Notas de Aula do Curso
ET101: Estatística 1 - Área 2

Leandro Chaves Rêgo, Ph.D.

2008.2

Prefácio

Estas notas de aula foram feitas para compilar o conteúdo de várias referências bibliográficas tendo em vista o conteúdo programático da disciplina ET101-Estatística 1 ministrada para os cursos de graduação em Engenharia na Área 2 da Universidade Federal de Pernambuco. Em particular, elas não contém nenhum material original e não substituem a consulta a livros textos. Seu principal objetivo é dispensar a necessidade dos alunos terem que copiar as aulas e, deste modo, poderem se concentrar em entender o conteúdo das mesmas.

Recife, fevereiro de 2008.
Leandro Chaves Rêgo, Ph.D.

Conteúdo

Capítulo 1

Introdução à Probabilidade

1.1 Definição de Conjuntos e Exemplos

Definição 1.1.1: Um *conjunto* é uma coleção de elementos distintos onde os elementos não são ordenados. ∎

Um conjuntos pode ser especificado, listando seus elementos dentro de chaves. Por exemplo,

$$A = \{0, 1, 2, 3, 5, 8, 13\},\ B = \{0, 1, 2, \ldots, 1000\}.$$

Alternativamente, um conjunto pode ser especificado por uma regra que determina os membros do conjunto, como em:

$$C = \{x : x \text{ é inteiro e positivo}\} \text{ ou } D = \{x : x \text{ é par}\}.$$

Como em um conjunto a ordem dos elementos não importa, temos:

$$\{1, 2, 3\} = \{2, 3, 1\}.$$

Se um dado elemento faz parte de um conjunto, dizemos que ele pertence ao conjunto e denotamos isso com o símbolo $\in$. Por exemplo, $2 \in D = \{x : x \text{ é par}\}$ ou $3 \in E = \{x : x \text{ é primo}\}$.

Por outro lado, se um dado elemento não faz parte de um conjunto, dizemos que ele não pertence ao conjunto e denotamos isso com o símbolo $\notin$. Por exemplo, $3 \notin D = \{x : x \text{ é par}\}$ ou $4 \notin E = \{x : x \text{ é primo}\}$.

Observação 1.1.2: Precisamos ter cuidado ao distinguir entre um elemento como 2 e o conjunto contendo somente este elemento $\{2\}$. Enquanto, temos $2 \in F = \{2, 3, 5\}$, $\{2\} \notin F = \{2, 3, 5\}$, pois o conjunto contendo somente o elemento 2 não pertence à F. ∎

O tamanho de um conjunto $||A||$ é a quantidade de elementos que ele possui, que é chamado de *cardinalidade*. Cardinalidades podem ser *finita*, *infinita enumerável*, ou *infinita não-enumerável*. Um conjunto finito quando existe uma função bijetiva cujo domínio é igual a este conjunto e a imagem é o conjunto dos inteiros não-negativos menores que um número

finito; seus elementos podem ser contados. Um conjunto infinito enumerável tem exatamente a mesma quantidade de elementos que os naturais, ou seja, existe uma função bijetiva cujo domínio é igual a este conjunto e a imagem é igual ao conjunto dos naturais. Um conjunto é enumerável se ele for finito ou infinito enumerável. Um conjunto é não-enumerável se ele não for enumerável. Por exemplo temos que os seguintes conjuntos são enumeráveis:

$$N_n = \{0, 1, 2, \ldots, n-1\},$$

$$\mathcal{Z} = \{x : x \text{ é um inteiro}\},$$

$$\mathcal{Z}^+ = \{x : x \text{ é um inteiro positivo}\},$$

$$\mathcal{Q} = \{x : x \text{ é racional}\}.$$

Por outro lado, os seguintes conjuntos são não-enumeráveis:

$$\mathbb{R} = \{x : x \text{ é um número real}\},$$

$$(a, b) = \{x : a < x < b\}, \text{ onde } a < b,$$

$$[a, b] = \{x : a \leq x \leq b\}, \text{ onde } a < b.$$

Existem dois conjuntos especiais que nos interessarão. Em muitos problemas nos dedicaremos a estudar um conjunto definido de objetos, e não outros. Por exemplo, em alguns problemas podemos nos interessar pelo conjunto dos números naturais; ou em outros problemas pelo conjuntos dos números reais; ou ainda por todas as peças que saem de uma linha produção durante um período de 24h, etc. O conjunto que contém todos os elementos que queremos considerar é chamado de *conjunto universo* e é denotado por Ω. Por outro lado, o conjunto especial que não possui elementos é chamado de *conjunto vazio* e é denotado por $\emptyset$. Este conjunto tem cardinalidade 0 e portanto é finito. Por exemplo,

$$\emptyset = \{\} = \{x : x \in \mathbb{R} \text{ e } x < x\} \text{ ou } \emptyset = (a, a).$$

Dois conjuntos A e B podem ser relacionados através da relação de inclusão (denotada por $A \subseteq B$, e lida A é um subconjunto de B ou B contém A) quando todo elemento de A é também elemento de B. Diz-se que A é um *subconjunto próprio* de B quando se tem $A \subseteq B$, $A \neq \emptyset$, e $A \neq B$. Diz-se que A e B são conjuntos iguais se, e somente se, $A \subseteq B$ e $B \subseteq A$. Se $A \subseteq B$, então nós também podemos dizer que $B \supseteq A$.

Identidade ou igualdade entre dois conjuntos A, B significa que eles tem precisamente a mesma coleção de elementos. *Um método básico para provar que $A = B$ é primeiro provar que $A \subseteq B$ e depois provar que $B \subseteq A$.*

1.2 Operações com Conjuntos

Queremos estudar a importante idéia de combinar conjuntos dados, a fim de formamos um novo conjunto. Conjuntos podem ser transformados através das seguintes operações Booleanas:

1. Complementação: $A^c = \{\omega \in \Omega : \omega \notin A\}$. Observe que de acordo com esta definição, para todo $\omega \in \Omega$ e todo conjunto A, não existe outra opção além de $\omega \in A$ ou $\omega \in A^c$, além disso não pode ser verdade que $\omega \in A$ e $\omega \in A^c$ simultaneamente.

2. União: $A \cup B = \{\omega : \omega \in A \text{ ou } \omega \in B\}$

3. Intersecção: $A \cap B = \{\omega : \omega \in A \text{ e } \omega \in B\}$

4. Diferença: $A - B = A \cap B^c = \{\omega : \omega \in A \text{ e } \omega \notin B\}$

Se $A \cap B = \emptyset$, então A e B não tem nenhum elemento em comum, e nós dizemos que A e B são *disjuntos*.

Exemplo 1.2.1: Seja $\Omega = \{0, 1, 2, 3, 4, 5, 6, 7\}$, $A = \{0, 1, 5\}$ e $B = \{1, 2, 3, 4\}$. Então segue que $A^c = \{2, 3, 4, 6, 7\}$, $A \cup B = \{0, 1, 2, 3, 4, 5\}$, $A \cap B = \{1\}$, $A - B = \{0, 5\}$. ∎

Exemplo 1.2.2: Sejam A, B, C, eD subconjuntos do conjunto universo Ω tal que $A \cup B = \Omega$, $C \cap D = \emptyset$, $A \subseteq C$ e $B \subseteq D$. Prove que $A = C$ e $B = D$.

Solução: Basta provar que $C \subseteq A$ e $D \subseteq B$. Seja $\omega \in C$, então como $C \cap D = \emptyset$, temos que $\omega \notin D$. Logo, como $B \subseteq D$, segue que $\omega \notin B$. Mas como $A \cup B = \Omega$, temos que $\omega \in A$. Portanto, $C \subseteq A$.

Para provar que $D \subseteq B$, seja $\omega \in D$, então como $C \cap D = \emptyset$, temos que $\omega \notin C$. Logo, como $A \subseteq C$, segue que $\omega \notin A$. Mas como $A \cup B = \Omega$, temos que $\omega \in B$. Portanto, $D \subseteq B$. ∎

Relações e propriedades das operações Booleanas incluem as seguintes:

1. Idempotência: $(A^c)^c = A$

2. Comutatividade (Simetria): $A \cup B = B \cup A$ e $A \cap B = B \cap A$

3. Associatividade: $A \cup (B \cup C) = (A \cup B) \cup C)$ e $A \cap (B \cap C) = (A \cap B) \cap C)$

4. Distributividade: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ e $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

5. Leis de De Morgan: $(A \cup B)^c = A^c \cap B^c$ e $(A \cap B)^c = A^c \cup B^c$.

   Prova: Suponha que $\omega \in (A \cup B)^c$. Então, $\omega \notin (A \cup B)$, o que por sua vez implica que $\omega \notin A$ e $\omega \notin B$. Logo, $\omega \in A^c$ e $\omega \in B^c$, ou seja, $\omega \in (A^c \cap B^c)$. Então, $(A \cup B)^c \subseteq (A^c \cap B^c)$. Agora suponha que $\omega \in (A^c \cap B^c)$. Então, $\omega \in A^c$ e $\omega \in B^c$, o que por sua vez implica que $\omega \notin A$ e $\omega \notin B$. Logo, $\omega \notin (A \cup B)$, ou seja, $\omega \in (A \cup B)^c$. Então, $(A^c \cap B^c) \subseteq (A \cup b)^c$. Portanto, $(A^c \cap B^c) = (A \cup b)^c$.

   A prova da outra Lei de Morgan é análoga e deixada como Exercício. ∎

Observe que as Leis de De Morgan permitem que possamos expressar uniões em termos de intersecções e complementos e intersecções em termos de uniões e complementos.

As noções de união e intersecção se estendem para coleções arbitrárias de conjuntos através de dois quantificadores: existe ($\exists$), e para todo ($\forall$).

Se temos uma coleção $\{A_{\alpha:\alpha\in\mathcal{I}}\}$ de subconjuntos de Ω indexados pelo conjunto de índices $\mathcal{I}$, então:

$$\cup_{\alpha\in\mathcal{I}}A_\alpha = \{\omega : (\exists\alpha\in\mathcal{I}, \omega\in A_\alpha)\} \text{ e } \cap_{\alpha\in\mathcal{I}}A_\alpha = \{\omega : (\forall\alpha\in\mathcal{I}, \omega\in A_\alpha)\}.$$

Por exemplo, se $\Omega = 0, 1, 2, \ldots$, $\mathcal{I}$ é o conjunto de inteiros positivos divisíveis por 3 e $A_\alpha = N_\alpha = \{0, 1, 2, \ldots, \alpha - 1\}$, então

$$\cup_{\alpha\in\mathcal{I}}N_\alpha = \Omega \text{ e } \cap_{\alpha\in\mathcal{I}}N_\alpha = N_3.$$

1.3 Produto Cartesiano

Definição 1.3.1: Produto Cartesiano. O produto Cartesiano $A \times B$ de dois conjuntos dados A e B é o conjunto de todos os pares ordenados de elementos, onde o primeiro pertence à A e o segundo pertence à B:

$$A \times B = \{(a, b) : a \in A, b \in B\}.$$

∎

Por exemplo, se $A = \{1, 2, 3\}$ e $B = \{c, d\}$, então:

$$A \times B = \{(1, c), (1, d), (2, c), (2, d), (3, c), (3, d)\}, \text{ e}$$
$$B \times A = \{(c, 1), (c, 2), (c, 3), (d, 1), (d, 2), (d, 3)\}.$$

A noção de produto cartesiano pode ser estendida da seguinte maneira: Se $A_1, \ldots, A_n$ forem conjuntos, então,

$$A_1 \times A_2 \times \ldots \times A_n = \{(a_1, a_2, \ldots, a_n) : a_i \in A_i\},$$

ou seja, o conjunto de todas as ênuplas ordenadas.

