Revisão de Probabilidade e Estatística

Preparatória para a disciplina de Econometria I do PPGECO/UFSC

Aishameriane Schmidt

Última revisão: 26 de fevereiro de 2019

Por favor, envie observações, comentários e correções para aishameriane@gmail.com

Sumário

Introdução

Ao longo destas notas iremos estudar tópicos de probabilidade e de estatística que são pré-requisitos para a disciplina de Econometria I do mestrado/doutorado no PPGEco/UFSC. As aulas são mais focadas em aspectos da teoria, porém na medida do possível exemplos computacionais em R serão apresentados, em conjunto com as aulas de R que serão ministradas no contraturno. **Essas notas não são totalmente autorais, muitas das definições foram apenas copiadas de livros, então por favor não distribua elas.**

Alguns exercícios tem solução em formato digitalizado e as soluções serão disponibilizadas conforme demanda. Caso você tenha alguma dúvida ou exercício que queira discutir, me procure. É possível que exercícios estejam repetidos. Nem todos conteúdos e nem todos exercícios serão vistos em sala de aula. *Estas aulas não são oficiais e nem contam nota, mas servem para moldar seu caráter. =P*

Qual a diferença entre probabilidade e estatística?
A probabilidade é, para a estatística, aquilo que o cálculo representa para as engenharias: uma ferramenta. A probabilidade é considerada por muitas pessoas como uma sub-área da matemática. Apesar de estar relacionada com a estatística através de uma certa *noção de incerteza*, a probabilidade não se ocupa de dados, isto é, não há uma amostra, nem uma população e muito menos um problema de inferência. Neste sentido, a estatística é formada de métodos para descrição de amostras, populações e de inferência do primeiro grupo para o segundo. Já a probabilidade acaba lidando com noções mais abstratas e, na sua forma axiomática, não irá tratar de dados coletados.

Iremos começar as aulas revisando as noções matemáticas de conjuntos para então fazer a revisão de probabilidade, que será a maior parte das aulas. O programa das aulas pode ser acessado aqui: https://github.com/aishameriane/msc-economics/tree/master/revisao-prob.

1 Noções de teoria dos conjuntos

Esta seção foi retirada de Lima (1982).

Um **conjunto** (ou coleção) é formado por objetos (que são os seus **elementos**). Eles têm como relação básica a relação de **pertinência**, que possibilita relacionar elementos com conjuntos. Quando x (objeto) é um elemento do conjunto A, dizemos que "x *pertence a* A" e denotamos por $x \in A$. Caso contrário, diremos que "x *não pertence a* A" e denotaremos por $x \notin A$. Um conjunto fica bem especificado (definido) quando há uma regra clara que permita avaliar se um elemento arbitrário pertence ou não ao conjunto.

Uma vez que existem conjuntos de tamanho muito grande, se torna difícil sempre que nos referirmos a um conjunto listarmos todos seus elementos. Sendo assim, podemos utilizar uma notação mais compacta que, através de uma regra de pertinência (ou propriedade), deixa o conjunto totalmente especificado. Por exemplo, podemos ter uma situação onde a propriedade P define totalmente o conjunto A: se um objeto x atende P, então $x \in A$, caso contrário, $x \notin A$. Então, podemos escrever:

$$A = \{x : x \text{ tem a propriedade } P\}$$

Os dois pontos na expressão acima fazem o papel da expressão "*tal que*". Outra notação usual, ao invés dos dois pontos, é a barra vertical "|" ou ainda o ponto e vírgula - notação utilizada por (Lima, 1982). Quando estamos falando de um subconjunto $B \subset A$ (B é subconjunto de A ou, equivalentemente, B está contido em A), podemos escrever:

$$B = \{x \in A : x \text{ tem a propriedade } P\}.$$

A expressão acima significa que o conjunto B são os elementos x do conjunto A que satisfazem a propriedade P. Por exemplo, se queremos nos referir aos números reais maiores que 10, podemos definir:

$$A = \{x \in \mathbb{R} : x > 10\},$$

e neste caso, A é um subconjunto do conjunto dos números reais, $\mathbb{R}$, isto é, $A \subset \mathbb{R}$.

Se A e B forem conjuntos, podemos compará-los através da relação de "*inclusão*" ($\subset$). Dizemos que B é subconjunto de A se todo elemento de B também[1] é elemento de A e denotamos por $B \subset A$. Neste caso, também são utilizados os termos: B é *parte* de A, B está *incluído* em A ou ainda B está *contido* em A. Em alguns livros encontramos as notações $B \subset A$ para indicar que B está contido em A (mas não é igual), $B \subseteq A$ para indicar que B está contido e pode ser igual a A ou ainda $B \subsetneq A$ para indicar B está contido mas não é igual a A. Nestas notas, iremos utilizar $B \subset A$ como "B está contido em A e eles podem ser iguais".

Quando não há elemento de A que satisfaça P, o conjunto B não tem nenhum elemento e é denominado *conjunto vazio* ($\emptyset$). Definimos $\emptyset$ da seguinte forma:

$$\forall x, x \notin \emptyset,$$

[1] Observe que não necessariamente a recíproca é verdadeira, não é necessário que todo elemento de A seja elemento de B. Quando a recíproca é de fato verdadeira, os dois conjuntos são iguais.

em que lê-se: qualquer que seja x, x não percence ao vazio. Por exemplo, $\{x \in \mathbb{N} | 1 < x < 2\} = \emptyset$. Perceba que uma implicação importante da relação de pertinência é que qualquer que seja o conjunto A, temos $\emptyset \subset A$. Isso é verdade pois o contrário implicaria que existiria um elemento do conjunto vazio que não pertence ao conjunto A. No entanto, o conjunto vazio não contém elementos, o que implica que ele deve ser subconjunto de todos demais conjuntos.

Uma coisa importante a se notar é que existe diferença entre $\emptyset$ e $\{\emptyset\}$. O primeiro é o conjunto vazio, o segundo é um conjunto cujo único elemento é o conjunto vazio. Para entender melhor a diferença, considere o seguinte exercício:

Exercício 1. Analise se são verdadeiras ou falsas as seguintes sentenças:

1. $\emptyset \in \emptyset$.
2. $\emptyset \in \{\emptyset\}$.
3. $\{\emptyset\} \in \emptyset$.
4. $\emptyset \subset \emptyset$.
5. $\emptyset \subset \{\emptyset\}$.
6. $\{\emptyset\} \subset \emptyset$.

Dado um conjunto A qulquer, podemos definir o conjunto formado pelas partes de A, que é denotado por $\mathscr{P}(A)$. Em outras palavras, se $B \subset A$, então $B \in \mathscr{P}(A)$. Note que este conjunto nunca é vazio pois teremos ao menos $\emptyset \in \mathscr{P}(A)$ e $A \in \mathscr{P}(A)$. Por exemplo:

$$A = \{a, b, c\} \Rightarrow \mathscr{P}(A) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

1.1 Eventos

Em probabilidade, utilizamos muito a teoria de conjuntos principalmente ao falar de probabilidade de eventos. Vamos introduzir algumas definições básicas e ver alguns exercícios (ainda sem probabilidade). Algumas definições desta seção foram retiradas de Stern and Izbicki (2016).

O modelo probabilístico tem dois "ingredientes" básicos: o *espaço amostral* (S) e a *lei de probabilidade*, para falar deles, precisamos entender a noção de *experimento*.

Por experimento podemos entender qualquer atividade cujo resultado final é desconhecido (porém sabemos quais são as possibilidades resultantes dele, apenas não sabemos a priori o que irá sair). Por exemplo, para o lançamento de uma moeda, podemos ter o resultado cara ou coroa e portanto $S = \{(cara), (coroa)\}$. Não sabemos, antes de lançar a moeda, mas sabemos que será[2] $cara$ ou $coroa$.

Definição 1.1.1. Espaço Amostral

O conjunto de todos os possíveis resultados de um experimento particular é chamado de *espaço amostral* e é denotado por Ω (lê-se *ômega*). Este conjunto pode ser: enumerável, finito ou infinito, se houver uma bijeção $f : \Omega \to \mathbb{N}$ ou ainda, pode ser não enumerável (por exemplo, no caso de $\Omega = \mathbb{R}$).

[2]Estamos abstraindo aqui a possibilidade de outros eventos, como por exemplo, um meteoro cair na terra ou um albatroz pegar a moeda antes de observarmos seu resultado.

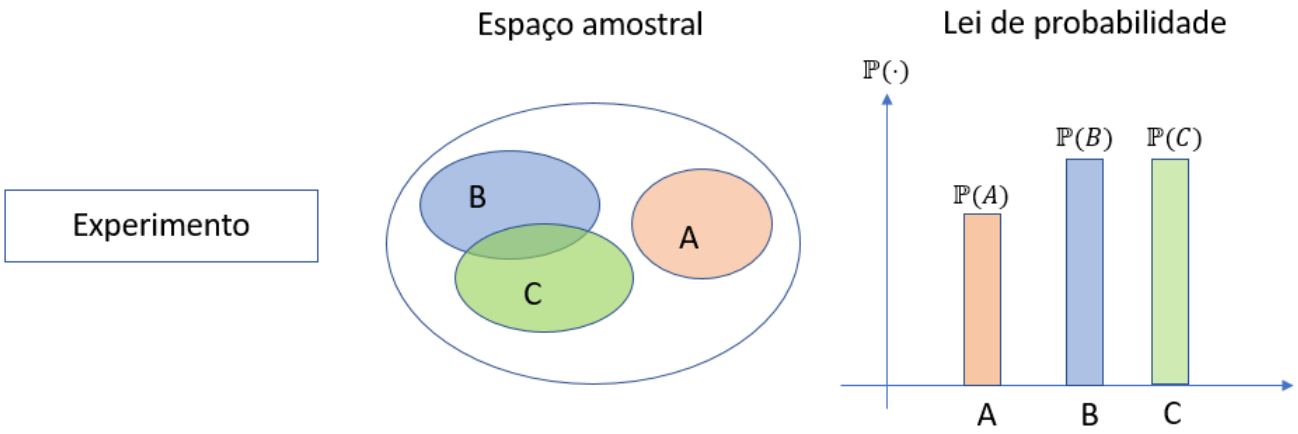


Figura 1: Representação de experimento, espaço amostral e lei de probabilidade.

Exemplo 1.1.2. Espaço amostral do lançamento de uma moeda Imagine que você faz o experimento de lançar uma moeda. Considere que *K* significa que o resultado foi cara e *C* significa que o resultado foi coroa. Então, $\Omega = \{K, C\}$ é o espaço amostral do experimento.

Exemplo 1.1.3. Espaço amostral do tempo até uma lâmpada queimar

Considere agora o seguinte experimento: você observa uma lâmpada e está interessado no tempo, em minutos, até a lâmpada queimar[3]. Então, $\Omega = [0, +\infty)$.

Definição 1.1.4. Evento Um *evento* é qualquer coleção de possíveis resultados de um experimento, isto é, qualquer subconjunto de Ω (incluindo o próprio Ω).

Definição 1.1.5. Relação entre eventos

$$A \subset B \iff x \in A \Rightarrow x \in B \tag{1}$$

$$A = B \iff A \subset B \quad e \quad B \subset A \tag{2}$$

A equação 1 significa, em palavras, que *o conjunto A está contido no conjunto B se e somente se qualquer elemento de A está pertence a B*. Já a equação 2 fala da igualdade de dois conjuntos e estabelece que *dois conjuntos A e B serão iguais se e somente se A está contido B e B está contido em A*. Se usarmos 1 em 2, chegamos à conclusão que A e B serão iguais se e somente se qualquer elemento de A está em B e qualquer elemento de B está em A.

Podemos definir ainda três operações de conjuntos: união, interseção e complementar.

Definição 1.1.6. União de dois eventos A união de dois eventos A e B, representada por $A \cup B$, é o conjunto de elementos que percentem a A, B ou ambos:

$$A \cup B = \{x : x \in A \quad ou \quad x \in B\} \tag{3}$$

De maneira mais geral, seja $(A_n)_{n\in\mathbb{N}}$ uma sequência de conjuntos. $x \in \Omega$ é um elemento da união de $(A_n)_{n\in\mathbb{N}}$, denotada por $\cup_{n\in\mathbb{N}} A_n$, se e somente se existe um $n \in \mathbb{N}$ tal que $x \in A_n$. Isto é, $\cup_{n\in\mathbb{N}} A_n = \{x \in \Omega : \textbf{existe } n \in \mathbb{N} \text{ tal que } x \in A_n\}$

Definição 1.1.7. Interseção de dois eventos A interseção de dois eventos A e B, representada por $A \cap B$, é o conjunto de elementos que percentem a A e, ao mesmo tempo, B:

[3] O intervalo será fechado em zero se considerarmos que é possível que a lâmpada já esteja queimada quando iniciamos o teste.

$$A \cap B = \{x : x \in A \quad e \quad x \in B\} \tag{4}$$

Dizemos ainda que A e B são eventos (ou conjuntos) disjuntos se, e somente se, $A \cap B = \emptyset$.

De maneira mais geral, seja $(A_n)_{n\in\mathbb{N}}$ uma sequência de conjuntos. $x \in \Omega$ é um elemento da interseção de $(A_n)_{n\in\mathbb{N}}$, denotada por $\cap_{n\in\mathbb{N}} A_n$, se e somente se para todo $n \in \mathbb{N}$, $x \in A_n$. Isto é, $\cap_{n\in\mathbb{N}} A_n = \{x \in \Omega : \textbf{para todo } n \in \mathbb{N}, x \in A_n\}$.

Definição 1.1.8. Complementar de um evento Seja A um conjunto. x é um elemento de A^c se e somente se $x \notin A$. Isto é, o complemento de A é definido formalmente como $A^c = \{x \in \Omega : x \notin A\}$.

Exemplo 1.1.9. Complementar do espaço amostral
O complementar do espaço amostral (Ω) é o conjunto vazio, $\emptyset$, pois:

- Como o conjunto $\emptyset$ não possui elementos, $\forall \omega \in \Omega$ temos que $\omega \notin \emptyset$;
- Uma vez que $\emptyset$ não possui elementos, não há elemento de $\emptyset$ que pertença a Ω (dizemos que isso ocorre por *vacuidade*).

O exemplo pode parecer confuso, pois também por vacuidade temos que $\emptyset \subset \Omega$. Para mais detalhes, aqui tem uma explicação boa: http://mathcentral.uregina.ca/qq/database/qq.09.06/narayana1.html. Em geral, o complementar não será subconjunto do conjunto original.

Teorema 1.1.10 (Leis de De Morgan)**.** *Seja $(A_n)_{n\in\mathbb{N}}$ uma sequência de subconjuntos de Ω. Então, para todo $n \in \mathbb{N}$,*

- $(\cup_{i=1}^n A_i)^c = \cap_{i=1}^n A_i^c$
- $(\cap_{i=1}^n A_i)^c = \cup_{i=1}^n A_i^c$

Além disso,

- $(\cup_{i\in\mathbb{N}} A_i)^c = \cap_{i\in\mathbb{N}} A_i^c$
- $(\cap_{i\in\mathbb{N}} A_i)^c = \cup_{i\in\mathbb{N}} A_i^c$

Definição 1.1.11 (Partição)**.** Seja $(A_n)_{n\in\mathbb{N}}$ uma sequência de conjuntos. Dizemos que $(A_n)_{n\in\mathbb{N}}$ particiona Ω se:

- para todo $i, j \in \mathbb{N}$ tal que $i \neq j$, A_i e A_j são disjuntos.
- $\cup_{n\in\mathbb{N}} A_n = \Omega$.

O último conceito que iremos ver antes de fazer alguns exercícios é o de álgebra e de σ-álgebra (lê-se *"sigma-algebra"*).

Definição 1.1.12. Álgebra[4]
Seja $\mathscr{F}$ uma família de conjuntos (eventos) de Ω. Diremos que $\mathscr{F}$ é uma *álgebra sobre* Ω se satisfaz as seguintes propriedades:

1. $\Omega \in \mathscr{F}$;
2. $A \in \mathscr{F} \Rightarrow A^c \in \mathscr{F}$ - fechada por complementar
3. $A, B \in \mathscr{F} \Rightarrow A \cup B \in \mathscr{F}$ - fechada por uniões finitas de seus elementos.

[4] Adaptado de https://www.ime.usp.br/~tassio/1-2012/probabilidade-1/notas-de-aula-probabilidade.pdf

Em geral, nos cursos de probabilidade regulares, não se vê a definição de σ-álgebra. Nós não iremos nos aprofundar muito nessas álgebras também, porém essa definição mais tarde vai nos ajudar a deixar algumas definições de probabilidade um pouco mais precisas. A σ-álgebra é um pouco mais geral que a álgebra, pois ela generaliza a propriedade 3 para o caso de uniões infinitas (arbitrárias).

Definição 1.1.13. **σ-algebra**[5]
Seja $\mathscr{F}$ uma família de conjuntos (eventos) de Ω. Diremos que $\mathscr{F}$ é uma *álgebra sobre* Ω se satisfaz as seguintes propriedades:

1. $\Omega \in \mathscr{F}$;
2. $A \in \mathscr{F} \Rightarrow A^c \in \mathscr{F}$;
3. $A_1, A_2, \ldots \in \mathscr{F} \Rightarrow \cup_{i=1}^{\infty} A_i \in \mathscr{F}$

1.2 Exercícios

Exercício 2. Prove as seguintes propriedades:

a.	Comutatividade	$A \cup B = B \cup A$
		$A \cap B = B \cap A$
b.	Associatividade	$A \cup (B \cup C) = (A \cup B) \cup C$
		$A \cap (B \cap C) = (A \cap B) \cap C$
c.	Leis distributivas	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
		$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
d.	Leis de De Morgan	$(A \cup B)^c = A^c \cap B^c$
		$(A \cap B)^c = A^c \cup B^c$

Dica: Para provar que dois conjuntos são iguais, é necessário utilizar a definição 2. É possível que até aqui você tenha trabalhado nessas provas utilizando os diagramas de Venn, porém eles não são considerados como uma prova formal. Por exemplo, vamos fazer a prova da lei distributiva da interseção $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Queremos mostrar duas coisas:
(1) $A \cap (B \cup C) \subset (A \cap B) \cup (A \cap C)$
(2) $A \cap (B \cup C) \supset (A \cap B) \cup (A \cap C)$

Observe que, utilizando notação de conjuntos,

$$A \cap (B \cup C) = \{x \in \Omega : x \in A \text{ e } x \in (B \cup C)\}$$
$$(A \cap B) \cup (A \cap C) = \{x \in \Omega : x \in (A \cap B) \text{ ou } x \in (A \cap C)\}$$

Vamos mostrar (1). Tome $x \in A \cap (B \cup C)$. Pela definição de interseção, temos que $x \in A$ (*) e, ao mesmo tempo, $x \in (B \cup C)$. Mas isso implica, pela propriedade de união, que $x \in B$ ou $x \in C$. Juntando com (*), temos que $x \in (A \cap B)$ ou $x \in (A \cap C)$. Logo, $x \in (A \cap B) \cup (A \cap C)$.

Agora, vamos mostrar (2).

Seja $x \in (A \cap B) \cup (A \cap C)$. Então, $x \in (A \cap B)$ ou $x \in (A \cap C)$. Abrimos então em dois casos:

Caso a: $x \in (A \cap B)$. Isso significa que $x \in A$ e, ao mesmo tempo, $x \in B$. Logo, x pertence à qualquer união de B, em

[5] Adaptado de https://www.ime.usp.br/~tassio/1-2012/probabilidade-1/notas-de-aula-probabilidade.pdf

particular, $x \in (B \cup C)$. Logo, $x \in A \cap (B \cup C)$.
Caso b: $x \in (A \cap C)$. Pelo mesmo argumento anterior, isso significa que $x \in A$ e, ao mesmo tempo, $x \in C$. Logo, x pertence à qualquer união de C, em particular, $x \in (C \cup B)$. Se você provou a propriedade de comutatividade da letra a, pode concluir que $x \in A \cap (B \cup C)$.

Vamos resolver o segundo item da parte (d) $(A \cap B)^c = A^c \cup B^c$.

- **Parte 1:** Queremos mostrar que $(A \cap B)^c \subset A^c \cup B^c$, isto é, mostrar que todo elemento de $(A \cap B)^c$ pertence a $A^c \cup B^c$.
  Tome $a \in (A \cap B)^c$. Então, $a \notin (A \cap B)^c$, pela definição de complementar, de forma que temos três casos possíveis.

Caso 1: $a \in A$ e $a \notin B$. Neste caso, $a \in B^c$ e portanto está em qualquer união de B^c, em particular, está em $A^c \cup B^c$.

Caso 2: $a \notin A$ e $a \in B$. Neste caso, $a \in A^c$ e portanto está em qualquer união de A^c, em particular, está em $A^c \cup B^c$.

Caso 3: $a \notin A$ e $a \notin B$. Neste caso, $a \in A^c$ e $a \in B^c$ e portanto está em $A^c \cup B^c$.

- **Parte 2:** Queremos mostrar que $(A \cap B)^c \supset A^c \cup B^c$, isto é, que todo elemento que está em $A^c \cup B^c$ pertence a $(A \cap B)^c$.
  Tome $a \in A^c \cup B^c$ (apesar de estar usando a mesma letra a, ele não é o mesmo elemento tomado na parte 1!). Então, por se tratar de uma união, novamente temos 3 casos possíveis.

Caso 1: $a \in A^c$. Então, $a \notin A \Rightarrow a \notin A \cap B$, isto é, o fato de a não pertencer ao conjunto A implica que ele não irá pertencer a qualquer interseção de A, pela própria definição de interseção. Podemos concluir, como a não está na interseção, que ele está no complementar, isto é, $a \in (A \cap B)^c$.

Caso 2: O caso 2 é análogo ao caso 1, supondo que $a \in B^c$.

Caso 3: O caso 3 decorre dos casos 1 e 2, isto é, suponha que $a \in A^c$ e $a \in B^c$.

Os outros exercícios são resolvidos de forma similar. Qualquer problema procure a Aishameriane mais próxima.

Exercício 3.

$$\bigcap_{n=1}^{\infty} \left(0, 1 + \frac{1}{n}\right) \triangle \left[\frac{1}{n}, 2 + \frac{1}{n}\right)$$

onde $A \triangle B = (A \cup B) \cap (A^c \cup B^c)$.

Exercício 4. Seja $\Omega = \{a, b, c, d\}$. Será que $\mathscr{F} = \{\emptyset, \{a\}, \{b, c\}, \Omega\}$ é uma σ-álgebra de Ω?

Exercício 5. Prove as leis de De Morgan do caso geral (teorema 1.1.10).

Exercício 6. Mostre que $A = (A \cap B) \cup (A \cap B^c)$. Este exercício é útil para conseguirmos escrever um conjunto como a união de conjuntos disjuntos.

Exercício 7. Considere A e B dois subconjuntos de Ω. A *diferença simétrica* entre A e B é o conjunto de todos elementos que estão em A ou em B mas que não estão em ambos. Escreva a diferença simétrica formalmente utilizando as operações de união, intersecção e/ou complementar. Mostre que a diferença simétrica entre A e B é igual à diferença simétrica entre A^c e B^c.

Exercício 8. Seja $A = \{2, 4, 6\}$, $B = \{3, 4\}$ com $\Omega = \{1, 2, 3, 4, 5, 6\}$. Encontre:

a. $A \cup B$

b. $A \cap B$

c. $A \cup B^c$

d. $A \cup A^c$ e $A \cap A^c$

e. $B \cup B^c$ e $B \cap B^c$

f. $(A \cup B)^c$

g. $(A \cap B)^c$

h. $A^c \cap B^c$

i. $A \cup \Omega$

j. $A^c \cup \Omega$

k. $B \cap \Omega$

l. $B^c \cap \Omega$

Exercício 9. Considere que você está analisando para um dia o evento *"A variação do dólar foi negativa em relação ao início do dia"* (N) e *"A variação do dólar foi positiva em relação ao início do dia"* (P). Estamos interessados na variação do dólar em dois dias consecutivos, isto é, se houveram duas variações negativas, duas positivas, etc.

a. Como você escreveria formalmente (notação de conjuntos) o espaço amostral Ω?

b. Como você escreveria formalmente *"Houveram dois dias consecutivos de variação negativa no dólar"*? Chame este evento de A.

c. Como você escreveria formalmente o evento *"Houve pelo menos um dia com variação negativa no dólar"*? Chame este evento de B.

d. Avalie se $A \subset B$.

e. Descreva com suas palavras quem é o conjunto B^c. Escreva-o formalmente.

f. A e B^c são disjuntos? Justifique.

Exercício 10. De um grupo de 25 alunos:

- 14 irão comprar itens para fazer churrasco no final de semana;
- 12 irão comprar um chocolate para o professor de Estatística;
- 5 irão fazer o churrasco e comprar o chocolate para o professor.

Quantos alunos não irão nem comprar os itens de churrasco nem o chocolate?

Exercício 11. Escreva o espaço amostral dos seguintes experimentos:

a. Duas moedas são lançadas simultâneamente e observa-se a sequência de caras e coroas obtida;

b. Um dado é lançado e observa-se a face virada para cima;

c. Duas cartas de um baralho são retiradas e observa-se a sequência dos naipes;

d. Dois dados são lançados e observa-se as faces viradas para cima;

e. Dois dados são lançados e observa-se a soma das duas faces;

f. Uma moeda e um dado são lançados e observa-se a face da moeda e do dado viradas para cima;

g. Três times A,B e C participam de um campeonato de *curling*. Inicialmente o time A joga contra o time B e o vencedor joga com o time C e assim por diante. O campeonato finaliza quando um time ganha duas vezes em seguida ou quando são disputadas, ao todo, quatro partidas. Você está interessado no espaço amostral: resultados possíveis das partidas do torneio.

Exercício 12. Seja o experimento da letra d. do exercício anterior. Defina os seguintes eventos:

- $A =$ a soma das faces dos dois dados é igual a 5;
- $B =$ a face do primeiro dado é menor ou igual a 2.

Determine os eventos A e B em termos de Ω encontrado anteriormente (isto é, explicite quais resultados de Ω pertencem a A e B) e determine os seguintes eventos:

- $A \cup B$;
- $A \cap B$;
- A^c.

2 Probabilidade

Nesta seção iremos introduzir os principais conceitos de probabilidade. Apesar do uso de algumas definições e teoremas, ainda não estamos fazendo uso das definições mais formais, como por exemplo, σ-álgebra. É comum encontrar nos livros específicos de probabilidade esse conteúdo de forma mais aprofundada, porém não é comum que se encontre dessa forma nos textos de estatística econômica. Para quem desejar se aprofundar no assunto de maneira mais rigorosa, recomenda-se a leitura do capítulo 1 de James (2010).

2.1 Definições básicas

Existem três definições de probabilidade que são comumente apresentadas nos livros texto: a probabilidade por frequência relativa, a definição subjetiva e a definição axiomática. As seguintes definições foram retiradas de Mittelhammer (2013):

Definição 2.1.1. Probabilidade por frequência relativa
Seja n o número de vezes que um experimento é repetido sob condições idênticas. Seja A o evento no espaço amostral Ω e defina n_A o número de vezes que o evento A ocorreu. Então, a probabilidade do evento A é igual a:

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{n_A}{n}$$

Apesar de ser uma definição que tem bases em experimentos *reais* e por isso ser mais familiar, ela não permite desenvolver formalmente muitos resultados, uma vez que não há garantias de que o limite irá convergir em todos os casos. Além disso, não é nem possível observar infinitas repetições para termos certeza deste limite.

Definição 2.1.2. Probabilidade subjetivista
Um número real, $\mathbb{P}(A)$, contido no intervalo $[0, 1]$ e escolhido para expressar o grau de crença pessoal na possibilidade de ocorrência de um evento A, sendo que 1 representa absoluta certeza.

A definição subjetiva está relacionada com a estatística Bayesiana, que incorpora essa subjetividade de maneira a não deixar toda a informação sobre os parâmetros apenas para a amostra. Voltaremos nessa definição na última aula.

A definição axiomática de probabilidade foi desenvolvida por Kolmogorov em 1956 e se baseia em *axiomas*. Axiomas são afirmações que consideramos como verdades (sem a necessidade de prová-los) e a partir disso os teoremas são provados. Por exemplo, temos na geometria Euclidiana, o axioma que diz que *dois pontos podem ser ligados por uma única reta*.

Definição 2.1.3. Probabilidade axiomática
Seja $\mathbb{P}$ uma função que tem como domínio o espaço amostral Ω e contradomínio em $\mathbb{R}$, isto é, temos $\mathbb{P} : \gamma \to \mathbb{R}$, onde γ é o conjunto de todos eventos em Ω, também denominado espaço de eventos. Diremos que $\mathbb{P} : \gamma \to \mathbb{R}$ é uma função de probabilidade (ou medida de probabilidade) se atender aos seguintes axiomas:

1. $\mathbb{P}(A) \geq 0, \ \forall \ A \subset \Omega$ (*Axioma da não-negatividade*);
2. $\mathbb{P}(\Omega) = 1$ (*Axioma da normalização*);
3. Se $(A_n)_{n\in\mathbb{N}}$ é uma sequência de conjuntos disjuntos em Ω, $\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_i)$. (*Axioma da aditividade enumerável*).

Iremos diferenciar a *função de probabilidade* da *probabilidade*, sendo que esta última é a imagem de um evento através da função $\mathbb{P}$.

Da definição axiomática decorrem imediatamente cinco lemas, que são comumente utilizados (e quase nunca demonstrados).

Lema 2.1.4. ***A probabilidade de nada ocorrer é zero***
$\mathbb{P}(\emptyset) = 0$

Demonstração. Tome $(\{A_n\})_{n\in\mathbb{N}}$ tal que $A_1 = \Omega$ e $A_n = \emptyset \ \forall \ n > 1$. Note que $\cap_{n\in\mathbb{N}} A_n = \emptyset$, portanto:

$$\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$$

Podemos então escrever:

$$\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \mathbb{P}(\Omega) + \sum_{n>1} \mathbb{P}(\emptyset) \tag{5}$$

Por outro lado, sabemos que $\cup_{n\in\mathbb{N}} A_n = \Omega \cup \emptyset \cup \emptyset \cup \ldots$, de forma que $\mathbb{P}(\cup_{n\in\mathbb{N}} A_n) = \mathbb{P}(\Omega)$. Assim, juntando com 5,

$$\begin{aligned}
\mathbb{P}(\Omega) &= \mathbb{P}(\Omega) + \sum_{n>1} \mathbb{P}(\emptyset) \\
\mathbb{P}(\Omega) - \mathbb{P}(\Omega) &= \sum_{n>1} \mathbb{P}(\emptyset) \\
0 = \sum_{n>1} \mathbb{P}(\emptyset) &= \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(\emptyset)
\end{aligned}$$

Note que $\sum_{k=1}^{n} P(\emptyset)$ é não decrescente pois a probabilidade é sempre não negativa. Como o limite é igual a zero e as somas parciais formam uma sequência não decrescente, as somas finitas são zero. Logo, $\mathbb{P}(\emptyset) = 0$. □

Lema 2.1.5. ***A probabilidade da união é a soma das probabilidades se os eventos forem disjuntos***
Se $A_1, A_2, \ldots, A_n$ são disjuntos, então $\mathbb{P}(A_1 \cup A_2 \ldots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i)$.

Demonstração. Tome $\{B_n\}_{n\in\mathbb{N}}$ tal que $B_i = A_i \; \forall \; i \in \{1, \ldots, n\}$ e $B_i = \emptyset$ para $i > n$. Por construção, os B_i's são disjuntos e portanto:

$$\begin{aligned}
\mathbb{P}(\cup_{i=1}^{n} A_i) &= \mathbb{P}(\cup_{i\in\mathbb{N}} B_i) \\
&= \sum_{i\in\mathbb{N}} \mathbb{P}(B_i) \\
&= \sum_{i=1}^{n} \mathbb{P}(B_i) + \sum_{k=n+1}^{\infty} \mathbb{P}(B_k) \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i) + \sum_{k=n+1}^{\infty} \mathbb{P}(\emptyset) \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i) + 0 \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i)
\end{aligned}$$

Juntando as duas extremidades, segue o resultado desejado. □

Lema 2.1.6. ***Probabilidade do complementar***
Para todo evento A, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

Demonstração. Sabemos que A e A^c são disjuntos pois $A \cap A^c = \emptyset$. Por outro lado, temos que $A \cup A^c = \Omega$. Então:

$$\mathbb{P}(\Omega) = 1 = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$$

Segue que $1 = \mathbb{P}(A) + \mathbb{P}(A^c) \Rightarrow \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$. □

Lema 2.1.7. ***Probabilidade da união***
Para todos os eventos A e B, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Demonstração. Note que podemos escrever o conjunto A como a união de dois conjuntos disjuntos:

$$A = (B^c \cap A) \cup (B \cap A)$$

Então:

$$\mathbb{P}(A) = \mathbb{P}[(B^c \cap A) \cup (B \cap A)] = \mathbb{P}(B^c \cap A) + \mathbb{P}(B \cap A)$$

De forma que:

$$\mathbb{P}(B^c \cap A) = \mathbb{P}(A) - \mathbb{P}(B \cap A) \tag{6}$$

Note que B é disjunto de $B^c \cap A$ e que $B \cup (B^c \cap A) = (B \cup B^c) \cap (B \cup A) = \Omega \cap (B \cup A) = (B \cup A)$, portanto,

$$\mathbb{P}(B \cup A) = \mathbb{P}(B) + \mathbb{P}(B^c \cap A) \tag{7}$$

Juntando 6 e 7, temos:

$$\mathbb{P}(B \cup A) = \mathbb{P}(B) + \mathbb{P}(B^c \cap A) = \mathbb{P}(B) + \mathbb{P}(A) - \mathbb{P}(A \cap B)$$

□

Lema 2.1.8. ***Probabilidade de subconjuntos***
Se $A \subset B$, então $\mathbb{P}(B) \geq \mathbb{P}(A)$.

