, que o conhecimento sobre o evento B (nível de educação elevado) pode transmitir informações sobre a probabilidade do evento A (o anúncio contém um texto "difícil").

1.8 Um pouco mais de formalidade

1.8.1 Modelos de Probabilidade

Quando discutimos modelos de probabilidade, nós falamos de **experimentos** aleatórios que produzem um dos vários **resultados** possíveis. Um **modelo de probabilidade** que descreve a incerteza de um experimento consiste em três elementos (descrevemos dois abaixo):

- O **espaço amostral**, geralmente denominado por Ω, representando o conjunto que contém todos os resultados possíveis.
- Uma **função de probabilidade** que atribui a um evento A um número não-negativo, $P[A]$, que representa a probabilidade de que o evento A ocorra como resultado do experimento.

Nós chamamos de $P[A]$ a **probabilidade** do evento A. Um evento A pode ser qualquer subconjunto do espaço amostral, não necessariamente um único resultado possível. As leis da probabilidade devem seguir uma série de regras, que são o resultado de um conjunto de axiomas que introduziremos agora.

1.9 Axiomas de Probabilidade

Dado um espaço amostral Ω para um experimento em particular, a **função de probabilidade** associada ao experimento deve satisfazer os seguintes axiomas.

- *Não-negatividade*: $P[A] \geq 0$ para qualquer evento $A \subset \Omega$.
- *Normalização*: $P[\Omega] = 1$. Ou seja, a probabilidade de todo o espaço é 1.
- *Aditividade*: Para eventos mutuamente exclusivos $E_1, E_2, \ldots$

$$P\left[\bigcup_{i=1}^{\infty} E_i\right] = \sum_{i=1}^{\infty} P[E_i]$$

.

Usando estes axiomas, muitas regras adicionais de probabilidade podem ser facilmente derivadas.

1.10 Regras de Probabilidade

- Dado um evento A e seu complemento, A^c, ou seja, os resultados em Ω que não estão em A, temos a regra do **complementar**:

$$P[A^c] = 1 - P[A]$$

- Em geral, para dois eventos A e B, temos a **regra da adição**:

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

Exercício: Prove que as relações acima são válidas usando os três axiomas de probabilidade. Prove também que se $A \subseteq B$, então $P(A) \leq P(B)$.

- Se A e B também são *disjuntos*, então temos:

$$P[A \cup B] = P[A] + P[B]$$

Se tivermos n eventos mutuamente exclusivos, $E_1, E_2, \ldots E_n$, então temos:

$$P\left[\bigcup_{i=1}^{n} E_i\right] = \sum_{i=1}^{n} P[E_i]$$

1.11 Regra de Bayes

Defina uma **partição** de um espaço amostral Ω como um conjunto de eventos disjuntos $A_1, A_2, \ldots, A_n$ cuja união é o espaço amostral Ω. Isso é

$$A_i \cap A_j = \emptyset$$

para todos $i \neq j$ e

$$\bigcup_{i=1}^{n} A_i = \Omega.$$

Seja $A_1, A_2, \ldots, A_n$ formam uma partição do espaço amostral, onde $P[A_i] > 0$ para todo i. Então para qualquer evento B com $P[B] > 0$ temos a **Regra de Bayes**:

$$P[A_i|B] = \frac{P[A_i]P[B|A_i]}{P[B]} = \frac{P[A_i]P[B|A_i]}{\sum_{i=1}^{n} P[A_i]P[B|A_i]}$$

O denominador da última igualdade é freqüentemente chamado de **lei da probabilidade total**:

$$P[B] = \sum_{i=1}^{n} P[A_i]P[B|A_i]$$

Dois eventos A e B são considerados **independentes** se satisfizerem

$$P[A \cap B] = P[A] \cdot P[B]$$

Uma coleção de eventos $E_1, E_2, \ldots E_n$ é considerada independente se

$$P\left[\bigcap_{i \in S} E_i\right] = \prod_{i \in S} P[E_i]$$

para cada subconjunto S de $1, 2, \ldots n$.

1.12 Regra de Bayes (Continuação)

- As probabilidades dos estados da natureza são alteradas de acordo com as informações obtidas. Sempre que coletarmos uma amostra e ela contiver informações relevantes, as nossas probabilidades serão revisadas. Quanto maior informação, menor a incerteza.
- Não sabemos qual é o **verdadeiro** estado da natureza (o que realmente ocorrerrá), mas temos uma avaliação das probabilidades (chamadas de probabilidades a priori, digamos, $P(S_j)$).
- Queremos calcular as probabilidades revisadas, digamos, $P(S_j|E_i)$, mas sabemos as **verossimilhanças** (likelihood, em inglês) $P(E_i|S_j)$ = probabilidade do resultado experimental (E_i) condicional aos estados de natureza (S_j).

STATE	PRIOR	LIKELIHOOD	JOINT: $P(S_j \& E_i)$	REVISED: $P(S_j \mid E_i)$
S_1	$P(S_1)$	$P(E_i \mid S_1)$	$P(S_1)P(E_i \mid S_1)$	$P(S_1 \& E_i)/P(E_i)$
S_2	$P(S_2)$	$P(E_i \mid S_2)$	$P(S_2)P(E_i \mid S_2)$	$P(S_2 \& E_i)/P(E_i)$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
S_n	$P(S_n)$	$P(E_i \mid S_n)$	$P(S_n)P(E_i \mid S_n)$	$P(S_n \& E_i)/P(E_i)$
			$P(E_i) = \Sigma_j$ $P(S_j)P(E_i \mid S_j)$	
		Known		Computed

1.13 Exercício: Problema de Diagnóstico Médico

1. Diagnóstico Médico: Suponha que 1% da população de um determinado lugar seja HIV positivo. O Departamento de Saúde Pública tem um teste de diagnóstico barato que é administrado às pessoas que pedem para serem testadas. O teste resulta em uma indicação positiva ou negativa e apresenta as seguintes características:

(a) O teste dá uma indicação positiva (verdadeira) 99% das vezes em que uma pessoa é realmente portadora do vírus. Infelizmente, o teste dá uma indicação negativa (falsa) para 1% das pessoas que portam o vírus.

(b) O teste dá uma indicação negativa (verdadeira) 95% das vezes em que uma pessoa não é portadora do vírus (e, portanto, dá uma indicação positiva (falsa) para 5% das pessoas).

- Suponha que um residente escolhido ao acaso tenha acabado de fazer o teste. Para seu choque, o teste resultou em uma indicação positiva. Qual é a probabilidade dele ser realmente portador do vírus?
- Suponha agora que o teste tenha resultado em uma indicação negativa. O residente deve ficar confiante de que ele não é realmente HIV positivo?
- Como as suas respostas mudam se o teste se tornar 10 vezes mais confiável, isto é, se as taxas de erro para falsos negativos e falsos positivos, respectivamente, forem reduzidas para 0.1% e 0.5%?