Um caso especial importante surge quando consideramos o produto cartesiano de um conjunto por ele próprio, isto é, $A \times A$. Exemplos disso surgem quando tratamos do plano euclideano, $\mathbb{R} \times \mathbb{R}$, onde $\mathbb{R}$ é o conjunto de todos os números reais, e do espaço euclideano tridimensional, representado por $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$.

1.4 Conjunto das Partes

Definição 1.4.1: Dado um conjunto qualquer A, pode-se definir um outro conjunto, conhecido como *conjuntos das partes de* A, e denotado por 2^A, cujos elementos são subconjuntos de A. ∎

Exemplo 1.4.2: Seja $A = \{1, 2, 3\}$, então temos que

$$2^A = \{\emptyset, A, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}\}.$$

∎

Pode-se provar que a cardinalidade do conjunto das partes de qualquer conjunto dado A é maior que a cardinalidade de A.

1.5 Partição

Definição 1.5.1: Dado um conjunto universo Ω, uma partição $\Pi = \{A_\alpha, \alpha \in \mathcal{I}\}$ de Ω é uma coleção de subconjuntos de Ω (neste caso, indexados por α que toma valores no conjunto de índices $\mathcal{I}$) e satisfaz:

P1. Para todo $\alpha \neq \beta$, $A_\alpha \cap A_\beta = \emptyset$;

P2. $\cup_{\alpha \in \mathcal{I}} A_\alpha = \Omega$.

∎

Deste modo os conjuntos de uma partição são disjuntos par a par e cobrem todo o conjunto universo. Portanto, cada elemento $\omega \in \Omega$ pertence a um, e somente um, dos conjuntos A_α de uma partição.

Exemplo 1.5.2: Se $\Omega = \{1, 2, 3, 4\}$, então $\{A_1, A_2\}$, onde $A_1 = \{1, 2, 3\}$ e $A_2 = \{4\}$, é uma partição de Ω. ∎

Exemplo 1.5.3: A coleção de intervalos $\{(n, n+1] : n \in Z\}$ é uma partição dos números reais $I\!R$. ∎

1.6 Função Indicadora

É sempre conveniente representar um conjunto A por uma função I_A tendo domínio (conjunto dos argumentos da função) Ω e contra-domínio (conjunto dos possíveis valores da função) binário $\{0, 1\}$.

Definição 1.6.1: Função Indicadora. A função indicadora $I_A : \Omega \to \{0, 1\}$ de um conjunto A é dada por

$$I_A(\omega) = \begin{cases} 1 & \text{se } \omega \in A, \\ 0 & \text{se } \omega \notin A. \end{cases}$$

∎

É fácil observar que $I_\Omega(\omega) = 1, \forall \omega \in \Omega$ e que $I_\emptyset(\omega) = 0, \forall \omega \in \Omega$. Note que existe uma correspondência 1-1 entre conjuntos e suas funções indicadoras:

$$A = B \Leftrightarrow (\forall \omega \in \Omega) I_A(\omega) = I_B(\omega).$$

O fato que conjuntos são iguais se, e somente se, suas funções indicadoras forem idênticas nos permitem explorar a aritmética de funções indicadoras:

$$I_{A^c} = 1 - I_A,$$

$$A \subseteq B \Leftrightarrow I_A \leq I_B,$$

$$I_{A \cap B} = \min(I_A, I_B) = I_A I_B,$$

$$I_{A\cup B} = \max(I_A, I_B) = I_A + I_B - I_{A\cap B},$$

$$I_{A-B} = \max(I_A - I_B, 0) = I_A I_{B^c},$$

para construir argumentos rigorosos no que se refere a relação entre conjuntos. Ou seja, nós transformamos proposições sobre conjuntos em proposições sobre funções indicadoras e podemos então utilizar nossa familiaridade com álgebra para resolver perguntas menos familiares sobre conjuntos.

Exemplo 1.6.2: Utilizando funções indicadoras, verifique que $A \subseteq B \Leftrightarrow B^c \subseteq A^c$.
Solução: Temos que

$$A \subseteq B \Leftrightarrow I_A \leq I_B \Leftrightarrow 1 - I_A \geq 1 - I_B \Leftrightarrow I_{A^c} \geq I_{B^c} \Leftrightarrow B^c \subseteq A^c.$$

∎

Exemplo 1.6.3: As seguintes questões não estão relacionadas umas com as outras.

a. Se $I_A I_B$ for identicamente igual a zero, o que sabemos a respeito da relação entre A e B?

b. Se $A \cap B^c = B \cap A^c$, o que sabemos a respeito da relação entre A e B?

c. Se $I_A^2 + I_B^2$ for identicamente igual a 1, o que podemos concluir sobre A e B?

Solução: Exercício. ∎

1.7 Experimento Aleatório

Um *experimento* é qualquer processo de observação. Em muitos experimentos de interesse, existe um elemento de incerteza, ou chance, que não importa quanto nós sabemos sobre o passado de outras performances deste experimento, nós essencialmente não somos capazes de predizer seu comportamento em futuras realizações. As razões para nossa falta de habilidade para predizer são varias: nós podemos não saber de todas as causas envolvidas; nós podemos não ter dados suficientes sobre as condições iniciais do experimento; as causas podem ser tão complexas que o cálculo do seu efeito combinado não é possível; ou na verdade existe alguma aleatoriedade fundamental no experimento. Estamos interessados em uma classe particular de experimentos, chamados *experimentos aleatórios*. Os seguintes traços caracterizam um experimento aleatório:

(a) Se for possível repetir as mesmas condições do experimento, os resultados do experimento em diferentes realizações podem ser diferentes. Por exemplo, jogar uma moeda diversas vezes com bastante cuidado para que cada jogada seja realizada da mesma maneira.

(b) Muito embora não sejamos capazes de afirmar que resultado particular ocorrerá, seremos capazes de descrever o conjunto de todos os possíveis resultados do experimento.

(c) Quando o experimento for executado repetidamente, os resultados individuais parecerão ocorrer de uma forma acidental. Contudo, quando o experimento for repetido um grande número de vezes, uma configuração definida ou regularidade surgirá. É esta regularidade que torna possível construir um modelo probabilístico. Por exemplo, pense nas repetidas jogadas de uma moeda, muito embora caras e coroas apareçam sucessivamente, em uma maneira arbitrária, é fato empírico conhecido que, depois de um grande número de jogadas, a proporção de caras e de coroas serão aproximadamente iguais (assumindo que a moeda é simétrica).

Os resultados de um experimento aleatório são caracterizados pelos seguintes componentes:

1. o conjunto de resultados possíveis Ω;
2. a coleção de conjuntos de resultados de interesse $\mathcal{A}$;
3. um valor numérico P da probabilidade de ocorrência de cada um dos conjuntos de resultados de interesse.

1.8 Espaço Amostral

O conjunto de possíveis resultados de um experimento aleatório é chamado de *espaço amostral.* Em um dado experimento aleatório a especificação do espaço amostral deve ser tal que este (1) liste todos os possíveis resultados do experimento sem duplicação e o faça em um nível de detalhamento suficiente para os interesses desejados, omitindo resultados que embora logicamente ou fisicamente possíveis não tenham nenhuma implicação prática para análise do experimento.

Por exemplo, uma única jogada de uma moeda pode ter o espaço amostral tradicional $\Omega = \{cara, coroa\}$, ou podemos considerar que a moeda pode fisicamente ficar equilibrada na borda $\Omega = \{cara, coroa, borda\}$. Uma outra possibilidade seria levar em consideração as coordenadas (x, y) do centro da moeda quando ela para após ser jogada no ar. Como vemos muito mais se sabe sobre o resultado de uma jogada de uma moeda que os simples resultados binários tradicionais *cara* e *coroa*. Nós ignoramos está informação adicional usando uma hipótese não mencionada que existe uma aposta com pagamentos que dependem apenas de qual lado da moeda cai para cima e não em outras informações.

1.9 Eventos e Coleção de Eventos

Um *evento* é um subconjunto do espaço amostral, ou seja, é um conjunto de resultados possíveis do experimento aleatório. Se ao realizarmos um experimento aleatório, o resultado pertence a um dado evento A, dizemos que A *ocorreu.* Estaremos interessados no estudo da ocorrência de combinações de eventos. Para tanto, utilizaremos as operações Booleanas de conjuntos (complementar, união, intersecção, diferença) para expressar eventos combinados de interesse.

Exemplo 1.9.1: Sejam A, B, e C eventos em um mesmo espaço amostral Ω. Expresse os seguintes eventos em função de A, B, e C e operações Booleanas de conjuntos.

(a) Pelo menos um deles ocorre. Resp.: $A \cup B \cup C$.

(b) Exatamente um deles ocorre. Resp.: $(A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C)$.

(c) Apenas A ocorre. Resp.: $(A \cap B^c \cap C^c)$.

(d) Pelo menos dois ocorrem. Resp.: $(A \cap B \cap C^c) \cup (A \cap B^c \cap C) \cup (A^c \cap B \cap C) \cup (A \cap B \cap C)$.

(e) No máximo dois deles ocorrem. Resp. $(A \cap B \cap C)^c$.

(f) Nenhum deles ocorrem. Resp. $(A^c \cap B^c \cap C^c)$.

(g) Ambos A e B ocorrem, mas C não ocorre. Resp. $(A \cap B \cap C^c)$.

∎

Embora possa-se pensar que, dado um espaço amostral, necessariamente é de interesse analisar todos os seus subconjuntos (e isto eventualmente é verdadeiro), temos três razões para esperar que estejamos apenas interessados em alguns subconjuntos do espaço amostral. Primeiro, o espaço amostral pode conter um grau de detalhamento superior ao que estamos interessados no momento. Por exemplo, ele pode representar uma única jogada de um dado com 6 elementos, mas nós apenas estamos interessados em saber se o resultado é par ou ímpar. Segundo, nós vamos querer associar cada evento A com uma probabilidade numérica $P(A)$. Como essas probabilidades estão baseadas em algum conhecimento sobre a tendência de ocorrer do evento ou no grau de nossa crença que determinado evento ocorrerá, nosso conhecimento sobre P pode não estender para todos os subconjuntos de Ω. A terceira (e técnica) razão para limitar a coleção de eventos de interesse é que condições impostas em P pelos axiomas de Kolmogorov, que estudaremos adiante, podem não permitir que P seja definida em todos os subconjuntos de Ω, em particular isto pode ocorrer quando Ω for não enumerável, mas não iremos demonstrar este fato que está fora do escopo deste curso.