Demonstração. Já vimos que $B = (A^c \cap B) \cup (A \cap B)$, que são disjuntos. Então:

$$\mathbb{P}(B) = \mathbb{P}(A^c \cap B) + \mathbb{P}(A \cap B) \tag{8}$$

Mas $A = (A \cap B^c) \cup (A \cap B) \Rightarrow \mathbb{P}(A) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) \Rightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A \cap B^c)$.

Juntando com 8, temos que $\mathbb{P}(B) = \mathbb{P}(A^c \cap B) + \mathbb{P}(A) - \mathbb{P}(A \cap B^c)$. Mas $A \subset B$, logo, $(A \cap B^c) = \emptyset$, de forma que $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$.

Como $\mathbb{P}(A^c \cap B) \geq 0$, temos $\mathbb{P}(B) \geq \mathbb{P}(A)$. □

2.1.1 Exercícios

Exercício 13. Mostre que $\mathbb{P}((A \cap B^c) \cup (A^c \cap B)) = \mathbb{P}(A) + \mathbb{P}(B) - 2\mathbb{P}(A \cap B)$.

Exercício 14. Prove que

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$$

Exercício 15. Seja Ω o espaço amostral e A, B e C eventos. Prove que:

a. $\mathbb{P}(A \cup B) = 1 - \mathbb{P}(A^c \cap B^c)$.

b. $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$.

c. $\max(\mathbb{P}(A), \mathbb{P}(B)) \leq \mathbb{P}(A \cup B) \leq \min(1, \mathbb{P}(A) + \mathbb{P}(B))$.

d. $\max(0, \mathbb{P}(A) + \mathbb{P}(B) - 1) \leq \mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$.

Exercício 16. Sejam $A_1, A_2, \ldots \in \gamma$. Mostre que:

$$\mathbb{P}(\cup_{n \geq 1} A_n) \leq \sum_{n \geq 1} \mathbb{P}(A_n)$$

2.2 Independência de eventos e probabilidade condicional

Quando trabalhamos com probabilidades, é muito comum nos depararmos com situações onde desejamos saber a probabilidade de algo ocorrer *dado* que sabemos algo sobre outro evento passado. Em outras palavras, queremos usar nosso conhecimento prévio para fazer alguma previsão sobre eventos futuros.

O reverendo Thomas Bayes (1701-1761) foi uma das primeiras pessoas a pensar neste tipo de situação e desenvolver uma solução matemática para o problema, que ficou conhecido como o problema das *probabilidades inversas*. Sua contribuição inicial deu origem a uma área inteira da estatística que atualmente é chamada, em sua homenagem, de *estatística bayesiana*[6].

Nesta seção iremos abordar somente o Teorema de Bayes e seu uso em probabilidade, sem explorar mais profundamente outros aspectos da área. Porém é importante destacar que atualmente muitos modelos macroeconométricos e microeconométricos estão baseados em metodologia bayesiana para estimação dos seus parâmetros e cada vez mais a metodologia bayesiana tem ganho espaço nas aplicações em economia. Um exemplo é o modelo atualmente utilizado pelo Banco Central do Brasil, o SAMBA (*Stochastic Analytical Model with a Bayesian Approach*). Para entender um pouco mais dessa relação entre estatística bayesiana e econometria, recomendo este texto do professor Christopher Sims (Nobel de Economia em 2011): http://sims.princeton.edu/yftp/UndrstndgNnBsns/GewekeBookChpter.pdf.

Considere para todos as definições e teoremas que existe um espaço de probabilidade $(\Omega, \gamma, \mathbb{P}[\cdot])$; isto é, existe um experimento aleatório para o qual o espaço amostral é Ω, uma coleção de eventos γ e uma função de probabilidade[7] $\mathbb{P}[\cdot]$ que estão bem definidos. As definições a seguir foram retiradas e/ou adaptadas das obras listadas nas referências.

Dados dois eventos, A e B, queremos definir a probabilidade condicional de A ocorrer dados que B ocorreu.

Definição 2.2.1. Probabilidade Condicional
Sejam A e B eventos em um espaço de probabilidade $(\Omega, \gamma, \mathbb{P}[\cdot])$. A *probabilidade condicional* do evento A dado B, denotada por $\mathbb{P}[A|B]$, é definida por:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \quad \text{, se} \quad \mathbb{P}(B) \neq 0 \tag{9}$$

e não está definida para $\mathbb{P}(B) = 0$.

Observação: 1 Um resultado direto da definição 9 é que $\mathbb{P}(A \cap B) = \mathbb{P}[A|B]\mathbb{P}(B) = \mathbb{P}[B|A]\mathbb{P}(A)$ se ambas probabilidades $\mathbb{P}(A)$ e $\mathbb{P}(B)$ são não nulas (faça a conta em um papel para conferir), de forma a relacionar as condicionais $\mathbb{P}[A|B]$ e $\mathbb{P}[B|A]$ com as probabilidades não-condicionadas de A e B.

Observação: 2 A definição 9 vai ao encontro da noção de probabilidade sob o enfoque frequentista. Suponha que foram observadas N ocorrências de um experimento aleatório (sendo N um número grande) para o qual os eventos A e B estão definidos. Então, $\mathbb{P}[A|B]$ representa a proporção de ocorrências do evento B onde o evento A também ocorreu, isto é:

$$\mathbb{P}[A|B] = \frac{N_{AB}}{N_B}$$

[6] Para saber mais sobre a origem da estatística bayesiana, veja McGrayne (2011)

[7] Embora toda a teoria aqui explicitada refira-se a eventos, pode ser extendida naturalmente para variáveis aleatórias tanto discretas como contínuas. Quando necessário, será especificado se a função é de densidade ou distribuição de probabilidade, mas considere como sendo possível adaptar para todos os casos.

Onde N_{AB} representa o número de ocorrências do evento $A \cap B$ e N_B representa o número de ocorrências do evento B nas N observações do experimento. Portanto, $\mathbb{P}[A \cap B] = \frac{N_{AB}}{N}$ e $\mathbb{P}(B) = \frac{N_B}{N}$, de forma que:

$$\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{N_{AB}/N}{N_B/N} = \frac{N_{AB}}{N_B} = \mathbb{P}[A|B]$$

O que é consistente com a definição dada.

Exemplo 2.2.2. Considere o lançamento de duas moedas. Seja $\Omega = \{(C,C),(C,K),(K,C),(K,K)\}$, onde C significa que a face observada foi cara e K significa que a face observada foi coroa. Assuma que as moedas são honestas. Vamos calcular 1) a probabilidade de saírem duas caras dado que saiu cara na primeira moeda e 2) a probabilidade de duas caras dado que pelo menos uma é cara.

Defina os eventos: $A_1 =$ cara na primeira moeda e $A_2 =$ cara na segunda moeda. Então, a probabilidade de saírem duas caras dado que saiu cara na primeira moeda é:

$$\mathbb{P}[A_1 \cap A_2|A_1] = \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} = \frac{1/4}{1/2} = \frac{1}{2}$$

A probabilidade de saírem duas caras dado que pelo menos uma das moedas é cara será igual a $\frac{1}{3}$ e fica sugerida como exercício.

Quando falamos de probabilidades condicionais do tipo $\mathbb{P}[A|B]$, o que fazemos é definir um novo espaço amostral, Ω_B, onde tomamos apenas as ocorrências do evento B e calculamos a probabilidade de A ocorrer. De certa forma, isso pode ser visto como uma *restrição* do espaço amostral original. No exemplo 2.2.2, temos $\Omega_B = \{(C,K),(C,C)\}$ para o primeiro item e $\Omega_B = \{(C,C),(C,K),(K,C)\}$. Em ambos casos, restringimos o espaço amostral para os eventos que já sabemos que ocorreram e com base nesse novo espaço amostral é que iremos calcular a probabilidade de interesse.

Surge então (ou deveria surgir) a pergunta: para um dado evento B para o qual $\mathbb{P}(B) > 0$, será que $\mathbb{P}[\,\cdot\,|B]$ é uma função de probabilidade que tem γ como seu domínio? Isto é, será que $\mathbb{P}[\,\cdot\,|B]$ satisfaz os três axiomas para ser considerada uma função de probabilidade? Observe que sim, pois:

(i) $\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}(B)} \geq 0 \quad \forall\, A \in \gamma$;

(ii) $\mathbb{P}[\Omega|B] = \frac{\mathbb{P}[\Omega \cap B]}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$;

(iii) Se $A_1, A_2, \cdots$ é uma sequência de eventos mutuamente exclusivos em γ e $\bigcup_{i=1}^{\infty} A_i \in \gamma$, então

$$\mathbb{P}\left[\bigcup_{i=1}^{\infty} A_i|B\right] = \frac{\mathbb{P}\left[\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right]}{\mathbb{P}(B)} = \frac{\mathbb{P}\left[\bigcup_{i=1}^{\infty} (A_i \cap B)\right]}{\mathbb{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}[A_i|B]$$

Então, $P[\,\cdot\,|B]$ para um dado B que satisfaz $\mathbb{P}(B) > 0$ é uma função de probabilidade, o que justifica chamá-la de probabilidade condicional. $\mathbb{P}[\,\cdot\,|B]$ também apresenta as mesmas propriedades que uma probabilidade não condicionada. Logo, podemos enunciar os seguintes resultados que são similares aos já obtidos para probabilidades não condicionais:

Teorema 2.2.3. *A probabilidade condicional do vazio é zero, isto é,* $\mathbb{P}[\,\emptyset\,|B] = 0$.

Demonstração.

$$\mathbb{P}(\emptyset|B) = \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = \frac{0}{\mathbb{P}(B)} = 0$$

□

Teorema 2.2.4. *Se $A_1, A_2, \cdots, A_n$ são eventos mutuamente exclusivos em γ, então*

$$\mathbb{P}[A_1 \cup \cdots \cup A_n|B] = \sum_{i=1}^{n} \mathbb{P}[A_i|B]$$

Demonstração.

$$\begin{aligned}
\mathbb{P}[A_1 \cup \cdots \cup A_n|B] = & && \text{(definição prop. condicional)} \\
&= \frac{\mathbb{P}[(A_1 \cup \cdots \cup A_n) \cap B]}{\mathbb{P}(B)} && \text{(distributiva)} \\
&= \frac{\mathbb{P}[(A_1 \cap B) \cup \cdots \cup (A_n \cap B)]}{\mathbb{P}(B)} && \text{(independência dos } A_i\text{'s)} \\
&= \frac{\mathbb{P}\left[\sum_{i=1}^{n}(A_i \cap B)\right]}{\mathbb{P}(B)} && \text{(definição prob. condicional)} \\
&= \sum_{i=1}^{n} \mathbb{P}[A_i|B]
\end{aligned}$$

□

Teorema 2.2.5. *Se A é um evento em γ, então*

$$\mathbb{P}[A^c|B] = 1 - \mathbb{P}[A|B]$$

onde A^c é o evento complementar de A.

Demonstração.

$$\begin{aligned}
\mathbb{P}[A^c|B] = & && \text{(def. prob. condicional)} \\
&= \frac{\mathbb{P}(A^c \cap B)}{\mathbb{P}(B)} && \text{(forma alternativa de } \mathbb{P}(B)\text{)} \\
&= \frac{\mathbb{P}(B) - \mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}(B)}{\mathbb{P}(B)} - \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\
&= 1 - \mathbb{P}[A|B]
\end{aligned}$$

□

Teorema 2.2.6. *Se A_1 e A_2 pertencem a γ, então*

$$\mathbb{P}[A_1|B] = \mathbb{P}[A_1 \cap A_2|B] + \mathbb{P}[A_1 \cap A_2^c|B]$$

Demonstração.

$$
\begin{aligned}
\mathbb{P}[A_1|B] = & && \text{(def. prob. cond.)}\\
&= \frac{\mathbb{P}[A_1 \cap B]}{\mathbb{P}(B)} && \text{(reescrevendo } A_1\text{)}\\
&= \frac{\mathbb{P}[((A_1 \cap A_2) \cup (A_1 \cap A_2^c)) \cap B]}{\mathbb{P}(B)} && \text{(distributiva)}\\
&= \frac{\mathbb{P}[(A_1 \cap A_2) \cap B \cup (A_1 \cap A_2^c) \cap B]}{\mathbb{P}(B)} && \text{(independência)}\\
&= \frac{\mathbb{P}[(A_1 \cap A_2) \cap B] + \mathbb{P}[(A_1 \cap A_2^c) \cap B]}{\mathbb{P}(B)} && \text{(reorganizando)}\\
&= \frac{\mathbb{P}[(A_1 \cap A_2) \cap B]}{\mathbb{P}(B)} + \frac{\mathbb{P}[(A_1 \cap A_2^c) \cap B]}{\mathbb{P}(B)} && \text{(def. prob. cond.)}\\
&= \mathbb{P}[A_1 \cap A_2|B] + \mathbb{P}[A_1 \cap A_2^c|B]
\end{aligned}
$$

□

Teorema 2.2.7. *Para quaisquer dois eventos A_1 e $A_2 \in \gamma$,*

$$\mathbb{P}[A_1 \cup A_2|B] = \mathbb{P}[A_1|B] + \mathbb{P}[A_2|B] - \mathbb{P}[A_1 \cap A_2|B]$$

Demonstração.

$$
\begin{aligned}
\mathbb{P}[A_1 \cup A_2|B] = & && \text{(def. prob. cond.)}\\
&= \frac{\mathbb{P}[(A_1 \cup A_2) \cap B]}{\mathbb{P}(B)} && \text{(distributiva)}\\
&= \frac{\mathbb{P}[(A_1 \cap B) \cup (A_2 \cap B)]}{\mathbb{P}(B)} && \text{(def. prob. união.)}\\
&= \frac{\mathbb{P}(A_1 \cap B) + \mathbb{P}(A_2 \cap B) - \mathbb{P}(A_1 \cap B \cap A_2 \cap B)}{\mathbb{P}(B)} && \text{(rearranjando)}\\
&= \frac{\mathbb{P}(A_1 \cap B)}{\mathbb{P}(B)} + \frac{\mathbb{P}(A_2 \cap B)}{\mathbb{P}(B)} - \frac{\mathbb{P}(A_1 \cap A_2 \cap B)}{\mathbb{P}(B)} && \text{(def. prob. cond.)}\\
&= \mathbb{P}[A_1|B] + \mathbb{P}[A_2|B] - \mathbb{P}[A_1 \cap A_2|B]
\end{aligned}
$$

□

Teorema 2.2.8. *Se A_1 e $A_2 \in \gamma$ com $A_1 \subset A_2$, então*

$$\mathbb{P}[A_1|B] \leq \mathbb{P}[A_2|B]$$

Demonstração. Esse resultado decorre do lema 2.1.8. □

Enquanto que os teoremas de 11 a 16 têm uma correspondência direta com as probabilidades não-condicionais, o seguinte teorema tem aplicação apenas no segundo caso:

Teorema 2.2.9. ***Teorema da probabilidade total***

Para um dado espaço de probabilidade $(\Omega, \gamma, \mathbb{P}[\cdot])$*, se* $B_1, B_2, \cdots, B_n$ *são uma coleção de eventos mutuamente disjuntos (ou exclusivos) em* γ *satisfazendo* $\Omega = \bigcup_{j=1}^{n} B_j$ *e* $\mathbb{P}(B_j) > 0$ *para* $j = 1, \cdots, n$[8]*, então, para todo* $A \in \gamma$*, vale que*

$$\mathbb{P}(A) = \sum_{j=1}^{n} \mathbb{P}[A|B_j]\mathbb{P}(B_j)$$

Demonstração. Observe que $A = \bigcup_{j=1}^{n} (A \cap B_j)$ e que os termos $(A \cap B_j)$ e $(A \cap B_i)$ são mutuamente exclusivos para $i \neq j$. Logo,

$$\mathbb{P}(A) = \mathbb{P}\left[\bigcup_{j=1}^{n}(A \cap B_j)\right] = \sum_{j=1}^{n} \mathbb{P}(A \cap B_j) = \sum_{j=1}^{n} \mathbb{P}[A|B_j]\mathbb{P}(B_j)$$

Na aula, eu fiz a versão acima, que é bem direta. Abaixo tem uma forma mais detalhada que pode ajudar a entender melhor o que está acontecendo.

Observe que

$$A = A \cap \Omega \tag{10}$$

Além disso, como $(B_j)_{j=1}^{n}$ é uma partição de Ω, temos

$$\Omega = \cup_{j=1}^{n}(B_j) \tag{11}$$

Usando a equação 11 na equação 10, obtemos:

$$A = A \cap (\cup_{j=1}^{n} B_j) \tag{12}$$
$$= \cup_{j=1}^{n}(A \cap B_j) \tag{13}$$

e portanto,

$$\mathbb{P}(A) = \mathbb{P}(\cup_{j=1}^{n}(A \cap B_j)) \tag{14}$$

Como $(B_j)_{j=1}^{n}$ é uma partição, é também uma sequência disjunta. Portanto, para todo $i \neq j$, $B_i \cap B_j = \emptyset$ e

$$(A \cap B_i) \cap (A \cap B_j) = A \cap (B_i \cap B_j) \tag{15}$$
$$= A \cap \emptyset \tag{16}$$
$$= \emptyset \tag{17}$$

Isto é, $(A \cap B_j)_{j=1}^{n}$ também é uma sequência disjunta. Assim, dos axiomas da probabilidade,

[8] Dizemos, neste caso, que os B_J formam uma *partição* de Ω.

$$\mathbb{P}(\cup_{j=1}^{n}(A \cap B_j)) = \sum_{j=1}^{n} \mathbb{P}(A \cap B_j) \tag{18}$$

Finalmente, do axioma da probabilidade condicional, temos para todo $j \in \{1, \ldots, n\}$, $\mathbb{P}(A \cap B_j) = \mathbb{P}(B_j)\mathbb{P}(A|B_j)$. Usando as equações 14 e 18, concluímos que:

$$\mathbb{P}(A) = \sum_{j=1}^{n} \mathbb{P}(B_j)\mathbb{P}(A|B_j) \tag{19}$$

□

O próximo resultado é uma implicação direta do teorema da probabilidade total:

Corolário 2.2.10. *Para um dado espaço de probabilidade $(\Omega, \gamma, \mathbb{P}[\cdot])$, seja $B \in \gamma$ satisfazendo $0 < \mathbb{P}(B) < 1$; então, para todo $A \in \gamma$:*

$$\mathbb{P}(A) = \mathbb{P}[A|B]\mathbb{P}(B) + \mathbb{P}[A|B^c]\mathbb{P}(B^c)$$

Demonstração. O resultado segue diretamente do teorema da probabilidade total. □

Finalmente podemos enunciar a fórmula de Bayes:

Teorema 2.2.11. ***Fórmula de Bayes***
Para um dado espaço de probabilidade $(\Omega, \gamma, \mathbb{P}[\cdot])$, se $B_1, B_2, \cdots, B_n$ são uma coleção de eventos mutuamente disjuntos (ou exclusivos) em γ satisfazendo $\Omega = \bigcup_{j=1}^{n} B_j$ e $\mathbb{P}(B_j) > 0$ para $j = 1, \cdots, n$, então, para todo $A \in \gamma$ para o qual $\mathbb{P}(A) > 0$ vale que:

$$\mathbb{P}[B_k|A] = \frac{\mathbb{P}[A|B_k]\mathbb{P}(B_k)}{\sum_{j=1}^{n} \mathbb{P}[A|B_j]\mathbb{P}(B_j)} \tag{20}$$

Demonstração.

$$\mathbb{P}[B_k|A] = \frac{\mathbb{P}[B_k \cap A]}{\mathbb{P}(A)} = \frac{\mathbb{P}[A|B_k]\mathbb{P}(B_k)}{\sum_{j=1}^{n} \mathbb{P}[A|B_j]\mathbb{P}(B_j)}$$

□

Corolário 2.2.12. *Para um dado espaço de probabilidade $(\Omega, \gamma, \mathbb{P}[\cdot])$, sejam A e $B \in \gamma$ satisfazendo $\mathbb{P}(A) > 0$ e $0 < \mathbb{P}(B) < 1$; então:*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}[A|B]\mathbb{P}(B) + \mathbb{P}[A|B^c]\mathbb{P}(B^c)}$$

Definição 2.2.13. Dois eventos A e B são independentes se $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Lema 2.2.14. *A e B são independentes se e somente se $\mathbb{P}(A|B) = \mathbb{P}(A)$. Em outras palavras, A e B são independentes se e somente se a incerteza sobre A não muda se assumirmos que B é verdadeiro.*

Demonstração. Assuma $\mathbb{P}(A|B) = \mathbb{P}(A)$. Da definição de probabilidade condicional, lembramos que $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$. Usando a suposição inicial, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ e A e B são independentes.

Assuma A e B independentes. Portanto, $\mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$. Juntando os dois extremos da desigualdade e dividindo por $\mathbb{P}(B)$, $\mathbb{P}(A|B) = \mathbb{P}(A)$. □

Teorema 2.2.15. *Se A e B são eventos independentes, então os eventos A e B^c, A^c e B e A^c e B^c também são independentes.*

Demonstração.

$$\begin{aligned}\mathbb{P}(A \cap B^c) &= \mathbb{P}(A) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) \\ &= \mathbb{P}(A)[1 - \mathbb{P}(B)] \\ &= \mathbb{P}(A)\mathbb{P}(B^c)\end{aligned}$$

$$\begin{aligned}\mathbb{P}(A^c \cap B) &= \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &= \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) \\ &= \mathbb{P}(B)[1 - \mathbb{P}(A)] \\ &= \mathbb{P}(A^c)\mathbb{P}(B)\end{aligned}$$

$$\begin{aligned}\mathbb{P}(A^c \cap B^c) &= \mathbb{P}(A \cup B)^c \\ &= 1 - \mathbb{P}(A \cup B) \\ &= 1 - [\mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)] \\ &= 1 - P(A) - \mathbb{P}(B) + \mathbb{P}(A)\mathbb{P}(B) \\ &= P(A^c) - \mathbb{P}(B)[1 - \mathbb{P}(A)] \\ &= P(A^c)[1 - \mathbb{P}(B)] \\ &= \mathbb{P}(A^c)\mathbb{P}(B^c)\end{aligned}$$

□

Definição 2.2.16. Independência Condicional

(surrupiado das notas de aula de Econometria Bayesiana) Seja C um evento qualquer, então dizemos que A e B são condicionalmente independentes se:

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B \cap C) \tag{21}$$

Podemos utilizar a definição de independência não condicional com a regra da multiplicação para obter:

$$\begin{aligned}\mathbb{P}(A \cap B|C) &= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} \quad \text{(prob. cond.)} \\ &= \frac{\mathbb{P}(C)\mathbb{P}(B \cap C)\mathbb{P}(A|B \cap C)}{\mathbb{P}(C)} \quad \text{(regra mult.)} \\ &= \mathbb{P}(B|C)\mathbb{P}(A|B \cap C)\end{aligned}$$

Igualando as duas pontas da equação acima, temos:

$$\mathbb{P}(A|C)\mathbb{P}(B|C) = \mathbb{P}(B|C)\mathbb{P}(A|B \cap C)$$
$$\mathbb{P}(A|C) = \frac{P(B|C)\mathbb{P}(A|B \cap B)}{\mathbb{P}(B|C)} = \mathbb{P}(A|B \cap C)$$

De maneira que $\mathbb{P}(A|C) = \mathbb{P}(A|B \cap C)$.

Note que isso não significa que A e B são independentes. Caso o C seja desconhecido, a nossa probabilidade é alterada. É possível que a informação que C tenha para A seja a mesma que B agregaria e por isso que, condicionado a C, a informação de B para A não seja relevante. Assim, *independência condicional não implica independência incondicional.*

2.3 Exercícios

Exercício 17. Independência do vazio

Na aula, conversamos sobre a independência e o que acontece quando um dos eventos é o evento impossível (ou tem probabilidade zero). O exercício então é mostrar que $\forall B \in \Omega$, B e $\emptyset$ são independentes.

Exercício 18. (Retirado de Viali (2004c)) Calcular a probabilidade de no lançamento de um dado equilibrado[9] obter-se os seguintes eventos:

a. A face observada (o resultado do dado, o número virado para cima) é igual a 5;

b. A face observada é um número ímpar;

c. A face observada é maior que 2;

d. A face observada é um número primo;

e. A face observada é menor que π;

f. A face observada é diferente de 3.

Exercício 19. (Adaptado de Stern and Izbicki (2016)) Um total de 4 bolas numeradas de 1 a 4 é distribuído aleatoriamente em n urnas, também numeradas de 1 a 4, uma de cada vez, de tal modo que cada bola tem exatamente a mesma probabilidade de cair em cada uma das urnas. Assim, cada urna pode receber entre 0 e 4 bolas. Qual é a probabilidade da primeira urna ficar vazia?

Qual é essa probabilidade se você fizer o mesmo experimento, porém com *n* bolas e *n* urnas?

[9] Dizemos que um dado é *equilibrado* quando a probabilidade das faces é sempre a mesma. Para um dado de 6 faces, teremos que a probabilidade de cada uma será $\frac{1}{6}$. Outras formas encontradas na literatura são: dado não viciado; dado não viesado; dado honesto. A mesma nomenclatura se aplica a situações envolvendo moedas.

Exercício 20. Na mochila de sua amiga Misty estão 12 *pokébolas*, sendo que 7 são de *pokémons* do tipo água e as outras 5 são de *pokémons* do tipo fada. Destes 12, um deles é o Togepi[10] e o outro é o Psyduck[11].

Misty irá lutar uma batalha de dois rounds, onde cada treinador irá utilizar 2 *pokémons*, sendo um para cada *round* (isto é, mesmo que vença a batalha, o treinador deve trocar de *pokémon*), selecionados aleatoriamente da mochila. Calcule as seguintes probabilidades:

a. Dos 2 *pokémons* que Misty seleciona serem do tipo água e do tipo fada, nessa ordem;

b. De Misty utilizar seu *pokémon* preferido, Togepi[12], na batalha;

c. De Misty não utilizar seu Psyduck[13] na batalha;

d. De Misty utilizar somente dois *pokémons* do tipo água;

e. De Misty utilizar Togepi e o Psyduck na batalha;

f. De Misty utilizar pelo menos um *pokémon* do tipo fada na batalha;

g. Do segundo *pokémon* escolhido ser de água dado que o primeiro foi o Togepi.

h. Resolva os itens a, b e g considerando agora que após a primeira batalha Misty guarda o *pokémon* utilizado de volta na *pokébola*, vai para o *pokémon* center para curá-lo e então retorna para o segundo round, selecionando o segundo pokémon dentre todas suas 12 *pokébolas*.

Exercício 21. (Retirado de Viali (2004c)) Suponha que A e B sejam eventos tais que $P(A) = a$, $P(B) = b$ e $P(A \cap B) = c$. Escreva as seguintes probabilidades em termos de a, b e c:

a. $P(A \cup B)$;

b. $P(A^c)$;

c. $P(B^c)$;

d. $P(A^c \cup B^c)$;

e. $P(A^c \cup B)$;

f. $P(A^c \cap B^c)$;

g. $P(A \cap B^c)$;

Exercício 22. (Retirado de Viali (2004c)) Uma amostra de 140 clientes de um banco revelou que 80 guardam seu dinheiro na poupança, 30 investem no tesouro direto (TD) e 10 tem aplicações tanto na poupança como no TD. Qual a probabilidade de que um cliente, que tenha sido escolhido ao acaso dos 140 entrevistados, tenha dinheiro na poupança ou no TD?

Exercício 23. (Retirado de Meyer (1973)) Sejam Ω um espaço amostral e A, B e C eventos em Ω. Demonstre as seguintes propriedades:

[10] Togepi é um pokémon do tipo fada a partir da geração VI.
[11] Psyduck é um pokémon do tipo água.
[12] Considere que Togepi está em uma das 12 *pokébolas* e não no colo da Misty.
[13] Desconsidere o fato do Psyduck fugir da *pokébola* em momentos inapropriados.

a. Se $\emptyset$ for o espaço vazio, então $P(\emptyset) = 0$;

b. Se A^c for o evento complementar de A, então $P(A) = 1 - P(A^c)$;

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

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

e. Se $A \subset B$, então $P(A) \leq P(B)$;

f. $P(A \cup B) = 1 - P(A^c \cap B^c)$;

g. $P(A) = P(A \cap B) + P(A \cap B^c)$. **Obs.:** Este resultado mostra como podemos escrever a probabilidade do evento A de uma forma diferente. Ele é muito utilizado em várias demonstrações e a sua ideia básica é utilizada no teorema de Bayes.

Exercício 24. (Retirado de Meyer (1973)) Mostre que

$$P[(A \cap B^c) \cup (A^c \cap B)] = P(A) + P(B) - 2P(A \cap B)$$

Obs.: *Este resultado trata da probabilidade que exatamente um dos eventos A ou B ocorra, enquanto que a letra c. do exercício anterior pode ser vista como a probabilidade de que pelo menos um deles ocorra. Para fazer a demonstração, você tem a opção de desenvolver um dos lados da igualdade e chegar no outro ou provar que os dois conjuntos são iguais.*

Exercício 25. (Retirado de Magalhães (2011)) Os eventos A e B são independentes. Sendo $P(A \cap B^c) = 0,3$ e $P(A^c \cap B) = 0,2$; calcule $P(A \cap B)$.

Exercício 26. (Retirado de Magalhães (2011)) Sejam os resultados de 3 lançamentos de uma moeda honesta e considere os seguintes eventos: $\alpha = \{$ocorrem pelo menos duas coroas$\}$ e $\beta = \{$ocorre coroa no 1^olançamento$\}$. Determine se α e β são eventos independentes.

Exercício 27. (Adaptado de Schmidt (2011)) Considere Ω, espaço amostral, e A, B eventos em Ω.

a. Calcule $P(A \cap B)$ considerando $P(A) = 0.5$, $P(B) = 0.25$ e A e B mutuamente exclusivos;

b. Caso você saiba que $A \subset B$, é verdade que $P(A|B) \leq P(A)? Justifique.$

c. Considerando $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$ e $P(A \cap B) = \frac{1}{4}$, calcule $P(A^c \cap B^c)$.

d. Se $P(A|B) = 0$, então A e B são independentes? Justifique.

2.4 Variáveis aleatórias

Até o momento trabalhamos apenas com probabilidades relacionadas a um espaço de probabilidade, $(\Omega, \gamma, \mathbb{P}(\cdot))$, onde Ω ou *espaço amostral*, é o conjunto de todos os possíveis resultados de um determinado experimento, c γ é o conjunto de todas coleção de eventos de Ω e $\mathbb{P}(\cdot)$ é uma função de probabilidade, que tem como domínio γ e contradomínio o intervalo $[0, 1]$.

Nós ainda estamos interessados nas probabilidades de eventos, porém com um grau de formalização maior. Por exemplo, seja o lançamento de uma moeda. Sabemos que $\Omega = \{C, K\}$, onde C representa "cara" e K representa "coroa" e, no caso de uma moeda honesta, $\mathbb{P}(C) = \mathbb{P}(K)$. Mas o que acontece se quisermos saber qual a probabilidade de saírem 5 caras em 7 lançamentos da moeda? Poderíamos ter as situações $\mathbb{P}(CCCCCKK)$, $\mathbb{P}(CKCCCCK)$, $\mathbb{P}(KKCCCCC)$ e assim por diante. Essa notação se torna pesada e pouco prática. Poderíamos criar um "código" onde $C = 1$ e $K = 0$, de maneira fazer a soma do número de caras. Por exemplo, na situação das 5 caras, estaríamos interessados em obter $\mathbb{P}(5)$. Essa noção é o fundamento das variáveis aleatórias.

Exemplo 2.4.1. (Juju)
Suponha que você tem uma moeda chamada Juju e que ela é uma moeda honesta[14], isto é, a probabilidade de sair cara é igual à probabilidade de sair coroa. Em notação matemática,

$$\mathbb{P}(\{\text{lançar Juju resultou em cara}\}) = \mathbb{P}(\{\text{lançar Juju resultou em coroa}\}) = \frac{1}{2}$$

Observe ainda que Ω, o espaço amostral do experimento *lançar a Juju*, é dado por $\Omega = \{(\text{cara}), (\text{coroa})\}$.