1.14 Exercício: Problema de Controle de Qualidade

2. Controle de Qualidade: A DataSafe produz pen drives. Seu processo de fabricação é direcionado para produzir defeitos a uma taxa de 1/10 de 1%, isto é, 0.1%. Às vezes, quando o processo é iniciado, ele desliza ligeiramente e produz defeitos a uma taxa de 2/10 de 1%. Como parar o maquinário para recalibrar o processo é muito caro, a DataSafe está disposta a operar com uma taxa de defeitos um pouco maior. Ocasionalmente, no entanto, o processo se desvia tanto que produz defeitos a uma taxa de 5/10 de 1%, o que é inaceitável para a gerência. Com base em experiências passadas, o engenheiro chefe da DataSafe, observou que, em qualquer dia, as probabilidades para a verdadeira taxa de defeitos são:

Condição Operacional	Boa	Ok	Ruim
Taxa de Defeito	.001	.002	.005
Probabilidade	.75	.24	.01

- No início de cada dia, ele retira uma amostra da produção inicial para deciri se a máquina precisa ser recalibrada. A sua regra operacional é recalibrar sempre que ele não puder ter pelo menos 90% de certeza de que a taxa de defeitos está abaixo de 5/10 de 1%.
- Suponha que o primeiro pen drive escolhido aleatoriamente esteja com defeito. A produção deveria ser parada para recalibração? E se os dois primeiros escolhidos aleatoriamente estiverem com defeito?

1.15 Solução para o problema do diagnóstico médico

Dada uma indicação positiva:

State	Prior	Likelihood	Joint	Revised
HIV +	0.01	0.99	0.01(0.99) = 0.0099	99/594 = 0.1667
HIV -	0.99	0.05	0.99(0.05) = 0.0495	495/594 = 0.8333
			P(+) = 0.0594	

Dada uma indicação negativa:

State	Prior	Likelihood	Joint	Revised
HIV +	0.01	0.01	0.01(0.01) = 0.0001	1/9406 = 0.0001
HIV -	0.99	0.95	0.99(0.95) = 0.9405	9405/9406 = 0.9999
			P(+) = 0.9406	

- Se o teste se tornar 10 vezes mais confiável.

Dada uma indicação positiva:

State	Prior	Likelihood	Joint	Revised
HIV +	0.01	0.999	0.01(0.999) = 0.00999	999/1494 = 0.6687
HIV -	0.99	0.005	0.99(0.005) = 0.00495	495/1494 = 0.3313
			P(+) = 0.01494	

Dada uma indicação negativa:

State	Prior	Likelihood	Joint	Revised
HIV +	0.01	0.001	0.01(0.001) = 0.00001	1/98506 = 0.00001
HIV -	0.99	0.995	0.99(0.995) = 0.98505	98505/98506 = 0.99999
			P(+) = 0.98506	

1.16 Solução para o Problema de Controle de Qualidade

States of Nature	Prior Probabilities
S_1: Good operating condition	P(Good) = 0.75
S_2: OK operating condition	P(OK) = 0.24
S_3: Bad operating condition	P(Bad) = 0.01

	Likelihoods		
Experimental outcomes	$P(E_j\|S_1)$	$P(E_j\|S_2)$	$P(E_j\|S_3)$
E_1: Selected diskette satisfactory	0.999	0.998	0.995
E_2: Selected diskette unsatisfactory	0.001	0.002	0.005

- Probabilidades revisadas dado que um pen drive defeituoso é selecionado:

State	Prior	Likelihood	Joint	Revised
Good	0.75	0.001	0.00075	75/128 = 0.5859
OK	0.24	0.002	0.00048	48/128 = 0.3750
Bad	0.01	0.005	0.00005	5/128 = 0.0391
			0.00128 = P(defective)	

- Então, P(Bom ou OK | defeituoso) = 0.9609 > 0.9. Ele não deve recalibrar.
- Note as implicações disso. Se o pen drive escolhido for defeituoso, ele não vai recalibrar. E se não estiver com defeito, ele certamente não irá recalibrar. Logo, amostrar apenas um pen drive é sem utilidade.
- Probabilidades revisadas dado que dois pen drives defeituosos são selecionados:

State	Prior	Likelihood	Joint	Revised
Good	0.75	0.000001	0.00000075	75/196 = 0.38625
OK	0.24	0.000004	0.00000096	96/196 = 0.48980
Bad	0.01	0.000025	0.00000025	25/196 = 0.12755
			0.00000196 = P(2 of 2 defective)	

- Então, P(Bom ou OK | 2 de 2 defeituosos) = 0.87245 < 0.9. Ele deve recalibrar.
- Deixo para você verificar que ele não irá recalibrar se nenhum pen drive amostrado estiver com defeito (seja calculando ou inferindo pelos que encontramos na parte anterior) ou se apenas um dos dois estiver com defeito.

1.17 Distribuição discreta

1.17.1 Exemplo: Dados das revistas

```
# Tabela contendo o percentual de anúncios em cada combinação dos níveis de 'words' e 'education'
tab <- tally( ~ words + education, data = magAds, format = "percent")
round(tab, 2) # Duas casas decimais
```

```
##            education
## words        high medium   low
##   [31,72]    7.41  11.11  9.26
##   (72,146]   9.26  11.11 14.81
##   (146,230] 16.67  11.11  9.26
```

- Os 9 eventos disjuntos acima (correspondentes as combinações de `words` e `education`) compõem todo o espaço amostral para as duas variáveis. As probabilidades empíricas de cada evento são dadas na tabela.

1.17.2 Distribuição discreta

- Em geral:

– Seja $A_1, A_2, \ldots, A_k$ uma subdivisão do espaço amostral em eventos disjuntos (par a par).
– As probabilidades $P(A_1), P(A_2), ..., P(A_k)$ (**distribuição discreta**) satisfazem

$$\sum_{i=1}^{k} P(A_i) = 1.$$

1.17.3 Exemplo: Três lançamentos de uma moeda

- **Variável aleatória**: Uma variável aleatória é simplesmente uma função Y que mapeia os resultados do espaço amostral para números reais, isto é, que mapeia os possíveis resultados do experimento em um número.
- Resultados possíveis de um experimento com 3 lançamentos de moedas:
  – 0 caras (KKK)
  – 1 cara (CKK, KCK, KKC)
  – 2 caras (CCK, CKC, KCC)
  – 3 caras (CCC)
- Os eventos acima são disjuntos e compõem todo o espaço amostral.
- Seja Y o número de caras no experimento: $Y(KKK) = 0, Y(CKK) = 1, \ldots$
- Assuma que cada resultado é igualmente provável, i.e. 1/8 de probabilidade para cada um deles. Então,
  – $P(\text{nenhuma cara}) = P(Y = 0) = P(KKK) = 1/8$.
  – $P(\text{uma cara}) = P(Y = 1) = P(CKK \text{ ou } KCK \text{ ou } KKC) = P(CKK) + P(KCK) + P(KKC) = 3/8$.
  – Similarmente para 2 ou 3 caras.
- Então, a distribuição de Y é

Número de caras, Y	0	1	2	3
Probabilidade	1/8	3/8	3/8	1/8

2 Distribuição de variáveis aleatórias

2.1 Distribuição de probabilidade

- Exemplo: Conduzimos um experimento no qual fazemos uma medição quantitativa Y (uma variável aleatória), por exemplo, contamos o número de palavras em um anúncio ou o tempo de espera em uma fila.
- De antemão, há muitos resultados possíveis para os experimentos, i.e. os valores de Y que irão acontecer em uma realização do experimento são incertos, mas nós podemos quantificá-los pela **distribuição de probabilidade** de Y.
- Uma definição não rigorosa, mas útil para transmitir a ideia é:

$$\text{distribuição} = \text{lista de possíveis } \textbf{valores} + \text{probabilidades } \textbf{associadas}$$

- Para qualquer intervalo (a, b), a distribuição indica a probabilidade de observar um valor da variável aleatória Y neste intervalo:

$$P(a < Y < b), \qquad -\infty < a < b < \infty.$$

- Se os possíveis valores de uma variável aleatória são discretos, isto é, se nós podemos enumerar todos os possíveis valores de Y, então a variável aleatória Y é chamada de **discreta**. Por exemplo, o número de palavras em um anúncio.