Estaremos interessados em uma coleção especial $\mathcal{A}$ de subconjuntos do espaço amostral Ω (note que $\mathcal{A}$ é um conjunto cujos elementos também são conjuntos!) que são eventos de interesse no que se refere ao experimento aleatório $\mathcal{E}$ e os quais temos conhecimento sobre a sua probabilidade. $\mathcal{A}$ é chamado de uma *σ-álgebra de eventos*. Como veremos adiante, o domínio de uma medida de probabilidade é uma σ-álgebra.

Definição 1.9.2: Uma álgebra de eventos $\mathcal{F}$ é uma coleção de subconjuntos do espaço amostral Ω que satisfaz:

1. não é vazia;
2. fechada com respeito a complementos (se $A \in \mathcal{F}$, então $A^c \in \mathcal{F}$);
3. fechada com respeito a uniões finitas (se $A, B \in \mathcal{F}$, então $A \cup B \in \mathcal{F}$).

Uma *σ-álgebra* $\mathcal{A}$ é uma álgebra de eventos que também é fechada com relação a uma união enumerável de eventos,

$$(\forall i \in Z) A_i \in \mathcal{A} \Rightarrow \cup_{i \in Z} A_i \in \mathcal{A}.$$

Pelas Leis de De Morgan, vemos que $\mathcal{A}$ é fechada com respeito a intersecções enumeráveis também.

Exemplo 1.9.3:

1. A menor álgebra de eventos é $\mathcal{A} = \{\emptyset, \Omega\}$;
2. A maior álgebra de eventos é o conjunto das partes de Ω;
3. Um exemplo intermediário, temos:

$$\Omega = \{1, 2, 3\},\ \mathcal{A} = \{\Omega, \emptyset, \{2\}, \{1, 3\}\}.$$

Se o espaço amostral for finito, toda álgebra é uma σ-álgebra, pois so existem um número finito de eventos diferentes. Se o espaço amostral for infinito, existem álgebras que não são σ-álgebras, como mostra o exemplo seguinte.

Exemplo 1.9.4: A coleção de conjuntos de números reais finitos e co-finitos é uma álgebra que não é uma σ-álgebra.

Exemplo 1.9.5: A σ-álgebra de Borel $\mathcal{B}$ de subconjuntos reais é, por definição, a menor σ-álgebra contendo todos os intervalos e é a σ-álgebra usual quando lidamos com quantidades reais ou vetoriais. Em particular, temos que uniões enumeráveis de intervalos (por exemplo, o conjunto dos números racionais), seus complementos (por exemplo, o conjunto dos números irracionais), e muito mais está em $\mathcal{B}$.

1.10 Freqüências Relativas

Resta-nos discutir o terceiro elemento para modelagem do raciocínio probabilístico, a associação de uma medida numérica a eventos que representam a probabilidade com que eles ocorrem. As propriedades desta associação são motivadas em grande parte pelas propriedades de freqüência relativas. Considere uma coleção de experimentos aleatórios $\mathcal{E}_i$ que possuem a mesma σ-álgebra de eventos $\mathcal{A}$ e tem resultados individuais não necessariamente numéricos $\{\omega_i\}$. Fixando uma dada seqüência de resultados $\{\omega_i\}$, se estamos interessados na ocorrência de um dado evento A, a freqüência relativa de A nada mas é que uma média aritmética da função indicadora de A calculada em cada um dos termos da seqüência $\{\omega_i\}$, ou seja,

Definição 1.10.1: A freqüência relativa de um evento A, determinada pelos resultados $\{\omega_1, \ldots, \omega_n\}$ de n experimentos aleatórios, é

$$r_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(\omega_i) = \frac{N_n(A)}{n}.$$

Propriedades chaves da freqüência relativa são:

FR0. $r_n : \mathcal{A} \to I\!R$.

FR1. $r_n(A) \geq 0$.

FR2. $r_n(\Omega) = 1$.

FR3. Se A e B são disjuntos, então $r_n(A \cup B) = r_n(A) + r_n(B)$.

Nós prosseguiremos como se existisse alguma base empírica ou metafísica que garanta que $r_n(A) \to P(A)$, embora que o sentido de convergência quando n cresce só será explicado pela Lei dos Grandes Números, que não será discutida em detalhes neste curso. Esta tendência da freqüência relativa de estabilizar em um certo valor é conhecida como *regularidade estatística*. Deste modo, P herdará propriedades da freqüência relativa r_n.

1.11 Interpretações de Probabilidade

Parece não ser possível reduzir probabilidade a outros conceitos; ela é uma noção em si mesma. O melhor que podemos fazer é relacionar probabilidade a outros conceitos através de uma interpretação. Os três mais comuns grupos de interpretação são os seguintes:

1. **Clássica:** baseada em uma enumeração de *casos igualmente prováveis*.
2. **Subjetiva:** se refere ao grau de crença pessoal na ocorrência do evento A e é medida através da interpretação comportamental de disposição a apostar ou agir.
3. **Freqüentista:** se refere ao limite da freqüência relativa de ocorrência do evento A em repetidas realizações não relacionadas do experimento aleatório $\mathcal{E}$. Note que limites de freqüência relativas são uma idealização, pois não se pode realizar infinitas realizações de um experimento.

1.12 Axiomas de Kolmogorov

Primeiro por razões técnicas, fora do escopo deste curso, temos que o domínio da medida formal de probabilidade é uma álgebra de eventos que também é fechada com relação a um número enumerável de uniões.

Os axiomas que descreveremos a seguir não descrevem um único modelo probabilístico, eles apenas determinam uma família de modelos probabilísticos, com os quais poderemos utilizar métodos matemáticos para descobrir propriedades que serão verdadeiras em qualquer modelo probabilístico. A escolha de um modelo específico satisfazendo os axiomas é feito pelo analista/estatístico familiar com o fenômeno aleatório sendo modelado.

Motivados pelas propriedades de freqüência relativa, impõe-se os primeiros quatro axiomas de Kolmogorov:

K0. **Inicial.** O experimento aleatório é descrito pelo espaço de probabilidade $(\Omega, \mathcal{A}, P)$ que consiste do espaço amostral Ω, de uma σ-álgebra $\mathcal{A}$, e de uma função de valores reais $P : \mathcal{A} \to I\!R$.

K1. **Não-negatividade.** $\forall A \in \mathcal{A}, P(A) \geq 0$.

K2. **Normalização Unitária.** $P(\Omega) = 1$.

K3. **Aditividade Finita.** Se A, B são disjuntos, então $P(A \cup B) = P(A) + P(B)$.

É fácil provar (tente!) utilizando indução matemática que K3 é válida para qualquer coleção finita de eventos disjuntos par a par, ou seja, se $A_i, i = 1, 2, \ldots, n$, são eventos disjuntos par a par, então $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.

Um último axioma embora não seja uma propriedade de limites de freqüência relativa nem tenha significado em espaços amostrais finitos, foi proposto por Kolmogorov para garantir um certo grau de continuidade da medida de probabilidade.

K4. **σ-aditividade.** Se $\{A_i\}$ é uma coleção enumerável de eventos disjuntos dois a dois, então

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

Note que para espaços amostrais finitos, somente existem um número finito de subconjuntos diferentes, logo para que tenhamos uma coleção enumerável de eventos disjuntos dois a dois, um número enumerável destes deve ser vazio. Como veremos adiante a probabilidade de um evento vazio é nula, o que implica que para espaços amostrais finitos K3 e K4 são equivalentes.

Definição 1.12.1: Uma função que satisfaz K0—K4 é chamada de uma medida de probabilidade.

1.12.1 Exemplos de Medidas de Probabilidade

Exemplo 1.12.2: Se Ω for um conjunto finito, então temos que a probabilidade clássica que assume que todos os resultados são igualmente prováveis, é um exemplo de uma medida de probabilidade. Neste caso, temos que

$$P(A) = \frac{||A||}{||\Omega||}$$

definido para qualquer subconjunto A de Ω. O fato que $0 \leq ||A|| \leq ||\Omega||$ e que

$$||A \cup B|| = ||A|| + ||B|| - ||A \cap B||,$$

permitem que verifiquemos que P satisfaz os axiomas de Kolmogorov.

Exemplo 1.12.3: Seja $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ um conjunto finito, e seja $P(\{\omega_i\}) = p_i$, onde $p_i \geq 0, i \geq 1$ e $\sum_{i=1}^{n} p_i = 1$, e $P(A) = \sum_{\omega_i \in A} P(\{\omega_i\})$. Neste caso, também é fácil verificar que P é uma medida de probabilidade verificando os axiomas.

1.12.2 Propriedades de uma Medida de Probabilidade

Teorema 1.12.4: *Se P é uma medida de probabilidade, então*

1. $P(A^c) = 1 - P(A)$.
2. $P(\emptyset) = 0$.
3. $P(A) \leq 1$.

Prova: Parte 1, segue do fato que $\Omega = A \cup A^c$, K2, e K3, pois

$$1 = P(\Omega) = P(A) + P(A^c).$$

Parte 2, segue da Parte 1, do fato que $\Omega^c = \emptyset$, e K2, K3, pois

$$P(\emptyset) = 1 - P(\Omega) = 0.$$

Parte 3, segue do fato que $1 = P(\Omega) = P(A) + P(A^c) \geq P(A)$, já que $P(A^c) \geq 0$ por K1. ∎

Teorema 1.12.5: Monotonicidade. *Se $A \subseteq B$, então $P(A) \leq P(B)$.*

Prova: Note que $B = A \cup (B - A)$, onde A e $B - A$ são disjuntos. Então K3 implica que $P(B) = P(A) + P(B - A)$. O resultado segue do fato que $P(B - A) \geq 0$. ∎

Corolário 1.12.6: $P(A \cup B) \geq \max(P(A), P(B)) \geq \min(P(A), P(B)) \geq P(A \cap B)$.

Teorema 1.12.7: *Uma expressão exata para a probabilidade de uma união não-disjunta é dada por*

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

Prova: Como $A \cup B = A \cup (B - A)$, e A e $B - A$ são disjuntos, K3 implica que $P(A \cup B) = P(A) + P(B - A)$. E como $B = (A \cap B) \cup (B - A)$, $A \cap B$ e $B - A$ são disjuntos, K3 implica que $P(B) = P(A \cap B) + P(B - A)$. Logo,

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

∎

Teorema 1.12.8: Probabilidade de Partições. *Se $\{A_i\}$ é uma partição enumerável de Ω feita de conjuntos em $\mathcal{A}$, então para todo $B \in \mathcal{A}$*

$$P(B) = \sum_i P(B \cap A_i).$$

Prova: Como $\{A_i\}$ é uma partição, segue que

$$B = B \cap \Omega = B \cap (\cup_i A_i) = \cup_i (B \cap A_i).$$

O resultado segue então por K4′. ∎

Teorema 1.12.9: Desigualdade de Boole. *Para n eventos arbitrários $\{A_1, \ldots, A_n\}$, a desigualdade de Boole é*

$$P(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i).$$

Prova: Omitida. ∎

Corolário 1.12.10: *Para n eventos arbitrários $\{A_1, \ldots, A_n\}$,*

$$P(\cap A_i) \geq \sum_{i=1}^n P(A_i) - (n-1).$$

Prova: Utilizando a Lei de De Morgan e a desigualdade de Boole para os eventos $\{A_1^c, \ldots, A_n^c\}$, temos

$$P(\cup_{i=1}^n A_i^c) = 1 - P(\cap A_i) \leq \sum_{i=1}^n P(A_i^c) = \sum_{i=1}^n (1 - P(A_i)).$$

Logo,

$$P(\cap A_i) \geq \sum_{i=1}^n P(A_i) - (n-1).$$

∎

O próximo teorema permite que possamos calcular de maneira exata a probabilidade $P(\cup_{i=1}^n A_i)$ para n eventos arbitrários.