Agora vamos criar uma função que será definida por:

$$f(\text{Juju}) = \begin{cases} 0 & \text{se o lançamento de Juju deu cara,} \\ 1 & \text{se o lançamento de Juju deu coroa.} \end{cases} \tag{22}$$

Vamos chamar essa função de X (ao invés de f). Observe que:

$$\mathbb{P}(X = 1) = \mathbb{P}(\{\text{lançar Juju resultou em coroa}\}) = \frac{1}{2}$$

E ainda:

$$\mathbb{P}(X = 0) = \mathbb{P}(\{\text{lançar Juju resultou em cara}\}) = \frac{1}{2}$$

Isto é, a probabilidade de que X assuma valor 1 é a mesma probabilidade de sair coroa quando lançamos a Juju e a probabilidade de X ser igual a 0 é a probabilidade de sair cara no lançamento da Juju.

Como X é uma função que "pega" um elemento de Ω e está associando um número real, então X preenche os requisitos para ser chamada de *variável aleatória*, que veremos a seguir.

2.5 Caracterização

Informalmente, *variável aleatória* (v.a.) é uma função que associa elementos do espaço amostral Ω ao conjunto dos números reais, $\mathbb{R}$. Formalmente, temos:

[14] Quando Juju não for honesta, diremos que Juju é ~~safada~~ desonesta, não equilibrada, viesada ou viciada.

Definição 2.5.1. (Variável Aleatória)
Para um dado espaço de probabilidade $(\Omega, \gamma, P(\cdot))$ uma variável aleatória, denotada por X ou $X(\cdot)$, é uma função com domínio em Ω e contradomínio em $\mathbb{R}$.

Analisando em termos dos conceitos já vistos de experimentos, temos Ω como sendo o conjunto de todos resultados possíveis de um experimento (conjunto de todos eventos) e $X(\cdot)$ é uma *função* que associa cada elemento $\omega \in \Omega$ a um número real. Observe que em nenhum momento utilizamos a função de probabilidade $P(\cdot)$.

Existe uma discussão sobre a nomenclatura *variável* para um objeto que na verdade é uma função, pois isso leva a muitas interpretações errôneas e até confusão por parte dos alunos. Mas o nome já está enraizado demais para que se volte atrás. Isso acaba se confundindo mais ainda pelo fato de, ao invés de usarmos $X(\dot{)}$ para denotar a variável aleatória X, utilizarmos apenas X (e iremos continuar perpetuando isso, desculpem). Ao longo do texto, usaremos letras latinas maiúsculas para denotar a variável aleatória (função) e letras minúsculas quando estivemos nos referindo a um valor específico que a função assumiu, ou, definindo formalmente:

Definição 2.5.2. Valor observado (outcome) de uma v.a.
A imagem $x = X(\omega)$ de um evento $\omega \in \Omega$ gerada por uma variável aleatória X é chamado de *valor observado* de X.

2.5.1 O espaço de probabilidade definido por uma v.a.

Um novo espaço de probabilidade será necessário para poder estabelecer probabilidades a subconjuntos do novo espaço amostral real definido pela imagem da v.a.. O espaço de probabilidade $\{\Omega, \gamma, \mathbb{P}(\cdot)\}$ nos permite determinar a probabilidade para eventos em S, porém qual é a probabilidade de que um resultado da v.a. X fique no subconjunto $A \subset R(X)$?

Como X é um mapa de Ω para a reta real, é possível definir o evento B em S tal que o evento B ocorre se e somente se $A \subset R(X)$ ocorre. Uma vez que os eventos A e B ocorrem simultaneamente, a probabilidade deles deve ser a mesma, ou seja, $\mathbb{P}_X(A) \equiv \mathbb{P}(B)$, onde $\mathbb{P}_X(\cdot)$ denota a medida de probabilidade que atribui a resultados de X suas probabilidades. Se dois eventos ocorrem sempre simultaneamente, eles são ditos equivalentes e ocorrem em espaços de probabilidades distintos, pois se ocorressem no mesmo espaço, eles seriam o mesmo evento. Logo, $\mathbb{P}_X(A) \equiv \mathbb{P}(B)$ para $B = \{w : X(w) \in A, w \in \Omega\}$. A figura 2 ilustra esta relação.

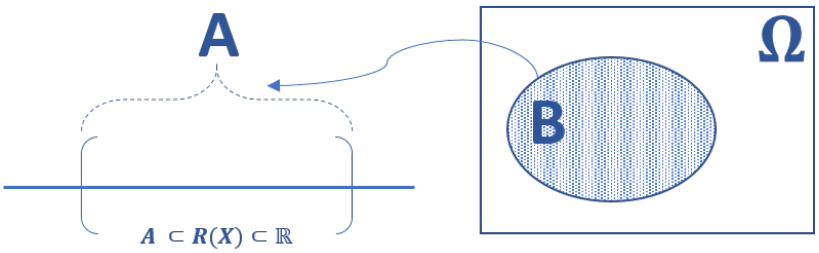


Figura 2: Correspondência entre $B \subset S$ e $A \subset R(X)$.

Probabilidades definidas para eventos em S são transferidas para eventos em $R(X)$ através da relação funcional que define uma v.a., $x_i = X(w_i)$. Então, sabendo que o domínio de $\mathbb{P}(\cdot)$ é γ (espaço de eventos), qual é o domínio de $\mathbb{P}_X(\cdot)$? Podemos dizer informalmente que γ_X é o espaço de eventos do espaço de probabilidade associado à variável aleatória X e é dado por todos os subconjuntos da imagem de X, $R(X)$.

Então, o espaço de probabilidade definido por X é $(R(X), \gamma_X, \mathbb{P}_X(\cdot)$. Lidar com espaços amostrais reais é muito mais conveniente pois nos permite utilizar os conhecimentos matematicos.

Definição 2.5.3. (Função distribuição acumulada de uma v.a.)
Seja X uma v.a.. A *função distribuição acumulada* de X, denotada por $F_X(\cdot)$, é definida como a função com domínio em $\mathbb{R}$ e contradomínio no intervalo fechado $[0, 1]$ que satisfaz

$$F_X(x) = \mathbb{P}(X \in (-\infty, x]) = \mathbb{P}(X \leq x) = \mathbb{P}[\{\omega : X(\omega) \leq x\}]$$

para todo número real x.

Em outras palavras, a função acumulada de X é a função que calcula as probabilidades de X assumir valores menores ou iguais a um valor específico x. Para tanto, ela avalia os valores ω do espaço amostral tais que $X(\omega)$ é menor ou igual que x. Essa função é importante pois ela define de forma única uma variável aleatória. Mesmo que X só assuma valores em um subconjunto dos reais, a função de distribuição é bem definida em toda a reta. Quando não houver a possibilidade de gerar confusão, a notação será apenas F ao invés de F_X.

Teorema 2.5.4. ***(Propriedades da função distribuição acumulada*** $F_X(\cdot)$***)***
As seguintes propriedades precisam ser atendidas para que possamos considerar uma função como sendo a função distribuição acumulada. Este teorema não será demonstrado, mas uma parte da prova pode ser encontrada em Mood and Graybill (1963) ou em Magalhães (2011).

i. $F_X(-\infty) \equiv \lim_{x \to -\infty} F_X(x) = 0$ *e* $F_X(+\infty) \equiv \lim_{x \to +\infty} F_X(x) = 1$*;*

ii. $F_X(\cdot)$ *é uma função monótona não-decrescente; isto é,* $F_X(a) < F_X(b)$ *para* $a < b$*;*

iii. $F_X(\cdot)$ *é contínua à direita; isto é,*

$$\lim_{h \to 0^+} F_X(x + h) = F_X(x)$$

As variáveis aleatórias se dividem entre discretas e contínuas. Iremos começar o estudo com as variáveis aleatórias discretas e na sequência veremos as contínuas.

2.5.2 Variáveis aleatórias discretas

Definição 2.5.5. Variável aleatória discreta
Uma variável aleatória é dita *discreta* se a sua imagem consiste de um número contável de elementos, isto é, sua imagem pode ser colocada em correspondência com um subconjunto (próprio ou não) de $\mathbb{N}$.

Definição 2.5.6. (Função massa de probabilidade de uma v.a. discreta)
Se X é uma v.a. discreta (X assume valores em um subconjunto de $\mathbb{N}$) que assume valores distintos $x_1, x_2, x_3, \cdots, x_n, \cdots$ então a função densidade de X, denotada $f_X(\cdot)$ é definida por:

$$f_X(x) = \begin{cases} \mathbb{P}(X = x_i) & \text{se } x = x_i,\ i = 1, 2, 3, \ldots, n, \ldots \\ 0 & \text{caso contrário.} \end{cases} \tag{23}$$

Outros nomes comumente dados à $f_X(x)$ são: função massa de probabilidade, função de frequência discreta e função de probabilidade. A notação $p_X(x)$ também é usada para diferenciar quando se trata de uma variável aleatória discreta e usualmente usa-se a notação $f_X(x)$ para as contínuas. Observação: Alguns autores partem da f.d.a para chegar na f.m.p, enquanto outros fazem o caminho inverso, mas a ideia é que a v.a. fique caracterizada de forma única.

Definição 2.5.7. f.d.a. de uma v.a. discreta (Retirado de Mittelhammer (2013))
A função distribuição acumulada de uma variável aleatória X, se X for discreta, é dada por

$$F_X(x) = \sum_{X \leq x,\ f(x)>0} f(x),\ x \in (-\infty, +\infty)$$

Algumas distribuições especiais são comumente utilizadas para modelar fenômenos e por seu uso frequente nas mais diversas aplicações, acabaram ganhando "nomes" e suas propriedades foram exaustivamente utilizadas. Uma lista bastante completa pode ser acessada aqui: https://en.wikipedia.org/wiki/List_of_probability_distributions. A estas distribuições, costumamos nos referir como sendo "distribuições tabeladas". Iremos ver brevemente a distribuição Binomial (e, portanto, a de Bernoulli), a distribuição Normal e algumas outras, se der tempo.

Imagine agora um experimento que tenha apenas dois possíveis desfechos A e A^c, onde este último denota o complementar do evento A. Imagine que a probabilidade do evento A seja p, com $0 \leq p \leq 1$ e, portanto, $\mathbb{P}(A^c) = 1 - \mathbb{P}(A) = 1 - p$. Ao definirmos a variável aleatória X, que assume dois valores, um para quando o evento A foi observado e outro para quando não foi observado, diremos que X segue uma *distribuição de Bernoulli de parâmetro* p.

Voltando ao exemplo da Juju, a variável aleatória X do exemplo segue uma distribuição de Bernoulli. Observe que se, ao fazer o experimento de lançar um dado, definirmos a variável aleatória Y como sendo 1 se o resultado foi um número par e 0 caso contrário, Y também é tem distribuição de Bernoulli, isto é, mesmo que o dado possa ter 6 diferentes resultados, podemos estar interessados em um evento que resulte em uma v.a. binária, ou dicotômica.

Definição 2.5.8. (Distribuição de Bernoulli)
Dizemos que uma variável aleatória discreta X segue uma distribuição de Bernoulli de parâmetro p se a função massa de probabilidade de X é dada por:

$$f_X(x) = p_X(x) = \begin{cases} p^x(1-p)^{1-x} & \text{para } x = 0 \text{ ou } 1 \\ 0 & \text{caso contrário.} \end{cases} \tag{24}$$

Onde o parâmetro p satisfaz $0 \leq p \leq 1$. A quantia $(1-p)$ é comumente denotada por q. Utilizamos a notação $X \sim Bernoulli(p)$ para dizer que X segue a distribuição Bernoulli de parâmetro p.

Observação: Se $X \sim Bernoulli(p)$, então a esperança ou valor esperado de X será dado por $\mathbb{E}[X] = p$ e a sua variância é $Var[X] = p(1-p)$.

Imagine agora que no experimento de Bernoulli de parâmetro p, ao invés de uma observação, sejam feitas n repetições independentes do experimento e estamos interessados no número de vezes em que um determinado desfecho ocorre nestas repetições. Se definirmos Y como a variável aleatória que denota esse número de "sucessos" em n tentativas, diremos que Y segue uma *distribuição Binomial de parâmetros* n *e* p.

Definição 2.5.9. (Distribuição Binomial)
Seja X uma v.a. discreta, tomando valores em $\{0, 1, 2, \cdots, n\}$. Dizemos que X tem distribuição Binomial de parâmetros n e p se sua função massa de probabilidade é dada por:

$$p_x(x) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & \text{para } x = 0, 1, \cdots, n \\ 0 & \text{caso contrário.} \end{cases} \tag{25}$$

onde $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ (lê-se a combinação de n x a x)[15] e $x!$ é o fatorial de x.

Observação 1: Se $X \sim Binomial(n, p)$, então a esperança ou valor esperado de X será dado por $\mathbb{E}[X] = np$ e a sua variância é $Var[X] = np(1-p)$.

Observação 2: Observe que a Bernoulli é um caso particular da Binomial, quando $n = 1$.

Exemplo 2.5.10. 2 **(O retorno de Juju)**
Imagine que agora você pretende lançar a Juju 3 vezes e contar o número de vezes em que o resultado foi cara. O espaço amostral do evento agora será:

$$\Omega = \{(CCC), (CCK), (CKC), (KCC), (CKK), (KCK), (KKC), (KKK)\}$$

Vamos definir Y como sendo o número de caras em $n = 3$ lançamentos da Juju. Logo, $Y = 0, 1, 2, 3$ (podem sair 0 caras, 1 cara e assim por diante). Observe que: $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 3) = \frac{1}{8}$, isto é, entre 8 resultados possíveis, somente em um deles a Juju apresentou 0 caras. E de forma análoga, somente em 1 situação ela apresentará 3 caras. Poderíamos ter chego no mesmo concluindo que os três lançamentos são independentes, de forma que:

$$\mathbb{P}(Y = 0) = \mathbb{P}\underbrace{(K \cap K \cap K)}_{\text{eventos independentes}} = \mathbb{P}(K)\mathbb{P}(K)\mathbb{P}(K) = \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{2^3} = \frac{1}{8}$$

Mas o que acontece quando calculamos $\mathbb{P}(Y = 1)$?

Note que observamos exatamente uma cara em três possíveis situações: (CKK), (KCK) ou (KKC). Como se tratam de eventos disjuntos, ao calcular a probabilidade da união, basta somar as probabilidades dos eventos individuais[16]:

$$\mathbb{P}(Y = 1) = \mathbb{P}(CKK) + \mathbb{P}(KCK) + \mathbb{P}(KKC) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

Que é justamente 3 resultados em 8. Mas como isso se relaciona com a equação 25?

Bom, para a Juju, já havíamos estabelecido que a probabilidade de sair cara, isto é, nosso "sucesso", é igual a $1/2$. Então, em 3 lançamentos, se queremos saber a probabilidade de exatamente 1 cara, ela será a probabilidade de sair cara, $p^1 = \left(\frac{1}{2}\right)^1$, vezes[17] a probabilidade de saírem duas coroas[18], que é $\frac{1}{2}\frac{1}{2} = \left(\frac{1}{2}\right)^2 = \left(\frac{1}{2}\right)^{3-1}$, o que irá resultar em $\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{3-1}$. Só que isso acontece em $\binom{3}{1} = \frac{3!}{1!(3-1)!} = \frac{3\cdot2\cdot1}{1(2\cdot1)} = 3$ vezes. Logo, chegamos em:

$$\mathbb{P}(Y = 1) = \binom{3}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^{3-1}$$

Que é justamente a aplicação da equação 25 para o caso onde $p = 1/2$, $n = 3$ e o valor x é 1. Tente calcular agora $\mathbb{P}(Y = 2)$ pelos dois métodos.

Para o exemplo dado, essa conta acaba parecendo mais complicada do que enumerar o espaço amostral inteiro, porém para 1000 lançamentos de Juju onde desejamos saber $\mathbb{P}(\{\text{foram observadas mais caras que coroas}\})$ esse cálculo acaba não sendo tão trivial e o uso da fórmula se justifica.

[15] Para uma revisão relâmpago de combinatoria, sugere-se a Apostila do professor Lorí Viali do departamento de Estatística da UFRGS, disponível em http://goo.gl/xxuygf. Último acesso em 28/02/2018.
[16] Caso isso não seja claro, revise a lista de probabilidade de eventos.
[17] O que estamos calculando é a probabilidade da interseção, mas como os eventos são independentes, será justamente a multiplicação das probabilidades.
[18] Lembre-se, para que se observe exatamente uma cara são necessárias duas coroas nos outros dois lançamentos

Quando trabalhamos com variáveis aleatórias nós "esquecemos" do espaço amostral original e dos eventos para lidar apenas com a função $X(\cdot)$. Uma discussão sobre o *espaço de probabilidade induzido por uma variável aleatória* pode ser visto em Mittelhammer (2013) (seção 2.2.1) e aborda em profundidade os fundamentos que nos permitem fazer isso. No momento, estamos apenas nos apropriando deste fato.

O seguinte lema a respeito da f.m.p. de uma v.a. discreta foi retirado de Stern and Izbicki (2016).

Lema 2.5.11. *Seja X uma variável aleatória discreta e p_X a fmp de X. Seja χ os valores possíveis de X.*

- *Para todo $x \in \chi$, $0 \leq p_X(x) \leq 1$.*
- $\sum_{x \in \chi} p_X(x) = 1$.

Demonstração.

- $p_X(x) = P(X = x) = P(\{w \in \Omega : X(w) = x\})$. Lembre que $0 \leq P(\{w \in \Omega : X(w) = x\}) \leq 1$.
- $\sum_{x \in \chi} p_X(x) = \sum_{x \in \chi} P(\{w \in \Omega : X(w) = x\})$. Observe que, para $x_1 \neq x_2$, os eventos $\{w \in \Omega : X(w) = x_1\}$ e $\{w \in \Omega : X(w) = x_2\}$ são disjuntos. Portanto, pela aditividade enumerável,

$$\sum_{x \in \chi} p_X(x) = P(\cup_{x \in \chi} \{w \in \Omega : X(w) = x\}) = P(\Omega) = 1$$

□

Por fim, precisamos da definição de independência de variáveis aleatórias discretas. Em estatística, iremos na maioria dos casos assumir que temos v.a.'s independentes para que os resultados se mantenham.

Definição 2.5.12. Independência de v.a.'s[19] discretas (Retirado de Stern and Izbicki (2016))
Seja $X_1, \ldots, X_n$ variáveis aleatórias discretas. Dizemos que elas são independentes se, para todo $x_1, \ldots, x_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) := \mathbb{P}(\cap_{i=1}^n X_i = x_i) = \prod_{i=1}^n \mathbb{P}(X_i = x_i)$$

Portanto, para todo $x_1, \ldots, x_n$, $\{\omega \in \Omega : X_i(\omega) = x_i\}$ são conjuntamente independentes.

2.5.3 Distribuição de Poisson

A distribuição de Poisson é utilizada em problemas com ocorrência mais rara e a variável aleatória de interesse assume valores discretos, por exemplo, número de eclipses, número de fragmentos de meteoros que atingem a Terra em um ano, número de pessoas que fazem aniversários em um mesmo dia, clientes chegando em uma agência bancária, número de vezes que a Aisha conseguiu fazer uma baliza sem arranhar o carro, etc.

Definição 2.5.13. Seja X uma v.a. discreta, tomando valores em $\mathbb{N}$. Dizemos que X tem distribuição de Poisson se sua função massa de probabilidade (f.m.p.) é dada por:

$$p_x = \mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad k \in \{1, 2, 3, \cdots\} \tag{26}$$

Onde k representa o número de ocorrências de interesse e λ a taxa média de ocorrência. Utilizamos a notação $X \sim Poisson(\lambda)$.

[19] v.a. = variável aleatória.

Lema 2.5.14. *Se X tem distribuição de Poisson com parâmetro λ, então $\mathbb{E}[X] = \lambda$ e $Var[X] = \lambda$.*

Demonstração. A demonstração será feita no exercício 1. □

Precisamos verificar ainda que a equação 39 de fato é uma distribuição de probabilidade. Note que p_x satisfaz $\mathbb{P}(X = k) \geq 0 \quad \forall \quad k$, uma vez que as funções exponencial e fatorial são sempre positivas e $\lambda > 0$ por ser uma taxa. Resta verificar que a soma das probabilidades é 1:

$$\sum_{k=0}^{+\infty} \mathbb{P}(X = k) = \sum_{k=0}^{+\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \frac{\sum_{k=0}^{+\infty} \lambda^k}{k!} \overset{20}{=} e^{-\lambda}e^{\lambda} = 1$$

Para que a distribuição de Poisson seja apropriada para modelar um fenômeno X, as seguintes condições devem ser atendidas:

a. X é o número de ocorrências em um determinado intervalo de tempo e assume valores em $\{1, 2, 3, \cdots\}$;

b. as ocorrências acontecem de forma independente, isto é, o fato de um carro já ter passado em um pedágio não altera a probabilidade do próximo carro passar;

c. a taxa λ com que os eventos ocorrem deve ser constante;

d. dois eventos não podem ocorrer no mesmo instante;

e. a probabilidade de que um evento ocorra em um intervalo de tempo é proporcional ao comprimento deste intervalo.

Exemplos de situações onde pode-se utilizar a distribuição de Poisson incluem: número de carros que passam em uma praça de pedágio; número de erros de digitação por página de livro; número de meteoros com mais de 1m de diâmetro que se chocam contra a Terra no período de 1 ano; o número de ligações telefônicas em uma central de telemarketing em uma hora; etc.

Exemplo 2.5.15. (Retirado de Viali (2004c)) Em um certo tipo de fabricação de fita magnética, ocorrem defeitos a uma taxa de 1 a cada 2000 metros. Qual a probabilidade de que um rolo com 2000 metros de fita magnética:

a. Não tenha defeitos?

b. Tenha no máximo 2 defeitos?

c. Tenha pelo menos dois defeitos?

Solução:

Temos:

$\lambda = 1$, taxa de defeitos a cada 2000 metros.

X é número de defeitos a cada 2000 metros.

$x = 0, 1, 2, \cdots$ os possíveis valores de X.

Então:

$p_x = \mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, para $x = 0, 1, 2, \cdots$

[20] Da expanção de Taylor de e^x, temos que:

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$$

a. $\mathbb{P}(X=0) = \frac{e^{-1}(-1)^0}{0!} = e^{-1} \approx 0,3679$

b. $\mathbb{P}(X \leq 2) = \mathbb{P}(X=0) + \mathbb{P}(X=1) + \mathbb{P}(X=2) = \frac{e^{-1}(-1)^0}{0!} + \frac{e^{-1}(-1)^1}{1!} + \frac{e^{-1}(-1)^2}{2!} \approx 0,9197$

c. $\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X < 2) = 1 - [\mathbb{P}(X=0) + \mathbb{P}(X=1)] = 1 - \left(\frac{e^{-1}(-1)^0}{0!} + \frac{e^{-1}(-1)^1}{1!}\right) \approx 0,2642$

Derivação da Poisson a partir da Binomial

O seguinte exemplo foi adaptado do Khan Academy e pode ser acessado no link: https://goo.gl/zMAzu2.

Imagine que você trabalha em uma praça de pedágio e quer dimensionar a quantidade de funcionários, catracas, etc. Seu problema é, então, saber quantos carros vão passar em um determinado período de tempo pelo local. Mais do que isso, você gostaria de saber quais as probabilidades, por exemplo, de que passem apenas 5 carros ou mais do que 100 carros em uma hora.

Comece definindo uma variável que representa a quantidade de interesse, isto é, defina X: número de carros que passam na praça do pedágio em uma hora. Precisamos agora pensar em quem é a f.m.p. de X, para então podermos calcular as probabilidades de interesse.

Precisamos de duas suposições para poder chegar na distribuição de Poisson: a primeira é que não há diferença do fluxo de veículos dependendo da hora do dia[21] e a segunda é que não há relação entre os carros que passam. Isso significa que se um montão de carros passou em determinado horário, não implica que menos carros irão passar na hora seguinte. Em probabilidade, essa suposição quer dizer que os eventos são independentes.

Vamos pensar agora em quem seria a média de X e em qual unidade ela seria mensurada. Se estamos interessados em carros por hora, podemos imaginar que a média da variável aleatória X é um número real λ que indica o valor esperado de carros que passam por hora na praça de pedágio.

Agora vamos tentar utilizar o que conhecemos da distribuição Binomial para aplicar no problema dos carros. Sabemos que se Y tem distribuição Binomial, a sua média é dada pelo número de repetições (por exemplo, número de lançamentos de uma moeda) vezes a probabilidade de sucesso, que para a moeda poderia ser a probabilidade de sair "cara". Se denotarmos o número de repetições por n e a probabilidade de sucesso por p, então a média da Binomial é simplesmente np. Se no exemplo dos carros nossa média é λ carros por minuto, podemos dizer que isso é igual a 60 minutos vezes a probabilidade de que um carro passe em um dado minuto, que é $\lambda/60$:

$$\lambda \frac{\text{carros}}{\text{hora}} = \underbrace{60 \frac{\text{minutos}}{\text{hora}}}_{n} \underbrace{\frac{\lambda}{60} \frac{\text{carros}}{\text{minuto}}}_{p}$$

Logo, se estamos modelando o fenômeno dos carros utilizando uma Binomial(n,p), a probabilidade de que exatamente k carros passem em uma hora será:

$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{60}{k} \left(\frac{\lambda}{60}\right)^k \left(1 - \frac{\lambda}{60}\right)^{60-k} \tag{27}$$

[21] É claro que essa é uma suposição bastante "forte" e obviamente existem mais veículos às 13h do que as 2h da manhã, mas para o exemplo vamos supor que o fluxo seja o mesmo independente do horário.

O que parece uma boa aproximação, exceto pelo fato de que se mais de um carro passar em um minuto, teríamos um problema. Observe que, como utilizamos $n = 60$, isso significa que a equação 27 comportaria no máximo 60 sucessos, isto é, um carro por minuto. Mas é perfeitamente possível que passem mais carros por minuto. Uma solução possível é dividir λ por segundos e reescrevendo 27 para fornecer a probabilidade de que um carro passe em um segundo:

$$\mathbb{P}(X = k) = \binom{3600}{k} \left(\frac{\lambda}{3600}\right)^k \left(1 - \frac{\lambda}{3600}\right)^{3600-k} \tag{28}$$

Mas ainda assim, como a praça de pedágio irá ter várias catracas, é possível que mais de um carro passe por segundo. O que precisamos fazer, então, é avaliar o que ocorre com a primeira parte da equação 27 quando $n \to +\infty$. Antes disso, vamos fazer uma pequena revisão matemática de limites. Vamos provar este resultado:

$$\lim_{x\to+\infty} \left(1 + \frac{a}{x}\right)^x = e^a \tag{29}$$

Definimos $\frac{1}{n} = \frac{a}{x}$, o que implica que $x = na$. Observe que se $x \to +\infty$, o mesmo ocorre com n, pois $n = \frac{x}{a}$. Fazendo as devidas substituições em 29, temos:

$$\lim_{n\to+\infty} \left(1 + \frac{1}{n}\right)^n a = \lim_{n\to+\infty} \left(\left(1 + \frac{1}{n}\right)^n\right)^a = \left(\underbrace{\lim_{n\to+\infty} \left(1 + \frac{1}{n}\right)^n}_{e}\right)^a = e^a \tag{30}$$

Para ver a definição do número e via limite, veja: https://goo.gl/qsf9pZ.

O segundo resultado matemático que iremos usar envolve fatoriais:

$$\frac{x!}{(x-k)!} \overset{22}{=} x(x-1)(x-2)\cdots(x-k+1) \tag{31}$$

Agora, vamos aplicar o limite na primeira parte da expressão 27:

$$\mathbb{P}(X = k) = \lim_{n\to+\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \lim_{n\to+\infty} \frac{n!}{k!(n-k)!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \tag{32}$$

Utilizando o resultado visto em 31 e trocando de posição os denominadores das duas primeiras frações, teremos:

$$\mathbb{P}(X = k) = \lim_{n\to+\infty} \frac{\overbrace{n(n-1)(n-2)\cdots(n-k+1)}^{k \text{ termos}}}{n^k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \tag{33}$$

Observe que a quantidade $\frac{\lambda^k}{k!}$ não depende de n e pode "sair" do limite. Vamos também utilizar o fato que $\lim_{x\to a} f(x)g(x) = \lim_{x\to a} f(x) \lim_{x\to a} g(x)$ (o limite do produto é o produto dos limites) e reescrever 33 como:

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} \underbrace{\lim_{n\to+\infty} \frac{\overbrace{n(n-1)(n-2)\cdots(n-k+1)}^{\text{polinômio do tipo } n^k + \text{outras coisas}}}{n^k}}_{1} \underbrace{\lim_{n\to+\infty} \left(1 - \frac{\lambda}{n}\right)^n}_{e^{-\lambda}} \underbrace{\lim_{n\to+\infty} \left(1 - \frac{\lambda}{n}\right)^{-k}}_{1} \tag{34}$$

De onde temos:

[22] Para um exemplo com números, veja o vídeo disponível em https://goo.gl/zMAzu2 a partir dos 10 minutos de gravação.

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{35}$$

que é justamente a expressão dada em 39. Observe que embora a distribuição de Poisson pareça abstrata, ela nada mais é do que o limite de uma distribuição Binomial quando $n \to \infty$ e $p = \frac{\lambda}{n}$.

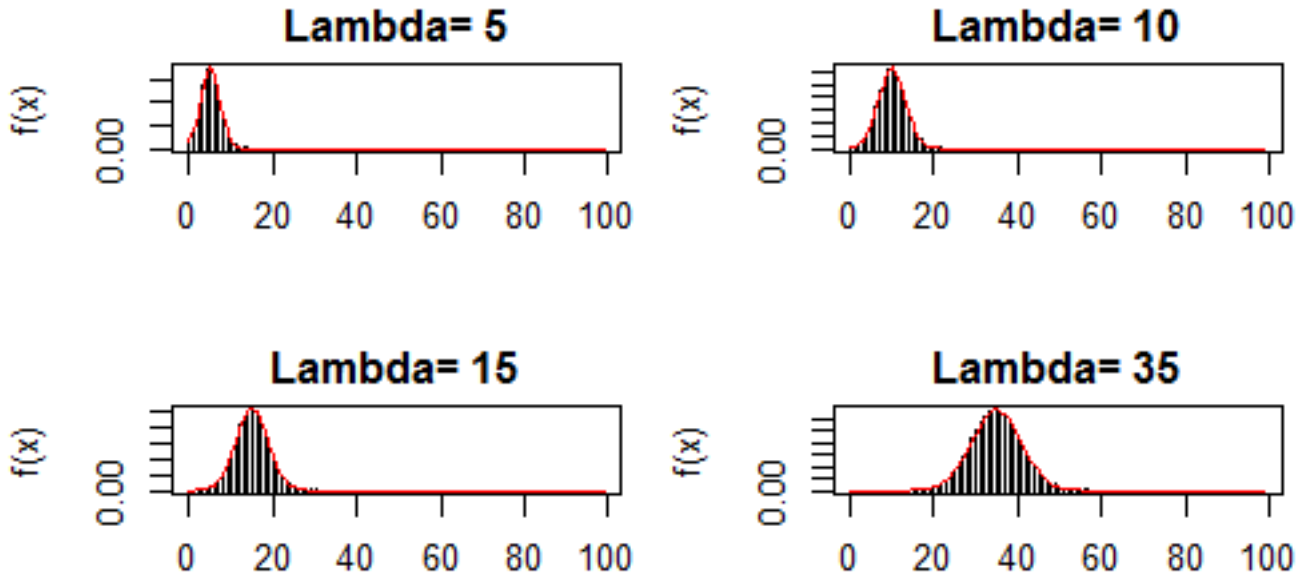
Aproximação da Poisson para a Normal

Assim como a Binomial pode ser aproximada por uma distribuição Normal, podemos aproximar a Poisson para valores suficientemente grandes de λ para uma distribuição normal. Se $X \sim Poisson(\lambda)$, podemos aproximá-la por uma variável $Y \sim Normal(\lambda, \sqrt{\lambda})$. Utilizando a padronização da Normal, teremos que:

$$\mathbb{P}(X \geq x) \approx \mathbb{P}(Y > \underbrace{x - 0,5}_{\text{correção de continuidade}}) = \mathbb{P}\left(Z > \frac{(x - 0,5) - \lambda}{\sqrt{\lambda}}\right) \tag{36}$$

Na figura 3 observamos a aproximação utilizando $\lambda = \{5, 10, 15, 35\}$. O histograma preto representa a f.m.p. da Poisson para valores de 1 a 100 e a linha vermelha a distribuição Normal com média λ e desvio padrão igual a $\sqrt{\lambda}$.

Figura 3: Aproximação da Poisson pela Distribuição Normal para diferentes valores de λ



O exemplo 2 foi retirado de https://onlinecourses.science.psu.edu/stat414/node/180 e mostra uma aplicação desta aproximação.