- Se os possíveis valores de uma variável aleatória são contínuos, isto é, Y pode assumir qualquer valor dentro de um intervalo, então a variável aleatória Y é chamada de **contínua**. Por exemplo, o tempo de espera em uma fila.

2.2 Variáveis Aleatórias Discretas

- A distribuição de uma variável aleatória discreta X é mais frequentemente especificada por uma lista de possíveis valores e uma função massa de probabilidade $p(x)$, isto é,

$$p(x) = p_X(x) = P[X = x].$$

Costumeiramente nós abandonamos o subscrito da notação mais correta $p_X(x)$ e simplesmente escrevemos $p(x)$. A variável aleatória relevante será discernida do contexto.

O exemplo mais comum de uma variável aleatória discreta é a variável aleatória com distribuição binomial. A função massa de uma variável aleatória X com distribuição binomial é dada por

$$P(X = x|n, p) = p_X(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n, \ n \in \mathbb{N}, \ 0 < p < 1.$$

A última linha contém uma grande quantidade de informação.

- A função $p_X(x|n, p)$ é a função massa. É uma função de x, os possíveis valores da variável aleatória X. Ela é condicional aos parâmetros n e p. Valores diferentes destes parâmetros especificam distribuições binomiais diferentes.
- $x = 0, 1, \ldots, n$ indica o espaço amostral de x, isto é, os possíveis valores da variável aleatória.

$n \in \mathbb{N}$ e $0 < p < 1$ especificam os espaços dos parâmetros. Estes são os valores possíveis dos parâmetros que fornecem uma distribuição binomial válida. Frequentemente, toda essa informação é simplesmente codificada escrevendo

$$X \sim \text{Bin}(n, p).$$

2.3 Variáveis Aleatórias Contínuas

- A distribuição de uma variável aleatória contínua X é mais frequentemente especificada pelo conjunto de possíveis valores e uma função densidade de probabilidade, $f(x)$. (A função de distribuição acumulada (densidade acumulada), $F(x)$ ou a função característica ou geratriz de momentos também seriam suficientes.)
- A probabilidade do evento $a < X < b$ é calculada por

$$P[a < X < b] = \int_a^b f(x)dx.$$

Note que as densidades não são probabilidades.

- O exemplo mais comum de uma variável aleatória contínua é uma variável aleatória com distribuição normal. A densidade de uma variável aleatória normal X, é dada por

$$f(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2p}} \cdot \exp\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \quad -\infty < x < \infty, \ -\infty < \mu < \infty, \ \sigma > 0.$$

- A função $f(x|\mu, \sigma^2)$ é a função de densidade. É uma função de x, os valores possíveis da variável aleatória X. Ela é condicional aos parâmetros μ e σ^2. Diferentes valores destes parâmetros especificam diferentes distribuições normal.
- $-\infty < x < \infty$ indica o espaço amostral de x. Neste caso, a variável aleatória pode assumir todo e qualquer valor real.
- $-\infty < \mu < \infty$ and $\sigma > 0$ especificam o espaço de parâmetros. Estes são os valores possíveis dos parâmetros que fornecem uma distriuição normal válida. Muitas vezes, toda essa informação é simplesmente codificada escrevendo

$$X \sim N(\mu, \sigma^2)$$

2.4 Independência entre Variáveis Aleatórias

Considere duas variáveis aleatórias X e Y. Nós dizemos que elas são independentes se

$$f(x, y) = f(x) \cdot f(y)$$

para todo x e y. Aqui $f(x, y)$ é a função densidade (massa) conjunta de X e Y. Nós chamamos de $f(x)$ e de $f(y)$ as funções densidade (massa) marginais de X e de Y, respectivamente.

- A função densidade (massa) conjunta $f(x, y)$ juntamente com os possíveis valores (x, y) especificam a distribuição conjunta de X e Y.

Noções similares existem para mais de duas variáveis aleatórias.

2.4.1 Amostra Aleatória

Nós realizamos um experimento n vezes, onde o resultado do i-ésimo experimento corresponde a uma medição de uma variável aleatória Y_i, onde assumimos que

- Os experimentos são **independentes**
- As variáveis $Y_1, \ldots, Y_n$ têm a **mesma distribuição**

2.5 Parâmetros populacionais

- Quando o tamanho da amostra aumenta, a média da amostra, $\overline{y}$, por exemplo, irá se estabilizar em torno de um valor fixo, μ, que é usualmente desconhecido. O valor μ é chamado de **média populacional**.
- Correspondentemente, o desvio padrão da amostra, s, irá se estabilizar em torno de um valor fixo, σ, que geralmente é desconhecido. O valor σ é chamado de **desvio padrão da população**.
- Notação:
  - μ (mu) denota a médida da população.
  - σ (sigma) denota o desvio padrão da população.

População	Amostra
μ	$\overline{y}$
σ	s

2.5.1 Distribuição de uma variável aleatória discreta

- Valores possíveis para Y: $\{y_1, y_2, \ldots, y_k\}$.
- A **distribuição** de Y é a probabilidade de cada valor possível: $p_i = P(Y = y_i), \quad i = 1, 2, \ldots, k$.
- A distribuição satisfaz: $\sum_{i=1}^{k} p_i = 1$.

2.6 Esperança

Para variáveis aleatórias discretas, nós definimos a **esperança** da função de uma variável aleatória X da seguinte maneira.

$$\mathbb{E}[g(X)] \triangleq \sum_x g(x)p(x)$$

Para variáveis aleatórias contínuas, nós temos uma definição semelhante.

$$\mathbb{E}[g(X)] \triangleq \int g(x)f(x)dx$$

Para funções específicas g, as esperanças recebem nomes

A **média** de uma variável aleatória X é dada por

$$\mu_X = \mathbb{E}[X].$$

Então, para uma variável aleatória discreta, nós temos

$$\mathrm{E}[X] = \sum_x x \cdot p(x)$$

Para uma variável aleatória contínua, simplesmente substituímos a soma por uma integral.