Teorema 1.12.11: Princípio da Inclusão-Exclusão. *Seja I um conjunto genérico de índices que é um subconjunto não-vazio qualquer de $\{1, 2, \ldots, n\}$. Para eventos arbitrários $\{A_1, \ldots, A_n\}$,*

$$P(\cup_{i=1}^n A_i) = \sum_{\emptyset \neq I \subseteq \{1,\ldots,n\}} (-1)^{||I||+1} P(\cap_{i \in I} A_i),$$

onde o somatório é sobre todos os $2^n - 1$ conjuntos de índices excluindo apenas o conjunto vazio.

No caso particular de $n = 3$, o princípio de inclusão-exclusão afirma que

$$P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3).$$

Exemplo 1.12.12: Professor Leônidas está tentando calcular a probabilidade $p = P(A)$ do evento A, e determinou que ela é uma raiz do seguinte polinômio de grau cinco:

$$(p-3)(p-3\sqrt{-1})(p+3\sqrt{-1})(p+0.3)(p-0.3) = 0.$$

Baseado nesta fato, qual é o valor de p?

Exemplo 1.12.13: Se $\Omega = \{a, b, c\}$, e a álgebra $\mathcal{A}$ é o conjunto das partes de Ω, e a medida de probabilidade P é parcialmente definida por

$$P(\{a, b\}) = 0.5,\ P(\{b, c\}) = 0.8,\ P(\{a, c\}) = 0.7,$$

então complete a especificação de P para todos os eventos em $\mathcal{A}$.

Exemplo 1.12.14: Se $\{A_i\}$ for uma partição enumerável de Ω e $P(A_i) = ab^i$, $i \geq 1$, então quais as condições que a e b devem satisfazer para que P seja uma medida de probabilidade?

Capítulo 2

Espaços Amostrais Finitos

2.1 Introdução

Verificamos no capítulo anterior que se $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ é um conjunto finito, então para determinar a probabilidade de qualquer evento A é suficiente especificar a probabilidade de cada eventos simples $\{\omega_i\}$, ou seja $P(\{\omega_i\}) = p_i$. É fácil ver que os axiomas de Kolmogorov implicam que $p_i \geq 0, i \geq 1$ e $\sum_{i=1}^{n} p_i = 1$, e $P(A) = \sum_{\omega_i \in A} P(\{\omega_i\})$.

Para determinarmos as probabilidades dos eventos simples, precisamos de algumas hipóteses adicionais. Por exemplo, se $\Omega = \{w_1, w_2, w_3\}$, $\{w_1\}$ for 3 vezes mais provável, que $\{w_2, w_3\}$, e $\{w_2\}$ for igualmente provável a $\{w_3\}$, temos que: $p_1 = 3(p_2 + p_3)$, $p_2 = p_3$. Logo, como $p_1 + p_2 + p_3 = 1$, temos que $p_3 = p_2 = \frac{1}{8}$, e $p_1 = \frac{3}{4}$.

Vimos também que de acordo com a interpretação clássica de probabilidade, onde o espaço amostral Ω é finito e os possíveis resultados do experimento são equiprováveis, então a probabilidade de qualquer evento $A \in \mathcal{A}$ é proporcional a sua cardinalidade, isto é, $P(A) = \frac{||A||}{||\Omega||}$. Portanto, é importante que saibamos contar a quantidade de elementos que um evento.

2.2 Métodos de Contagem

Nesta seção estudaremos alguns métodos de contagem, também conhecidos como métodos de análise combinatória. Embora conjuntos pequenos possam ser contados exaustivamente (força-bruta), mesmo conjuntos com tamanho moderado podem ser difíceis de contar sem a utilização de técnicas matemáticas.

2.2.1 Regra da Adição

Suponha que um procedimento, designado por 1, possa ser realizado de n_1 maneiras. Admita-se que um segundo procedimento, designado por 2, possa ser realizado de n_2 maneiras. Além disso, suponha que não seja possível que ambos os procedimentos 1 e 2 sejam realizados em conjunto. Então, o número de maneiras pelas quais poderemos realizar ou 1 ou 2 será $n_1 + n_2$.

Esta regra também pode ser estendida da seguinte maneira: Se existirem k procedimentos e o i-ésimo procedimento puder ser realizado de n_i maneiras, $i = 1, 2, \ldots, k$, então, o número

de maneiras pelas quais poderemos realizar ou o procedimento 1, ou o procedimento 2, ..., ou o procedimento k, é dado por $n_1 + n_2 + \ldots + n_k$, supondo que dois quaisquer deles não possam ser realizados conjuntamente.

Exemplo 2.2.1: Suponha que estejamos planejando uma viagem e devamos escolher entre o transporte por ônibus ou por trem. Se existirem três rodovias e duas ferrovias, então existirão $3 + 2 = 5$ caminhos disponíveis para a viagem. ∎

2.2.2 Regra da Multiplicação

Suponha que um procedimento designado por 1 possa ser executado de n_1 maneiras. Admita-se que um segundo procedimento, designado por 2, possa ser executado de n_2 maneiras. Suponha também que cada maneira de executar 1 possa ser seguida por qualquer maneira para executar 2. Então o procedimento formado por 1 seguido de 2 poderá ser executado de $n_1 \cdot n_2$ maneiras.

Obviamente, esta regra pode ser estendida a qualquer número finito de procedimentos. Se existirem k procedimentos e o i-ésimo procedimento puder ser executado de n_i maneiras, $i = 1, 2, \ldots, k$, então o procedimento formado por 1, seguido por 2,..., seguido pelo procedimento k, poderá ser executado de $n_1 \cdot n_2 \cdots n_k$ maneiras.

Exemplo 2.2.2: Quantos divisores inteiros e positivos possui o número 360? Quantos desses divisores são pares? Quantos são ímpares? Quantos são quadrados perfeitos?

Solução: $360 = 2^3 \times 3^2 \times 5$. Os divisores inteiros e positivos de 360 são os números da forma: $2^a \times 3^b \times 5^c$, onde $a \in \{0, 1, 2, 3\}$, $b \in \{0, 1, 2\}$, e $c \in \{0, 1\}$. Portanto, existem $4 \times 3 \times 2 = 24$ maneiras de escolher os expoentes a, b, c. Logo há 24 divisores.

Para o divisor ser par, a não pode ser zero. Então, existem $3 \times 3 \times 2 = 18$ divisores pares. Por outro lado, para o divisor ser ímpar, a tem que ser zero. Logo, existem $1 \times 3 \times 2 = 6$ divisores ímpares. Por fim para o divisor ser quadrado perfeito, os expoentes tem que ser pares. Logo, existem $2 \times 2 \times 1 = 4$ divisores quadrados perfeitos. ∎

Exemplo 2.2.3: De quantos modos o número 720 pode ser decomposto em um produto de dois inteiros positivos? Aqui consideramos, naturalmente, 8×90 como sendo o mesmo produto que 90×8. E o número 144?

Solução: $720 = 2^4 \times 3^2 \times 5$. Os divisores inteiros e positivos de 720 são os números da forma: $2^a \times 3^b \times 5^c$, onde $a \in \{0, 1, 2, 3, 4\}$, $b \in \{0, 1, 2\}$, e $c \in \{0, 1\}$. Portanto, existem $5 \times 3 \times 2 = 30$ maneiras de escolher os expoentes a, b, c. Logo há 30 divisores. Observe que como 720 não é um quadrado perfeito, para cada divisor x de 720 existe um outro divisor $y \neq x$ de 720 tal que $x \times y = 720$. Portanto, cada produto contém dois divisores diferentes de 720. Como existem 30 divisores, existem 15 produtos diferentes.

$144 = 2^4 \times 3^2$. Seguindo o mesmo raciocínio anterior, temos $5 \times 3 = 15$ divisores de 144. Note que $144 = 12^2$ e este constitui um produto de inteiros positivos que é igual a 144. Para os demais produtos sempre temos que eles contém dois inteiros positivos diferentes que são divisores de 144. Como existem 14 divisores de 144 diferentes de 12, temos que existem 7 produtos envolvendo estes divisores. Logo, temos um total de 8 produtos diferentes. ∎

Exemplo 2.2.4: O conjunto A possui 4 elementos e, o conjunto B, 7 elementos. Quantas funções $f : A \to B$ existem? Quantas delas são injetoras?

Solução: Note que para cada elemento de A temos 7 opções de valores diferentes. Como A contém 4 elementos, existem $7 \times 7 \times 7 \times 7 = 7^4$ funções diferentes. Recorde que uma função é injetora se $f(a) \neq f(b)$ sempre que $a \neq b$. Portanto, não podemos repetir o mesmo elemento de B como imagem de dois elementos de A, logo existem $7 \times 6 \times 5 \times 4 = 840$ funções injetoras. ∎

Exemplo 2.2.5: Em uma banca há 5 exemplares iguais da "Veja", 6 exemplares iguais da "Época" e 4 exemplares iguais da "Isto é". Quantas coleções não-vazias de revistas dessa banca podemos formar?

Solução: Note que cada coleção de revistas vai ser composta por a revistas Veja, b revistas Época, e c revistas Isto é, onde $0 \leq a \leq 5$, $0 \leq b \leq 6$, $0 \leq c \leq 4$, e pelo menos 1 de a, b, ou c é diferente de zero. Então, temos $6 \times 7 \times 5 - 1 = 210 - 1 = 209$ diferentes coleções não-vazias destas revistas. ∎

Amostragem com Reposição

Dado um conjunto com n elementos distintos, o número $\mu_{n,r}$ de maneiras de selecionar uma seqüência distinta de comprimento r escolhida desse conjunto com repetidas seleções do mesmo elemento sendo permitida (*amostragem com repetição*) é dada por n^r, já que estamos repetindo o mesmo procedimento r vezes, e cada procedimento tem n maneiras de ser executado.

Este resultado também se aplica ao número de resultados possíveis em r jogadas de uma moeda ($n = 2$), ou de um dado ($n = 6$), ou o número de bytes ($r = 8, n = 2$) (Um byte é uma seqüência ordenada de comprimento 8 de 0's e 1's).