Exemplo 2.5.16. 2 O número anual de terremotos com pelo menos 2.5 pontos na Escala Richter e com epicentro de até 40 milhas do centro de Memphis segue uma distribuição de Poisson com $\lambda = 6.5$. Qual a probabilidade de que pelo menos 9 terremotos desses atinjam a região no próximo ano?

Solução: Poderíamos utilizar a f.m.p. da Poisson para obter a probabilidade exata:

$$\mathbb{P}(X \geq 9) = 1 - \mathbb{P}(X \leq 8) = 1 - [\mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \cdots + \mathbb{P}(X = 8)] \approx 0,208$$

Usando a normal, obtemos o seguinte resultado:

$$\mathbb{P}(X \geq 9) \approx \mathbb{P}(Y \geq 9 - 0,5) = \mathbb{P}\left(Z \geq \frac{8,5 - 6,5}{\sqrt{6,5}}\right) = \mathbb{P}(Z \geq 0,78) \approx 0,218$$

O nosso erro, neste exemplo, fica na segunda casa decimal. Quanto maior o valor de λ, menor será esse erro de aproximação.

2.5.4 Principais distribuições discretas

O seguinte resumo foi retirado de Stern and Izbicki (2016).

1. **Variável Aleatória Binomial** - $X \sim \text{Binomial}(n, p)$.

   X conta o número de sucessos em n experimentos independentes (Bernoulli), cada um com probabilidade de "sucesso" p. Para uma animação sobre a distribuição Binomial, acesse: https://istats.shinyapps.io/BinomialDi

2. **Variável Aleatória Geométrica** - $X \sim \text{Geom}(p)$.

   Considere uma série de experimentos independentes (Bernoulli), cada um com probabilidade de "sucesso" p. X é o número de experimentos até o primeiro "sucesso".

3. **Variável Aleatória Binomial Negativa** - $X \sim \text{Binomial Negativa}(r, p)$.

   Considere uma série de experimentos independentes (Bernoulli), cada um com probabilidade de "sucesso" p. X é o número de experimentos até obtermos r "sucessos". Por definição, Binomial Negativa$(1, p)$ é o mesmo que $Geom(p)$.

4. **Variável Aleatória Hipergeométrica** - $X \sim \text{Hipergeométrica}(N, n, k)$.

   Uma amostra de tamanho n é escolhida (sem respoição) de um grupo de tamanho N que tem k "sucessos" (e $N - k$ "falhas"). X é o número de "sucessos" na amostra.

5. **Variável Aleatória Poisson** - $X \sim \text{Poisson}(\lambda)$.

   A família de distribuições de Poisson frequentemente dá um bom modelo para o número de eventos (em particular, eventos raros) que ocorrem num período de tempo fixo (ou outra unidade fixa). Por exemplo, o número de clientes chegando em uma hora, número de reivindicações de seguro em um mês, número de terremotos num ano, número de erros de digitação numa página. λ é o número médio de eventos.

Exercícios:

Exercício 28. (Retirado de Stern and Izbicki (2016))
Seja X uma v.a. tal que $X \in \mathbb{N}$ e $p_X(i) = c \cdot 2^{-i}$. Encontre o valor de c.

Resposta: Exercício resolvido na aula, $c = 1/2$.

Exercício 29. (Retirado de Stern and Izbicki (2016))
Considere que você lança 3 vezes uma moeda com probabilidade de cara igual a p. Seja X a variável aleatória que representa o número de caras observadas. Encontre a $f.m.p.$ de X.

Exercício feito em aula.

Exercício 30. (Retirado de Stern and Izbicki (2016))
Você lança duas vezes um dado de quatro lados. X_1 e X_2 são v.a.'s que denotam os resultados do primeiro e segundo

lançamento, respectivamente. Seja $Y = X_1 + X_2$. Encontre a f.m.p. de Y e de $2X_1$. Observe que a f.m.p. da soma dos dois lançamentos é diferente da f.m.p. obtida ao multiplicar o resultado do primeiro lançamento por 2.

Ainda sem gabarito.

Exercício 31. (Retirado de Stern and Izbicki (2016))
Sejam X e Y duas variáveis aleatórias discretas independentes. Verifique se e^X e e^Y também são independentes.

Sim.

Exercício 32. (Retirado de Stern and Izbicki (2016))
Uma moeda com probabilidade p de cara é lançada 5 vezes. Qual a probabilidade de se obter 3 caras?

Resposta: $10p^3(1-p)^2$**.**

Exercício 33. (Retirado de Stern and Izbicki (2016))
Uma moeda honesta é lançada 7 vezes. Calcule a probabilidade de se obter 1 ou mais caras.

Resposta: Genericamente (usando como probabilidade p**), temos** $1-(1-p)^7$**.**

Exercício 34. (Retirado de Stern and Izbicki (2016))
Considere que uma caixa contém $\theta \geq 1$ bolas laranjas e $p \geq 1$ bolas rosas. Considere que duas bolas são removidas sem reposição da caixa. Seja X o número de bolas rosas em duas retiradas. X tem distribuição binomial? Se sim, informe os parâmetros.

Não.

Exercício 35. (Retirado de Viali (2004a))
Se $X \sim Poisson(\alpha)$ e $\mathbb{P}(X = 0) = 0,2$, calcule $\mathbb{P}(X > 2)$.

Ainda sem gabarito

Exercício 36. (Adaptado de Schmidt (2011))
Suponha que o número de Pokémons que uma pessoa captura utilizando o novo *app Pokémon GO* seja modelado por uma variável aleatória com distribuição de Poisson com média igual a 2 pokémons por semana. Adicionalmente, suponha o uso do item *incenso* aumente o número esperado de pokémons capturados para 4 por semana, mas ele somente funciona para 80% dos jogadores, sem ter efeito nos 20% restantes. Suponha ainda que todos os jogadores em uma dada semana estão utilizando o incenso. Julgue as seguintes afirmativas como verdadeiras ou falsas, justificando suas respostas:

a. A probabilidade de um indivíduo não se beneficia do incenso capturar 2 Pokémons na semana é de $4e^{-2}$.

b. A probabilidade de um indivíduo que se beneficia do incenso caputar dois Pokémons em uma semana é de $8e^{-2}$.

c. A probabilidade de um indivíduo que não se beneficia do incenso pegar no máximo dois pokémons em uma semana é de $13e^{-4}$.

d. Suponha que um jogador escolhido aleatoriamente tenha capturado dois pokemóns na semana em que ele utilizou o incenso. A probabilidade dele fazer parte do grupo de jogadores que se beneficiam dos efeitos do item é de $(1+e^{-4})^{-1}$

Exercício 37. (Adaptado de Ross (2010)) Suponha que o número de capivaras que atravessam determinada rua possa ser modelado por uma distribuição de Poisson com parâmetro $\lambda = 3$.

a. Encontre a probabilidade de que 3 ou mais capivaras atravessem a rua hoje.

b. Repita o item (a) supondo que ao menos uma capivara já atravessou a rua hoje.

Exercício 38. 3 (adaptado de Schmidt (2011)) Seja Y_i, $i = 1, 2, \cdots, n$ uma v.a. tal que $Y_i = 1$ com probabilidade $1 - p$ e $Y_i = 0$ com probabilidade p. Defina $X = \sum_{i=1}^{n} Y_i$. Responda se cada uma das afirmativas abaixo é verdadeira, falsa ou indeterminada, justificando suas respostas:

a. Y_i, $i = 1, 2, \cdots, n$ possui distribuição Poisson com parâmetro $\lambda = np$.

b. X possui distribuição binomial com parâmetros n e p.

c. A distribuição assintótica de X é Poisson(λ), com $\lambda = np$.

d. $Var[Y_i] = \frac{Var[X]}{n} = (1 - p)p$.

e. Se $n \to +\infty$ e p permanece fixo, então $\frac{X - np}{\sqrt{np(1-p)}}$ converge para a distribuição normal padrão.

f. $E[Y_i^2] = p^2$

g. Se $\bar{X} = \frac{X}{n}$, então $E[\bar{X}] = p$.

2.5.5 Variáveis aleatórias contínuas

Definição 2.5.17. Variável aleatória contínuas (Retirado de James (2010))
A variável aleatória X é dita contínua se existe uma função $f(x) \geq 0$ tal que:

$$F_X(x) = \int_{\infty}^{x} f(t)dt, \quad \forall\, x \in \mathbb{R}$$

As variáveis aleatórias contínuas são tais que a função densidade em um ponto é igual a zero.

Uma definição alternativa é (retirado de Stern and Izbicki (2016)):

Definição 2.5.18. Seja X uma variável aleatória contínua. Denotamos a função de densidade de probabilidade de X por $f_X : \mathbb{R} \to \mathbb{R}$. Ela satisfaz as seguintes propriedades:

1. $f_X(x) \geq 0$.
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$.
3. $\int_a^b f_X(x)dx = \mathbb{P}(a \leq X \leq b)$.

Como propriedade, temos que as probabilidades de uma variável aleatória contínua são as integrais sob a curva $f(\cdot)$ em determinados intervalos.

Definição 2.5.19. f.d.a. de uma v.a. contínua (Retirado de Mittelhammer (2013))
A função distribuição acumulada de uma variável aleatória X, se X for contínua, é dada por

$$F_X(x) = \int_{-\infty}^{x} f(x)dx,\ x \in (-\infty, +\infty)$$

Lema 2.5.20. *Seja X uma variável aleatória contínua com a função de distribuição acumulada F_X. Para $b \geq a$, $F_X(b) - F_X(a) = \mathbb{P}(a \leq X \leq b)$.*

Demonstração. Usando a definição de fda

$$\begin{aligned} F(b) - F(a) &= \int_{-\infty}^{b} f_X(x)dx - \int_{-\infty}^{a} f_X(x)dx \\ &= \int_{a}^{b} f_X(x)dx = \mathbb{P}(a \leq X \leq b) \end{aligned}$$

□

As funções densidade/massa e distribuição acumulada tem uma relação direta, como é possível perceber nas definições, tanto para o caso discreto como para o caso contínuo. Os próximos dois resultados formalizam essa relação e foram retirados de Mittelhammer (2013):

Teorema 2.5.21. *Sejam $x_1 < x_2 < x_3$ um conjunto contável de resultados possíveis da variável aleatória discreta X. Então, a função massa de probabilidade de X pode ser definida por:*

$$\begin{aligned} f(x_1) &= F(x_1) \\ f(x_i) &= F(x_i) - F(x_{i-1}),\ i = 1, 2, 3, \ldots \\ f(x) &= 0 \ \textit{para } x \textit{ que não esteja na imagem de } X \end{aligned}$$

Demonstração. A demonstração segue diretamente das definições das funções massa e acumulada de variáveis aleatórias. □

Teorema 2.5.22. *Sejam $f(x)$ e $F(x)$ as f.d.p. e f.d.a. de uma variável aleatória contínua X. A função densidade de X pode ser definida como*

$$f(x) = \frac{d}{dx}[F(x)]$$

em todo ponto onde $f(x)$ é contínua e será igual a zero em todos outros pontos.

Demonstração. Pelo teorema fundamental do cálculo, temos que:

$$\frac{dF(x)}{dx} = \frac{d\int_{-\infty}^{x} f(t)dt}{dx} = f(x)$$

Sempre que $f(x)$ for contínua, o que prova a primeira parte do teorema.

Uma vez que X é uma v.a. contínua, então $\mathbb{P}(X \leq b) = F(b) = \int_{-\infty}^{b} f(x)dx$ está definida para qualquer valor de b. Como mudar o valor deste integrando nos pontos de descontinuidade não irá afetar o valor de $F(b)$, então $f(x)$ pode ser definida de forma arbitrária nestes pontos[23]. □

2.5.6 Principais modelos contínuos

1. **A Variável Aleatória Uniforme** - $X \sim \mathrm{U}(a, b)$.

   Todos os subconjuntos de (a, b) com o comesmo comprimento são equiprováveis. A distribuição é normalmente usada quando todos os pontos em (a, b) são "igualmente prováveis".

[23] Uma discussão mais formal a respeito disso está na página 67 de Mittelhammer (2013)

2. **A Variável Aleatória Exponencial** - $X \sim \text{Exp}(\lambda)$.

   Este distribuição é usada, frequentemente, para modelar o tempo até um certo evento ocorrer. É a única distribuição contínua com propriedade de perda de memória.

3. **A Variável Aleatória Gama** - $X \sim \text{Gama}(k, \lambda)$.

   A Gama(1, λ) é uma Exponencial(λ) e, assim, a distribuição Gama é a generaliação da Exponencial. Se k é um número natural, $X = \sum_{i=1}^{k} Y_i$, onde Y_i são variáveis aleatórias independentes e $Y_i \sim \text{Exponencial}(\lambda)$.

4. **A Variável Aleatória Beta** - $X \sim \text{Beta}(\alpha, \beta)$.

   Como a distribuição Beta assume valores em $(0, 1)$, ela é frequentemente usada para modelar frequências e razões.

5. **A Variável Aleatória Normal (Gaussiana)** - $X \sim \text{N}(\mu, \sigma^2)$.

   Usando o Teorema do Limite Central, a distribuição Normal é frequentemente usada para aproximar a distribuição da média de uma sequência de variáveis aleatórias independentes e identicamente distribuídas. Para uma animação sobre a distribuição Normal padrão, acesse: https://istats.shinyapps.io/NormalDist/.

6. **A variável aleatória t** - $X \sim \text{t}_p$

   A distribuição t recebe este nome em homenagem ao mestre cervejeiro William Gosset, que publicou seus achados utilizando o pseudônimo de Student[24]. A distribuição t com p graus de liberdade é definida como o quociente de uma normal padrão pela raiz de uma qui-quadrado com p graus de liberdade, onde o numerador e o denominador são independentes.

   Note que a estatística

   $$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

   xom S sendo o desvio padrão amostral, definido como $S = \sqrt{S^2}$, e $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$, segue uma distribuição t com $n - 1$ graus de liberdade, pois podemos reescrevê-la como:

   $$\frac{(\bar{X} - \mu)/(\sigma/n)}{\sqrt{S^2/\sigma^2}}$$

   O numerador é de fato uma normal padrão e o denominador é $\sqrt{\chi^2_{n-1}/(n-1)}$ e é independente do numerador. Para uma demonstração sobre a distribuição do denominador, recomendo ver Casella and Berger (2002). Para uma animação em R sobre a distribuição t, acesse: https://istats.shinyapps.io/tdist/. A distribuição t converge para a distribuição normal (isso significa que os valores ficam cada vez mais próximos) quando a amostra aumenta. Como regra de bolso, costuma-se dizer que para valores de n acima de 30, é indiferente utilizar a normal padrão ou a t. Porém, para valores amostrais pequenos, a t costuma ter mais área nas suas caudas e por isso às vezes usamos a expressão "caudas gordas" para nos referir a ela. A Cauchy é um caso particular da t: t com 1 grau de liberdade (ou seja, $n = 2$) é uma Cauchy.

7. **A variável aleatória χ^2 (qui-quadrado)** - $X \sim \chi_p$

[24] Para saber mais sobre essa história - depois que passar o semestre - veja: https://priceonomics.com/the-guinness-brewer-who-revolutionized-statistics/.

Se $X \sim \mathcal{N}(0,1)$, então $Y = X^2$ terá distribuição qui-quadrado com 1 grau de liberdade. Além disso, se $W \sim Gamma(\alpha, \beta)$ com $\alpha = p/2$ (p inteiro) e $\beta = 2$, então $W \sim \chi^2(p)$. Assim como a t, a qui-quadrado tem apenas um parâmetro, que é o número de graus de liberdade. Para uma animação sobre a qui-quadrado, acesse: https://istats.shinyapps.io/ChisqDist/.

8. **A variável aleatória com distribuição F de Snedecor** - $X \sim F_{p,q}$.

   Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população com distribuição $\mathcal{N}(\mu_X, \sigma_X^2)$ e $Y1, \ldots, Y_m$ uma amostra aleatória de uma população com $\mathcal{N}(\mu_Y, \sigma_Y^2)$. A variável aleatória $F = (S_X^2/\sigma_X^2)/(S_Y^2/\sigma_Y^2)$ tem distribuição *F de Snedecor* com $n-1$ e $m-1$ graus de liberdade. Para uma animação da distribuição F, acesse: https://istats.shinyapps.io/FDist/.

   O seguinte teorema é útil:

   Teorema 2.5.23. *a) Se $X \sim F_{p,q}$, então $1/X \sim F_{q,p}$; isto é, o recíproco de uma v.a. com distribuição F também tem distribuição F, invertendo os graus de liberdade do numerador e do denominador;*

   b) Se $X \sim t_q$, então $X^2 \sim F_{1,q}$ - uma variável aleatória com distribuição t quando elevada ao quadrado é uma variável aleatória com distribuição F com 1 grau de liberdade no numerador e q graus de liberdade no denominador;

   c) Se $X \sim F_{p,q}$, então $(p/q)X/(1+(p/q)X) \sim Beta(p/2, q/2)$.

2.6 Distribuição Normal (retirado de Stern and Izbicki (2016))

Definição 2.6.1. Dizemos que X tem distribuição Normal com parâmetros (μ, σ^2) e denotamos por $X \sim \mathrm{N}(\mu, \sigma^2)$ se a densidade de X é

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Se $\mu = 0$, e $\sigma = 1$, dizemos que X tem distribuição normal padrão.

Definição 2.6.2. Seja $Z \sim \mathrm{N}(0,1)$.

$$\phi(z) = \mathbb{P}(Z \leq z) = F_Z(z)$$

ϕ não tem solução analítica. As duas formas de obter valores aproximados de ϕ são:

1. Muitas linguagens de programação têm livrarias estatísticas com estes valores. Por exemplo, R, python, C, ...
2. Consultar uma tabela da normal padrão, disponível em muitos livros de Estatística.

Lema 2.6.3. *Se $X \sim N(\mu, \sigma^2)$, então $\frac{(X-\mu)}{\sigma} \sim N(0,1)$.*

Demonstração. Seja $Z = \frac{(X-\mu)}{\sigma}$.

$$\begin{aligned} F_Z(z) = \mathbb{P}(Z \leq z) &= \mathbb{P}\left(\frac{(X-\mu)}{\sigma} \leq z\right) \\ &= \mathbb{P}(X \leq \sigma z + \mu) = F_X(\sigma z + \mu) \end{aligned}$$

Assim,

$$
\begin{aligned}
f_Z(z) = \frac{\partial F_Z(z)}{\partial z} &= \frac{\partial F_X(\sigma z + \mu)}{z} && \text{(Lema ??)} \\
&= \frac{\partial F_X(\sigma z + \mu)}{\partial(\sigma z + \mu)} \cdot \frac{\partial(\sigma z + \mu)}{\partial z} && \text{Regra da Cadeia} \\
&= f_X(\sigma z + \mu) \cdot \sigma \\
&= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\sigma z + \mu - \mu)^2}{\sigma^2}} \cdot \sigma = \frac{1}{\sqrt{2\pi}} e^{-z^2}
\end{aligned}
$$

□

Segue do Lema 2.6.3 que qualquer parametrização da distribuição normal pode ser obtida com mudança apropriada de escala da normal padrão. A Figura 4 apresenta a densidade da distribuição normal padrão.

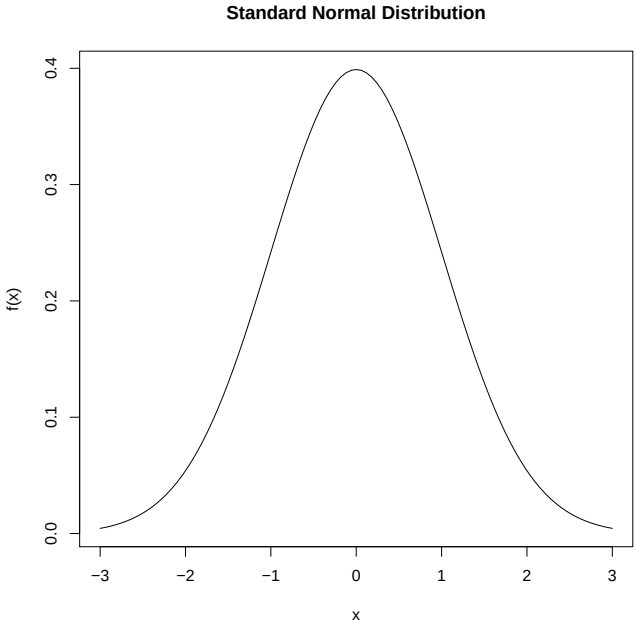


Figura 4

Lema 2.6.4. *Se* $X \sim N(\mu, \sigma^2)$*, então* $\mathbb{E}[X] = \mu$ *e* $Var[X] = \sigma^2$.

Demonstração. Considere que $X \sim \mathrm{N}(0, 1)$.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}} dx$$

Como xe^{-x^2} é uma função impar, $\mathbb{E}[X] = 0$. **Observação:** Na aula vimos como calcular essa integral, quem precisar, pode pedir pra Aisha.

Lembramos que $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Como $\mathbb{E}[X] = 0$, $Var[X] = \mathbb{E}[X^2]$.

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}} dx \\
&= -\frac{1}{2\sqrt{2\pi}} x e^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx && \text{integração por partes} \\
&= 0 + 1 && \text{fdp da normal padrão (Definições, 2.5.18 2.6.1)}
\end{aligned}$$

A seguir, considere que $X \sim \mathrm{N}(\mu, \sigma^2)$,

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\sigma \cdot \frac{X-\mu}{\sigma} + \mu\right] \\
&= \sigma \cdot \mathbb{E}\left[\frac{X-\mu}{\sigma}\right] + \mu \\
&= \sigma \cdot 0 + \mu = \mu && \text{(Lema 2.6.3)}
\end{aligned}$$

$$\begin{aligned}
Var[X] &= Var\left[\sigma \cdot \frac{X-\mu}{\sigma} + \mu\right] \\
&= \sigma^2 \cdot Var\left[\frac{X-\mu}{\sigma}\right] \\
&= \sigma^2 \cdot 1 = \sigma^2 && \text{(Lema 2.6.3)}
\end{aligned}$$

□

Lema 2.6.5 (Aproximação da Binomial e Teorema do Limite Central)**.** *Considere que $X \sim Binomial(n, p)$ e que n é "grande" ($n = 30$ frequentemente é suficiente para um bom resultado).*

$$\mathbb{P}\left(\frac{X - np}{\sqrt{np(1-p)}} \leq z\right) \approx \phi(z)$$

Esta é uma instância específica de uma regra mais genérica. Se $X_1, X_2, \ldots$ é uma sequência de variáveis aleatórias independentes com a mesma distribuição, $\mathbb{E}[X_i] = \mu$ e $Var[X_i] = \sigma^2$, então, para n grande,

$$\mathbb{P}\left(\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma} \leq z\right) \approx \phi(z)$$

2.6.1 Distribuição Exponencial

A distribuição exponencial é utilizada para modelagem de fenômenos onde a variável de interesse assume valores estritamente positivos com decaimento exponencial, como por exemplo tempo até a falha de equipamentos, tempo médio de vida de medicamentos, etc. Muitas distribuições acabam se relacionando com a Exponencial, como a distribuição Gamma, a Normal e outras.

Caracterização

Definição 2.6.6. Seja X uma v.a.[25] contínua, tomando valores em $\mathbb{R}_+^*$. Dizemos que X tem distribuição Exponencial se sua função densidade de probabilidade (f.d.p.[26]) é dada por:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{se } x > 0. \\ 0 & \text{caso contrário.} \end{cases} \tag{37}$$

Observe que a equação 37 de fato representa uma densidade pois sua integral no domnínio onde está definida é 1:

Demonstração.

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{0}^{+\infty} \lambda e^{-\lambda x} dx$$

Resolvendo a integral imprópria:

$$\int \lambda e^{-\lambda x} dx \Rightarrow \begin{cases} u & = -\lambda x \\ du & = -\lambda dx \end{cases}$$

Fazendo a integral por substituição:

$$-\int e^{\overbrace{-\lambda x}^{u}} \underbrace{(-\lambda)dx}_{du} = -\int e^u du = -e^u = -e^{-\lambda x}$$

Por fim, avaliamos o resultado nos limites de integração original:

$$-e^{-\lambda x}\Big|_{x=0}^{x=+\infty} = -e^{-\infty} - (-e^0) = 0 - (-1) = +1$$

□

Lema 2.6.7. *Se X tem distribuição Exponencial com parâmetro λ, então $\mathbb{E}[X] = \frac{1}{\lambda}$ e $Var[X] = \frac{1}{\lambda^2}$.*

Demonstração. A demonstração fica como exercício. □

Uma outra forma de parametrizar a distribuição exponencial, bastante utilizada na literatura, é escrevendo 37 com $\lambda = \frac{1}{\beta}$, de forma que agora seu valor esperado é β e a variância é dada por β^2. Neste caso, escrevemos 37 como:

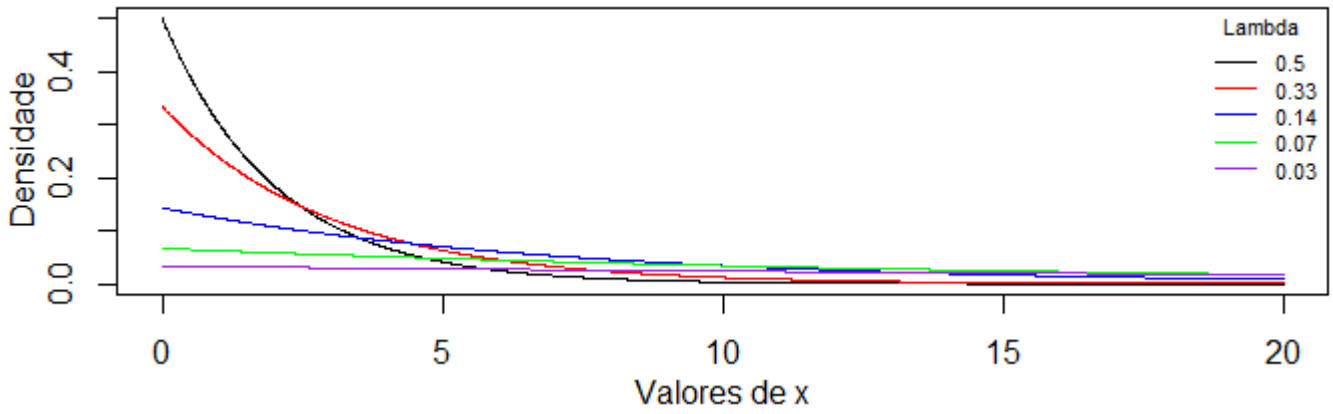
$$f(x) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & \text{se } x > 0. \\ 0 & \text{caso contrário.} \end{cases} \tag{38}$$

A figura 5 tem o gráfico da distribuição exponencial avaliado no intervalo $x \in (0, 20)$ para diferentes valores de λ. É possível observar que quanto menor o parâmetro λ, mais "achatada"será a distribuição. Já para valores maiores, a distribuição fica mais inclinada, tendendo a zero mais rapidamente.

[25] v.a. = variável aleatória

[26] Em inglês, a f.d.p. é chamada de p.d.f. (*Probability Density Function*).

Figura 5: Função densidade da distribuição exponencial para diferentes valores de λ



Lema 2.6.8. *Se $X \sim Exp(\lambda)$, então sua função densidade acumulada (fda) é dada por $F_X(x) = 1 - e^{-\lambda x}$, se $x > 0$.*

Demonstração. A demonstração fica de exercício. □

Observe que como a $F_X(x)$ representa, por definição, $\mathbb{P}(X \leq x)$, então a probabilidade de seu complementar será simplesmente

$$\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - F_X(x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}$$

Que é uma quantidade de interesse em muitas aplicações, como por exemplo, a probabilidade de que determinada lâmpada dure mais que 5 anos.

Lema 2.6.9. ***(Falta de memória da distribuição exponencial)***

Se $X \sim Exp(\lambda)$, então X não tem memória. Isto é, para $t, s > 0$,

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s)$$

Demonstração. Para a demonstração, vamos utilizar a definição de probabilidade condicional, isto é: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B}{\mathbb{P}(B)}$.

$$\begin{aligned}
\mathbb{P}(X > t + s | X > t) =& \frac{\mathbb{P}(X > t + s \cap X > t)}{\mathbb{P}(X > t)} \\
=& \frac{\mathbb{P}(X > t + s)}{\mathbb{P}(X > t)} \\
=& \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} \\
=& e^{-\lambda(t)} e^{-\lambda(s)} e^{+\lambda(t)} \\
=& e^{-\lambda(t)+\lambda(t)-\lambda(s)} \\
=& e^{-\lambda(s)}
\end{aligned}$$

□

Essa propriedade é mais facilmente compreendida em termos de "tempo de espera". Suponha que o tempo até um cliente chegar em uma agência bancária tem distribuição exponencial de parâmetro λ. A probabilidade de um cliente chegar em

30 *minutos após a abertura da agência ou* 30 *minutos após as primeiras três horas em que ela está em funcionamento é a mesma. Isso significa que na prática, se queremos saber a probabilidade de um evento ocorrer nas próximas* s *unidades de tempo, obteremos o mesmo valor se considerarmos que já se passaram* t *unidades de tempo ou começamos do zero.*

As únicas duas distribuições que apresentam falta de memória são a geométrica e a exponencial. É possível demonstrar, inclusive, que se uma variável aleatória contínua apresenta falta de memória, necessariamente ela segue uma distribuição exponencial (para uma discussão mais detalhada, veja Meyer (1973)).

Relação entre a Exponencial e a Poisson

A distribuição de Poisson apresenta a probabilidade de ocorrência de um evento discreto em um intervalo contínuo de tempo t[27] que apresenta λ ocorrências médias. Sua f.m.p. é dada por:

$$p_x = \mathbb{P}(X = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \quad k \in \{1, 2, 3, \cdots\} \tag{39}$$

Observe que se queremos saber a probabilidade de **não** haver ocorrências no intervalo, queremos $P(X = 0)$, temos:

$$p_x = \mathbb{P}(X = 0) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t} \tag{40}$$

E isso é $P(X > t)$, como mostrado no lema 2. Isso significa que, se temos uma variável aleatória de Poisson, a probabilidade de não haverem ocorrências em um intervalo de tempo t será calculada utilizando a distribuição exponencial e calculando a probabilidade de que o tempo até a próxima ocorrência seja maior que t.

Exemplo 2.6.10. (Retirado de Freeman et al. (2017)) Suponha que o número de carros que chegam a um lava-rápido durante uma hora seja descrito por uma distribuição de probabilidade de Poisson. Logo, a probabilidade de x carros chegarem em por hora é dado por:

$$\mathbb{P}(X = x) = \frac{10^x e^{\lambda}}{x!}$$

Uma vez que o número médio de carros que chegam ao lava-rápido é 10 por hora, o tempo médio entre a chegada de cada carro é $\frac{1 \text{ hora}}{10 \text{ carros}} = 0,1$ hora/carro. Desse modo, a distribuição exponencial correspondente que descreve o tempo entre as chegadas dos carros tem uma média de $0,1$ hora por carro. Como a média da exponencial é dada por $\frac{1}{\text{parâmetro}}$, temos uma exponencial de parâmetro $\mu = \frac{1}{0,1} = 10$[28]. A f.d.p. correspondente será dada por:

$$f(x) = 10e^{-10x} \tag{41}$$

Exemplo 2.6.11. (Retirado de Viali (2004b)) Suponha que um componente eletrônico tenha um tempo de vida T (em unidades de 1000 horas) que segue uma distribuição exponencial de parâmetro $\lambda = 1$. Suponha que o custo de fabricação seja R$2,00 e o preço de venda R$5,00. O fabricante garante total devolução do dinheiro se o aparelho estragar antes de 900 horas. Qual o lucro esperado por item vendido?

Solução: Neste caso, tem-se que a f.d.p. da v.a. será dada por:

$$f(x) = \lambda^x e^{-\lambda x} = 1^x e^{-1x} = e^{-x}$$

Logo, se queremos $\mathbb{P}(T < 0,9)$, temos:

[27] Até então tratamos a Poisson como tendo parâmetro λ, considerando que $t = 1$ unidade de tempo. Em 39 consideramos que ainda há a média λ, mas agora por cada t unidades de tempo sendo que t assume valores em R_+^*

[28] Aqui utilizamos μ para representar o parâmetro da exponencial para não confundir com a Poisson.

$$\int_0^{0,9} e^{-x}dx = -e^{-x}\Big|_{x=0}^{x=0,9} = -e^{-(0,9)} - (-e^0) = 1 - e^{-0,9} = 0,5934 \tag{42}$$

O lucro do fabricante, por unidade, é dado por uma variável aleatória Bernoulli, onde o seu lucro é -R$2,00 com probabilidade $0,5934$ e R$3,00 com probabilidade $1 - 0,5934 = 0,4066$. Logo, seu lucro esperado será:

$$\mathbb{E}[lucro] = -2(0,5934) + 3(0,4066) = 0,03$$

2.7 Distribuição Gama (retirado de Stern and Izbicki (2016))

Definição 2.7.1. $\Gamma : \mathbb{R}^+ \to \mathbb{R}^+$ é chamada de função Gama e é tal que:

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$

Lema 2.7.2. *A função Gama ($\Gamma(\cdot)$) satisfaz as seguintes propriedades:*

1. *Para $a \geq 1$, $\Gamma(a) = (a-1)\Gamma(a-1)$*
2. *Se $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$*

Demonstração.