A variância de uma variável aleatória X é dada por

$$\sigma_X^2 = \mathrm{var}[X] \triangleq \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

O desvio padrão de uma variável aleatória X é dada por

$$\sigma_X = \mathrm{sd}[X] \triangleq \sqrt{\sigma_X^2} = \sqrt{\mathrm{var}[X]}.$$

A covariância de variáveis aleatórias X e Y é dada por

$$\mathrm{cov}[X, Y] \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

2.7 Verossimilhança (Likelihood)

Considere n variáveis aleatórias iid $X_1, X_2, \ldots X_n$. Nós definimos a função de verossimilhança por

$$\mathcal{L}(\theta \mid x_1, x_2, \ldots x_n) = \prod_{i=i}^{n} f(x_i; \theta)$$

onde $f(x_i; \theta)$ é a função densidade (ou massa) da variável aleatória X_i avaliada em x_i com parâmetro θ.

Enquanto a probabilidade é uma função de um possível valor (ou intervalo) observado, dados determinados valores dos parâmetros, a verossimilhança é o "oposto": é uma função dos valores possíveis dos parâmetros dada a amostra, i.e., verossimilhança é uma medida da evidência que uma amostra fornece para valores específicos dos parâmetros em um modelo paramétrico.

A maximização da verossimilhança é uma técnica comum para ajustar um modelo aos dados.

2.8 Valor esperado (média) para uma distribuição discreta

- O **valor esperado** ou **média (populacional)** de Y é

$$\mu = \sum_{i=1}^{k} p_i y_i$$

- Uma propriedade importante do valor esperado é que ele tem a mesma unidade de medida das observações (por exemplo, metro).

2.8.1 Exemplo: número de caras em 3 lançamentos de uma moeda

- Lembre-se da distribuição de Y (número de caras):

y (número de caras)	0	1	2	3
$P(Y = y)$	1/8	3/8	3/8	1/8

- Então o valor esperado é

$$\mu = 0\frac{1}{8} + 1\frac{3}{8} + 2\frac{3}{8} + 3\frac{1}{8} = 1.5.$$

Observe que o valor esperado não precisa ser um resultado possível do próprio experimento.

2.9 Variância e desvio padrão de uma distribuição discreta

- A **variância (populacional)** de Y é

$$\sigma^2 = \sum_{i=1}^{k} (y_i - \mu)^2 p_i$$

- O **desvio padrão (populacional)** é $\sigma = \sqrt{\sigma^2}$.
- Nota: Se as observações forem medidas em metro, a **variância** terá unidade metro^2 o que comumente não estamos acostumados a interpretar. O **desvio padrão**, por outro lado, tem a mesma unidade de medida que as observações.

2.9.1 Exemplo: número de caras em 3 lançamentos de uma moeda

A distribuição da variável aleatória 'número de caras em 3 lançamentos de uma moeda' tem variância

$$\sigma^2 = (0-1.5)^2\frac{1}{8} + (1-1.5)^2\frac{3}{8} + (2-1.5)^2\frac{3}{8} + (3-1.5)^2\frac{1}{8} = 0.75.$$

e desvio padrão

$$\sigma = \sqrt{\sigma^2} = \sqrt{0.75} = 0.866.$$

2.10 A distribuição binomial

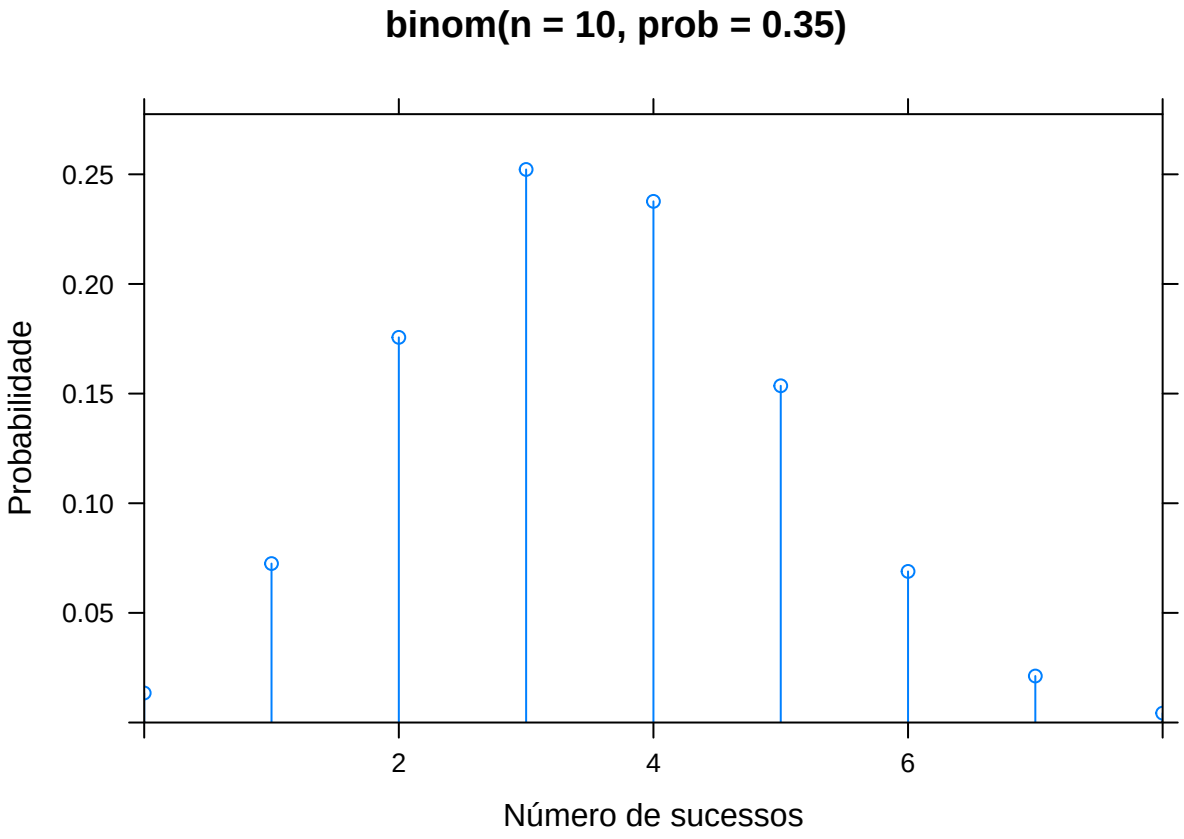
- A **distribuição binomial** é uma distribuição discreta.
- Seja Y a variável aleatória que representa o número de sucessos obtidos em n experimentos aleatórios (independentes). Assuma que cada experimento tem apenas dois resultados possíveis, denominados **sucesso** e **fracasso** e que cada experimento tem a mesma probabilidade p de sucesso.
- Dizemos que Y tem uma **distribuição binomial** com parâmetros n e p.