Exemplo 2.2.6: Número de Seqüências Binárias ou Subconjuntos. O número de seqüências binárias de comprimento r é igual a 2^r pois neste caso temos para cada posição i da seqüência $n_i = 2$. O número de subconjuntos de um dado conjunto $||A|| = r$ pode ser determinado enumerando $A = \{a_1, a_2, a_3, \ldots, a_r\}$ e descrevendo cada subconjunto B de A por uma seqüência binária

$$(b_1, b_2, \ldots, b_r)$$

, onde $b_i = 1$ se $a_i \in B$ e $b_i = 0$, caso contrário. Como existem 2^r destas seqüências, então existem 2^r subconjuntos de um conjunto de r elementos. Portanto, se $||A|| = r$, o conjunto das partes de A, possui 2^r elementos, o que explica a notação exponencial do conjunto das partes. ∎

Amostragem sem Reposição

Dado um conjunto com n elementos distintos, o número $(n)_r$ de maneiras de selecionar uma seqüência distinta de comprimento r escolhida desse conjunto com repetidas seleções do mesmo elemento não sendo permitida (*amostragem sem repetição*) é dada por

$$(n)_r = n(n-1)\cdots(n-r+1) = \prod_{i=0}^{r-1}(n-i),$$

já que no primeiro procedimento (escolha do primeiro elemento da seqüência) temos n maneiras de executá-lo, no segundo procedimento (escolha do segundo elemento da seqüência) temos $n-1$ maneiras de executá-lo, ..., e no r-ésimo e último procedimento (escolha do r-ésimo elemento da seqüência) temos $n-r+1$ maneiras de executá-lo. Este número de seqüências é também chamado na literatura do número de arranjos quando temos n elementos distintos e queremos escolher r deles onde a ordem de escolha é importante.

Um caso particular de amostragem sem reposição é quando queremos saber o número de permutações de um conjunto de n elementos distintos. Neste caso temos que $r = n$, então o número de permutações é dado por

$$n! = (n)_n = n(n-1)\cdots 1,$$

onde $n!$ é conhecida como *função fatorial*. Em termos, de função fatorial, nós podemos escrever:

$$(n)_r = \frac{n!}{(n-r)!}.$$

Propriedades da função fatorial $n!$ incluem as seguintes:

$$0! = 1! = 1 \text{ e } n! = n(n-1)!.$$

Exemplo 2.2.7: Se A é um conjunto de n elementos, quantas são as funções $f : A \to A$ bijetoras?

Solução: Temos que garantir que cada elemento de A tem uma imagem diferente. Como A é finito e tem n elementos, garante-se deste modo que f também é sobrejetora e, portanto, bijetora. Então, o primeiro elemento de A tem n opções, o segundo $n-1$ opções, até que o último elemento de A tem somente uma opção disponível. Portanto, existem $n!$ funções bijetoras $f : A \to A$. ∎

Exemplo 2.2.8: De quantos modos é possível colocar r rapazes e m moças em fila de modo que as moças permaneçam juntas?

Solução: Primeiro temos $r+1$ opções de escolher o lugar das moças. Em seguida, temos $r!$ maneiras de escolher a posição dos rapazes entre si, e $m!$ maneiras de escolher a posição das moças entre si. Portanto, temos $(r+1)r!m!$ modos diferentes de escolha. ∎

Exemplo 2.2.9: Quantas são as permutações simples dos números $1, 2, \ldots, 10$ nas quais o elemento que ocupa o lugar de ordem k, da esquerda para a direita, é sempre maior que $k-3$?

Solução: Comecemos escolhendo os números da direita para esquerda. Observe que o número no lugar de ordem 10, tem que ser maior que 7, portanto existem 3 opções. O número no lugar de ordem 9, tem que ser maior que 6, existem, portanto, 3 opções visto que um dos números maiores que 6 já foi utilizado na última posição. De maneira similar pode-se ver que existem 3 opções para os números que ocupam do terceiro ao oitavo lugar. O número no lugar de ordem 2, tem somente 2 opções, pois oito números já foram escolhidos anteriormente. Finalmente, resta apenas um número para o lugar de ordem n. Portanto, existem 2×3^8 permutações deste tipo. ∎

Enumeração de Conjuntos: Coeficientes Binomiais

O número de conjuntos, ou coleções não ordenadas, de tamanho r escolhidas de um conjunto universo de tamanho n, onde, como apropriado para conjuntos, não é permitido a duplicação de elementos (amostragem sem repetição), é dado pelo *coeficiente binomial*:

$$\binom{n}{r} = \frac{(n)_r}{r!} = \frac{n!}{(n-r)!r!}.$$

Para verificar isto, note que o número de coleções ordenadas de tamanho r sem repetição é $(n)_r$. Como os elementos de cada seqüência de comprimento r são distintos, o número de permutações de cada seqüência é $r!$. Porém, utilizando a regra da multiplicação, o procedimento de escolhermos uma coleção ordenada de r termos sem repetição é igual a primeiro escolher uma coleção não-ordenada de r termos sem repetição e depois escolhermos uma ordem para esta coleção não ordenada, ou seja, temos que

$$(n)_r = \binom{n}{r} \cdot r!,$$

de onde segue o resultado.

O coeficiente binomial tem as seguintes propriedades:

$$\binom{n}{r} = \binom{n}{n-r}, \binom{n}{0} = 1, \binom{n}{1} = n, \binom{n}{r} = 0 \text{ se } n < r.$$

Note que o coeficiente binomial é também igual ao número de subconjuntos de tamanho r que pode ser formado de um conjunto de n elementos. Como já vimos que, o número total de subconjuntos de um conjunto de tamanho n é 2^n, temos que

$$2^n = \sum_{r=0}^{n} \binom{n}{r}.$$

Os números $\binom{n}{r}$ são chamados de coeficientes binomiais, porque eles aparecem como coeficientes na expressão binomial $(a+b)^n$. Se n for um inteiro positivo, $(a+b)^n = (a+b)(a+b)\cdots(a+b)$. Quando a multiplicação tiver sido executada, cada termo será formado de k elementos de a e de $(n-k)$ elementos de b, para $k = 0, 1, 2, \ldots, n$. Mas quantos termos da forma $a^k b^{n-k}$ existirão? Simplesmente contaremos o número de maneiras possíveis de escolher k dentre os n elementos a, deixando de lado a ordem (onde o i-ésimo elemento a corresponde ao i-ésimo fator do produto acima). Mas isto é justamente dado por $\binom{n}{k}$. Daí obtém-se o que é conhecido como o *Teorema Binomial*:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Exemplo 2.2.10: Dentre oito pessoas, quantas comissões de três membros podem ser escolhidas, desde que duas comissões sejam a mesma comissão se forem constituídas pelas mesmas pessoas (não se levando em conta a ordem em que sejam escolhidas)? A resposta é dada por $\binom{8}{3} = 56$ comissões possíveis. ∎

Exemplo 2.2.11: Com oito bandeiras diferentes, quantos sinais feitos com três bandeiras diferentes se podem obter? Este problema parece-se muito com o exemplo anterior, mas neste caso a ordem acarreta diferença e por isso temos $(8)_3 = 336$ sinais. ∎

Exemplo 2.2.12: Um grupo de oito pessoas é formado de cinco homens e três mulheres. Quantas comissões de três pessoas podem ser constituídas, incluindo exatamente dois homens? Aqui deveremos fazer duas coisas, escolher dois homens (dentre cinco) e escolher duas mulheres (dentre três). Daí obtemos como número procurado $\binom{5}{2}\binom{3}{1} = 30$ comissões. ∎

Exemplo 2.2.13: Quantos seqüências binárias de comprimento n contém no máximo três números 1? Neste caso, temos quatro casos possíveis: todas seqüencias que não contém 1, todas seqüências que contém apenas um número 1, todas seqüências que contém dois números 1, e todas as seqüências que contém três números 1. Para $0 \leq r \leq n$, temos que existem exatamente $\binom{n}{r}$ seqüências binárias com r números 1. Portanto, pela regra da adição temos que existem

$$\binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \binom{n}{3}$$

seqüências binárias de comprimento n contendo no máximo três números 1. ∎

Exemplo 2.2.14: Quantas seqüências de cara e coroa de comprimento n contém pelo menos 1 cara? Neste caso, note que apenas uma seqüência não contém nenhuma cara (a seqüência que contém apenas coroa). Como o número total de seqüências de cara e coroa de comprimento n é igual a 2^n, temos então $2^n - 1$ seqüências de comprimento n contendo pelo menos uma cara. ∎

Exemplo 2.2.15: Determine o coeficiente de x^3 no desenvolvimento de $(x^4 - \frac{1}{x})^7$.

Solução: O termo genérico do desenvolvimento é

$$\binom{7}{k}(x^4)^k(-\frac{1}{x})^{7-k} = (-1)^{7-k}\binom{7}{k}x^{5k-7}.$$

Portanto, temos o termo x^3 se $5k - 7 = 3$, o que implica que $k = 2$. Logo, o coeficiente de x^3 é $(-1)^5\binom{7}{2} = -21$. ∎

Contagem Multinomial

Considere que temos r tipos de elementos e n_i cópias indistinguíveis do elemento do tipo i. Por exemplo, a palavra *probabilidade* tem duas cópias de cada uma das letras *a,b,d,i* e uma cópia de cada uma das letras *l,p,r,o,e*. O número de seqüências ordenadas de comprimento $n = \sum_{i=1}^{r} n_i$ é dado por:

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots 1 = \frac{n!}{\prod_{i=1}^{r} n_i!}.$$

Esta quantidade é conhecida como *coeficiente multinomial* e denotada por:

$$\binom{n}{n_1\ n_2\ \ldots\ n_r},$$

onde $n = \sum_{i=1}^{r} n_i$.

Para verificar esta contagem, note que das n posições na seqüência de comprimento n, nós podemos escolher n_1 posições para os n_1 elementos indistinguíveis do tipo 1 de $\binom{n}{n_1}$ maneiras. Das $n - n_1$ posições restantes na seqüência, podemos escolher n_2 posições para os n_2 elementos indistinguíveis do tipo 2 de $\binom{n-n_1}{n_2}$ maneiras. Finalmente, após repetir este processo $r - 1$ vezes, restam-nos n_r posições na seqüência para os n_r elementos do tipo r, que só podem ser escolhidas de uma única maneira. Utilizando o método da multiplicação, o número total de seqüências possíveis é produto do número de maneiras que podemos colocar os r tipos de elementos.

O coeficiente multinomial também calcula o número de partições de um conjunto n elementos em r subconjuntos com tamanhos dados $n_1, n_2, \ldots, n_r$. Aplicando-se o mesmo argumento que utilizamos para demonstrar o Teorema Binomial, pode-se provar a seguinte generalização conhecida como *Teorema Multinomial*:

$$(x_1 + x_2 + \ldots + x_r)^n = \sum_{i_1=0}^{n} \sum_{i_2=0}^{n-i_1} \cdots \sum_{i_{r-1}=0}^{n-\sum_{j<r-1} i_j} \binom{n}{i_1\ i_2\ \ldots\ i_r} \prod_{k=1}^{r} x_k^{i_k},$$

onde $i_r = n - \sum_{j<r} i_j$.