1.

$$\begin{aligned}
\Gamma(a) &= \int_0^\infty t^{a-1}e^{-t}dt \\
&= -t^{a-1}e^{-t}\big|_0^\infty - \int_0^\infty -(a-1)t^{a-2}e^{-t}dt && \text{Integração por partes} \\
&= 0 + (a-1)\int_0^\infty t^{a-2}e^{-t}dt = (a-1)\Gamma(a-1) && \text{Função Gama (Definição 2.7.1)}
\end{aligned}$$

2. Usando o item anterior e iterando em n, observe que

$$\Gamma(n) = (n-1)\cdot(n-2)\ldots 2\cdot 1\cdot\Gamma(1) = (n-1)!\cdot\Gamma(1)$$

Assim, basta mostrar que $\Gamma(1) = 1$.

$$\Gamma(1) = \int_0^\infty e^{-t}dt = -e^{-t}\big|_0^\infty = 1$$

□

Definição 2.7.3. Dizemos que uma variável aleatória X tem distribuição Gama com parâmetros(k, λ) e denotamos por $X \sim$ Gama(k, λ) se a densidade de X é

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(k)\lambda^k}x^{k-1}e^{-\frac{x}{\lambda}} & \text{, if } x > 0 \\ 0 & \text{, caso contrário} \end{cases}$$

A figura 6 mostra algumas densidades possíveis para a distribuição Gama.

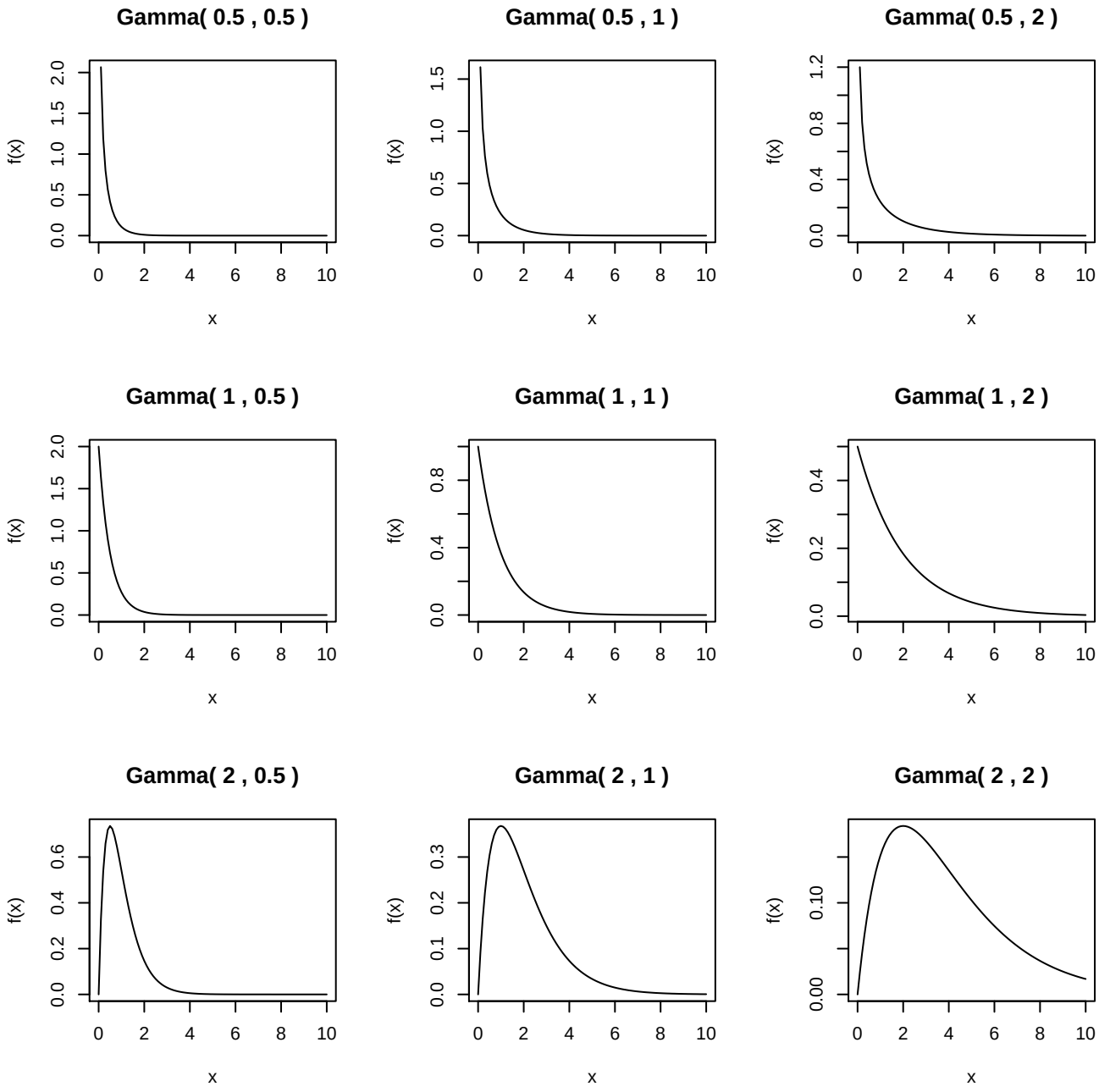


Figura 6: densidades de algumas distribuições Gama

Lema 2.7.4. *A densidade da distribuição Gama integra* 1.

Demonstração.

$$\begin{aligned}\int_0^\infty \frac{1}{\Gamma(k)\lambda^k} x^{k-1} e^{-\frac{x}{\lambda}} dx &= \int_0^\infty \frac{1}{\Gamma(k)\lambda^k} (\lambda t)^{k-1} e^{-t} \lambda dt && \text{Fazemos } x = \lambda t \\ &= \int_0^\infty \frac{1}{\Gamma(k)} t^{k-1} e^{-t} dy = 1 && \text{Função Gama (Definição 2.7.1)}\end{aligned}$$

□

Observe que, no caso especial em que $k = 1$, a distribuição Gama corresponde à distribuição Exponencial.

Lema 2.7.5. *Para todo* $a > 0$, $\mathbb{E}[X^a] = \frac{\Gamma(k+a)\lambda^a}{\Gamma(k)}$. *If* $X \sim$ *Gama*(k, λ), *então* $\mathbb{E}[X] = k\lambda$ *e* $Var[X] = k\lambda^2$.

Demonstração.

$$\begin{aligned}\mathbb{E}[X^a] &= \int_0^\infty x^a \frac{1}{\Gamma(k)\lambda^k} x^{k-1} e^{-\frac{x}{\lambda}} dx \\ &= \int_0^\infty \frac{1}{\Gamma(k)\lambda^k} x^{k+a-1} e^{-\frac{x}{\lambda}} dx \\ &= \frac{\Gamma(k+a)\lambda^a}{\Gamma(k)} \int_0^\infty \frac{1}{\Gamma(k+a)\lambda^{k+a}} x^{k+a-1} e^{-\frac{x}{\lambda}} dx && \text{Densidade de uma Gama}(k+a, \lambda) \\ &= \frac{\Gamma(k+a)\lambda^a}{\Gamma(k)} && \text{Propriedades da função Gama (Lema 2.7.2)}\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X] &= \frac{\Gamma(k+1)\lambda^1}{\Gamma(k)} = \frac{k\Gamma(k)\lambda^1}{\Gamma(k)} = k\lambda && \text{Use } a = 1 \text{ no item anterior} \\ \mathbb{E}[X^2] &= \frac{\Gamma(k+2)\lambda^2}{\Gamma(k)} = \frac{k(k+1)\Gamma(k)\lambda^2}{\Gamma(k)} = k(k+1)\lambda^2 && \text{Use } a = 2 \text{ no item anterior} \\ Var[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = k(k+1)\lambda^2 - k^2\lambda^2 = k\lambda^2\end{aligned}$$

□

2.8 Distribuição Beta (retirado de Stern and Izbicki (2016))

Definição 2.8.1. Dizemos que X tem distribuição Beta com parâmetros (α, β) e denotamos por $X \sim$ Beta(α, β) se a densidade de X é

$$f_X(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1} & \text{, if } 0 < x < 1 \\ 0 & \text{, caso contrário} \end{cases}$$

Lema 2.8.2. *A densidade na Definição 2.8.1 é uma fdp válida. Em particular,*

$$\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} dx = 1$$

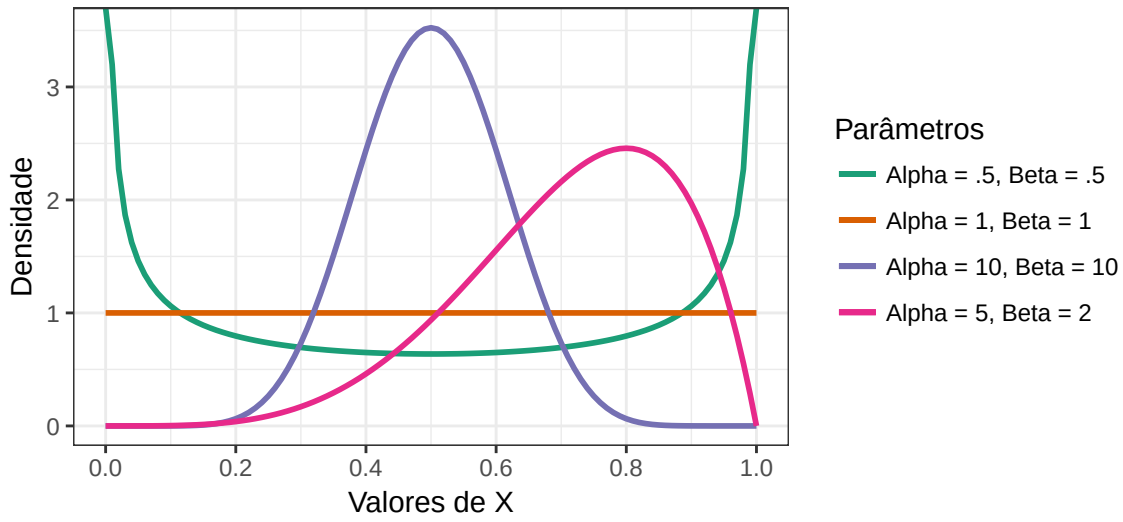


Figura 7: Densidades de algumas distribuições Beta

Demonstração.

$$\begin{aligned}
\Gamma(\alpha)\Gamma(\beta) &= \int_0^\infty x^{\alpha-1}e^{-x}dx \int_0^\infty y^{\beta-1}e^{-y}dy && \text{(Definição 2.7.1)}\\
&= \int_0^\infty \int_0^\infty x^{\alpha-1}y^{\beta-1}e^{-(x+y)}dxdy \\
&= \int_0^\infty \int_0^1 (tu)^{\alpha-1}(t(1-u))^{\beta-1}e^{-t}t\cdot dudt && t=x+y,\ u=x(x+y)^{-1}\\
&= \int_0^\infty t^{\alpha+\beta-1}e^{-t}dt \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du \\
&= \Gamma(a+b)\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du && \text{(Definição 2.7.1)}
\end{aligned}$$

Assim, $\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(a+b)}$ e $\int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx = 1$. □

Como a distribuição Beta assume valores em $(0,1)$, ela é frequentemente usada para modelar frequências relativas e razões. É uma distribuição flexível e pode assumir muitas formas diferentes. A Figura 7 mostra algumas densidades que a distribuição Beta pode assumir.

Lema 2.8.3. *Se $X \sim Beta(\alpha,\beta)$, então para todo $c,d \geq 0$ $\mathbb{E}[X^c(1-X)^d] = \frac{\Gamma(\alpha+c)\Gamma(\beta+d)}{\Gamma(\alpha+\beta+c+d)}$. Portanto, $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ e $Var[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.*

Demonstração.

$$\begin{aligned}
\mathbb{E}[X^c(1-X)^d] &= \int_{-\infty}^{\infty} x^c(1-x)^d f_X(x)dx \\
&= \int_0^1 x^c(1-x)^d \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+c)\Gamma(\beta+d)}{\Gamma(\alpha+\beta+c+d)}\int_0^1 \frac{\Gamma(\alpha+\beta+c+d)}{\Gamma(\alpha+c)\Gamma(\beta+d)}x^{\alpha}(1-x)^{\beta-1}dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\cdot\frac{\Gamma(\alpha+c)\Gamma(\beta+d)}{\Gamma(\alpha+\beta+c+d)}\cdot 1 && \text{(Lema 2.8.2)}
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}[X] &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta+0)}{\Gamma(\alpha+\beta+1+0)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \cdot \frac{\alpha\Gamma(\alpha)}{(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta} && \text{(Lema 2.7.2)} \\
\mathbb{E}[X^2] &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+2)\Gamma(\beta+0)}{\Gamma(\alpha+\beta+2+0)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \cdot \frac{\alpha(\alpha+1)\Gamma(\alpha)}{(\alpha+\beta+1)(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} && \text{(Lema 2.7.2)} \\
Var[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{(\alpha+\beta)(\alpha+1)\alpha - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta+1)(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}
\end{aligned}
$$

□

2.9 Exercícios

Exercício 39. Seja $0 \leq X \leq 1$ uma v.a. contínua tal que

$$
f_X(x) = \begin{cases} c, \text{ para } 0 \leq x \leq 1 \\ 0, \text{ caso contrário.} \end{cases}
$$

Encontre o valor de c.

Resposta: $c = 1$.

Exercício 40. Seja $X \geq 0$ uma variável aleatória contínua tal que $f_X(x) = ce^{-x}$ para $x \geq 0$. Encontre o valor de c.

Resposta: $c = 1$.

Exercício 41. Considere $X \geq 1$. Pode haver c tal que a densidade de probabilidade de X seja $f_X(x) = ce^{-x}$?

Resposta: Não.

Exercício 42. Determinar a área limitada pela curva normal padrão em cada um dos casos abaixo:

a. Entre $z = 0$ e $z = 1,2$;

b. Entre $z = -0,68$ e $z = 0$;

c. Entre $z = 0,46$ e $z = 2,21$;

d. Entre $z = -0,81$ e $z = 1,94$;

e. À esquerda de $z = -0,6$;

f. À direita de $z = -1,23$;

g. À direita de $z = 2,05$ e à esquerda de $z = 1,44$;

h. Entre $z = -1$ e $z = +1$;

i. Entre $z = -1,96$ e $z = +1,96$;

j. Entre $z = -2,56$ e $z = +2,56$.

Respostas:

a. $0,3848$

b. $0,2517$

c. $0,3092$

d. $0,7648$

e. $0,2743$

f. $0,8907$

g. $0,9453$

h. $0,6826$

i. $0,95$

j. $0,9896$

Exercício 43. A altura dos indivíduos de uma população distribui-se normalmente com média de $1,56$m e desvio padrão de $0,09$m. Qual a porcentagem, nessa população, de indivíduos com altura de $1,8$m ou mais?

Resposta: $0,39\%$

Exercício 44. O peso médio das reses que se encontram num curral de uma determinada fazenda é de 200kg e o desvio padrão é de 10kg. Em 120 animais retirados ao acaso do curral, quantos pesarão mais de 185kg? Considere que o peso das reses tenha distribuição normal.

Resposta: 111 **bois**

Exercício 45. Foi feito um estudo sobre a altura de plantas de milho de certo híbrido, observando-se que ela se distribui normalmente com média igual a $2,20$m e desvio padrão igual a $0,20$m. Qual a porcentagem de plantas com altura:

a. Entre $2,15$m e $2,25$m;

b. Entre $2,0$m e $2,4$m;

c. Acima de $2,3$m.

Respostas:

a. $0,1974$

b. $0,6826$

c. $0,3085$

Exercício 46. Na distribuição normal de média μ e desvio padrão σ^2, encontre:

a. $\mathbb{P}(X < \mu + 2\sigma)$;

b. $\mathbb{P}(|X - \mu| < \sigma)$;

c. O número a tal que $\mathbb{P}(\mu - a\sigma < X < \mu + a\sigma) = 0,9$;

c. O número a tal que $\mathbb{P}(X > a) = 0,95$.

Respostas:

a. $0,9772$

b. $0,6826$

c. $1,645$

d. $\mu - 1,645\sigma$

Exercício 47. Seja $X \sim N(0,1)$. Determine:

a. $\mathbb{P}(-1 < X < 1)$;

b. $\mathbb{P}(-2 < X < 2)$;

c. $\mathbb{P}(-3 < X < 3)$.

Seja $Y \sim N(5,10)$. Determine:

a. $\mathbb{P}(-5 < Y < 15)$;

b. $\mathbb{P}(-15 < Y < 25)$;

c. $\mathbb{P}(-25 < Y < 35)$.

Seja $W \sim N(\mu,\sigma)$. Determine:

a. $\mathbb{P}(\mu - \sigma < W < \mu + \sigma)$;

b. $\mathbb{P}(\mu - 2\sigma < W < \mu + 2\sigma)$;

c. $\mathbb{P}(\mu - 3\sigma < W < \mu + 3\sigma)$.

Compare os resultados obtidos para X, Y e W. Que conclusão você pode tirar?

Respostas:

a. $0,6826$

b. $0,9544$

c. $0,9974$

As probabilidades de valores distantes da média em desvios-padrões se mantém. Essa propriedade da normal é muito usada em controle estatístico de qualidade. Para quem quiser saber mais, pode procurar por *six-sigma*.

Exercício 48. Prove que se $X \sim Exponencial(\lambda)$ então $E[X] = \frac{1}{\lambda}$ $Var[X] = \frac{1}{\lambda^2}$.

Exercício 49. (Adaptado de Schmidt (2011)) Sejam $X_1, X_2, \cdots, X_n$ variáveis aleatórias com distribuição Exponencial dada por $f(x) = \frac{1}{\lambda} e^{\frac{-x}{\lambda}}$, $0 < x < \infty$. Considere $\hat{\lambda} = \frac{c}{n} \sum_{i=1}^{n} X_i$, onde c é um número real. Determine se as seguintes afirmativas são verdadeiras ou falsas, justificando suas respostas:

a. $E[\hat{\lambda}] = \lambda$.

b. $Var[\hat{\lambda}] = \frac{c}{\lambda}$.

c. $Var[\hat{\lambda}] + (\lambda - E[\hat{\lambda}])^2 = \lambda^2(2c^2 - 2c + 1)$ e essa quantidade é minimizada quando $c = 0,5$.

d. Se c=1, $E[\hat{\lambda}] = \lambda$.

e. Se c=1, $E[\hat{\lambda}] \neq \lambda$ e $Var[\hat{\lambda}] + (\lambda - E[\hat{\lambda}])^2 = \lambda^2$.

2.10 Esperança, variância e covariância de v.a.'s

2.10.1 Definições básicas

Definição 2.10.1. Seja X uma variável aleatória discreta com função massa de probabilidade denotada por p_X e que assume valores $x \in \chi$. O *valor esperado* ou *esperança matemática* ou *média* de X é definida por:

$$\mathbb{E}[X] = \sum_{x \in \chi} x_i p_X(x_i) \tag{43}$$

Se X é uma variável aleatória contínua com função densidade de probabilidade denotada por f_X, sua esperança será dada por:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f_X(x) dx \tag{44}$$

Definição 2.10.2. O *k-ésimo momento* da variável aleatória X é dado pela esperança de X elevada à potência k, isto é, $\mathbb{E}[X^k]$ (desde que essa quantidade esteja bem definida), para $k \in \{1, 2, \cdots\}$. Se a esperança de X for um número finito μ, isto é, se $\mathbb{E}[X] = \mu < \infty$, então definimos $\mathbb{E}[(X - \mu)^k]$ como o *k-ésimo momento central de* X, desde que essa quantidade esteja bem definida.

Definição 2.10.3. Seja X uma variável aleatória com média finita denotada por μ. Sua variância é dada pelo momento central de ordem 2 de X:

$$Var[X] = \mathbb{E}[(X - \mu)^2] \tag{45}$$

Definição 2.10.4. Sejam X e Y duas variáveis aleatórias definidas no mesmo espaço de probabilidade. A *covariância* entre elas será dada por:

$$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{46}$$

2.10.2 Propriedades

As seguintes propriedades podem ser demonstradas a partir das definições dadas na página anterior e a maioria pode ser adaptada para o caso de variáveis aleatórias contínuas. Neste caso, ao invés de somatórios, teremos as integrais correspondentes. Para algumas das demonstrações, consulte https://www.overleaf.com/read/nfzbfjsrwspf e https://1drv.ms/b/s!AlHDLj_7OjaL1HyxBhvFIbEMhUx0.

Tabela 1: Propriedades do valor esperado

1	Esperança da soma	$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$	É a soma das esperanças
2	Esperança do produto	$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ se X e Y indep.	É o produto das esperanças, desde que X e Y sejam independentes
3	Esperança de um escalar	$\mathbb{E}[\alpha] = \alpha$, $\alpha \in \mathbb{R}$	É o próprio escalar
4	Esperança de X vezes um escalar	$\mathbb{E}[\alpha X] = \alpha\mathbb{E}[X]$, $\alpha \in \mathbb{R}$	É o escalar vezes a esperança de X
5	Esperança de um escalar mais X	$\mathbb{E}[\beta + X] = \beta + \mathbb{E}[X]$, $\beta \in \mathbb{R}$	É escalar mais a esperança de X
6	Esperança de uma função de X Lei do Estatístico Inconsciente	$\mathbb{E}[g(X)] = \sum_{x\in\chi} g(x_i)p_X(x_i)$ se X é discreta	$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)f_X(x)dx$ se X é contínua
7	Lei das Expectativas Iteradas	$\mathbb{E}[\mathbb{E}[X\|Y]] = \mathbb{E}[X]$	O valor esperado da esperança de X dado Y é a esperança de X.
8	Forma alternativa da esperança	$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}(X \geq i)$	Se X assume valores positivos.
9	Lei da esperança total	$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X\|A_i] \cdot \mathbb{P}(A_i)$	Se $A_1, \ldots, A_n$ uma partição de Ω e X uma v.a. discreta.

Propriedades da Variância e Covariância

Propriedades da Esperança

Tabela 2: Propriedades da variância e Covariância

1	Forma alternativa da variância 1	$Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$	Variância de X é a esperança de X ao quadrado menos a média de X ao quadrado
2	Forma alternativa da variância 2	$Var[X] = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2$	
3	Forma alternativa da variância 3	$Var[X] = \mathbb{E}[(X-d)^2] - (\mathbb{E}[X]-d)^2,\ d \in \mathbb{R}$	se $Var[X] < \infty$
4	Variância da soma	$Var[X+Y] = Var[X] + Var[Y]$ Se X e Y indep.	É a soma das variâncias, se X e Y independentes
5.1	Variância da soma (caso geral)	$Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$	É a soma das variâncias mais duas vezes a covariância
5.2	Variância da diferença (caso geral)	$Var[X-Y] = Var[X] + Var[Y] - 2Cov[X,Y]$	É a soma das variâncias menos duas vezes a covariância
6	Variância de um escalar	$Var[\alpha] = 0,\ \alpha \in \mathbb{R}$	É zero pois um número é sempre ele mesmo (não há variação)
7	Variância de X vezes um escalar	$Var[\beta X] = \beta^2 Var[X],\ \beta \in \mathbb{R}$	É o escalar ao quadrado vezes a variância de X
8	Variância de um escalar mais um escalar vezes X	$Var[\alpha + \beta X] = \beta^2 Var[X],\ \alpha, \beta \in \mathbb{R}$	
9	Forma alternativa da covariância	$Cov[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$	A covariância de X e Y é a esperança do produto menos o produto das esperanças.
10	Covariância de X com X	$Cov[X,X] = Var[X]$	Logo, se X é constante, $Cov[X,X] = 0$
11	Simetria da covariância	$Cov[X,Y] = Cov[Y,X]$	
12	Covariância de um escalar vezes X	$Cov[\alpha X, Y] = \alpha Cov[X,Y]$ $Cov[\alpha X, \beta Y] = \alpha\beta Cov[X,Y]$ $\alpha, \beta \in \mathbb{R}$	
13	Linearidade na 1^a entrada	$Cov[\alpha X + \beta Y, Z] = \alpha Cov[X,Z] + \beta Cov[Y,Z]$	
13.1	Linearidade na 1^a entrada e multiplicação por escalar	$Cov[\alpha X + \beta Y, \gamma Z + \delta W] = \alpha\gamma Cov[X,Z] + \alpha\delta Cov[X,W] + \beta\gamma Cov[Y,Z] + \beta\delta Cov[Y,W]$	
14	Covariância de sequências de variáveis aleatórias	$Cov[\sum_{i=1}^n \alpha_i X_i, \sum_{j=1}^m \beta_i Y_i] = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j Cov[X_i, Y_j]$	Suponha $\{X_1, \cdots, X_n\}$ e $\{Y_1, \cdots, Y_m\}$ duas sequências de variáveis aleatórias e $\{\alpha_1, \cdots, \alpha_n\}$ $\{\beta_1, \cdots, \beta_n\}$ sequências de constantes.

As propriedades de esperança são mais difíceis de provar pois requerem que constantemente o espaço amostral com seus eventos seja invocado e a matemática envolvida no processo é um pouco mais cheia de detalhes. A partir do momento que todas as propriedades da esperança estão demonstradas, as da variância e da covariância seguem de maneira mais natural. Algumas das propriedades se encontram demonstradas em Stern and Izbicki (2016). Em especial, a lei do estatístico inconsciente pode ser encontrada também em Magalhães (2011) ou ainda uma versão em vídeo aqui: https://www.youtube.com/watch?v=uDEHBrsLjzE&t=430s (a notação do menino é um pouquinho diferente, mas é a mesma ideia).

Antes de avançar, deixo aqui um resultado de covariância (que comentei por cima na terça-feira) que se relaciona com os conceitos de álgebra linear, que é pouco abordado nas aulas tradicionais de estatística e probabilidade (e vai ficar mais claro porque o coeficiente de correlação de Pearson está entre 0 e 1). As próximas definições e resultados foram retirados de Stern and Izbicki (2016).

Definição 2.10.5. Seja Ω um conjunto enumerável, P a função probabilidade em Ω.

1. Defina $\mathcal{V} = \{X : \Omega \to \mathbb{R} \text{ tal que } \mathbb{E}[X] = 0\}$.
2. Para todo $X \in \mathcal{V}$ e $Y \in \mathcal{V}$, nós definimos $(X + Y) : \Omega \to \mathbb{R}$ tal que $(X + Y)(w) = X(w) + Y(w)$.
3. Para todo $X \in \mathcal{V}$ e $a \in \mathbb{R}$, nós definimos $(aX) : \Omega \to \mathbb{R}$ tal que $(a \cdot X)(w) = aX(w)$.

Lema 2.10.6. *$(\mathcal{V}, +, \cdot)$ é um espaço vetorial sobre $\mathbb{R}$.*

Demonstração. Para provar isso, é suficiente demonstrar os itens abaixo:

1. Como $\mathbb{E}[0] = 0$, $0 \in \mathcal{V}$.
2. Se $X \in \mathcal{V}$ e $Y \in \mathcal{V}$, $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 0$. Potanto, $X + Y \in \mathcal{V}$.
3. Se $a \in \mathbb{R}$ e $X \in \mathcal{V}$, $\mathbb{E}[aX] = a\mathbb{E}[X] = 0$. Portanto, $aX \in \mathcal{V}$.

□

Lema 2.10.7. *Cov é um produto interno em $\mathcal{V}$. Portanto, $\sqrt{Var[X]}$ é uma norma em $\mathcal{V}$.*

Demonstração. Da álgebra linear, sabemos que $\langle X, Y\rangle$ é produto interno se e somente se:

1. $\langle X, X\rangle \geq 0$ e $\langle X, X\rangle = 0 \iff x = 0$ (no caso da covariância, isso só vale para os elementos que percentem a $\mathcal{V}$, pois outros elementos podem ter covariância negativa);
2. $\langle X, Y\rangle = \langle Y, X\rangle$;
3. $\langle aX + bY, Z\rangle = a\langle X, Z\rangle + b\langle Y, Z\rangle$

Pelas propriedades estudadas na aula passada, segue que a covariância é um produto interno em $\mathcal{V}$.

Sabemos ainda que a norma é a raiz do produto interno, isto é, $\|X\| = \sqrt{\langle X, Y\rangle}$. Logo, $\|X\| = \sqrt{Cov(X, X)} = \sqrt{Var(X)} = DP(X)$, o que significa que o desvio padrão de X é uma norma em $\mathcal{V}$. □

Lema 2.10.8 (Cauchy-Schwarz paravariáveis aleatórias)**.**

$$|Cov[X, Y]| \leq \sqrt{Var[X]}\sqrt{Var[Y]}.$$

A igualdade ocorre se e somente se existem $a, b \in \mathbb{R}$ com $a \neq 0$ tais que $Y = aX + b$.

Demonstração. Seja $V = X - \mathbb{E}[X]$ e $W = Y - \mathbb{E}[Y]$. Como $V, W \in \mathcal{V}$ e, como pelo Lema 2.10.7, covariância é um produto interno, segue que

$$\begin{aligned} |Cov[X, Y]| = |Cov[V, W]| &\leq \sqrt{Cov[V, V]}\sqrt{Cov[W, W]} && \text{(desig. de C-S: https://goo.gl/DYDLbD)} \\ &= \sqrt{Var[V]}\sqrt{Var[W]} && (Var(X - E(X)) = Var(X)) \\ &= \sqrt{Var[X]}\sqrt{Var[Y]} \end{aligned}$$

Agora, da desigualdade de Cauchy-Schwarz sabemos que a igualdade ocorre se e somente se existe $b \in \mathbb{R}$, $b \neq 0$, tal que $W = bV$. Em outras palavras, a igualdade ocorre se e somente se existem $b \in \mathbb{R}$, $b \neq 0$, tais que $Y - \mathbb{E}[Y] = b(X - \mathbb{E}[X])$, o qe conclui a demonstração. □

Definição 2.10.9 (Correlação)**.**

$$Corr[X, Y] = \frac{Cov[X, Y]}{\sqrt{Var[X]}\sqrt{Var[Y]}}$$

Seja $\langle \cdot, \cdot \rangle$ um produto interno e $\| \cdot \|$ a norma gerada pelo produto interno. Lembramos das aulas de álgebra linear que $\frac{<v_1, v_2>}{\|v_1\|\|v_2\|}$ é o cosseno do ângulo entre v_1 e v_2. Portanto, usando o Lema 2.10.7, podemos interpretar a $Corr[X, Y]$ como o cosseno do ângulo entre variáveis aleatórias X e Y. Em outras palavras, $Corr[X, Y]$ é a medida da associação linear entre X e Y.

Lema 2.10.10.

$$|Corr[X, Y]| \leq 1$$

Demonstração. Segue diretamente da aplicação do Lema 2.10.8 à Definição 2.10.9. □

Lema 2.10.11. *Seja X a variável aleatória discreta.*

a. Seja $b \in \mathbb{R}$, $Cov[b, X] = 0$.

b. Seja $a \neq 0$,

$$Corr[aX + b, X] = \begin{cases} 1 & , \text{se } a > 0 \\ -1 & , \text{se } a < 0 \end{cases}$$

Demonstração.

a. $Cov[b, X] = E[bX] - E[b]E[X] = bE[X] - bE[X] = 0$

b.

$$\begin{aligned} Corr[aX + b, X] &= \frac{Cov[aX + b, X]}{\sqrt{Var[aX + b]}\sqrt{Var[X]}} \\ &= \frac{aCov[X, X] + Cov[b, X]}{\sqrt{Var[aX + b]}\sqrt{Var[X]}} \\ &= \frac{aVar[X]}{\sqrt{a^2 Var[X]}\sqrt{Var[X]}} \\ &= \frac{a}{|a|} \end{aligned}$$

□

2.10.3 Exercícios

Exercício 50. Demonstre todas as propriedades da tabela 2 (para variáveis aleatórias discretas) exceto a propriedade (14) da covariância.

Exercício 51. (Adaptado de Stern and Izbicki (2016))
Calcule a esperança e a variância de X dos exercícios 37 e 38. Para o exercício (38), calcule ainda $\mathbb{P}(X \geq 1)$.

Exercício 52. (Retirado de Schmidt (2011))
Suponha que X seja uma v.a. com densidade dada por:

$$f_X(x) = \begin{cases} 2(1-x), \text{ para } 0 \leq x \leq 1 \\ 0, \text{ caso contrário.} \end{cases}$$

Sendo $Y = 6X + 10$, calcule $Var(Y)$.

Resposta: 2.

Exercício 53. Mostre que se X e Y são independentes, então $Cov[X, Y] = 0$.

Exercício 54. Dê um exemplo de duas variáveis X e Y que não sejam independentes mas tenham covariância igual a zero.