- Neste caso, podemos mostrar que

$$p_Y(y) = P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y},$$

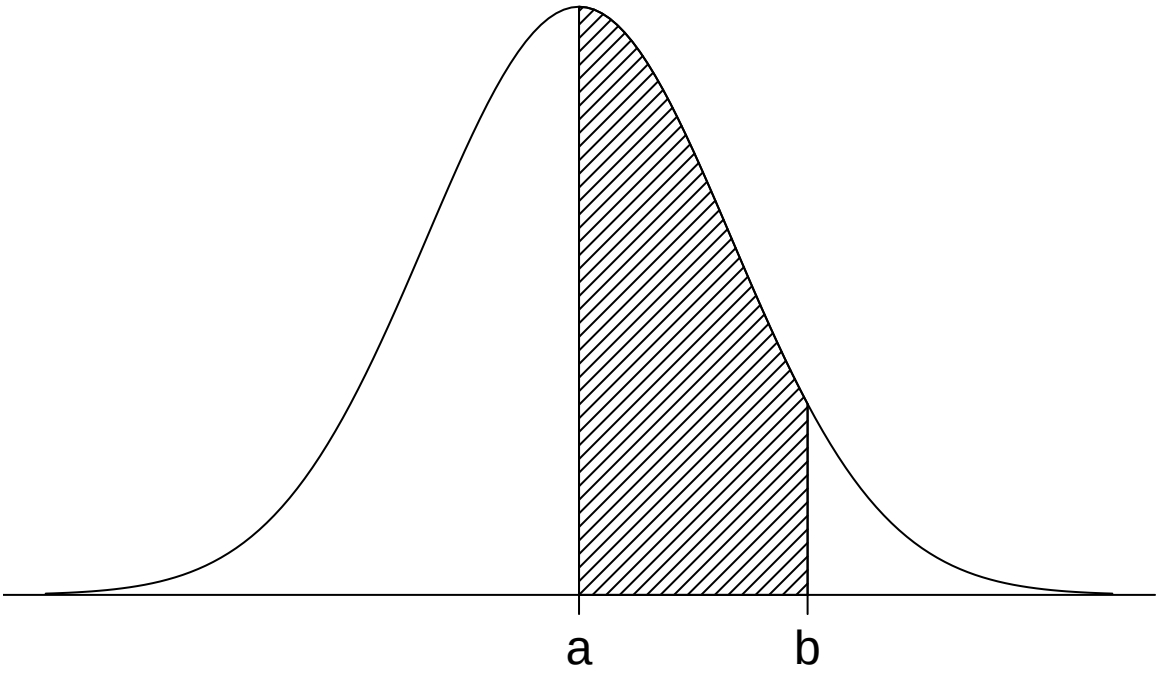
  onde $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ e $m!$ é o produto dos primeiros m inteiros.
- Valor esperado: $\mu = np$.
- Variância: $\sigma^2 = np(1-p)$.
- Desvio padrão: $\sigma = \sqrt{np(1-p)}$.

```
# A distribuição binomial com n = 10 e p = 0.35:
plotDist("binom", size = 10, prob = 0.35,
         ylab = "Probabilidade", xlab = "Número de sucessos", main = "binom(n = 10, prob = 0.35)")
```



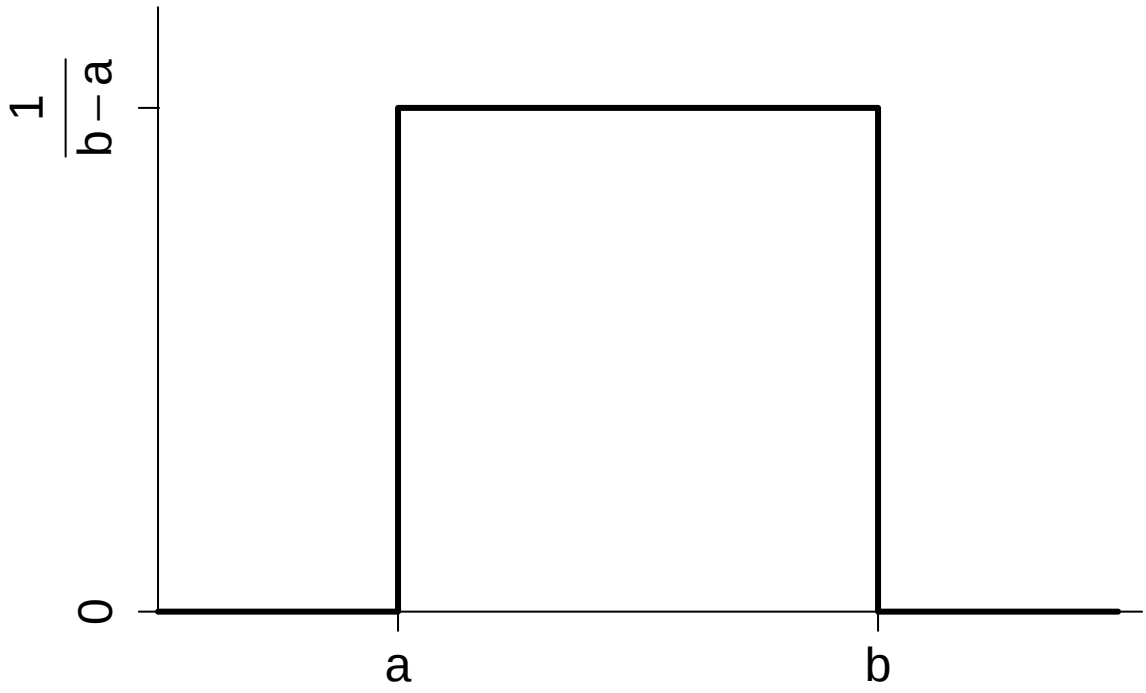
2.11 Distribuição de uma variável aleatória contínua

- A distribuição de uma variável aleatória contínua Y é caracterizada pela função densidade de probabilidade f_Y.



- A área sob o gráfico da função densidade de probabilidade entre a e b é igual a probabilidade de uma observação neste intervalo.
- $f_Y(y) \geq 0$ para todos os números reais y.
- A área sob o gráfico de f_Y é igual a 1.
- Por exemplo, a **distribuição uniforme** de a até b:

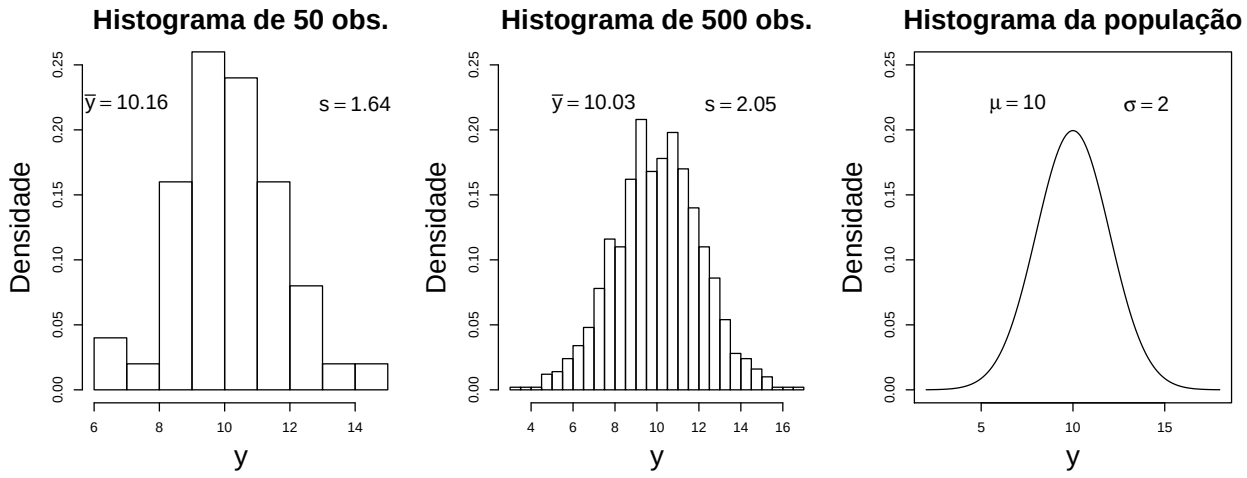
$$f_Y(y) = \begin{cases} \frac{1}{b-a} & a < y < b \\ 0 & \text{otherwise} \end{cases}$$