Exemplo 2.2.16: Um monitor tendo resolução de $n = 1.280 \times 854$ pixels, com $r = 3$ cores possíveis (verde, azul, e vermelho) para cada pixel, pode mostrar $\binom{n}{i_1\ i_2\ i_3}$ imagens tendo i_1 pixels verdes, i_2 pixels azuis, e i_3 pixels vermelhos. O número total de imagens que pode ser exibida por este monitor para qualquer composição de cores de ver, azul, e vermelho pode ser obtido utilizando o Teorema Multinomial fazendo $x_1 = x_2 = \ldots = x_r = 1$, dando o resultado de r^n possíveis imagens. ∎

Exemplo 2.2.17: Determine o coeficiente de x^9y^4 no desenvolvimento de $(x^3 + 2y^2 + \frac{5}{x^2})^5$.

Solução: O termo genérico do desenvolvimento é

$$\binom{5}{i_1\ i_2\ 5 - i_1 - i_2} (x^3)^{i_1} (2y^2)^{i_2} (\frac{5}{x^2})^{5-i_1-i_2} =$$
$$(2)^{i_2} (5)^{5-i_1-i_2} \binom{5}{i_1\ i_2\ 5 - i_1 - i_2} x^{3i_1-10+2i_1+2i_2} y^{2i_2}. \tag{2.1}$$

Portanto, temos o termo x^9y^4 se $5i_1 + 2i_2 - 10 = 9$ e $2i_2 = 4$, o que implica que $i_2 = 2$ e $i_1 = 3$. Logo, o coeficiente de x^9y^4 é $(2)^2(5)^0\binom{5}{3\ 2\ 0} = 40$. ∎

Capítulo 3

Probabilidade Condicional

3.1 Probabilidade Condicional

Como vimos no capítulo anterior, existem várias possíveis interpretações de probabilidade. Por exemplo, pode-se interpretar probabilidade de um evento A como um limite das freqüências relativas de ocorrência do evento A em realizações independentes de um experimento. Por outro lado, a interpretação subjetiva de probabilidade associa a probabilidade de um evento A com o grau de crença pessoal que o evento A ocorrerá. Em ambos os casos, probabilidade é baseada em informação e conhecimento. Revisão desta base de informação ou conhecimento pode levar a revisão do valor da probabilidade. Em particular, conhecimento que determinado evento ocorreu pode influenciar na probabilidade dos demais eventos.

Considerando-se a interpretação freqüentista de probabilidade, suponha que estejamos interessados em saber qual a probabilidade de um dado evento A, visto que sabe-se que um dado evento B ocorreu. Suponha que realizasse um experimento n vezes das quais o evento A (resp., B e $A \cap B$) ocorre N_A (resp., $N_B > 0$ e $N_{A \cap B}$) vezes. Seja $r_A = N_A/n$ a freqüência relativa do evento A nestas n realizações do experimento. A probabilidade condicional de A dado que sabe-se que B ocorreu segundo esta interpretação freqüentista, sugere que ela deve ser igual ao limite das freqüências relativas condicionais do evento A dado o evento B, isto é, ela deve ser o limite da razão $N_{A \cap B}/N_B$ quando n tende ao infinito. É fácil provar que esta razão é igual a $r_{A \cap B}/r_B$, que por sua vez segundo a interpretação freqüentista de probabilidade é aproximadamente igual a $P(A \cap B)/P(B)$ para valores grandes de n.

Considerando-se uma interpretação mais subjetiva suponha que a incerteza de um agente é descrita por uma probabilidade P em $(\Omega, \mathcal{A})$ e que o agente observa ou fica sabendo que o evento B ocorreu. Como o agente deve atualizar sua probabilidade $P(\cdot|B)$ de modo a incorporar esta nova informação? Claramente, se o agente *acredita* que B é verdadeiro, então parece razoável requerer que

$$P(B^c|B) = 0 \tag{3.1}$$

Em relação aos eventos contidos em B, é razoável assumir que sua chance relativa permaneça inalterada se tudo que o agente descobriu foi que o evento B ocorreu, ou seja, se

$A_1, A_2 \subseteq B$ com $P(A_2) > 0$, então

$$\frac{P(A_1)}{P(A_2)} = \frac{P(A_1|B)}{P(A_2|B)} \tag{3.2}$$

Segue que (3.1) e (3.2) determinam completamente $P(\cdot|B)$ se $P(B) > 0$.

Teorema 3.1.1: *Se $P(B > 0)$ e $P(\cdot|B)$ é uma medida de probabilidade em Ω que satisfaz (3.1) e (3.2), então*

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

Prova: Como $P(\cdot|B)$ é uma medida de probabilidade e satisfaz $P(B^c|B) = 0$, nós temos que $P(B|B) = 1 - P(B^c|B) = 1$. Considerando $A_1 = A$ e $A_2 = B$ em (3.2), temos então $P(A|B) = \frac{P(A)}{P(B)}$ para $A \subseteq B$. Se A não é um subconjunto de B, temos que $A = (A \cap B) \cup (A \cap B^c)$. Como $(A \cap B)$ e $(A \cap B^c)$ são eventos disjuntos, temos $P(A|B) = P(A \cap B|B) + P(A \cap B^c|B)$. Como $A \cap B^c \subseteq B^c$ e $P(B^c|B) = 0$, temos que $P(A \cap B^c|B) = 0$. Como $A \cap B \subseteq B$, usando o caso anterior

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

∎

Deste modo as interpretações freqüentista e subjetivista de probabilidade justificam a seguinte definição.

Definição 3.1.2: Seja $(\Omega, \mathcal{A}, P)$ um espaço de probabilidade. Se $A, B \in \mathcal{A}$ e $P(B) > 0$ a *probabilidade condicional de A dado B* é definida por

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

Vamos provar que para um evento fixo B que satisfaz $P(B) > 0$, $P(\cdot|B)$ satisfaz os axiomas K1-K4 acima e realmente é uma medida de probabilidade. Para provar $K1$, note que para todo $A \in \mathcal{A}$, como $P(A \cap B) \geq 0$, nós temos

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \geq 0.$$

Para provar $K2$, note que $\Omega \cap B = B$, então

$$P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

Finalmente, para provar K4 (que implica K3), note que se $A_1, A_2, \ldots$ são mutuamente exclusivos $A_1 \cap B, A_2 \cap B, \ldots$ também o são, então

$$P(\cup_i A_i|B) = \frac{P((\cup_i A_i) \cap B)}{P(B)} = \frac{P(\cup_i (A_i \cap B))}{P(B)}$$
$$= \frac{\sum_i P(A_i \cap B)}{P(B)} = \sum_i P(A_i|B).$$

A probabilidade condicional também satisfaz as seguintes propriedades:

1. $P(B|B) = 1$;
2. $P(A|B) = P(A \cap B|B)$;
3. se $A \supseteq B$, então $P(A|B) = 1$;
4. $P(A \cap B|C) = P(A|B \cap C)P(B|C)$.

Fazendo $C = \Omega$ na propriedade 4 acima, temos que:

$$P(A \cap B) = P(A|B)P(B).$$

Utilizando indução matemática, pode-se facilmente provar que

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2|A_1)\ldots P(A_n|A_1 \cap \ldots \cap A_{n-1}).$$

Um método de se obter uma probabilidade (incondicional) de uma probabilidade condicional é utilizando o Teorema da Probabilidade Total. Antes de enunciar este teorema precisamos recordar o que é uma *partição* do espaço amostral. Uma seqüência de eventos $A_1, A_2, A_3, \ldots$ é uma partição do espaço amostral Ω se estes eventos são mutuamente exclusivos e contém todos os elementos de Ω ($\cup_i A_i = \Omega$).

Teorema 3.1.3: *Seja a seqüência de eventos $B_1, B_2, \ldots$ uma partição de Ω, então para todo $A \in \mathcal{A}$*

$$P(A) = \sum_{i:P(B_i)\neq 0} P(A|B_i)P(B_i)$$

Prova:

Como $B_1, B_2, \ldots$ é uma partição de Ω, temos que

$$A = A \cap \Omega = A \cap (\cup_i B_i) = \cup_i (A \cap B_i).$$

Como os eventos B_i's são mutuamente exclusivos, os eventos $(A \cap B_i)$'s também são mutuamente exclusivos. Então axioma $K3$ implica que

$$P(A) = P(\cup_i (A \cap B_i)) = \sum_i P(A \cap B_i)$$
$$= \sum_{i:P(B_i)\neq 0} P(A \cap B_i) = \sum_{i:P(B_i)\neq 0} P(A|B_i)P(B_i).$$

∎

Se nós interpretarmos a partição $B_1, B_2, \ldots$ como possíveis causas e o evento A corresponda a um efeito particular associado a uma causa, $P(A|B_i)$ especifica a relação estocástica entre a causa B_i e o efeito A.

Por exemplo, seja $\{D, D^c\}$ uma partição do espaço amostral, onde o evento D significa que um dado indivíduo possui uma certa doença. Seja A o evento que determinado teste para

o diagnóstico da doença deu positivo. Então, $P(A|D^c)$ descreve a probabilidade do exame dá positivo mesmo que o paciente esteja saudável, é a chamada probabilidade de *falso positivo*. $P(A^c|D)$ é a probabilidade do exame dá negativo mesmo que o paciente esteja doente, é a chamada probabilidade de *falso negativo*. Estas probabilidades determinam a qualidade do teste, quanto menores as probabilidades de falso negativo e falso positivo melhor a qualidade do teste. Caso as probabilidades $P(D), P(A|D), P(A|D^c)$ sejam conhecidas pode-se usando o Teorema da Probabilidade Total obter a probabilidade incondicional de determinado exame dar positivo $P(A)$. Porém geralmente, o que se busca é saber que dado que o resultado de um exame deu positivo qual a probabilidade de que o indivíduo esteja doente. Pode-se obter esta probabilidade utilizando a famosa *fórmula de Bayes*:

$$P(D|A) = \frac{P(A \cap D)}{P(A \cap D) + P(A \cap D^c)} = \frac{P(A|D)P(D)}{P(A|D)P(D) + P(A|D^c)P(D^c)}.$$

Mais geralmente, quando temos uma partição $B_1, B_2, \ldots$, temos que a fórmula de Bayes é dada por:

$$\begin{aligned} P(B_i|A) &= \frac{P(A \cap B_i)}{\sum_j P(A \cap B_j)} = \frac{P(A \cap B_i)}{\sum_{j:P(B_j)\neq 0} P(A \cap B_j)} \\ &= \frac{P(A|B_i)P(B_i)}{\sum_{j:P(B_j)\neq 0} P(A|B_j)P(B_j)}. \end{aligned}$$

É fácil de provar esta fórmula usando o Teorema da Probabilidade Total. As probabilidades $P(B_i)$ são usualmente chamadas de probabilidades *a priori* e as probabilidades condicionais $P(B_i|A)$ são chamadas de probabilidades *a posteriori*. O seguinte exemplo ilustra uma aplicação da fórmula de Bayes.