Exercício 55. (Retirado de Stern and Izbicki (2016)) Considere que você lança 200 vezes uma moeda com probabilidade p de ocorrer cara. Seja X o número total de caras observadas. Determine $\mathbb{E}[X]$. Dica: Denote por H_i o evento em que o i-ésimo lançamento resulta em cara. Observe que $X = \sum_{i=1}^{200} I_{H_i}$, onde I_{H_i} é a função indicadora de H_i e assume valor 1 se H_i ocorreu e 0 caso contrário.

Exercício 56. (Retirado de Stern and Izbicki (2016)) Seja $X_1, \ldots, X_n$ uma variável aleatória discreta tal que, para todo $i \in \{1, \ldots, n\}$, $\mathbb{E}[X_i] = \mu \in \mathbb{R}$.

- Seja $p_i \geq 0$ tal que $\sum_{i=1}^n p_i = 1$. Encontre, $\mathbb{E}[\sum_{i=1}^n p_i X_i]$.
- Seja $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$, usualmente chamado "média amostral". Encontre $\mathbb{E}[\bar{X}]$.

Exercício 57. (Retirado de Stern and Izbicki (2016)) Seja X o número de caras observadas em dois lançamentos de uma moeda com probabilidade p de cara. Seja H_i o evento em que o i-ésimo lançamento resulta em cara. Calcule $Var[X]$.

Exercício 58. (Adaptado de Schmidt (2011)) Determine se os itens de (a) a (e) são verdadeiros ou falsos, justificando suas respostas. Considere X e Y duas variáveis aleatórias quaisquer com médias μ_X e μ_Y, respectivamente.

a. $\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \frac{1}{2}[\mathbb{E}[(X - \mu_X)]^2 + \mathbb{E}[(Y - \mu_Y)^2] - \mathbb{E}[(X + Y - \mu_X - \mu_Y)^2]]$

b. $Var(Y - X) = Var(Y) - Var(X) - 2Cov(Y, X)$

c. Se $Var(Y + X) = Var(Y) + Var(X)$, então X e Y são independentes.

d. Se Y e X são independentes, seu coeficiente de correlação linear de Pearson é zero.

e. Se $Cov(X, Y) = 0$ e se X e Y têm distribuição conjunta Normal, então X e Y são independentes.

Exercício 59. (Retirado de Davidson and MacKinnon (2009)) Uma variável aleatória calculada como a razão entre duas variáveis aleatórias independentes com distribuição normal padrão tem distribuição de **Cauchy**. Pode-se demonstrar que a densidade dessa distribuição é dada por

$$f(x) = \frac{1}{\pi(1 + x^2)} \tag{47}$$

a. Mostre que a distribuição de Cauchy não tem o primeiro momento, o que significa que sua média não está definida.

b. **Exercício computacional (para depois da ANPEC):** Gere amostras de tamanho $n = \{10, 100, 1000, 10000\}$ de uma distribuição de Cauchy (e quantas amostras de tamanhos intermediários entre esses valores você conseguir). Para cada amostra, calcule a média amostral. Essas amostras parecem convergir para algum valor conforme o tamanho da amostra aumenta? Repita o exercício para amostras retiradas de uma distribuição normal padrão. A média amostral parece convergir para qual valor?[29]

Exercício 60. (Adaptado de Schmidt (2011)) Considere um treinador que tenha no seu time apenas dois Pokémons: Magikarp e Metapod. Determine se as seguintes afirmativas são verdadeiras ou falsas, justificando suas respostas com cálculos:

a. Sabendo que, em média, Magikarp vence 10% das partidas que luta e Metapod vence, também em média, 5% de suas partidas e que o treinador utiliza o Magikarp em 40% das partidas e Metapod nas restantes, então o valor esperado do percentual de lutas ganhas é $7,5\%$;

b. Supondo que o fato de um Pokémon vencer uma batalha não influencia o resultado do outro (isto é, as batalhas são independentes) e que a variância das vitórias de Magikarp é 10 e de Metapod é 20, então a variância das vitórias deste time[30] é de $8,8$;

c. Supondo agora que ambos Pokémons apresentem mesma variância de número de vitórias, o fato do treinador utilizar os dois Pokémons somente reduz sua variância total se a correlação das vitórias de Magikarp com Metapod for negativa.

Exercício 61. (Adaptado de Schmidt (2011)) Muambildo esteve em *Ciudad del Leste* onde comprou vários artigos *importados* para revender em sua banca no Camelódromo de Florianópolis. O preço médio pago por Muambildo foi de US$ $15,00$, com desvio padrão de um dólar. Sabendo que Muambildo utilizou uns dólares que havia comprado no passado a uma taxa de câmbio de R$ 3,00 por dólar, julgue se as seguintes afirmativas são corretas, justificando suas respostas:

a. Fazendo a conversão para a moeda brasileira, o preço médio dos produtos foi de quarenta e cinco reais;

b. O desvio padrão é R$$3,00$;

c. Se Muambildo adicionar ao preço original uma margem de lucro de R$10,00, o novo preço médio será de R$$55,00$ com desvio padrão de R$ $6,00$;

d. Se a margem de lucro de Muambildo é de 20% sobre o preço da mercadoria em reais, o novo preço médio será de R$$54,00$ e o novo desvio padrão é de R$$3,60$;

e. O coeficiente de variação[31], calculado em reais, devido à taxa de câmbio, será 3 vezes maior do que aquele calculado utilizando-se os valores em dólar.

Exercício 62. (Adaptado de Stern and Izbicki (2016))

a. Juju é lançada 1001 vezes. Encontre a probabilidade de que se observem mais caras do que coroas.

   efina X como a v.a. que representa o número de caras em 1001 lançamentos da Juju. Então, $X \sim Binomial(n = 1001, p = \frac{1}{2})$.

[29] Uma solução no R pode ser encontrada aqui: https://www.overleaf.com/read/ndjgtmtkmkmy. Tente entender o código e reproduzir no Excel, por exemplo.

[30] A variância das vitórias de um time pode ser calculada como a soma das variâncias das vitórias individuais.

[31] Seja X uma variável aleatória com média μ_X e desvio-padrão σ_X. Então, o coeficiente de variação de X, denotado por CV_X, será $CV_X = \frac{\sigma_X}{\mu_X}$

Estamos interessados em $X \geq 501$, isto é, que o número de caras seja maior ou igual a 501. Logo, queremos calcular $\mathbb{P}(X \geq 501)$.

$$\begin{aligned}
\mathbb{P}(X \geq 501) &= \sum_{x=501}^{1001} \binom{1001}{x} \left(\frac{1}{2}\right)^{x} \left(\frac{1}{2}\right)^{1001-x} \\
&= \sum_{x=501}^{1001} \binom{1001}{x} \left(\frac{1}{2}\right)^{1001-x+x} \\
&= \sum_{x=501}^{1001} \frac{1001!}{x!(1001-x)!} \left(\frac{1}{2}\right)^{1001}
\end{aligned}$$

Chame $Y = 1001 - x$ e troque o somatório:

$$\sum_{y=0}^{500} \binom{1001}{y} \left(\frac{1}{2}\right)^{1001}$$

Observe que isso é igual à $\mathbb{P}(X < 500)$. Logo, $\mathbb{P}(X \geq 501) = \mathbb{P}(X < 500)$.

Por outro lado, temos:

$$\underbrace{\mathbb{P}(X \geq 501)}_{A} + \underbrace{\mathbb{P}(X < 501)}_{A^c} = 1$$

Assim,

$$\begin{aligned}
\mathbb{P}(X \geq 501) + \mathbb{P}(X > 500) &= 1 \\
\mathbb{P}(X \geq 501) + \mathbb{P}(X \geq 501) &= 1 \\
2\mathbb{P}(X \geq 501) &= 1 \\
\mathbb{P}(X \geq 501) &= \frac{1}{2}
\end{aligned}$$

b. Repita o item (a) porém considerando apenas 1000 lançamentos.

eja Y uma v.a. que representa o número de caras da Juju em 1000 lançamentos. Então, $Y \sim Binomial(1000, \frac{1}{2})$. Estamos interessados em determinar $\mathbb{P}(Y \geq 501)$. Com o raciocínio do item (a), concluímos que $\mathbb{P}(Y \geq 501) = \mathbb{P}(Y \leq 499)$, de forma que resta calcular $\mathbb{P}(Y = 500)$, então:

$$\mathbb{P}(Y \geq 501) + \mathbb{P}(Y = 500) + \underbrace{\mathbb{P}(Y \leq 499)}_{\mathbb{P}(Y \geq 501)} = 1$$

$$2\mathbb{P}(Y \geq 501) + \mathbb{P}(Y = 500) = 1$$

$$\mathbb{P}(Y \geq 501) = \frac{1 - \mathbb{P}(Y = 500)}{2}$$

$$\mathbb{P}(Y \geq 501) = \frac{1 - \binom{1000}{500}\left(\frac{1}{2}\right)^{1000}}{2}$$

Exercício 63. 5 (Adaptado de Stern and Izbicki (2016))
Seja X uma variável aleatória discreta assumindo valores em $\{1, 2, \cdots\}$. Definimos a função geradora de probabilidades de X como $G_X(z) = \mathbb{E}[z^X]$.

a. Mostre que $\left.\frac{d^k G_X(z)}{dz^k}\right|_{z=0} = k!\mathbb{P}(X = k), \quad \forall \quad k \in \{1, 2, \cdots\}$. Lembre-se que 0^w vale 1 se $w = 0$ e 0 se $w \neq 0$.

align*
$\partial^k[G_x(z)]\partial^k z = \frac{\partial^k[\mathbb{E}[z^x]]}{\partial^k z} = \frac{\partial^k}{\partial^k z}[\sum_{i=0}^k z^i f_x(i)]$

A primeira derivada, quando $i = 0$ será $i \cdot z^{i-1} f_X(i) = 0$.

Quando $i > 0$, teremos a soma das derivadas nos pontos $i = 1$, $i = 2$, ..., $i = k$. Note que cada vez que uma derivada é feita, o expoente de z diminui e será eventualmente zero se $i > k$.

Note ainda que

$$\frac{\partial^k}{\partial^k z}[\sum_{i=0}^k z^i f_x(i)] = \sum_{i=0}^k \frac{\partial^k}{\partial^k z}[z^i f_x(i)] = i!z^{i-k} f_X(i) =$$
$$= 0 + 1z^{1-k} f_X(1) + 2 \cdot 1z^{2-k} f_x(2) + \ldots + k(k-1)!z^0 f_x(k)$$

Mas se $z = 0$, todos os termos com z na expressão acima serão iguais a zero e com isso teremos apenas o último termo, $k(k-1)!f_X(k)$ que é exatamente $k!\mathbb{P}(X = k)$.

b. Você acha que $G_X(z)$ caracteriza a distribuição de X de maneira única, isto é, você acha que dada uma $G_X(z)$ específica é possível dizer qual a distribuição e os parâmetros de X? Justifique.

f.g.m. é única para cada $F_X(x)$ e por isso caracteriza uma v.a. . Como a função geradora de probabilidades está definida apenas para v.a.'s discretas ela também define unicamente a variável aleatória. Note que esse não seria o caso se estivesse definida para variáveis aleatórias contínuas, pois neste caso, uma vez que a integral no ponto é sempre 0, podemos ter infinitas f_X que levam em uma mesma F_X.

c. Mostre que a função geradora de probabilidades de $Y \sim Binomial(n, p)$ é $(1 - p + pz)^n$.

amos mostrar a geradora de probabilidades de uma Bernoulli e usar o fato de que a soma de v.a.'s Bernoulli é uma Binomial.

Seja $X_i \sim Bernoulli(p)$. Então:

$$G_{Xi}(z) = \mathbb{E}[z^{X_i}] = z^0\mathbb{P}(X_i = 0) + z^1\mathbb{P}(X_i = 1)$$
$$(1 - p) + zp$$

Seja $X = \sum_{i=1}^{n} X_i$. Então, $X \sim Binomial(n, p)$. Assim:

$$\begin{aligned}\mathbb{E}[z^X] = & \\ &= \mathbb{E}\left[z^{\sum_{i=1}^{n} X_i}\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{n} z^{X_i}\right] \\ &= \prod_{i=1} \mathbb{E}[z^{X_i}] \\ &= ((1 - p) + zp)^n\end{aligned}$$

2.11 Distribuição de Vetores Aleatórios

Definição 2.11.1. Sejam X e Y duas variáveis aleatórias discretas. A fmp conjunta de X e Y é $p_{X,Y}(x,y) := \mathbb{P}(X = x, Y = y)$. De forma similar, se X e Y são duas variáveis aleatórias com distribuições contínuas, a fdp conjunta de X e Y é $f_{X,Y}(x,y)$. Seja A um subconjunto de $\mathbb{R}^2$,

$$\mathbb{P}((X,Y) \in A) = \mathbb{P}(\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\}) = \begin{cases} \sum_{(x,y)\in A} p_{X,Y}(x,y) & \text{, se } X \text{ e } Y \text{ são discretas.} \\ \int_{(x,y)\in A} f_{X,Y}(x,y)d(x,y) & \text{, se } X \text{ e } Y \text{ são contínuas.} \end{cases}$$

Exemplo 2.11.2. Sejam X e Y variáveis aleatórias contínuas, com a seguinte fdp

$$f_{X,Y}(x,y) = \begin{cases} cxy & \text{, if } x > 0, y > 0, x + y < 1 \\ 0 & \text{, caso contrário} \end{cases}$$

Qual é o valor de c? Observe que $\mathbb{P}((X,Y) \in \mathbb{R}^2) = 1$. Além disso, observe que

$$\begin{aligned} \mathbb{P}((X,Y) \in \mathbb{R}^2) &= \int_{(x,y)\in\mathbb{R}^2} f_{X,Y}(x,y)d(x,y) \\ &= \int_0^1 \int_0^{1-x} cxydydx \\ &= c\int_0^1 x \int_0^{1-x} ydy \\ &= c\int_0^1 x\frac{(1-x)^2}{2}dx \\ &= \frac{c}{2}\int_0^1 (x - 2x^2 + x^3)dx \\ &= \frac{c}{2}\left(\frac{1}{2} - \frac{2}{3} + \frac{1}{4}\right) \\ &= \frac{c}{2}\cdot\frac{1}{12} = \frac{c}{24} \end{aligned}$$

Conclua que $c = 24$.

Exemplo 2.11.3. Considere que X e Y are são variáveis aleatórias discretas com a fmp conjunta dada na Tabela 3.

$$\sum_{x\in\{0,1\}}\sum_{y\in\{0,1\}} p_{X,Y}(x,y) = 0.2 + 0.4 + 0.3 + 0.1 = 1$$

Tabela 3: Tabela de $p_{X,Y}(x,y)$

X/Y	0	1
0	0.2	0.4
1	0.3	0.1

Também podemos calcular $\mathbb{P}(X = 1)$.

$$\begin{aligned}\mathbb{P}(X = 1) &= \sum_{\{(x,y)\in\{0,1\}^2 : x=1\}} p_{X,Y}(x, y) \\ &= \sum_{y\in\{0,1\}} p_{X,Y}(1, y) = 0.3 + 0.1 = 0.4\end{aligned}$$

Observe que $X \sim$ Bernoulli(0.4). De forma similar, $Y \sim$ Bernoulli(0.5).

Exemplo 2.11.4. Considere Exemplo 2.11.2. Também podemos calcular $\mathbb{P}(Y < X)$.

$$\begin{aligned}\mathbb{P}(Y < X) &= \int_{\{(x,y)\in\mathbb{R}^2 : y<x\}} f_{X,Y}(x, y)d(x, y) \\ &= \int_0^{0.5}\int_0^x 24xydydx + \int_{0.5}^1\int_0^{1-x} 24xydydx \\ &= \int_0^{0.5} 24x\frac{x^2}{2}dx + \int_{0.5}^1 24x\frac{(1-x)^2}{2}dx \\ &= 3x^4\Big|_0^{0.5} + \int_{0.5}^1 12(x - 2x^2 + x^3)dx \\ &= \frac{3}{16} + 12(\frac{x^2}{2} - \frac{2x^3}{3} + \frac{x^4}{4})\Big|_{0.5}^1 \\ &= \frac{3}{16} + 1 - 3(\frac{1}{2} - \frac{1}{3} + \frac{1}{16}) = \frac{1}{2}\end{aligned}$$

Lema 2.11.5. *Se X e Y são variáveis aleatórias discretas com distribuição conjunta $p_{X,Y}(x, y)$, então a função de massa marginal de X é $p_X(x) = \sum_{y\in Im[Y]} p_{X,Y}(x, y)$. De forma similar, se X e Y são variáveis aleatórias contínuas com densidade conjunta $f_{X,Y}(x, y)$, então a densidade marginal de X é $f_X(x) = \int_{Im[Y]} f_{X,Y}(x, y)dy$.*

Demonstração. Se X e Y são discretos,

$$\begin{aligned}p_X(x) &= \mathbb{P}(X = x) \\ &= \mathbb{P}(X = x, Y \in Im[Y]) \\ &= \sum_{\{(a,b):a=x,b\in Im[Y]\}} p_{X,Y}(a, b) \qquad \text{(Definição 2.11.1)} \\ &= \sum_{b\in Im[Y]} p_{X,Y}(x, b)\end{aligned}$$

A demonstração do caso contínuo é deixada como exercício. □

Exemplo 2.11.6. Considere que

$$f_{X,Y}(x, y) = \begin{cases} c & \text{, se } 0 \le x \le 1 \text{ e } 0 \le y \le 1 \\ 0 & \text{, caso contrário} \end{cases}$$

Observe que

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx &= \int_0^1\int_0^1 cdydx \\ &= \int_0^1 cy|_0^1 dx \\ &= \int_0^1 cdx = cx|_0^1 = c\end{aligned}$$

Portanto, $c = 1$. Observe que, if $0 < x < 1$

$$\begin{aligned}F_X(x) &= \mathbb{P}(X \leq x) \\ &= \mathbb{P}(X \leq x \cap Y \in \mathbb{R}) \\ &= \int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx \\ &= \int_0^x\int_0^1 1dydx\end{aligned}$$

De forma similar, se $x < 0$, então $F_X(x) = 0$ e se $x > 1$, então $F_X(x) = 1$. Portanto $X \sim \text{Uniform}(0,1)$.

Note que a mesma conclusão segue do Lema 2.11.5:

$$f_X(x) = \int_{Im[Y]} f_{X,Y}(x,y)dy = \int_0^1 1dy = 1,$$

que é a densidade de uma variável aleatória uniforme em $(0,1)$.

Definição 2.11.7. Sejam X e Y variáveis aleatórias. A distribuição acumulada conjunta de (X,Y) é

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y)$$

Lembramos da definição que X e Y são independentes se e somente se $p_{X,Y}(x,y) = p_X(x)p_Y(y)$. A independência é definida de forma similar para variáveis aleatórias contínuas:

Definição 2.11.8. Duas variáveis aleatórias contínuas X e Y são independentes se e somente se $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Lema 2.11.9. *Sejam X e Y duas variáveis aleatórias. X e Y são independentes se e somente se para todos $A, B \subset \Re$, $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$.*

Lema 2.11.10. *Duas variáveis contínuas, X e Y, são independentes se e somente se a distribuição conjunta pode ser escrita como $f_{X,Y}(x,y) = g(x)h(y)$ para alguma função g e h.*

Demonstração. Se X e Y são independentes por definição $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, que prova a condição suficiente. Além disso, se $f_{X,Y}(x,y) = g(x)h(y)$, então $f_X(x) = \int f_{X,Y}(x,y)dy = \int g(x)h(y)dy = g(x)C$, onde C não depende de x. Como $g(x)C$ deve integrar para um, segue que $g(x)$ é proporcional à densidade de X. De forma similar, $h(y)$ deve ser proporcional a densidade de Y. Segue que $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ e portanto X e Y são independentes. **Observação:** Fizemos em aula uma versão mais detalhada dessa demonstração, se alguém precisar pode pedir para a Aisha. □

Em geral, um vetor aleatório é uma função $\mathbf{X} : \Omega \longrightarrow \mathbb{R}^d$ que é tipicamente representada como $\mathbf{X} = (X_1, \ldots, X_d)$. Isto é, cada coordenada é uma variável aleatória. Um vetor aleatório é caracterizado quanto pela sua função de distribuição acumulada, tanto pela sua função de densidade (multivariada):

Definição 2.11.11. A distribuição acumulada conjunta de $\mathbf{X} = (X_1, \ldots, X_d)$ é

$$F_{\mathbf{X}}(\mathbf{x}) := \mathbb{P}(X_1 \leq x_1, \ldots, X_d \leq x_d),$$

onde $\mathbf{x} = (x_1, \ldots, x_d)$. Se as componentes de $\mathbf{X}$ são contínuas, a fdp de $\mathbf{X}$, $f_{\mathbf{X}}(\mathbf{x})$, é uma função não negativa tal que

$$\mathbb{P}(\mathbf{X} \in B) = \int_B f_{\mathbf{X}}(x_1, \ldots, x_d) dx_1 \ldots dx_d$$

para qualquer conjunto $B \subset \mathbb{R}^d$. Além disso, se as componentes de $\mathbf{X}$ são discretas, a fmp de $\mathbf{X}$, $p_{\mathbf{X}}(\mathbf{x})$, é uma função não negativa tal que

$$\mathbb{P}(\mathbf{X} \in B) = \sum_{(x_1,\ldots,x_d) \in B} p_{\mathbf{X}}(x_1, \ldots, x_d)$$

para qualquer conjunto $B \subset \mathbb{R}^d$.

Temos a seguinte versão da lei do estatístico inconsciente para vetores aleatórios:

Lema 2.11.12. *Seja $\mathbf{X} = (X_1, \ldots, X_d)$ um vetor aleatório. Definimos a função $g : \Re^d \longrightarrow \Re$. Então $g(\mathbf{X})$ é uma variável aleatória com esperança*

$$\mathbb{E}[g(\mathbf{X})] = \sum_{\mathbf{x} \in \chi} g(\mathbf{x}) p_{\mathbf{X}}(\mathbf{x})$$

se $\mathbf{X}$ tem distribuição discreta (onde $\chi \subset \mathbb{R}^d$ denotam os valores que $\mathbf{x}$ assume), e

$$\mathbb{E}[g(\mathbf{X})] = \int \ldots \int g(\mathbf{x}) f_{\mathbf{X}}(x_1, \ldots, x_d) dx_1 \ldots dx_n$$

Demonstração. Vamos provar o lema para o caso discreto. Da definição de esperança temos que

$$\begin{aligned}
\mathbb{E}[g(\mathbf{X})] &= \sum_{w \in \Omega} g(\mathbf{X}(w)) \mathbb{P}(\{w\}) = \\
&= \sum_{\mathbf{x} \in \chi} \sum_{w : \mathbf{X}(w) = \mathbf{x}} g(\mathbf{X}(w)) \mathbb{P}(\{w\}) \\
&= \sum_{\mathbf{x} \in \chi} g(\mathbf{x}) \sum_{w : \mathbf{X}(w) = \mathbf{x}} \mathbb{P}(\{w\}) \\
&= \sum_{\mathbf{x} \in \chi} g(\mathbf{x}) \cdot p_{\mathbf{X}}(\mathbf{x})
\end{aligned}$$

□

Definição 2.11.13. Se X e Y são variáveis aleatórias discretas, definimos a *massa de probabilidade condicional* de X dado que $Y = y$ por

$$p_{X|Y}(x|y) = \mathbb{P}(X = x | Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)},$$

para todos os valores de y tais que $\mathbb{P}(Y = y) > 0$.

Note que, para um dado y, $g(x) = p_{X|Y}(x|y)$ é uma fmp n sentido em que a estudamos antes.

Exemplo 2.11.14. No Exemplo 2.11.3,

$$p_{X|Y}(0|0) = \frac{0.2}{0.5} = 0.4 = 1 - p_{X|Y}(1|0)$$

e

$$p_{X|Y}(0|1) = \frac{0.4}{0.5} = 0.8 = 1 - p_{X|Y}(1|1).$$

Remark: Podemos extender de forma trivial a Definição 2.11.13 para vetores aleatórios.

Definição 2.11.15. Se X e Y são variáveis aleatórias contínuas com densidade conjunta $f(x, y)$, definimos a *densidade de probabilidade condicional* de X dado $Y = y$ por

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)},$$

para todos os valores de y tais que $f_Y(y) > 0$.

Note que, para um dado y, $g(x) = f_{X|Y}(x|y)$ é uma fdp no sentido em que a estudamos antes.

Exemplo 2.11.16. Seja densidade conjunta de X e Y definida por

$$f(x, y) = \begin{cases} \frac{15}{2}x(2 - x - y) & \text{se } 0 < x, y < 1 \\ 0 & \text{caso contrário} \end{cases}$$

Para calcular a densidade condicional de $X|Y = y$ quando $0 < y < 1$, primeiro observamos que

$$f_Y(y) = \int_0^1 f(x, y)dx = \int_0^1 \frac{15}{2}x(2 - x - y)dx = \frac{15}{2}(\frac{2}{3} - \frac{y}{2}).$$

Segue da Definição 2.11.15 que, para $0 < x, y < 1$,

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{x(2 - x - y)}{\frac{2}{3} - \frac{y}{2}} = \frac{6x(2 - x - y)}{4 - 3y}.$$

Distribuições condicionais podem ser usadas para checar independência:

Lema 2.11.17. *Sejam X e Y duas variáveis aleatórias discretas. X é independente de Y se e somente se $\mathbb{P}(X = x|Y = y) = \mathbb{P}(X = x)$ para todos x e y. Analogamente, sejam X e Y duas variáveis aleatórias contínuas. X é independente de Y se e somente se $f(x|y) = f(x)$ para todos x e y.*

O Teorema de Bayes também pode ser usado para variáveis discretas e contínuas:

Lema 2.11.18. *Sejam X e Y variáveis aleatórias contínuas, e N e M variáveis aleatórias discretas. Então valem as seguintes versões do Teorema de Bayes:*

1. $$\mathbb{P}(N = n|M = m) = \frac{\mathbb{P}(M = m|N = n)\mathbb{P}(N = n)}{\sum_n \mathbb{P}(M = m|N = n)\mathbb{P}(N = n)}$$

2. $$f(x|y) = \frac{f(y|x)f(x)}{\int_{\mathbb{R}} f(y|x)f(x)dx}$$

Também é possível falar sobre distribuições condicionais de vetores aleatórios que são parcialmente contínuos e parcialmente discretos. Assim, se X é contínua e N é discreta, então

$$f(x|N=n) = \frac{\mathbb{P}(N=n|X=x)f(x)}{\int_{\mathbb{R}} \mathbb{P}(N=n|X=x)f(x)dx}$$

e

$$\mathbb{P}(N=n|X=x) = \frac{f(x|N=n)\mathbb{P}(N=n)}{\sum_n f(x|N=n)\mathbb{P}(N=n)}.$$

Observe que $\mathbb{P}(N=n) = \int_{\mathbb{R}} \mathbb{P}(N=n|X=x)f(x)dx$ e $f(y) = \int_{\mathbb{R}} f(y|x)f(x)dx$ são versões contínuas da Lei das Probabilidades Totais.

Como as probabilidades condicionais também são probabilidades, também vale que, para toda variável aleatória Z, $\mathbb{P}(N=n|Z=z) = \int_{\mathbb{R}} \mathbb{P}(N=n|Z=z, X=x)f(x|Z=z)dx$ e $f(y|Z=z) = \int_{\mathbb{R}} f(y|Z=z,x)f(x|Z=z)dx$.

Exemplo 2.11.19. Assuma que escolhemos um número $U \sim \text{Unif}(0,1)$. Então lançamos uma moeda com probabilidade de cara U. Seja X o indicador de cara. Temos que

$$\mathbb{P}(X=1) = \int_0^1 \mathbb{P}(X=1|u)f(u)du = \int_0^1 u \cdot 1 du = 1/2.$$

Segue do Teorema de Bayes que

$$f(u|X=1) = \frac{\mathbb{P}(X=1|u)f(u)}{1/2} = 2u.$$

Isto é, $U|X=1 \sim \text{Beta}(2,1)$

Observação: Note que na verdade não precisamos calcular o denominador no Teorema de Bayes para encontrar a distribuição de $U|X=1$ no último exemplo. Isso ocorre pois o Teorema afirma que $f(u|x) = C\mathbb{P}(X=x|u)f(u)$, onde C é constante em u (mais precisamente, é a única constante que faz com que $f(u|x)$ seja uma densidade genuina que integra para um). segue que $f(u|x) = c \cdot u$, e a única constante que faz com que esta densidade integre para um é 2.

Também é possível definir independência condicional de variaveis aleatórias:

Definição 2.11.20. Dizemos que duas variáveis aleatórias discretas X e Y são independentes dado uma variável aleatória Z se $\mathbb{P}(X=x, Y=y|Z=z) = \mathbb{P}(X=x|Z=z)\mathbb{P}(Y=y|Z=z)$

A independência condicional é definida de forma análoga para vetores aleatórios contínuos.

Exemplo 2.11.21. Considere Exemplo 2.11.19 novamente, mas agora lançando a moeda duas vezes. Sejam X_1 e X_2 as indicadoras de que o resultado é caras. É razoável assumir que X_1 é independente de X_2 dado que a probabilidade de caras é $U=u$. Segue que

$$\mathbb{P}(X_2=1|X_1=1) = \int_0^1 \mathbb{P}(X_2=1|X_1=1,u)f(u|X_1=1)du.$$

Segue do Exercício 2.11.19 que

$$\mathbb{P}(X_2=1|X_1=1) = \int_0^1 u2udu = \frac{2}{3} \neq \frac{1}{2} = \int_0^1 \mathbb{P}(X_2=1|u)f(u)du = \mathbb{P}(X_2=1),$$

i.e., X_1 e X_2 não são independentes.

2.11.1 Esperança Condicional de Variáveis Aleatórias

Aplicando a Definição a variáveis aleatórias, temos que a esperança condicional de uma variável aleatória discreta X que assume valores em χ dado que outra variável aleatória discreta Y assume o valor y é dada por

$$\begin{aligned}\mathbb{E}[X|Y=y] &:= \sum_{w\in\Omega} X(w)P(\{w\}|Y=y) = \sum_{x\in\chi}\sum_{w:X(w)=x} X(w)P(\{w\}|Y=y)\\ &= \sum_{x\in\chi} x \sum_{w:X(w)=x} P(\{w\}|Y=y) = \sum_{x\in\chi} xP(X=x|Y=y)\\ &\sum_{x\in\chi} xp_{X|Y}(x|y)\end{aligned}$$

De forma análoga, quando X e Y são contínuas,

$$\mathbb{E}[X|Y=y] = \int_{x\in\chi} xf_{X|Y}(x|y)dx$$

Exemplo 2.11.22. Assuma que a densidade conjunta de X e Y é dada por $f_{X,Y}(x,y) = e^{-x/y}e^{-y}/y$ para $0 < x, y < \infty$. A densidade condicional de X dado que $Y = y$ é dada por

$$f_{X|Y}(x|y) = \frac{e^{-x/y}e^{-y}/y}{\int_0^\infty e^{-x/y}e^{-y}/ydx} = \frac{e^{-x/y}e^{-y}/y}{e^{-y}/y\int_0^\infty e^{-x/y}dx} = 1/ye^{-x/y}.$$

Segue que $\mathbb{E}[X|Y=y] = \int_0^\infty x1/ye^{-x/y}dx = y$. Note que $X|Y=y \sim \text{Exp}(1/y)$.

Usamos a notação $\mathbb{E}[X|Y]$ para denotar a função em Y dada por $g(y) = \mathbb{E}[X|Y=y]$. Como Y é uma variável aleatória, $g(Y) = \mathbb{E}[X|Y]$ também o é. Um resultado muito útil para calcular esperanças é o seguinte

Teorema 2.11.23. *Sejam X e Y duas variáveis aleatórias. Então $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$.*

Demonstração. Provaremos o teorema no caso de X discreto. Lembramos que

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X|A_i]\cdot\mathbb{P}(A_i)$$

onde $(A_i)_i$ é a partição de Ω. Agora, seja $A_i = \{\omega\in\Omega : Y(\omega) = i\}$ para todo $i\in\chi$, a imagem de Y. Como $(A_i)_i$ é uma partição de Ω,

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X|A_i]\cdot\mathbb{P}(A_i) = \sum_{i\in\chi}\mathbb{E}[X|Y=i]p_Y(i),$$

o que conclui a prova (pense um pouco a respeito.). □

Exemplo 2.11.24. N pessoas decidem doar dinheiro a uma instituição. Seja X_i a quantia doada pela i-ésima pessoa. Assumindo que X_i é binomial com parâmetros n_B e p_B, que N é geométrica com parâmetro p_G, e que todas as variáveis aleatórias são independentes, podemos calcular a soma esperada de dinheiro que a instituição vai receber, $T := \sum_{i=1}^N X_i$ usando o Teorema 2.11.23. Primeiro, determinamos a esperança condicional

$$\mathbb{E}[T|N=n] = \mathbb{E}[\sum_{i=1}^N X_i|N=n] = \mathbb{E}[\sum_{i=1}^n X_i|N=n] = \sum_{i=1}^n \mathbb{E}[X_i|N=n] = nn_Bp_B.$$

Assim $\mathbb{E}[T|N] = Nn_Bp_B$. Segue que

$$\mathbb{E}[T] = \mathbb{E}[\mathbb{E}[T|N]] = n_Bp_B\mathbb{E}[N] = \frac{n_Bp_B}{p_G}$$

Definição 2.11.25. (Função Geradora de Momentos) A função geradora de momentos de uma variável aleatória X é uma função $M_X : \mathbb{R} \to \mathbb{R}$ definida por:

$$M_X(t) = \mathbb{E}[e^{tX}] \tag{48}$$

Isto é, a função geradora de momentos é calculada através da esperança da função e^{tX}.