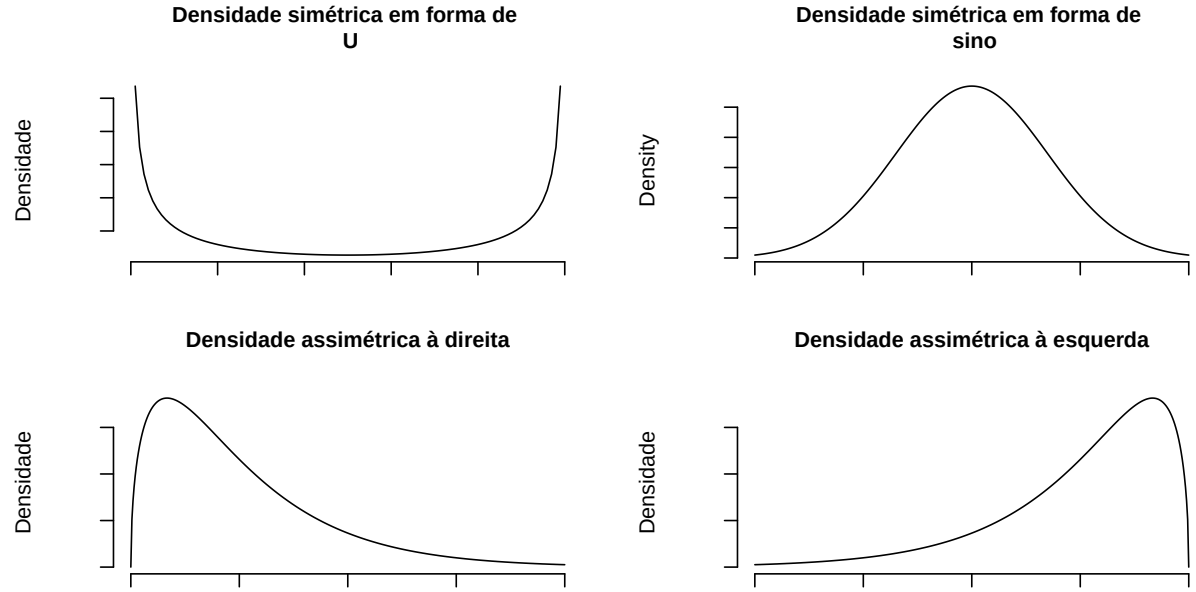
2.12 Função Densidade

2.12.1 Aumentando o número de observações

- Outra maneira de pensar sobre a densidade é em termos do histograma.
- Se desenharmos um histograma para uma amostra onde a área de cada caixa corresponde a frequência relativa de cada intervalo, a área total será 1.
- Quando o número de observações (tamanho da amostra) aumenta nós podemos fazer intervalos menores e obter um histograma mais suave.
- Com um número infinito de observações, nós poderíamos produzir uma curva suave, onde a área embaixo dela é 1. Uma função derivada dessa forma é o que nós chamamos de **função densidade de probabilidade**.



2.12.2 Formatos das densidades



2.13 Distribuição Normal

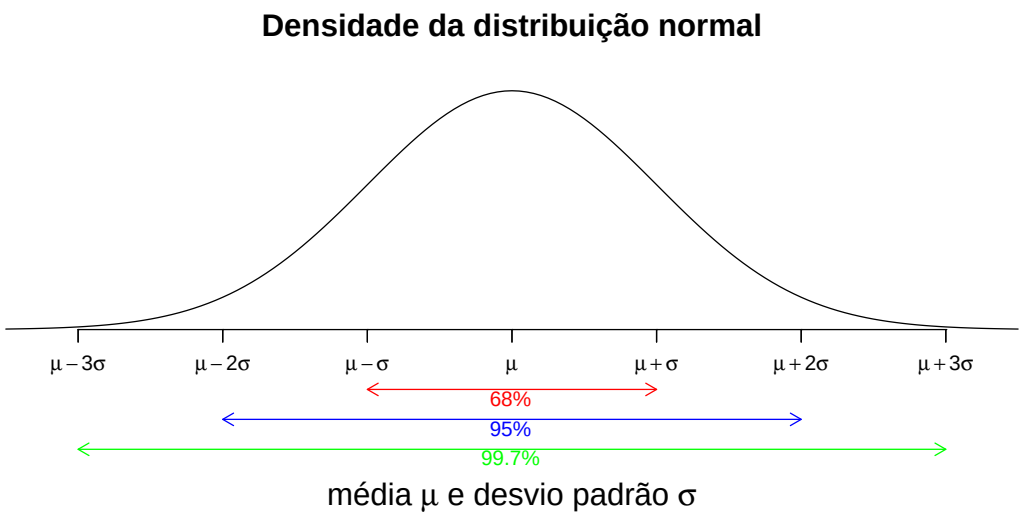
- A distribuição Normal é uma distribuição contínua determinada por 2 parâmetros:
  - μ: a **média** (valor esperado), que determina onde a distribuição é centrada.
  - σ^2 a **variância**, que determina a dispersão da distribuição em torno da média.
- A distribuição tem uma função densidade de probabilidade em forma de sino:

$$f_Y(y; \mu, \sigma^2) = \frac{1}{\sqrt{2p\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$$

- Quando uma variável aleatória Y segue uma distribuição normal com média μ e variância σ^2, então nós escrevemos que $Y \sim \mathtt{N}(\mu, \sigma^2)$.

- Chamamos de distribuição **Normal padrão** uma distribuição Normal com média 0 e variância 1 e notamos isto por $Z \sim \mathtt{N}(0,1)$.

2.13.1 Alcance da distribuição normal



Interpretação:

- $\approx 68\%$ da população está dentro de um desvio padrão da média.
- $\approx 95\%$ da população está dentro de 2 desvios padrão da média.
- $\approx 99.7\%$ da população está dentro de 3 desvios padrão da média.