Exemplo 3.1.4: Considere uma imagem formada por $n \times m$ pixels com a k-ésima linha contendo $d_k (\leq m)$ pixels defeituosos. No primeiro estágio do experimento uma linha é escolhida ao acaso e nós não sabemos qual foi a escolha. Nós então examinamos um pixel selecionada ao acaso nesta linha e descobrimos que o pixel é defectivo (chamamos este evento de D). Qual a probabilidade de que este pixel defeituoso esteja na linha k? Seja $R = k$ o evento que este pixel pertencia a k-ésima linha da imagem. A fórmula de Bayes nos permite determinar que dado que

$$P(R = k) = \frac{1}{n} \quad \text{e} \quad P(D|R = k) = \frac{d_k}{m},$$

nós temos que

$$P(R = k|D) = \frac{\frac{1}{n}\frac{d_k}{m}}{\sum_{i=1}^{n} \frac{1}{n}\frac{d_i}{m}} = \frac{d_k}{\sum_{i=1}^{n} d_i}.$$

Então, mesmo que a linha tenha inicialmente sido escolhida ao acaso, dado o evento que encontramos ao acaso um pixel defectivo nesta linha, agora é mais provável que seja uma linha contendo um número grande de pixels defectivos d_k.

Exemplo 3.1.5: Uma urna contém 4 bolas brancas e 6 bolas pretas. Sacam-se, sucessivamente e sem reposição, duas bolas dessa urna. Determine a probabilidade da primeira bola ser branca sabendo que a segunda bola é branca.

Solução: Sejam B_1 e B_2 os eventos a primeira bola é branca e a segunda bola é branca, respectivamente. Queremos calcular $P(B_1|B_2)$. Utilizando a fórmula de Bayes, temos

$$P(B_1|B_2) = \frac{P(B_2|B_1)P(B_1)}{P(B_2|B_1)P(B_1) + P(B_2|B_1^c)P(B_1^c)}.$$

Mas $P(B_2|B_1) = \frac{3}{9}$, $P(B_2|B_1^c) = \frac{4}{9}$, $P(B_1) = \frac{4}{10}$ e $P(B_1^c) = \frac{6}{10}$. Logo,

$$P(B_1|B_2) = \frac{\frac{3}{9} \cdot \frac{4}{10}}{\frac{3}{9} \cdot \frac{4}{10} + \frac{4}{9} \cdot \frac{6}{10}} = \frac{\frac{2}{15}}{\frac{2}{5}} = \frac{1}{3}.$$

Embora probabilidade condicional seja bastante útil, ela sofre de alguns problemas, em particular quando se quer tratar de eventos de probabilidade zero. Tradicionalmente, se $P(B) = 0$, então $P(A|B)$ não é definida. Isto leva a um número de dificuldades filosóficas em relação a eventos com probabilidade zero. São eles realmente impossíveis? Caso contrário, quão improvável um evento precisa ser antes de ele ser atribuído probabilidade zero? Deve um evento em algum caso ser atribuído probabilidade zero? Se existem eventos com probabilidade zero que não são realmente impossíveis, então o que significa condicionar em eventos de probabilidade zero? Por exemplo, considere o espaço de probabilidade $([0,1], \mathcal{B}, \mu)$ onde $\mathcal{B}$ é a σ-álgebra de Borel restrita a eventos contidos em $[0,1]$ e μ é uma medida de probabilidade na qual todo intervalo em $[0,1]$ possui probabilidade igual ao seu comprimento. Seja $B = \{1/4, 3/4\}$ e $A = \{1/4\}$. Como $P(B) = 0$, $P(A|B)$ não é definida. Porém parece razoável assumir que neste caso $P(A|B) = 1/2$ já que μ intuitivamente implica que todos os estados são equiprováveis, mas a definição formal de probabilidade condicional não nos permite obter esta conclusão.

Alguns dos problemas mencionados no parágrafo anterior podem ser tratados considerando-se probabilidades condicionais (e não probabilidade incondicionais) como a noção fundamental, porém a discussão destes modelos está fora do escopo deste curso.

Exemplo 3.1.6: Se $P(C|D) = 0,4$ e $P(D|C) = 0,5$, que evento é mais provável C ou D?
Solução:

Exemplo 3.1.7: Se $P(E) = 0,4$ e $P(F) = 0,7$, o que pode-se concluir sobre $P(E|F)$?
Solução: Por definição, temos que:

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Porém, sabemos que $\max(P(E) + P(F) - 1, 0) \leq P(E \cap F) \leq \min(P(E), P(F))$. Logo, $0,1 \leq P(E \cap F) \leq 0,4$, portanto

$$\frac{0,1}{0,7} \leq P(E|F) \leq \frac{0,4}{0,7}.$$

Exemplo 3.1.8: (Paradoxo de Monty Hall) Monty Hall foi um popular apresentador de programa de jogos em TV cujo jogo começava mostrando ao participante 3 portas fechadas d_1, d_2, d_3, e atrás de apenas uma delas havia um prêmio valioso. O participante selecionava uma porta, por exemplo, d_1, mas antes que a porta fosse aberta, Monty Hall, que sabia em que porta estava o prêmio, por exemplo, d_2, abria a porta restante d_3, que não continha o prêmio. O participante tinha então permissão para ficar com sua porta original, d_1, ou escolher a outra porta fechada. A pergunta é se é melhor ficar com a porta original ou trocar de porta. Vamos agora utilizar a fórmula de Bayes para analisar este problema. Seja G uma porta escolhida aleatoriamente para conter o prêmio; Y a porta que o participante escolhe primeiro; e M a porta que Monty Hall abre. O participante não tem nenhum conhecimento a priori sobre a localização do prêmio, ou seja ele considera todas as portas equiprováveis, e isto pode ser modelado por:

$$P(G = d_i|Y = d_j) = \frac{1}{3};$$

todas as portas tem a mesma probabilidade de conter o prêmio não importa qual porta o participante escolhe. Se o participante escolher uma porta que não contém o prêmio, Monty Hall necessariamente terá de abrir a porta que não contém o prêmio, isto pode ser modelado por:

$$P(M = d_{i_1}|Y = d_{i_2}, G = d_{i_3}) = 1,$$

onde $i_1, i_2, i_3 \in \{1, 2, 3\}$ e são distintos. Se o participante escolher corretamente, por exemplo, $Y = G = d_{i_2}$, então assumimos que Monty Hall escolhe aleatoriamente entre as outras duas outras portas:

$$P(M = d_{i_1}|Y = G = d_{i_2}) = \frac{1}{2}, \text{ para } d_{i_1} \neq d_{i_2}.^{1}$$

Para determinar se o participante deve trocar de porta, devemos calcular

$$\begin{aligned} P(G = d_1|Y = d_2, M = d_3) &= \frac{P(G = d_1, Y = d_2, M = d_3)}{P(Y = d_2, M = d_3)} \\ &= \frac{P(M = d_3|G = d_1, Y = d_2)P(G = d_1|Y = d_2)P(Y = d_2)}{P(M = d_3|Y = d_2)P(Y = d_2)} \\ &= \frac{P(M = d_3|G = d_1, Y = d_2)P(G = d_1|Y = d_2)}{P(M = d_3|Y = d_2)} \\ &= \frac{1/3}{P(M = d_3|Y = d_2)} \end{aligned}$$

Para determinar o valor de $P(M = d_3|Y = d_2)$ utilizamos o Teorema da Probabilidade Total

[1] A solução depende como resolvemos este caso.

e a definição de probabilidade condicional:

$$
\begin{aligned}
&P(M=d_3|Y=d_2) = \frac{P(Y=d_2, M=d_3)}{P(Y=d_2)} \\
&= \frac{P(Y=d_2, M=d_3, G=d_1) + P(Y=d_2, M=d_3, G=d_2) + P(Y=d_2, M=d_3, G=d_3)}{P(Y=d_2)} \\
&= \frac{P(M=d_3|Y=d_2, G=d_1)P(G=d_1|Y=d_2)P(Y=d_2)}{P(Y=d_2)} \\
&+ \frac{P(M=d_3|Y=d_2, G=d_2)P(G=d_2|Y=d_2)P(Y=d_2)}{P(Y=d_2)} \\
&+ \frac{P(M=d_3|Y=d_2, G=d_3)P(G=d_3|Y=d_2)P(Y=d_2)}{P(Y=d_2)} \\
&= P(M=d_3|Y=d_2, G=d_1)P(G=d_1|Y=d_2) \\
&+ P(M=d_3|Y=d_2, G=d_2)P(G=d_2|Y=d_2) \\
&+ P(M=d_3|Y=d_2, G=d_3)P(G=d_3|Y=d_2) \\
&= 1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + 0 = \frac{1}{2}.
\end{aligned}
$$

Logo, $P(G=d_1|Y=d_2, M=d_3) = \frac{2}{3}$, e o participante deve trocar de porta de sua escolha original d_2 para d_1!

Exemplo 3.1.9: Seja D o evento que um indivíduo selecionado ao acaso de uma população tem uma doença particular, D^c seu complemento. A probabilidade que um indivíduo selecionado ao acaso nesta população tenha determinada dença é p_d. Existe um teste para diagnóstico desta doença que sempre acusa presença da doença quando o indivíduo tem a doença. Contudo, quando o indivíduo não tem a doença, o teste reporta falsamente que o indivíduo tem a doença com probabilidade p_t. Seja TP o evento que o teste reporta positivamente que o indivíduo tem a doença. Formalmente, temos:

$$P(D) = p_d, P(TP|D) = 1, P(TP|D^c) = p_t.$$

Um indivíduo deve estar interessado em saber a probabilidade $P(D|TP)$ que ele tenha a doença dado que o teste deu positivo. Se, por exemplo, a doença for rara e $p_d = 0,001$, e o teste reportar falsamente com probabilidade pequena $p_t = 0,05$, veremos que apesar desta pequena probabilidade do teste da um resultado errado, a probabilidade do indivíduo ter a doença é pequena. Pela fórmula de Bayes

$$P(D|TP) = \frac{P(TP|D)P(D)}{P(TP|D)P(D) + P(TP|D^c)P(D^c)} = \frac{p_d}{p_d + p_t(1-p_d)} = 0,02.$$

Exemplo 3.1.10: Sabemos que os eventos $\{B_1, B_2, B_3\}$ são disjuntos par a par e que sua união é igual ao espaço amostral. Estes eventos tem as seguintes probabilidades $P(B_1) = 0,2$ e $P(B_2) = 0,3$. Existe um outro evento A que sabemos que $P(A|B_1) = 0,3$; $P(A|B_2) = 0,4$; e $P(A|B_3) = 0,1$. Calcule:

(a) $P(A)$

(b) $P(B_2|A)$

Exemplo 3.1.11: Suponha que todos os bytes tenham a mesma probabilidade. Seja W o número de 1's em um byte. Considere os seguintes eventos:

$$A = \{\text{O primeiro e o segundo bit são iguais a 1, e}\}$$

$$B = \{W \text{ é um número ímpar.}\}$$

Calcule:

(a) $P(A)$

(b) $P(B)$

(c) $P(B|A)$

(d) $P(A|B)$

Solução:

$$P(A) = \frac{||A||}{||\Omega||} = \frac{2^6}{2^8} = \frac{1}{4}.$$

$$P(B) = \frac{||B||}{||\Omega||} = \frac{\binom{8}{1} + \binom{8}{3} + \binom{8}{5} + \binom{8}{7}}{2^8} = \frac{1}{2}.$$

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

onde $P(A \cap B) = \frac{||A \cap B||}{\Omega} = \frac{\binom{6}{1}+\binom{6}{3}+\binom{6}{5}}{2^8} = \frac{1}{8}$. Portanto,

$$P(B|A) = \frac{\frac{1}{8}}{\frac{1}{4}} = \frac{1}{2}.$$

$$P(A|B) = \frac{P(A \cap B)}{B} = \frac{\frac{1}{8}}{\frac{1}{2}} = \frac{1}{4}.$$

Exemplo 3.1.12: Se jogarmos dois dados um após o outro e observamos o evento que a soma dos dois dados é igual a 9, então qual a probabilidade do primeiro dado ter dado resultado 4?