Lembrete - esperança de funções de variáveis aleatórias:

Seja X uma variável aleatória discreta com função massa de probabilidade p_X e que assume valores em χ, então:

$$\mathbb{E}[f(X)] = \sum_{x\in\chi} f(x)p_X(x) \tag{49}$$

E dada uma variável aleatória contínua X com densidade $f_X(x)$, o valor esperado de $g(X)$ é:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x)f_X(x)dx \tag{50}$$

Voltando à função geradora de momentos, a f.m.g. de uma variável aleatória $X \sim Bernoulli(p)$ é dada por $(1-p)+e^tp$[32].

Demonstração.

$$\mathbb{E}[e^{tX}] = \sum_{i=\{0,1\}} e^{tx_i}p_X(x_i) =$$

$$= e^{t0} * \mathbb{P}(X = 0) + e^{t1} * \mathbb{P}(X = 1) =$$

$$= 1 * (1 - p) + e^t * p = (1 - p) + e^tp$$

□

Sabemos que se $Y \sim Binomial(n, p)$ então podemos escrever Y como a soma de n variáveis com distribuição Bernoulli, isto é:

$$Y = \sum_{i=1}^{n} X_i \tag{51}$$

com $X_i \sim Bernoulli(p), \forall i \in \{1, ..., n\}$.

Então podemos calcular a f.g.m. de Y apenas utilizando propriedades de esperança e da função exponencial[33].

[32] É possível demonstrar que a função geradora de momentos está definida de forma única para cada variável aleatória, sendo assim, se duas variáveis aleatórias tem a mesma f.g.m., obrigatoriamente elas são a mesma variável. A f.g.m. então é uma forma de caracterizar variáveis aleatórias, assim como as funções massa, densidade e acumuladas. Uma limitação é que nem sempre a função geradora de momentos está bem definida, como no caso da distribuição de Cauchy, mas este é um caso que não será abordado nessa disciplina.

[33] É possível também calcular utilizando a definição, porém é um pouco mais trabalhoso.

Demonstração.

$$\mathbb{E}[e^{tY}] \overset{(51)}{=} \mathbb{E}[e^{t\sum_{i=1}^{n} X_i}] =$$
$$\overset{\text{Propriedade Soma}}{=} \mathbb{E}[e^{\sum_{i=1}^{n} tX_i}] =$$
$$\overset{\text{Propriedade Exponencial}}{=} \mathbb{E}[\prod_{i=1}^{n} e^{tX_i}] =$$
$$\overset{\text{Propriedade Esperança}}{=} \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] =$$
$$\overset{\text{f.g.m Bernoulli}}{=} \prod_{i=1}^{n} M_X(t) =$$
$$\overset{\text{Prop. Produtório}}{=} (M_X(t))^n =$$
$$\overset{\text{Exemplo anterior}}{=} ((1-p) + e^t p)^n$$

□

Agora considere $Z \sim Normal(0,1)$. Vamos calcular a f.g.m. de Z.

Demonstração.

$$\begin{aligned}
M_Z(t) = \mathbb{E}[e^{tz}] &= \\
&= \int_{-\infty}^{+\infty} e^{tz} f_Z(z) dz = \\
&= \int_{-\infty}^{+\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = && \text{(Densidade da Normal)} \\
&= e^{\frac{t^2}{2}} e^{-\frac{t^2}{2}} \int_{-\infty}^{+\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = && \text{(T.M.Q.D.C.)} \\
&= e^{\frac{t^2}{2}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2} + \frac{2zt}{2} - \frac{t^2}{2}} dz = && \text{(Propriedade Exponencial)} \\
&= e^{\frac{t^2}{2}} \underbrace{\int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{-(z-t)^2}{2}} dz}_{\text{Densidade da Normal(t,1)}} = \\
&= e^{\frac{t^2}{2}}
\end{aligned}$$

□

Exercícios (justifique suas respostas apresentando os devidos cálculos):

a. Seja $X \sim Normal(\mu, \sigma^2)$. Calcule $M_X(t)$.[34]

b. Mostre, usando função geradora de momentos, que se $X \sim Normal(\mu_1, \sigma_1^2)$ e $Y \sim Normal(\mu_2, \sigma_2^2)$ e X e Y são independentes, então $X + Y \sim Normal(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

c. (Adaptado de ANPEC, 2016) Verdadeiro ou Falso? *Sejam Y_1 e Y_2 variáveis aleatórias independentes, cada uma delas com distribuição normal padrão. Então podemos dizer que a v.a. $X = Y_1 + Y_2$ tem distribuição normal padrão.*

[34] Dica: Se $X \sim Normal(\mu, \sigma^2)$ então $\frac{X-\mu}{\sigma} \sim Normal(0,1)$.

d. (Adaptado de ANPEC, 2015) Verdadeiro ou Falso? *Sejam X_1 e X_2 variáveis aleatórias independentes com médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2, respectivamente, e considere a combinação linear $Y = aX_1 + bX_2$, onde a e b são constantes reais. Então, $Y \sim Normal(a\mu_1 + b\mu_2, \sigma_1^2 + \sigma_2^2)$.*

e. Seja $X \sim Poisson(\lambda)$, isto é, $\mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$. Encontre $M_X(t)$.[35] Se $X_i \sim Poisson(\lambda_i)$ para $i = 1, ..., n$, e os X_i são independentes, qual a distribuição de $\sum_{i=1}^n X_i$?

f. Sejam $X_1, ..., X_n$ independentes e identicamente distribuídas. Mostre que $M_y(t) = (M_X(t))^n$, onde $Y = \sum_{i=1}^n X_i$.

Soluções

a. Se $X \sim N(\mu, \sigma^2)$, sabemos que $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$. Queremos calcular $E[e^{tx}]$.

Demonstração.

$$
\begin{aligned}
\mathbb{E}[e^{tx}] &= && \text{(Soma e subtrai } \mu\text{)} \\
\mathbb{E}[e^{t(x-\mu+\mu)}] &= && \text{(Multiplica e divide por } \sigma\text{)} \\
\mathbb{E}[e^{t(\frac{(x-\mu)\sigma}{\sigma}+\mu)}] &= && \text{(Substitui } Z = \frac{X-\mu}{\sigma}\text{)} \\
\mathbb{E}[e^{t(Z\sigma+\mu)}] &= && \text{(Propriedade da função exponencial)} \\
\mathbb{E}[e^{t(Z\sigma)}e^{t\mu}] &= && \text{(Propriedade do valor esperado)} \\
e^{t\mu}\mathbb{E}[e^{t(Z\sigma)}] &= && \text{(Definição)} \\
e^{t\mu}M_Z(t\sigma) &= && \text{(Exemplo anterior)} \\
e^{t\mu}e^{\frac{(t\sigma)^2}{2}} &= && \text{(Propriedade da exponencial)} \\
e^{t\mu+\frac{(t\sigma)^2}{2}} & &&
\end{aligned}
$$

□

b. Considere $W = X + Y$. Queremos calcular $M_W(t)$.

[35] Dica: $\sum_{i=0}^{+\infty} \frac{x^i}{i!} = e^x$

Demonstração.

$$\begin{aligned}
\mathbb{E}[e^{tW}] &= && \text{(Substitui W por } X+Y\text{)}\\
\mathbb{E}[e^{t(X+Y)}] &= && \text{(Propriedade distributiva)}\\
\mathbb{E}[e^{tX+tY}] &= && \text{(Propriedade função exponencial)}\\
\mathbb{E}[e^{tX}e^{tY}] &= && \text{(Independência de } X \text{ e } Y\text{)}\\
\mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}] &= && \text{(Definição)}\\
M_x(t)M_Y(t) &= && \text{(Exercício } \textbf{a.}\text{)}\\
e^{t\mu_1+\frac{(t\sigma_1)^2}{2}}e^{t\mu_2+\frac{(t\sigma_2)^2}{2}} &= && \text{(Propriedade f. exponencial)}\\
e^{t\mu_1+\frac{(t\sigma_1)^2}{2}+t\mu_2+\frac{(t\sigma_2)^2}{2}} &= && \text{(Propriedade f. exponencial)}\\
e^{t\mu_1+t\mu_2+\frac{(t\sigma_1)^2}{2}+\frac{(t\sigma_2)^2}{2}} &= && \text{(Propriedade f. exponencial)}\\
e^{t(\mu_1+\mu_2)+\frac{t^2(\sigma_1^2+\sigma_2^2)}{2}} &= && \text{(Manipulação no expoente)}\\
e^{t(\mu_1+\mu_2)+\frac{\left(t\sqrt{\sigma_1^2+\sigma_2^2}\right)^2}{2}} & &&
\end{aligned}$$

□

Note que essa é a f.g.m. de uma variável com distribuição normal com média $\mu_1+\mu_2$ e variância $\sigma_1^2+\sigma_2^2$ (isso decorre diretamente do fato de que a função geradora de momentos caracteriza unicamente uma v.a., conforme falado anteriormente). Como as variáveis são independentes, esse resultado condiz com o que esperaríamos, pois lembre-se que para quaisquer v.a.'s $X_1,\cdots,X_n$ (independentes ou não), teremos:

$$\mathbb{E}[X_1+X_2+\cdots+X_n]=\mathbb{E}[X_1]+\mathbb{E}[X_2]+\cdots+\mathbb{E}[X_n] \quad \text{(linearidade do valor esperado)}$$

E ainda, se $X_1,\cdots,X_n$ forem independentes, teremos:

$$Var[X_1+X_2+\cdots+X_n]=Var[X_1]+Var[X_2]+\cdots+Var[X_n]$$

(as provas para o caso discreto com duas variáveis são sugeridas como exercício).

c. A alternativa é falsa.
Sabemos que $X\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$ se $X=Y_1+Y_2$ com $Y_1\sim N(\mu_1,\sigma_1^2)$, $Y_2\sim N(\mu_2,\sigma_2^2)$ e Y_1 é independente de Y_2. Se ambas variáveis tem distribuição normal-padrão, então $\mu_1=\mu_2=0$ e $\sigma_1=\sigma_2=1$. Logo, $\mu_1+\mu_2=0$ como o exercício sugere, porém $\sigma_1^2+\sigma_2^2=1+1=2\neq 1$ e, portanto, $Y_1+Y_2\sim N(0,2)$.

d. A alternativa é falsa.
Das propriedades de valor esperado e variância, temos:

$$\mathbb{E}[Y]=\mathbb{E}[aX_1+bX_2]=\mathbb{E}[aX_1]+\mathbb{E}[bX_2]=a\mathbb{E}[X_1]+b\mathbb{E}[X_2]=a\mu_1+b\mu_2$$

$$Var[Y]=Var[aX_1+bX_2]\overset{\text{Indep.}}{=}Var[aX_1]+Var[bX_2]=a^2Var[X_1]+b^2Var[X_2]=a^2\sigma_1^2+b^2\sigma_2^2$$

Sabemos ainda, dos exercícios anteriores, que a soma de variáveis aleatórias independentes com distribuição normal tam-

bém tem distribuição normal.

Então, $Y \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2$.

e.

$$
\begin{aligned}
\mathbb{E}[e^{tX}] = & \quad \text{(Def. esperança)} \\
\sum_{x=0}^{+\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} = & \quad \text{(Reescrevendo)} \\
e^{-\lambda} \sum_{x=0}^{+\infty} (e^t)^x \frac{\lambda^x}{x!} = & \quad \text{(Juntando termos)} \\
e^{-\lambda} \sum_{x=0}^{+\infty} \frac{(\lambda e^t)^x}{x!} = & \quad \text{(Expansão de Taylor de } e^{\lambda e^t}\text{)} \\
e^{-\lambda} e^{\lambda e^t} = & \quad \text{(Propr. Exponencial)} \\
e^{-\lambda + \lambda e^t} = & \quad \text{(Manip. algébrica)} \\
e^{\lambda(e^t - 1)} &
\end{aligned}
$$

Agora considere $W = \sum_{i=1}^n X_i$. Então,

$$
\begin{aligned}
\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t\sum_{i=1}^n X_i}] = \mathbb{E}[e^{\sum_{i=1}^n tX_i}] = \\
\mathbb{E}[\prod_{i=1}^n e^{tX_i}] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] = \prod_{i=1}^n M_{Xi}(t) = \\
\prod_{i=1}^n (e^{\lambda_i(e^t-1)}) = (e^{\sum_{i=1}^n \lambda_i(e^t-1)})
\end{aligned}
$$

Logo, $W \sim Poisson(\sum_{i=1}^n \lambda_i)$.

f.

$$
\begin{aligned}
\mathbb{E}[e^{tX}] = \mathbb{E}[e^{t\sum_{i=1}^n X_i}] = \mathbb{E}[e^{\sum_{i=1}^n tX_i}] = \\
\mathbb{E}[\prod_{i=1}^n e^{tX_i}] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] = \prod_{i=1}^n M_{Xi}(t) = \\
\prod_{i=1}^n M_X(t) = (M_X(t))^n
\end{aligned}
$$

Como os X_i são indenticamente distribuídas, temos que $M_{X1}(t) = M_{X2}(t) = \cdots = M_{Xn}(t)$, isto é, $M_X(t)$ não depende de i. Isso é diferente do exercício anterior onde os parâmetros λ_i eram diferentes para cada X_i.

Um último tópico que eu comentei na aula de quinta feira mas ainda não estava formalizado aqui é sobre *núcleo de distribuições*. Na aula, vimos como utilizar o núcleo de uma $Gama(k + a, \lambda)$ para fazer aparecer a densidade e com isso simplificar a integral que estava no meio das nossas contas. Essa operação, de multiplicar e dividir pela constante para que

uma densidade apareça, é bastante comum.

De acordo com Bauwens et al. (2000), a noção de núcleo foi introduzida por Schlaifer and Raiffa (1961). Informalmente, o núcleo de uma distribuição e/ou de uma verossimilhança é tudo que resta da função após remover as contantes. Formalmente, temos a definição a seguir:

Definição 2.11.26. Núcleo de uma distribuição (Adaptado de Bauwens et al. (2000))
O núcleo (em inglês *Kernel*) de uma densidade $f_x(\cdot)$ é uma função $K(x)$ tal que

$$f_X(\cdot) = \frac{K(x)}{\int K(x)dx} \tag{52}$$

O núcleo inclui todos os fatores da densidade $f_X(\cdot)$ que dependente de x. Uma implicação direta da definição 52 é que $\frac{K(x)}{f_X(\cdot)}$ é uma constante em relação a x. Será conveniente utilizar o símbolo de proporcionalidade ($\propto$) para escrever:

$$f_X(\cdot) \propto K(x) \tag{53}$$

(significa "$f_X(\cdot)$ é prorcional a $K(x)$)". A notação em (53) nos ajuda a explicitar a relação funcional entre a densidade e o núcleo sem precisar ficar "carregando" junto a constante de integração. Note que dependendo do que se inclui no núcleo, é possível que ele não seja único, uma vez que na definição 52, se multiplicarmos numerador e denominador pela mesma constante, a igualdade se mantém. Então, para não haver nenhuma ambiguidade, vamos falar de núcleo sem que hajam constantes desnecessárias nele (e dessa forma teremos que o núcleo é um só).

2.12 *Alguns Modelos Multivariados

E o último conceito de probabilidade que ficou pendente é o das distribuições multivariadas tabeladas. As definições abaixo foram retiradas de Stern and Izbicki (2016). Tal como no caso univariado, há muitos modelos de probabilidade multivariados bem conhecidos. Aqui, vamos ilustrar dois deles.

Definição 2.12.1. A esperança de um vetor aleatório $\mathbf{X} = (X_1, \ldots, X_d)$ é definida por $\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])$.

Definição 2.12.2. A matriz de variâncias e covariâncias de um vetor aleatório $\mathbf{X} = (X_1, \ldots, X_d)$, $\mathbb{V}[\mathbf{X}]$, é a matriz $d \times d$ cujo componente (i, j) é dado por $Cov(X_i, X_j)$.

Lema 2.12.3. *Seja* $\mathbf{X} = (X_1, \ldots, X_d)$ *e* $\vec{a}$ *um vetor* d *dimensional. Então* $\mathbb{E}[\vec{a} \cdot \mathbf{X}] = \vec{a} \cdot \mathbb{E}[\mathbf{X}]$ *e* $VAR[\vec{a} \cdot \mathbf{X}] = \vec{a} \times VAR[\mathbf{X}] \times \vec{a}^t$*, onde "*$\cdot$*"representa o produto escalar entre dois vetores, e* $\times$ *é o produto matricial.*

2.12.1 Distribuição Normal Multivariada

Definição 2.12.4. Seja $\boldsymbol{\Sigma}$ uma matriz $d \times d$ não negativa definida e $\boldsymbol{\mu} \in \mathbb{R}^d$. Dizemos que $\mathbf{X} \sim$ Normal Multivariada$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ se, para todo $\mathbf{X} \in \mathbb{R}^d$,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})'\right).$$

Lema 2.12.5. *Se* $\mathbf{X} \sim$ *Normal Multivariada*$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, então* $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ *e* $Var[\mathbf{X}] = \boldsymbol{\Sigma}$.

2.12.2 Distribuição Multinomial

A distribuição multinomial é a generalização da distribuição binomial. Suponha que lançamos um dado com d faces n vezes. Se os lançamentos são independentes e cada face tem probabilidade p_i, $i = 1, \ldots, d$, então $(X_1, \ldots, X_d) \sim$ Multinomial$(n, (p_1, \ldots, p_d))$, onde X_i é o número de vezes que a face i é observada.

Definição 2.12.6. Seja $\boldsymbol{p} = (p_1, \ldots, p_d)$ tal que $p_i > 0$ para todo i e $\sum_{i=1}^{d} p_i = 1$ e $n \in \mathbb{N}$. Dizemos que $\mathbf{X} = (X_1, \ldots, X_d) \sim$ Multinomial$(n, \boldsymbol{p})$ se

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \frac{n!}{x_1! \ldots x_n!} \prod_{i=1}^{d} p_i^{x_i},$$

para todo $x_i \geq 0$ com $\sum_{i=1}^{n} x_i = n$.

Lema 2.12.7. *Se* $\mathbf{X} = (X_1, \ldots, X_d) \sim$ *Multinomial*$(n, \boldsymbol{p})$*, então* $\mathbb{E}[\mathbf{X}] = n\boldsymbol{p}$ *e a componente* (i, j) *de* $Var[\mathbf{X}]$ *é* $-np_ip_j$ *para* $i \neq j$ *e* $np_i(1 - p_i)$ *caso contrário.*

Exemplo 2.12.8. Há três candidatos a uma eleição, A, B e C. Se 10% da população vota em A, 40% vota em B e 50% vota em C, então a probabilidade de que em uma amostra de tamanho $n = 10$, retirada desta população com reposição, obteremos 2 eleitores de A, 3 de B e 5 de C é dada por

$$\frac{10!}{2!3!5!}(0.1)^2(0.4)^3(0.5)^5.$$

3 Noções de Estatística

Até o momento nos ocupamos de probabilidades. Como dito no início da revisão, probabilidade é considerada uma área da matemática e é para a estatística uma ferramenta. Como este não é propriamente um curso de estatística, não vai ser o foco aqui ficar explicando quais são os propósitos e motivações da estatística. O que precisa ficar claro é que estaremos lidando com estatística frequentista (ou clássica) e que nosso objetivo é tirar conclusões a respeito de informações de uma população de interesse.

Na teoria estatística clássica[36] (ou frequentista), quando lidamos com amostras provenientes de uma população de interesse, um dos problemas que surgem é o de encontrar o melhor estimador para os parâmetros.

Suponha, por exemplo, que você quer investigar as diferenças de salários entre homens e mulheres na cidade de Florianópolis, mas você não tem recursos (financeiros e de tempo) para entrevistar todas as pessoas (isto é, não dá para fazer um censo). Sendo assim, você irá estudar uma *amostra* dessa população, ou seja, entrevistar apenas alguns homens e algumas mulheres e com base nas respostas deles irá *estimar* o valor do salário para todas as pessoas da cidade.

O primeiro problema que surge é o da amostragem: quantas pessoas devem ser entrevistadas? De que forma que elas devem ser selecionadas? Mas como isso é assunto de outra disciplina, vamos supor, por simplicidade, que a sua amostra já foi selecionada e que você utilizou os melhores métodos possíveis e que essa amostra é de fato representativa da população. Então resta lidar com o problema da inferência, tanto da estimação pontual (encontrar as médias salariais) como o teste de hipóteses (comparar as médias salariais). Em estatística, são as noções de amostragem que vão conectar todo o conhecimento adquirido em probabilidade com a parte de inferência estatística, onde esta última busca relacionar fenômenos observados em uma amostra com a população de interesse.

De acordo com Mood and Graybill (1963)[37], *"O problema da estimação pode ser descrito como: assuma que alguma característica dos elementos de uma população pode ser representada como uma variável aleatória X cuja densidade é $f_x(\cdot;\theta) = f(\cdot;\theta)$, onde a forma da densidade é assumida como conhecida, exceto pelo parâmetro desconhecido θ (se θ fosse conhecido, a função densidade seria completamente especificada e não haveria necessidade de fazer inferências). Adicionalmente, assumimos que os valores $x_1, x_2, \cdots, x_n$ de uma amostra aleatória $X_1, X_2, \cdots, X_n$[38] de $f(\cdot;\theta)$ pode ser observada. Com base nos valores amostrais observados $x_1, \cdots, x_n$, desejamos estimar o valor desconhecido de θ ou de alguma função desse parâmetro, que poderia ser $\tau(\theta)$. Essa estimativa pode ser feita de duas formas: a primeira é chamada de estimação pontual, onde o valor de alguma estatística[39], por exemplo, $t(X_1, \cdots, X_n)$ representa ou estima o valor desconhecido $\tau(\theta)$ - a estatística $t(X_1, \cdots, X_n)$ é chamada de estimador pontual. O segundo método, chamado de método estimação por intervalo, envolve definir duas estatísticas, digamos $t_1(X_1, \ldots, X_n,)$ e $t_2(X_1, \ldots, X_n,)$ onde $t_1(X_1, \ldots, X_n,) < t_2(X_1, \ldots, X_n,)$ de forma que $(t_1(X_1, \ldots, X_n,), t_2(X_1, \ldots, X_n,))$ é um intervalo (aleatório) para o qual podemos calcular a probabilidade que contenha o valor desconhecido $\tau(\theta)$.*

Para entender o que o trecho do livro diz, considere $X_1, \cdots, X_n$ uma a.a. (amostra aleatória - será definida daqui a pouco)

[36] Existem duas "escolas de pensamento" na inferência estatística: a estatística clássica e a estatística bayesiana. A primeira é aquela comumente utilizada nos cursos atuais e chamada simplesmente de estatística e será o foco deste texto.

[37] Tradução livre.

[38] Fazemos a diferenciação entre valores em letra maiúscula para indicar valores não observados dos observados, que são denotados por letras minúsculas. Então, sempre que forem usadas letras minúsculas, a amostra já foi observada.

[39] Define-se como *estatística* uma função que depende apenas do resultado de uma amostra. Um valor particular de uma estatística é chamado de estimativa. Por exemplo, a média amostral ($\bar{X}$) definida como $\frac{\sum_{i=1}^{n} X_i}{n}$ é uma estatística.

de densidade $f(\cdot, \underset{\sim}{\theta})$, onde a forma da densidade é conhecida, mas não sabemos o valor de $\underset{\sim}{\theta}$. Repare que $\underset{\sim}{\theta}$ é um *vetor de parâmetros* que desejamos estimar. Queremos, então, montar uma estatística (função da amostra) que possa ser usada para encontrar $\tau(\theta)$, nossa função do parâmetro desconhecido.

Por exemplo, no caso da distribuição Normal de parâmetros μ e σ^2, teremos $\underset{\sim}{\theta} = (\mu, \sigma^2)$, vetor de parâmetros desconhecidos. Precisamos encontrar uma função t que, aplicada nos valores da amostra, produz uma estimativa para $\tau(\underset{\sim}{\theta})$. Uma proposta *natural* para estimar a média populacional é usar a média da amostra. Conforme veremos, a média amostral, no caso da Normal, é um estimador não-viesado e consistente para μ, porém a variância $\hat{\sigma^2}$ não é um bom estimador para a variância populacional por apresentar o que chamamos de *viés*.

Alguns termos serão utilizados daqui para frente então as seguintes definições irão auxiliar para uniformizar a nomenclatura. Elas foram adaptadas de Casella and Berger (2002) e Mood and Graybill (1963).

Definição 3.0.1. (Amostra Aleatória) Sejam as variáveis aleatórias $X_1, X_2, \ldots, X_n$ com densidade conjunta dada por $f_{X_1,X_2,\cdots,X_n}(\cdot, \ldots, \cdot)$ que pode ser escrita da seguinte forma:

$$f_{X_1,X_2,\cdots,X_n}(\cdot, \ldots, \cdot) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

onde $f(\cdot)$ é a densidade de cada X_i. Então, $X_1, X_2, \ldots, X_n$ é definida como sendo *uma amostra aleatória* de tamanho *n* de uma população com densidade $f(\cdot)$.

Definição 3.0.2. Amostra aleatória (a.a.) - outra definição
As variáveis aleatórias $X_1, X_2, \ldots, X_n$ formam uma *amostra aleatória de tamanho n da população com densidade $f_X(x)$* se $X_1, X_2, \ldots, X_n$ são mutualmente independentes e cada uma das densidades marginais de $X_1, \ldots, X_n$ é a própria $f_X(x)$. Neste caso, também se diz que $X_1, X_2, \ldots, X_n$ são variáveis aleatórias independentes e identicamente distribuídas[40] com f.m.p. ou f.d.p. $f_X(x)$.

Em outras palavras, se tivermos uma população cujos elementos tem uma densidade $f(\cdot)$ (que pode ser normal, bernoulli, exponencial ou até mesmo uma densidade não tabelada), diremos que os elementos dessa população, quando selecionados com reposição e seguindo todo o "ritual"de amostragem probabilística (logo, serão independentes entre si e portanto sua densidade conjunta é o produto das densidades individuais) formam uma amostra aleatória da população de onde foram retirados.

Uma observação importante é a notação: a variável aleatória X_i é a *representação* do que o valor do i-ésimo elemento amostrado poderá vir a assumir. Enquanto a amostra não é coletada, esse valor é desconhecido e aleatório. Após a observação da amostra, o X_i irá receber o valor x_i. Somente faz sentido falar em distribuição quando ainda não observamos uma amostra específica, depois disso ela passa a a ser uma constante.

Por exemplo, suponha que desejamos montar uma amostra aleatória com $n = 10$ com as as alturas dos alunos da sala. *Antes* de observação dos dez valores, nós sabemos[41] que as alturas seguem uma distribuição aproximadamente normal, que sua média estará aproximadamente em torno de $1,75m$, mas ainda não observamos nada. Então, fazemos o sorteio de 10 alunos e anotamos suas alturas. Essas alturas coletadas serão uma *realização* ou *observação* do fenômeno e já não são mais aleatórias. Só que poderíamos ter feito um outro sorteio e observados outros 10 valores distintos, que seriam outra

[40] É comum encontrar a abreviação *iid*.
[41] Na verdade assumimos que segue uma distribuição - a menos que fossem números gerados artificialmente...

realização. É neste sentido que a amostra é aleatória antes da coleta: ela pode receber qualquer valor daquela população, mas uma vez observada, já não é mais aleatória.

Exemplo 3.0.3. a.a. com $n = 2$ de uma Bernoulli

Suponha que X só pode assumir dois valores, 0 e 1, com probabilidades p e $q = 1 - p$, respectivamente. Isto é, X é uma variável aleatória discreta com distribuição de Bernoulli:

$$p_X(x) = p^x q^{1-x} \mathbb{I}_{\{0,1\}}(x)$$

Onde $\mathbb{I}$ é a função *indicadora*, que será igual a 1 se $x = 0$ ou $x = 1$ e será igual a 0 em todos os outros casos[42].

A função densidade conjunta para uma amostra aleatória da $f(\cdot)$ que tenha 2 valores é:

$$f_{X_1,X_2}(x_1, x_2) = f(x_1)f(x_2) = p^{x_1+x_2} q^{2-x_1-x_2} \mathbb{I}_{\{0,1\}}(x_1)\mathbb{I}_{\{0,1\}}(x_2)$$

Observe que isso *não* é igual à montar a distribuição do número de sucessos na retirada com reposição de dois elementos de uma distribuição de Bernoulli. Observe ainda que não é possível ter uma amostra aleatória quando o processo de amostragem é sem reposição, pois os sorteios não seriam independentes entre si. Uma forma alternativa de nomenclatura é dizer que os $X_1, X_2, \ldots, X_n$ são *independentes e identicamente distribuídos* (**i.i.d**).

Um problema que surge é o de como saber qual a distribuição da amostra aleatória ou então qual a distribuição de uma estatística[43]. específica. Uma ferramenta útil para conseguir avaliar funções de variáveis aleatórias é a *função geradora de momentos* (f.g.m.) que foi vista anteriormente.

Note que, nas definições de a.a. dadas acima, uma vez que as variáveis envolvidas na amostra aleatória são independentes, é necessário que o processo de amostragem seja feito *com reposição* se a população for finita (se ela for infinita, tanto faz). Caso contrário, o fato de sortear uma unidade amostral afetaria a probabilidade de sorteio das unidades ainda não selecionadas.

Da definição (3.0.2), temos que a densidade conjunta[44] de $X_1, \ldots, X_n$ será dada por:

$$f(x_1, \ldots, x_n) = f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_n) = \prod_{i=1}^{n} f(x_i)$$

A densidade conjunta pode ser utilizada para calcular probabilidades referentes à amostra. Note que a notação do produtório pôde ser utilizada pois todas as densidades são iguais, para todo $i \in \{1, \ldots, n\}$. Em particular, se a densidade populacional depende de um vetor de parâmetros[45], por exemplo, $\boldsymbol{\theta}$ com f.m.p. ou f.d.p. dada por $f(\boldsymbol{x}|\boldsymbol{\theta})$, então a densidade conjunta pode ser denotada por:

$$f(x_1, \ldots, x_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

No contexto da estatística, assumimos que a população em estudo segue uma determinada distribuição porém o verdadeiro valor de $\boldsymbol{\theta}$ é desconhecido. Pode-se, então, calcular como é o comportamento de amostras aleatórias para diferentes

[42] Para saber mais: https://en.wikipedia.org/wiki/Indicator_function

[43] **Estatística** é uma função da amostra - e como é função de variável aleatória, ela por si só é uma variável aleatória - que não depende de nenhum parâmetro desconhecido. *E é por isso que não falamos que a Aisha é estatística e sim uma estatista!*

[44] A partir de agora irei omitir o subscrito das funções densidade, isto é, ao invés de escrever $f_{X_1}(x_1)$, irei usar apenas $f(x_1)$.

[45] Sempre que possível, os vetores estarão denotados em **negrito** - onde possível depende da Aisha lembrar de colocar assim.

populações.