2.13.2 Escore z

- Se $Y \sim \mathtt{N}(\mu, \sigma^2)$ então o escore-z correspondente é

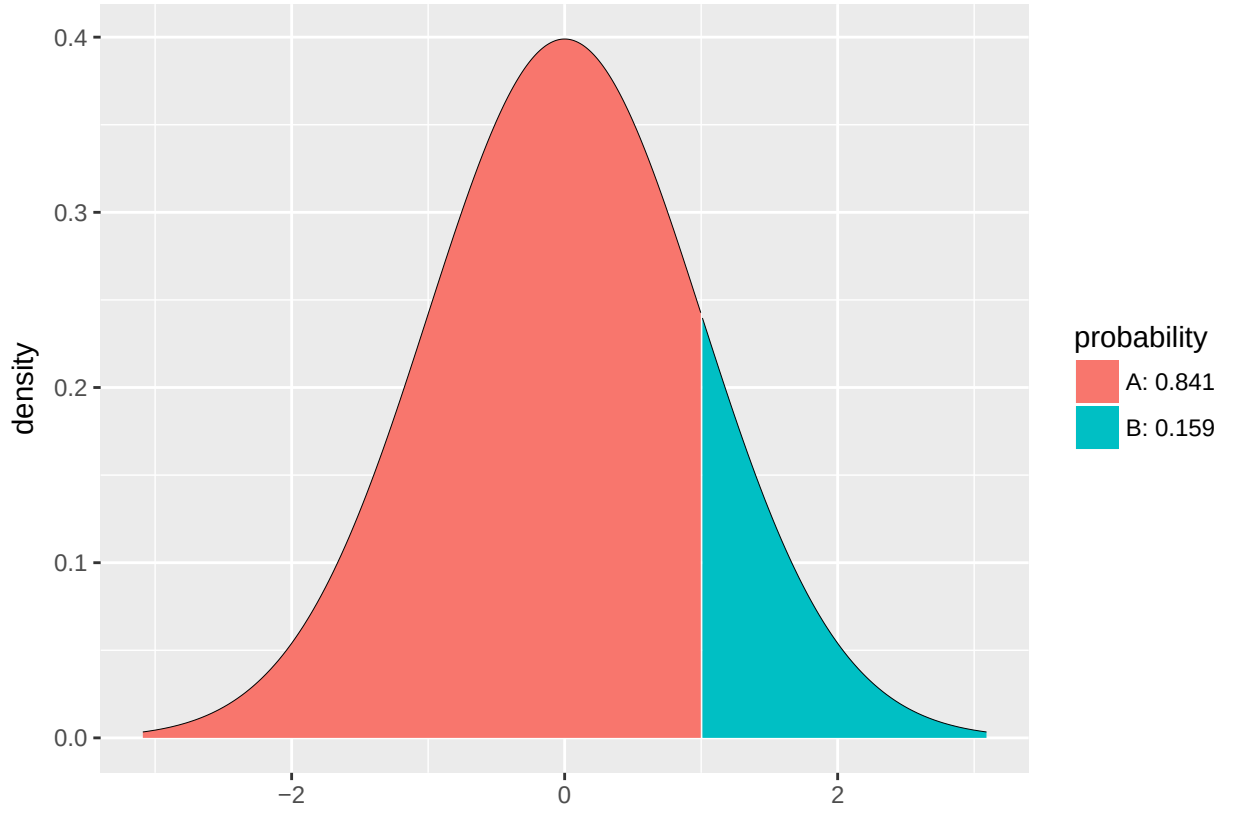
$$Z = \frac{Y - \mu}{\sigma} = \frac{\mathtt{observação} - \mathtt{média}}{\mathtt{desvio\ padrão}}$$

- I.e. Z representa o número de desvios padrão da observação em relação à média.
- $Z \sim \mathtt{N}(0,1)$, i.e. Z tem média zero e variância 1.
- Isto implica que
  - Z situa-se entre -1 e 1 com probabilidade de 0.6826
  - Z situa-se entre -2 e 2 com probabilidade de 0.9544
  - Z situa-se entre -3 e 3 com probabilidade de 0.9973
- Isto também implica que:
  - A probabilidade de Y estar entre $\mu - z\sigma$ e $\mu + z\sigma$ é igual a probabilidade de Z estar entre $-z$ e z.

2.13.3 Calculando probabilidades na distribuição normal padrão

- A função `pdist` produz a área à esquerda do valor z (quantil/percentil) que informamos (variável `q` na função), i.e. mostra a probabilidade de obter um valor menor do que z. O primeiro argumento de `pdist` denota a distribuição que estamos considerando.

```
#Para uma distribuição normal padrão, a probabilidade de obter um valor menor que 1 é:
left_prob <- pdist("norm", q = 1, mean = 0, sd = 1)
```




```
left_prob
```

```
## [1] 0.8413447
```

```
right_prob <- 1 - left_prob
right_prob
```

```
## [1] 0.1586553
```

- Para $z = 1$ nós temos uma probabilidade à direita de $p = 0.1587$, então a probabilidade de uma observação entre -1 e 1 é $1 - 2 \cdot 0.1587 = 0.6826 = 68.26\%$ devido a simetria.

2.13.4 Exemplo

A escala de inteligência de Stanford-Binet é calibrada para ser aproximadamente normal com média 100 e desvio padrão 16.

Qual é o 99-percentil dos escores de QI?

- O correpondente escore-z é $Z = \frac{IQ-100}{16}$, o que significa que $QI = 16Z + 100$.
- O 99-percentil dos escores-z é 2.326 (pode ser calculado usando `qdist`).
- Então, o 99-percentil dos escores de QI é:

$$QI = 16 \cdot 2.326 + 100 = 137.2.$$

- Então, esperamos que uma a cada cem pessoas tenha um QI superior a 137.

3 Distribuição da estatística amostral

3.1 Estimativas e sua variabilidade

Temos uma amostra $y_1, y_2, \ldots, y_n$.

- A média amostral $\bar{y}$ é a estimativa mais comum da média populacional μ.
- O desvio padrão amostral, s, é a estimativa mais comum do desvio padrão populacional σ. Note que há uma incerteza (de amostra para amostra) conectada a estas estatísticas e, portanto, estamos interessados em descrever a sua **distribuição**.

3.2 Distribuição da média amostral

- Temos uma amostra $y_1, y_2, \ldots, y_n$ de uma população com média μ e desvio padrão σ.
- A média amostral
$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n)$$
tem distribuição
  - com média μ,
  - e desvio padrão $\frac{\sigma}{\sqrt{n}}$ (também chamado de **erro padrão**), e
  - quando n cresce, a distribuição se aproxima de uma distribuição normal. Este resultado é provado usando o **teorema central do limite**.

3.2.1 Teorema Central do Limite

- Os pontos anteriores podem ser resumidos por
$$\bar{y} \sim \mathtt{N}\left(\mu, \frac{\sigma^2}{n}\right),$$
i.e. $\bar{y}$ tem distribuição aproximadamente normal com média μ e erro padrão $\frac{\sigma}{\sqrt{n}}$.
- Quando a amostra é suficientemente grande (para que a aproximação seja boa) isto nos permite fazer as seguintes observações:
  - Nós estamos 95% certos de que $\bar{y}$ está no intervalo de $\mu - 2\frac{\sigma}{\sqrt{n}}$ a $\mu + 2\frac{\sigma}{\sqrt{n}}$.
  - Estamos quase completamente certos de que $\bar{y}$ está no intervalo $\mu - 3\frac{\sigma}{\sqrt{n}}$ a $\mu + 3\frac{\sigma}{\sqrt{n}}$.

3.2.2 Ilustração do TCL

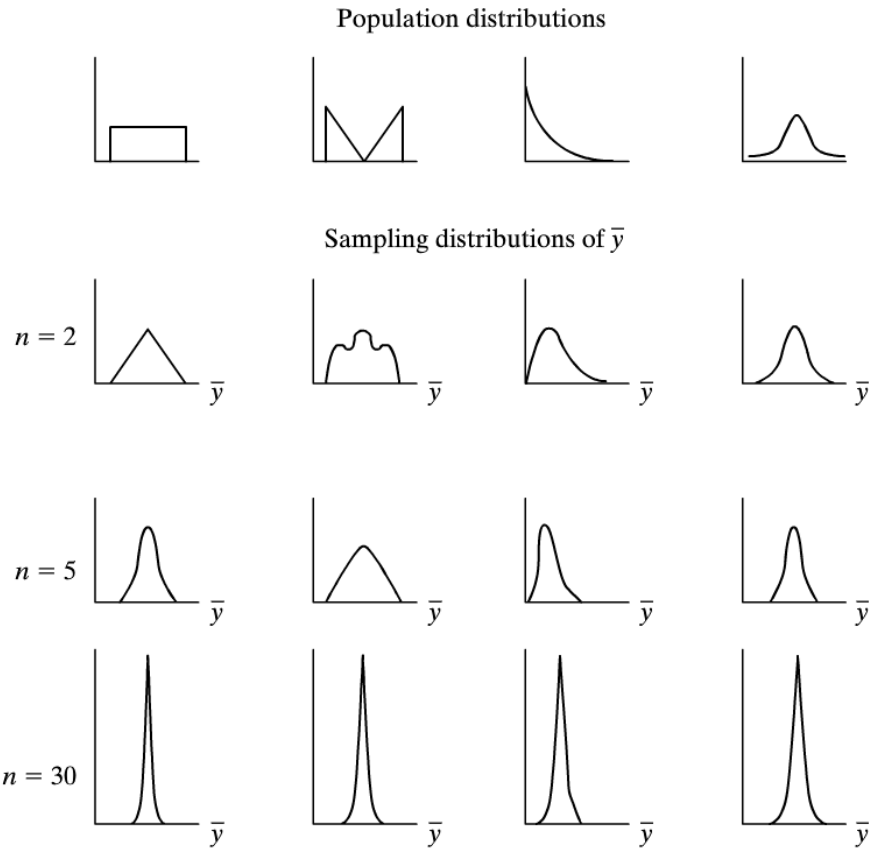


FIGURE 4.14: Four Different Population Distributions and the Corresponding Sampling Distributions of $\bar{y}$. As n increases, the sampling distributions get narrower and have more of a bell shape.

3.2.3 Exemplo

- Índice de Massa Corporal (IMC) de pessoas do norte da europa (2010) tem média $\mu = 25.8$ kg/m^2 e desvio padrão 4.8 kg/m^2.
- Uma amostra aleatória de $n = 100$ clientes de uma hamburgueria teve um IMC médio dado por $\bar{y} = 27.2$.
- Se comer hamburguer "não influencia" o IMC (e a amostra é representativa da população de interesse), então

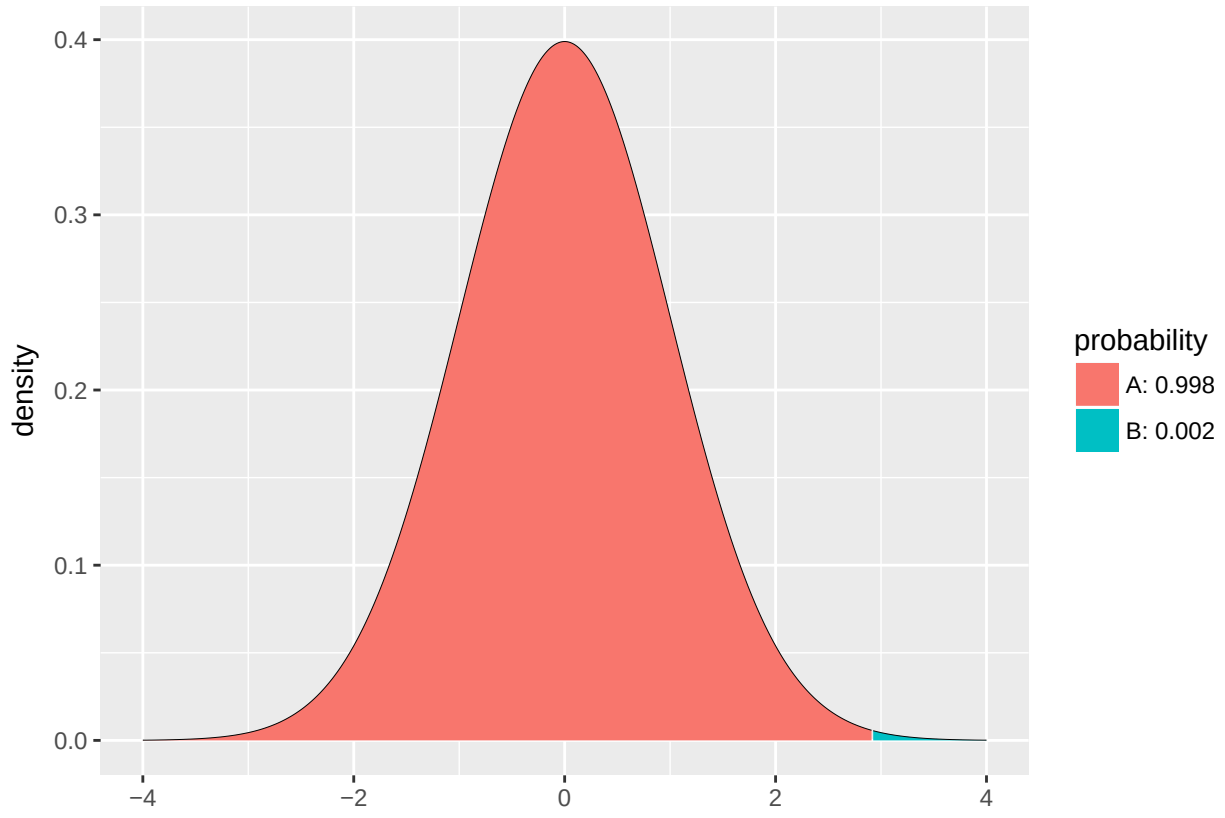
$$\bar{y} \approx \mathtt{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathtt{N}(25.8, 0.48^2).$$

- Para amostra o escore-z observado

$$z_{obs} = \frac{27.2 - 25.8}{0.48} = 2.92$$

- Lembrando que o escore-z é (aproximadamente) normal padrão, a probabilidade de obter uma pontuação mais alta que o escore-z é:

```
1 - pdist("norm", mean = 0, sd = 1, q = 2.92, xlim = c(-4, 4))
```




```
## [1] 0.001750157
```

- Assim, é altamente imporvável obter uma amostra aleatória com um escore-z tão alto. Há evidências de que os clientes da hamburgueira tem um IMC médio maior que a média populacional.