Exemplo 3.0.4. Distribuição amostral da exponencial (Retirado de Casella and Berger (2002))
Seja $X_1, \ldots, X_n$ uma a.a. de uma população com distribuição exponencial de parâmetro λ. Então, a densidade conjunta é dada por:

$$\begin{aligned} f(x_1, \ldots, x_n|\lambda) &= \prod_{i=1}^{n} f(x_i|\lambda) \\ &= \prod_{i=1}^{n} \frac{1}{\lambda} e^{-\frac{x_i}{\lambda}} \\ &= \frac{1}{\lambda^n} e^{-\frac{x_1+x_2+\ldots+x_n}{\lambda}} \\ &= \frac{1}{\lambda^n} e^{-\frac{1}{\lambda}\sum_{i=1}^{n} x_i} \end{aligned}$$

Esta densidade conjunta pode ser utilizada para responder questões a respeito da amostra, como *qual a probabilidade de que todos X_i sejam maiores que 2?*

$$\begin{aligned} \mathbb{P}(X_1 > 2, \ldots, X_n > 2) = & \\ &= \int_2^{+\infty} \ldots \int_2^{+\infty} \prod_{i=1}^{n} \frac{1}{\lambda} e^{-\frac{x_i}{\lambda}} dx_1 dx_2 \ldots dx_n \\ &= e^{-\frac{2}{\lambda}} \int_2^{+\infty} \ldots \int_2^{+\infty} \prod_{i=2}^{n} \frac{1}{\lambda} e^{-\frac{x_i}{\lambda}} dx_2 \ldots dx_n \\ &\vdots \\ &= (e^{-\frac{2}{\lambda}})^n \\ &= e^{-2n/\lambda} \end{aligned}$$

Este mesmo cálculo pode ser feito sem a necessidade de resolver n integrais, apenas utilizando o resultado de probabilidade de eventos independentes:

$$\begin{aligned} \mathbb{P}(X_1 > 2, \ldots, X_n > 2) = & \\ &= \mathbb{P}(X_1 > 2) \cdot \mathbb{P}(X_2 > 2) \cdot \ldots \cdot \mathbb{P}(X_n > 2) \\ &= [\mathbb{P}(X_1 > 2)]^n \\ &= (e^{-\frac{2}{\lambda}})^n \\ &= e^{-2n/\lambda} \end{aligned}$$

Quando discutimos probabilidade, havíamos visto que uma função de variável aleatória também é uma v.a. . Agora, na estatística, este conceito será utilizado também, para definirmos o que é uma *estatística*:

Definição 3.0.5. Estatística
Seja $X_1, \ldots, X_n$ uma a.a. de tamanho n de uma população e seja $T(X_1, \ldots, X_n)$ uma função real (ou um vetor de funções

reais) cujo domínio inclui o espaço amostral de $(X_1, \ldots, X_n)$. Então a v.a. ou o vetor aleatório $Y = T(X_1, \ldots, X_n)$ é chamado de *estatística*. A função densidade de probabilidade de uma estatística Y é chamada de *distribuição amostral de* Y.

Observação: Note que a definição de estatística é bastante abrangente e não necessariamente Y irá ser uma função do parâmetro populacional θ.

Uma vez que estatísticas são v.a., faz sentido nos perguntarmos a respeito de suas distribuições, como no exemplo a seguir.

Exemplo 3.0.6. Distribuição amostral de $\bar{X}$ para a distribuição Normal
Considere uma a.a. de tamanho n de uma população com distribuição Normal de média μ e variância σ^2, isto é, $X_1, \ldots, X_n$ são i.i.d. com $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Defina a estatística $\bar{X}$ como sendo:

$$\bar{X} := \sum_{i=1}^{n} \frac{X_i}{n}$$

Utilizando a função geradora de momentos, podemos encontrar qual a densidade de $\bar{X}$:

$$\begin{aligned}
\mathbb{E}[e^{t\bar{X}}] &= \mathbb{E}[e^{t\sum\limits_{i=1}^{n}\frac{X_i}{n}}] \\
&= \prod_{i=1}^{n} \mathbb{E}[e^{t\frac{X_i}{n}}] \\
&= \prod_{i=1}^{n} \mathbb{E}[e^{\frac{t}{n}X_i}] \\
&= \prod_{i=1}^{n} M_X(t/n) \\
&= \prod_{i=1}^{n} e^{t/n\mu + \frac{(t/n\sigma)^2}{2}} \\
&= \left(e^{t/n\mu + \frac{(t/n\sigma)^2}{2}}\right)^n \\
&= e^{t\mu + \frac{(t\sigma/\sqrt{n})^2}{2}}
\end{aligned}$$

Que é a f.g.m. de uma v.a. com distribuição Normal com média μ e variância σ^2/n.

Para uma animação sobre a distribuição amostral da média, acesse: https://istats.shinyapps.io/sampdist_cont/.

Em uma população com densidade $f(\boldsymbol{x}|\boldsymbol{\theta})$, o conhecimento a respeito de $\boldsymbol{\theta}$ é crucial para poder fazer afirmações a respeito da população sendo estudada. Porém, conforme já mencionado, em geral este valor, apesar de fixo, é desconhecido e na maioria dos casos fazer um censo para descobrir seu valor está fora de cogitação. Sendo assim, é natural buscar informações na amostra que auxiliem no conhecimento sobre os parâmetros populacionais.

Definição 3.0.7. (Estimador pontual)
Um *estimador pontual* é qualquer função $W(X_1, \ldots, X_n)$ de uma amostra, isto é, qualquer estatística é um estimador pontual.

A definição de estimador dada por Casella and Berger (2002) é, assim como a definição de estatística, bastante abrangente. Isso se deve ao fato de que os autores buscaram não eliminar da definição nenhum potencial estimador.

Observação: Assim como na probabilidade usávamos letras maiúsculas para denotar a variável aleatória (enquanto função) e letras minúsculas para denotar seus valores, agora usaremos letras maiúsculas para denotar o estimador de maneira genérica e letras minúsculas para representar os resultados obtidos em uma determinada amostra. Quando a amostra é coletada e seus valores inseridos em um estimador, obtemos uma *estimativa*.

Note que o estimador será, por construção, uma variável aleatória e portanto, terá uma distribuição. Uma vez que um estimador é uma forma de calcular uma estimativa (pois ele é uma função), a cada nova amostra, teremos uma nova estimativa. Quando dizemos que $\bar{X}$ do exemplo (3.0.6) segue uma distribuição normal com média μ, isso significa que se tomarmos muitas amostras e para cada uma delas calcular o valor de $\bar{X}$ e após calcular a média destes valores, o valor que esperaríamos encontrar é μ. E, uma vez que a cada amostra o valor de $\bar{X}$ está mudando, a sua variância representa este fato. Note ainda que quando maiores forem as amostras, menor será a variância de $\bar{X}$.

A seguir, vamos ver um dos métodos mais comuns de encontrar estimadores, utilizando a função de verossimilhança. Outros métodos conhecidos são o método dos momentos e o método dos mínimos quadrados (embora existam outros além destes), porém não irão ser discutidos neste texto.

3.1 Métodos para encontrar estimadores

3.1.1 Estimador de Máxima Verossimilhança

Suponha[46] que você está fazendo um estudo nacional de acompanhamento de municípios e em um determinado mês, 400 cidades decretaram estado de emergência. Passado um mês, 72 precisaram pedir ajuda ao governo federal e 328 haviam conseguido se reestruturar com recursos próprios.

Faz sentido, então, nos perguntar qual a estimativa da proporção de municípios que precisam de auxílio federal 1 mês após decretar estado de emergência.

Sob o enfoque frequentista, o primeiro passo é definir a população alvo. Podemos imaginar que a população de interesse são os municípios de determinada região que decretaram estado de emergência em um dado mês, municípios do Brasil que decretam estado de emergência como um todo, etc. Observe que de qualquer modo, o que observamos não é uma amostra aleatória da população de referência, porque os municípios não têm todos a mesma probabilidade de "seleção". Essa dificuldade em definir populações e extrair amostras aleatórias das mesmas é uma das fragilidades da estatística frequentista, mas iremos abstrair esse problema por enquanto e seguir com a análise.

Partindo para a nossa estimativa a ser obtida, podemos pensar que cada um dos municípios vem de uma distribuição de Bernoulli com parâmetro desconhecido θ. Vamos denotar o i-ésimo município por Y_i, de forma a usar a notação $Y_i \sim Bernoulli(\theta)$, onde $\sim$ significa *segue uma distribuição*. Observe que poderíamos usar a letra p para o parâmetro da distribuição, porém para ressaltar que este valor é o valor desconhecido que queremos estimar, vamos usar a letra θ.

Então a probabilidade de que $Y_i = 1$, isto é, a probabilidade de que um município que decretou estado de emergência pedir ajuda ao governo federal, é igual[47] a θ. Denotaremos isso por $\mathbb{P}(Y_i = \theta)$. Observe que isso significa que estamos

[46] Exemplo adaptado do curso *Bayesian Statistics*, disponível em: https://www.coursera.org/learn/bayesian-statistics

[47] Caso isso não esteja claro para você, aqui tem uma introdução às variáveis aleatórias e fala da distribuição de Bernoulli: https://www.overleaf.com/read/cxzghqvktbpt

definindo como “sucesso” da nossa distribuição o fato do município pedir auxílio federal.

A densidade conjunta de todos os municípios indexados 1 até n será a probabilidade de que o vetor de municípios, representados com a notação vetorial $\underset{\sim}{Y}$ assuma os valores do vetor $\underset{\sim}{y}$ dado um valor de θ, isto é:

$$\mathbb{P}(\underset{\sim}{Y} = \underset{\sim}{y}|\theta) = \mathbb{P}(Y_1 = y_1 \cap Y_2 = y_2 \cap \cdots \cap Y_n = y_n|\theta) \tag{54}$$

Assumindo que todos os municípios são independentes, teremos que 54 pode ser escrita como:

$$\mathbb{P}(\underset{\sim}{Y} = \underset{\sim}{y}|\theta) = \mathbb{P}(Y_1 = y_1|\theta)\mathbb{P}(Y_2 = y_2|\theta)\cdots\mathbb{P}(Y_n = y_n|\theta) = \prod_{i=1}^{n}\mathbb{P}(Y_i = y_i|\theta) \tag{55}$$

Uma vez que definimos que a distribuição é de Bernoulli, então 55 será:

$$\mathbb{P}(\underset{\sim}{Y} = \underset{\sim}{y}|\theta) = \prod_{i=1}^{n}\theta^{y_i}(1-\theta)^{1-y_i} \tag{56}$$

Vamos agora pensar na expressão 56 como uma função de θ (por enquanto ela ainda é uma probabilidade de $\underset{\sim}{Y}$ condicional ao valor de θ). Esse é justamente o conceito de verossimilhança ou de *função de verossimilhança*. Ela é uma função de θ dado os valores encontrados na amostra e será denotada por $L(\theta|\underset{\sim}{y})$. Para o nosso exemplo dos municípios, teremos:

$$L(\theta|\underset{\sim}{y}) = \prod_{i=1}^{n}\theta^{y_i}(1-\theta)^{1-y_i} \tag{57}$$

As funções em 56 e 57 parecem iguais, mas enquanto a primeira é uma probabilidade condicional de $\underset{\sim}{Y}$ dado θ, a segunda é uma função (e não é densidade de probabilidade) de θ dado os valores observados $y_1, \cdots, y_n$. Podemos utilizar a função de verossimilhança para encontrar o valor mais “plausível” (ou verossímil) para θ, que será o valor de θ que maximiza 57. Este $\hat{\theta}$ que maximiza a função de verossimilhança será chamado de *estimativa de máxima verossimilhança de* θ (EMV ou MLE, em inglês, de *maximum likelihood estimate*).

O próximo exemplo foi retirado de Mood and Graybill (1963) e também serve para dar a intuição do EMV.

Exemplo 3.1.1. Suponha que uma urna contém um número de bolas pretas e brancas, e suponha que razão entre elas é de $3/1$, mas não se sabe se existem mais bolas brancas ou mais bolas pretas. Em outras palavras, a probabilidade de retirar uma bola preta é de $1/4$ ou é de $3/4$. Se n bolas são retiradas com reposição, a distribuição de X, número de bolas pretas, é dada pela distribuição binomial:

$$f(x|p) = \binom{n}{x}p^x(1-p)^{n-x} \quad \text{para } x = 0, 1, 2, \ldots, n$$

Então, vamos retirar uma amostra de tamanho $n = 3$, com reposição e tentar estimar o parâmetro desconhecido p, que representa a probabilidade de retirar uma bola preta da urna. Como somente temos duas possibilidades para p (pelo enunciado), precisamos apenas calcular as probabilidades de X ser igual a $0, 1, 2, 3$ para os dois valores possíveis de p. Os valores calculados estão na tabela (4):

Neste exemplo, se $x = 0$ (isto é, em três retiradas de bolas, não houver nenhuma bola preta), a estimativa de $p = 1/4$ é preferível a $p = 3/4$ porque a probabilidade de $x = 0$ no primeiro caso é de $27/64$ e no segundo é de apenas $1/64$. Em outras palavras, é muito mais provável encontrar um resultado de $x = 0$ em uma distribuição com o parâmetro $p = 1/4$ do que se fosse $p = 3/4$. De maneira geral, neste exemplo, iremos preferir $p = 0.25$ se $x = 0$ ou $x = 1$ e $p = 0.75$ caso contrário.

Tabela 4: Probabilidades de retirada de bolas pretas para diferentes valores de p

x	0	1	2	3
$f\left(x\middle\vert p=\frac{3}{4}\right)$	$1/64$	$9/64$	$27/64$	$27/64$
$f\left(x\middle\vert p=\frac{1}{4}\right)$	$27/64$	$27/64$	$9/64$	$1/64$

O estimador então pode ser definido como:

$$\hat{p} = \hat{p}(x) = \begin{cases} .25, & \text{para } x = 0, 1 \\ .75, & \text{para } x = 2, 3 \end{cases}$$

De maneira que nosso estimador seleciona, para cada valor possível de x, o valor de p que maximiza a probabilidade de x, isto é, para cada x encontra-se o valor $\hat{p}$ tal que $f(x|\hat{p}) > f(x|p')$.

Os exemplos anteriores usaram a densidade de uma maneira não convencional, pois enxergamos ela como uma função do parâmetro dada uma amostra fixa observada. A próxima definição formaliza isto.

Definição 3.1.2. Função de Verossimilhança
Seja $f(\boldsymbol{x}|\boldsymbol{\theta})$ a densidade conjunta de uma amostra $\boldsymbol{X} = (X_1, \ldots, X_n)$. Então, dado que $\boldsymbol{X} = \boldsymbol{x}$ foi observada, a função de $\boldsymbol{\theta}$ definida como

$$\mathcal{L}(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta)$$

é chamada de *função de verossimilhança*.

Em particular, se $X_1, \ldots, X_n$ é uma amostra aleatória, então:

$$L(\theta|\boldsymbol{x}) = f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$$

Além de ser usada no estimador de máxima verossimilhança, a função de verossimilhança tem uma importância crucial na estatística clássica e na bayesiana. Em muitas situações ela vai aparecer e por isso merece um pouco mais da nossa atenção. A primeira observação a ser feita é que a verossimilhança, apesar de parecer uma densidade conjunta, *não é uma densidade*. A notação $L(\theta|x)$ deixa claro que estamos falando de uma função de θ condicional ao vetor de dados x. Neste sentido, a função de verossimilhança não precisa, por exemplo, integrar 1.

Exemplo 3.1.3. Verossimilhança da Bernoulli
Se X tem distribuição Binomial, sabemos que:

$$p_x(x) = \mathbb{P}(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{se } x = 1, 2, \ldots, \\ 0, \text{caso contrário.} \end{cases}$$

Definição 3.1.4. Estimador de máxima verossimilhança
Para cada ponto da amostra $\boldsymbol{x}$, seja $\hat{\theta}(\boldsymbol{x})$ o valor do parâmetro no qual $L(\theta|\boldsymbol{x})$ atinge seu máximo como função de θ (e $\boldsymbol{x}$ fixo). O estimador de máxima verossimilhança (EMV) do parâmetro θ baseado na amostra $\boldsymbol{X}$ é $\hat{\theta}(\boldsymbol{X})$.

Observação: Se a função de verossimilhança é diferenciável em θ_i, os possíveis candidatos a EMV são os valores de $(\theta_1, \ldots, \theta_k)$ que são solução de

$$\frac{\partial}{\partial \theta_i} L(\boldsymbol{\theta}|\boldsymbol{x}) = 0, \quad i = 1, \ldots, k$$

Lembre-se que pontos onde a primeira derivada é zero podem ser mínimos locais/globais, máximos locais/globais ou ainda pontos de inflexão. Estamos interessados apenas no caso onde se trata de um máximo global.

Na maioria dos casos é mais fácil trabalhar com $log(L(\theta|\boldsymbol{x}))$. Como a função logaritmica é estritamente crescente em $(0, +\infty)$, seu máximo coincide com o máximo da função de verossimilhança.

Exemplo 3.1.5. EMV da Bernoulli
Visto em aula.

Exemplo 3.1.6. EMV da Poisson
Visto em aula.

Exemplo 3.1.7. EMV do modelo Normal com variância conhecida
Ver Meyer (1973).

Exemplo 3.1.8. EMV do modelo Normal com variância desconhecida
Ver no capítulo 7 de Casella and Berger (2002).

Existe uma propriedade chamada *propriedade da invariância dos estimadores de máxima verossimilhança* que estabelece que se $\hat{\theta}$ é EMV de θ, então para qualquer função $\tau(\theta)$, o EMV de $\tau(\theta)$ é $\tau(\hat{\theta})$.

Além disso, o EMV é: consistente, eficiente e assintoticamente normal (sob algumas condições de regularidade). Note que nada foi dito sobre o viés, de fato existem EMV que são viesados. Aqui tem algumas demonstrações do EMV: https://www.overleaf.com/read/kzrcvjdmngrp .

3.2 Propriedades de Estimadores

Agora que vimos um método de encontrar estimadores, iremos ver medidas para avaliar a qualidade dos mesmos.

Definição 3.2.1. 1 Seja $W(X_1, X_2, \cdots, X_n)$ um estimador pontual para o parâmetro desconhecido θ. Dizemos que W é não-viesado caso satisfaça $\mathbb{E}[W] = \theta$. Caso contrário, dizemos que W é um estimador viesado para θ e seu viés é dado por $B_\theta[W] = \mathbb{E}[W] - \theta$.

Definição 3.2.2. 2 Uma sequência de variáveis aleatórias $X_1, X_2, \cdots, X_n$ converge em probabilidade para X se, $\forall \quad \varepsilon > 0$, temos

$$\lim_{n \to +\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0 \tag{58}$$

Notação: $X_n \xrightarrow{\mathbb{P}} X$

Definição 3.2.3. 3 Seja $W(X_1, X_2, \cdots, X_n)$ um estimador pontual para o parâmetro desconhecido θ. Dizemos que W é consistente[48] se ele converge em probabilidade para θ.

Observe que, enquanto o viés diz respeito ao estimador independente do tamanho amostral n, a consistência avalia se o estimador está se aproximando do valor do parâmetro à medida que n aumenta. Esses dois conceitos se referem à coisas diferentes: um estimador pode ser viesado e não consistente, viesado e consistente, não viesado e não consistente e não viesado e consistente. Para uma discussão a respeito desses dois conceitos, veja: http://eranraviv.com/

[48] Em economia, ao dizermos que um estimador W é consistente, dizemos que $plim[W] = \theta$

bias-vs-consistency/.

Em Gujarati (2006), é feita a consideração de que uma condição suficiente para que W seja consistente é que tanto o viés como a variância tendam para zero à medida que a amostra aumenta[49], isto é, para verificar se W é consistente, é suficiente mostrar que $\lim_{n\to\infty} \mathbb{E}[W] = \theta$ e $\lim_{n\to\infty} Var[W] = 0$.

Definição 3.2.4. 4 Seja $W(X_1, X_2, \cdots, X_n)$ um estimador pontual para o parâmetro desconhecido θ. O *erro quadrático médio* (EQM) de W é dado por $EQM[W] = \mathbb{E}[(W-\theta)^2]$. É possível mostrar que $EQM[W] = Var[W] + B_\theta[W]^2$.

De acordo com Gujarati (2006), podemos usar o EQM para avaliar a consistência de W: uma outra condição suficiente para W ser consistente é que seu EQM tenda a zero quando $n \to \infty$, o que fica evidente se usarmos a equação alternativa do EQM e combinarmos com a condição vista anteriormente. Observe ainda que um estimador não viesado terá EQM igual à sua variância. Essa explicação está melhor detalhada em Casella and Berger (2002) e utiliza a desigualdade de Chebychev para mostrar o resultado (páginas 233 - exemplo para s^2 - e 469).

Definição 3.2.5. 5 Um estimador W^* é dito ser um estimador de variância mínima para θ se $Var[W^*] \leq Var[W']$, onde W' representa qualquer outro estimador para θ.

A definição 5 é muito útil quando queremos comparar, por exemplo, dois estimadores não viesados. Uma forma de escolher entre os dois é optar por aquele que tem a menor variância. Existe ainda a classe dos estimadores não viesados de variância uniformemente mínima (ENVUM), que são, dos estimadores não viesados, aqueles que têm variância menor que todos os demais.

Exercícios

Os exercícios a seguir foram retirados de Schmidt (2011) e são de provas anteriores da *ANPEC* sobre propriedades de estimadores[50]. Para cada exercício, julgue se as alternativas são verdadeiras ou falsas, justificando suas respostas. Esses exercícios eu tenho resolvidos integralmente, se alguém quiser pode me pedir por email que eu mando fotos.

Exercício 64. Seja X uma variável aleatória com distribuição de probabilidade que dependa do parâmetro desconhecido θ, tal que $\mathbb{E}[X] = \theta$. Seja também $X_1, X_2, \cdots, X_n$ uma amostra aleatória de X.

a. Para amostras suficientemente grandes, o estimador de máxima verossimilhança de θ, caso exista, segue uma distribuição Normal.

b. Se $\hat{\theta} = \sum_{i=1}^n c_i X_i$ é um estimador de θ, este não será viciado desde que $\sum_{i=1}^n c_i = 1$. Além do mais, $\hat{\theta}$ terá variância mínima se $c_i = \frac{1}{n} \forall \quad i$.

c. Se $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$ é um estimador não viciado de θ, então $\hat{\theta}^2$ também será um estimador não viciado de θ^2.

d. Se $\hat{\theta}_1$ e $\hat{\theta}_2$ são dois estimadores do parâmetro θ em que $\mathbb{E}[\hat{\theta}_1] = \theta$ e $\mathbb{E}[\hat{\theta}_2] \neq \theta$ mas $Var(\hat{\theta}_2) < Var(\hat{\theta}_1)$, então o estimador $\hat{\theta}_2$ deve ser preferível a $\hat{\theta}_1$.

Exercício 65. Sejam $X_1, X_2, \cdots, X_n$ variáveis aleatórias independentes e normalmente distribuídas com média μ e variância σ^2; $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. É correto afirmar:

[49] Página 726 - apêndice A

[50] Para exemplos de estimadores de máxima verossimilhança e pelo método dos momentos, é sugerido que se olhe em Meyer (1973) e Casella and Berger (2002)

a. $\bar{X}$ é um estimador tendencioso para a média μ.

b. $s^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ é um estimador tendencioso da variância σ^2.

c. $n\bar{X}$ é uma variável aleatória normalmente distribuída com média $n\mu$ e variância σ^2.

Exercício 66. São corretas as afirmações

a. Um estimador não tendencioso pode não ser consistente.

b. Um estimador consistente pode não ser eficiente.

Exercício 67. Considere uma amostra aleatória de n variáveis $X_1, X_2, \cdots, X_n$ normalmente distribuídas com média μ e variância σ^2. Sejam $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ e $s^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. É correto afirmar:

a. $\bar{X}$ e s^2 são estimadores de máxima verossimilhança de μ e σ^2, respectivamente.

b. $\bar{X}$ e s^2 são não viesados.

c. $\bar{X}$ e s^2 são consistentes.

d. Apenas $\bar{X}$ é consistente.

e. Apenas $\bar{X}$ é não viesado.

Exercício 68. Sejam $X_1, X_2, \cdots, X_n$ n variáveis aleatórias independentes, igualmente distribuídas, com distribuição de Poisson dada por

$$p_X(x) = \begin{cases} \frac{e^{\lambda}\lambda^x}{x!} & \text{, se } x \in 0, 1, 2, \cdots \\ 0 & \text{, caso contrário.} \end{cases} \tag{59}$$

Julgue as afirmativas.

a. Suponha que $n > 5$. $T = \frac{1}{5}\sum_{i=1}^{5} X_i + \frac{1}{n-5}\sum_{i=6}^{n} X_i$ é um estimador consistente de $\mathbb{E}[X_i]$.

b. $T = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 - \frac{1}{n}\sum_{i=1}^{n} X_i$ é um estimador tendencioso de λ^2.

c. $T = \frac{1}{n}\sum_{i=1}^{n} X_i$ é o estimador de máxima verossimilhança do parâmetro λ.

Exercício 69. Suponha que Y_1 e Y_2 sejam variáveis aleatórias independentes, com média μ e variâncias $Var[Y_1] = 75$ e $Var[Y_2] = 25$. O valor de μ é desconhecido e é proposto estimar μ por uma média ponderada de Y_1 e Y_2, isto é, por: $\alpha Y_1 + (1 - \alpha)Y_2$. Qual o valor de α que produz o estimador com a menor variância possível na classe dos estimadores não viesados? Multiplique o resultado por 100.

Gabarito

Exercício 62

a. V

b. V

c. F

d. F

Exercício 63

a. F

b. V

c. F

Exercício 64

a. V

b. V

Exercício 65

a. V

b. F

c. V

d. F

e. V

Exercício 66

a. F

b. V, se $n > 1$.

c. V

Exercício 67 25

3.3 Testes de Hipóteses

Sugiro a leitura neste link, para uma visão mais intuitiva: http://www.lbd.dcc.ufmg.br/colecoes/erad/2017/002.pdf. Fora isso, acho que o material do livrinho de questões é suficiente. Eu não tenho nenhuma lista de exercícios pronta, sugiro fazer os da ANPEC. Aqui tem uma animação para entender a relação entre poder e significância: http://rpsychologist.com/d3/NHST/.

3.4 Intervalos de Confiança

Sobre intervalos de confiança, gostaria de lembrá-los o que vimos na aula: parâmetro é sempre fixo (uma constante) e portanto não tem probabilidade. O intervalo de confiança, enquanto a amostra ainda não foi coletada, será um intervalo composto de funções de variáveis aleatórias e portanto ele mesmo é aleatório. Porém, após a amostra ser observada, isso já não vale mais.

O próximo exemplo ajuda a entender como que testes de hipóteses se relacionam com intervalos de confiança.

Exemplo 3.4.1. Invertendo um teste normal

Seja $X_1, \ldots, X_n$ iid com distribuição $\mathcal{N}(\mu, \sigma^2)$ e considere testar $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. Para um nível α fixo, um teste adequado tem região de rejeição $\{\boldsymbol{x} : |\bar{x} - \mu_0| > z_{\alpha/2}\sigma/\sqrt{n}\}$. Note que H_0 não é rejeitada para valores amostrais onde $|\bar{x} - \mu_0| \leq z_{\alpha/2}\sigma/\sqrt{n}$ ou, equivalentemente:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Uma vez que o teste tem tamanho α, isso significa que $\mathbb{P}(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = \mathbb{P}(\text{Rejeitar } H_0 | \mu = \mu_0) = \alpha$, ou, de outra forma, $\mathbb{P}(\text{Aceitar } H_0 | \mu = \mu_0) = 1 - \alpha$. Combinando isso com a região de aceitação definida acima, podemos escrever:

$$\mathbb{P}\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} | \mu = \mu_0\right) = 1 - \alpha$$

Mas a afirmação acima é válida para todo μ_0, de forma que podemos omitir a condicional e teremos uma condição ainda válida.

$$\mathbb{P}_\mu\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

O intervalo $[\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}; \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}]$ obtido através da *inversão* da região de aceitação de um teste de nível α, é chamado de *intervalo de confiança* $1 - \alpha$.

Esse site tem uma animação que ajuda a entender mais sobre o que quer dizer um intervalo com $1 - \alpha\%$ de confiança: http://rpsychologist.com/d3/CI/.

Os exercícios a seguir foram baseados em questões de provas passadas da *ANPEC*.

Exercício 70. Para os itens **(a)** a **(e)** considere X uma variável aleatória tal que $X \sim N(\mu, 1)$ de onde é extraída uma a.a.[51] de tamanho n, representada por $X_1, X_2, \cdots, X_n$. Avalie se cada item é verdadeiro ou falso, justificando suas respostas. Esses exercícios eu tenho resolvidos integralmente, se alguém quiser pode me pedir por email que eu mando fotos.

a. $\bar{X} \sim N(\mu, \frac{1}{n})$, onde $\bar{X} = \sum_{i=1}^n \frac{X_i}{n}$;

b. A probabilidade de que o intervalo $[\bar{X} - \frac{1{,}96}{n}; \bar{X} + \frac{1{,}96}{n}]$ contenha o verdadeiro valor de μ é de 95%;

c. A probabilidade de que o intervalo $[\bar{X} - \frac{1{,}96}{n}; \bar{X} + \frac{1{,}96}{n}]$ contenha $\bar{X}$ é de 95%;

d. O intervalo de confiança de $(1 - \alpha)\%$ para μ não depende do tamanho amostral;

e. Em um intervalo de confiança de 95% para μ, esperamos que, extraindo-se muitas amostras de mesmo tamanho n dessa população, esse intervalo conterá μ em 95% das vezes.

Exercício 71. Justifique se as sentenças abaixo são verdadeiras ou falsas.

a. O valor esperado da estatística

$$s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$$

é igual a $(\frac{n-1}{n})\sigma^2$, onde σ^2 representa a variância populacional. Um estimador não viciado para σ^2 é dado por $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$;

[51] a.a. = amostra aleatória

b. Suponha que $X \sim N(\mu, \sigma^2)$ com σ^2 desconhecido. O I.C. para μ com 95% de confiança será dado por:

$$[\bar{X} - z_{0,025}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{0,025}\frac{\sigma}{\sqrt{n}}]$$

Exercício 72. Seja $\{X_1, X_2, \cdots, X_n\}$ uma a.a. de uma população normal de parâmetros μ e σ^2 para a média e variância, respectivamente. Julgue as seguintes alternativas (verdadeiro/falso), justificando suas respostas.

a. O I.C. com 95% de confiança para μ é dado por

$$[\bar{X} - 1,96\frac{\sqrt{\sigma^2 n}}{n}; \bar{X} + 1,96\frac{\sqrt{\sigma^2 n}}{n}]$$

b. Se σ^2 é desconhecido, o I.C. de 95% de confiança para μ será dado por

$$[\bar{X} - t_c\frac{s}{\sqrt{n}}; \bar{X} + t_c\frac{s}{\sqrt{n}}]$$

onde s é o desvio padrão amostral, t_c é calculado de forma que $\mathbb{P}(|t| < t_c) = 0,95$ e t segue uma distribuição de Student com n graus de liberdade;

c. Se construirmos vários I.C.'s para μ com m amostras de tamanho n da população do enunciado, todos os intervalos terão a mesma amplitude;

d. Se a a.a. $\{Y_1, Y_2, \cdots, Y_n\}$ não vem de uma população normal, não se pode construir um I.C. para a sua média populacional, nem mesmo quando $n \to \infty$.

Gabarito:
Exercício 68

a. V

b. F

c. F

d. F

e. V

Exercício 69

a. V

b. F

Exercício 70

a. V

b. F

c. V

d. F

Referências

Luc Bauwens, Michel Lubrano, and Jean-Francois Richard. *Bayesian inference in dynamic econometric models*. OUP Oxford, 2000. 75

George Casella and Roger Berger. *Statistical inference*. Duxbury, 2nd edition, 2002. 39, 78, 80, 82, 85, 86

Russell Davidson and James Gordon. MacKinnon. *Econometric theory and methods*. Oxford University Press, 2009. 58

Jim Freeman, Edward Shoesmith, Dennis J. Sweeney, David Ray Anderson, and Thomas Arthur Williams. *Statistics for business and economics*. Cengage Learning, 2017. 45

Damodar Gujarati. *Econometria Básica*. Editora Campus, 4th edition, 2006. 86

Barry R. James. *Probabilidade: um curso em nivel intermediario*. IMPA, 2010. 11, 37

Elon Lages Lima. *Curso de analise: volume 1*, volume 1. Instituto de Matematica Pura e Aplicada, 1982. 4

Marcos Nascimento Magalhães. *Probabilidade e variáveis aleatórias*. Edusp, 2011. 24, 27, 55

Sharon Bertsch McGrayne. *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines and emerged triumphant from two centuries of Controversy*. Yale University Press, 2011. 15

P. L. Meyer. *Probabilidade aplicacoes a estatistica*. Livros Tecnicos e Cientificos, 1973. 23, 24, 45, 85, 86

Ron Mittelhammer. *Mathematical statistics for economics and business*. Springer, 2013. 11, 28, 30, 37, 38

Alexander M. Mood and Franklin A. Graybill. *Introduction to the theory of statistics*. 1963. 27, 77, 78, 83

Sheldon M. Ross. *A first course in probability*. Pearson Prentice Hall, 2010. 36

Robert Schlaifer and Howard Raiffa. *Applied statistical decision theory*. 1961. 75

Cristiane Alkmin Junqueira Schmidt. *Estatística - Questões comentadas dos concursos de 2002 a 2011. Coleção ANPEC*. Elsevier, 2011. 24, 36, 37, 52, 58, 59, 86

Rafael Stern and Rafael Izbicki. *Introducao à Teoria das Probabilidades e Processos Aleatorios*. UFSCAR, 2016. 2, 5, 22, 30, 35, 36, 37, 40, 46, 48, 55, 56, 57, 58, 59, 61, 75

Lori Viali. *Probabilidade Univariada I - Enfoque Exatas (Apostila)*. 2004a. 36

Lori Viali. *Probabilidade Univariada II - Enfoque Exatas (Apostila)*. 2004b. 45

Lorí Viali. *Probabilidade - Introdução (Apostila)*. Universidade Federal do Rio Grande do Sul (UFRGS), 2004c. 22, 23, 31