Solução:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{36}}{\frac{4}{36}} = \frac{1}{4}.$$

Exemplo 3.1.13: Em um teste de múltipla escolha, a probabilidade do aluno saber a resposta da questão é p. Havendo m escolhas, se ele sabe a resposta ele responde corretamente com probabilidade 1; se não sabe ele responde corretamente com probabilidade $\frac{1}{m}$.

(a) Qual a probabilidade que a pergunta foi respondida corretamente?

(b) Qual a probabilidade que o aluno sabia a resposta dado que a pergunta foi respondida corretamente?

Solução: Para a parte (a), usamos o Teorema da Probabilidade Total:

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) = 1 \cdot p + \frac{1}{m}(1-p).$$

Para a parte (b), usamos a fórmula de Bayes

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} = \frac{1 \cdot p}{1 \cdot p + \frac{1}{m}(1-p)}$$

3.2 Independência

O que exatamente significa que dois eventos são independentes? Intuitivamente, isto significa que eles não têm nada haver um com o outro, eles são totalmente não relacionados; a ocorrência de um não tem nenhuma influência sobre o outro. Por exemplo, suponha que duas diferentes moedas são lançadas. A maioria das pessoas viria os resultados desses lançamentos como independentes. Portanto, a intuição por trás da frase "o evento A é independente do evento B" é que nosso conhecimento sobre a tendência para A ocorrer dado que sabemos que B ocorreu não é alterada quando ficamos sabendo que B ocorreu. Então, usando probabilidades condicionais podemos formalizar esta intuição da seguinte forma, A é independente de B se $P(A|B) = P(A)$. Mas usando a definição de probabilidade condicional, chega-se a seguinte conclusão A é independente de B se $P(A \cap B) = P(A)P(B)$. Como esta última expressão é definida inclusive para o caso de $P(B) = 0$, ela é a expressão adotada como a definição de independência entre eventos.

Definição 3.2.1: O evento A é independente do evento B se $P(A \cap B) = P(A)P(B)$.

Note que esta definição de independência implica que independência é um conceito simétrico em teoria da probabilidade, isto é, A é independente de B se e somente se B é independente de A. Note que esta definição também implica que eventos A e B são independentes se $P(A) = 0$ ou $P(B) = 0$, o que pode gerar algumas conclusões não intuitivas se de fato $P(A) = 0$ ou $P(B) = 0$. Por exemplo, se $P(A) = 0$, então A é independente dele mesmo, porém A certamente não é não relacionado consigo mesmo. Similarmente, é fácil provar que se $P(A) = 1$, A é independente dele mesmo. O seguinte teorema prova que estes são os únicos casos em que um evento é independente dele mesmo.

Teorema 3.2.2: *A é independente dele mesmo se e somente se $P(A) = 0$ ou $P(A) = 1$.*

Prova:

$$P(A \cap A) = P(A) = P(A)P(A) \Leftrightarrow P(A) = 0 \text{ ou } P(A) = 1.$$

∎

Intuitivamente, se A é independente de B o fato que B não ocorreu, ou seja que B^c ocorreu, não deve alterar a probabilidade de A. Portanto, é de se esperar que se A e B são independentes, então A e B^c também são. O seguinte teorema prova que esta intuição é verdadeira.

Teorema 3.2.3: *Se A e B são eventos independentes, A e B^c (resp., A^c e B, A^c e B^c) também o são.*

Prova: Note que

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

Então, como $A \cap B$ e $A \cap B^c$ são mutuamente exclusivos, axioma K3 implica que

$$P(A) = P(A \cap B) + P(A \cap B^c).$$

Como A e B são independentes, nós temos

$$P(A) = P(A)P(B) + P(A \cap B^c).$$

Rearrajando os termos e utilizando o fato que $P(B^c) = 1 - P(B)$, temos $P(A \cap B^c) = P(A)P(B^c)$, como queríamos demonstrar. ∎

O conceito de independência também se aplica a uma coleção arbitrária de eventos $\{A_i\}_{i \in \mathcal{I}}$, onde $\mathcal{I}$ é um conjunto de índices. Neste caso, têm-se duas definições.

Definição 3.2.4: Uma coleção de eventos $\{A_i\}_{i \in \mathcal{I}}$ é *independente par a par* se para todo $i \neq j \in I$, A_i e A_j são eventos independentes.

Definição 3.2.5: Uma seqüência finita de eventos $A_1, A_2, \ldots, A_n$, $n \geq 1$, é *mutuamente independente* se para todo $I \subseteq \{1, \ldots, n\}$,

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

E uma coleção de eventos $\{A_i\}_{i \in \mathcal{I}}$ é mutuamente independente se para todo $J \subseteq \mathcal{I}$ finito, $\{A_i\}_{i \in J}$ é mutuamente independente.

Considere os seguintes exemplos que ilustram o conceito de independência.

Exemplo 3.2.6: Se $\Omega = \{1, 2, 3, 4\}$ e $P(\{w\}) = 1/4$, então $A = \{1, 2\}$, $B = \{1, 3\}$, e $C = \{2, 3\}$ são eventos independentes par a par. Pode-se verificar isto pelo fato que

$$P(A \cap B) = P(\{1\}) = \frac{1}{4} = \frac{1}{2}\frac{1}{2} = P(A)P(B).$$

Similarmente, pode-se provar o mesmo resultado para os outros pares. Contudo, a probabilidade

$$P(A \cap B \cap C) = P(\emptyset) = 0 \neq P(A)P(B)P(C) = \frac{1}{8}.$$

Então, A, B, e C não são mutuamente independentes.

Exemplo 3.2.7: Se $\Omega = \{1,2,3,4,5,6\}$, $A = \{1,2,4\}$, e $B = \{2,3,5\}$, então construa uma medida de probabilidade em Ω tal que A e B sejam independentes.

Solução: Seja p_i a probabilidade do elemento $i \in \Omega$. Então, para que A e B sejam indeendentes devemos ter:

$$P(A \cap B) = p_2 = P(A)P(B) = (p_1 + p_2 + p_4)(p_2 + p_3 + p_5).$$

Por exemplo, podemos escolher $p_1 = p_2 = p_3 = p_6 = \frac{1}{4}$ e $p_4 = p_5 = 0$. Deste modo temos, $P(A \cap B) = \frac{1}{4}$ e $P(A) = P(B) = \frac{1}{2}$.

Exemplo 3.2.8: O evento F de um determinado sistema falhar ocorre se os eventos A_1 ou A_2 ocorrerem, mas o evento A_3 não ocorrer. Se A_1, A_2, A_3 são mutumente independetes e $P(A_1) = 0,4$, $P(A_2) = 0,35$, e $P(A_3) = 0,1$, então calcule $P(F)$.

Solução: O evento F é igual ao evento $(A_1 \cup A_2) \cap A_3^c$. Logo sua probabilidade é igual a:

$$P(F) = P((A_1 \cup A_2) \cap A_3^c) = P(A_1 \cup A_2)P(A_3^c)$$
$$= (P(A_1) + P(A_2) - P(A_1)P(A_2))(1 - P(A_3)) = (0,4 + 0,35 - 0,4 \cdot 0,35)(0,9) = 0,549.$$

Exemplo 3.2.9: Assuma que $A_1, \ldots, A_n$ são eventos mutuamente independentes e que $P(A_i) = p_i$. Nós calculamos as probabilidades dos seguintes eventos:

- O evento A é o evento que todos estes eventos ocorrem, então

$$P(A) = P(\cap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i) = \prod_{i=1}^n p_i$$

- O evento B é o evento que nenhum desses eventos ocorre, então

$$P(B) = P(\cap_{i=1}^n A_i^c) = \prod_{i=1}^n P(A_i^c) = \prod_{i=1}^n (1 - p_i)$$

- O evento C é o evento que pelo menos um desses eventos ocorre, então $C = B^c$

$$P(C) = P(B^c) = 1 - P(B) = 1 - \prod_{i=1}^n (1 - p_i)$$

Exemplo 3.2.10: João e José disputam um jogo com uma moeda equilibrada. Cada jogador lança a moeda duas vezes e vence o jogo aquele que primeiro obtiver dois resultados iguais. João começa jogando e se não vencer passa a moeda para José e continuam alternando jogadas. Qual a probabilidade de João vencer o Jogo?

Solução: Seja A_k o evento dois resultados iguais são obtidos na k-ésima tentativa. Note que $P(A_k) = \frac{1}{2}$. Seja B_k o evento João ganha na sua k-ésima jogada. Então,

$$B_1 = A_1;\ B_2 = A_1^c \cap A_2^c \cap A_3;\ B_3 = A_1^c \cap A_2^c \cap A_3^c \cap A_4^c \cap A_5,$$

em geral,

$$B_k = A_1^c \cap A_2^c \cap \cdots \cap A_{2k-2}^c \cap A_{2k-1}.$$

Portanto,

$$P(B_k) = P(A_1^c \cap A_2^c \cap \cdots \cap A_{2k-2}^c \cap A_{2k-1}) = P(A_1^c)P(A_2^c)\cdots P(A_{2k-2}^c)P(A_{2k-1}) = (\frac{1}{2})^{2k-1},$$

onde a penúltima igualdade se deve ao fato dos lançamentos serem independentes. Logo,

$$P(\text{João vencer}) = P(\cup_{k=1}^{\infty} B_k) = \sum_{k=1}^{\infty} P(B_k) = \sum_{k=1}^{\infty} (\frac{1}{2})^{2k-1} = \frac{2}{3}.$$

Capítulo 4

Variáveis Aleatórias

4.1 Introdução

Suponha que uma moeda é lançada cinco vezes. Qual é o número de caras? Esta quantidade é o que tradicionalmente tem sido chamada de *variável aleatória*. Intuitivamente, é uma variável porque seus valores variam, dependendo da seqüência de lançamentos da moeda realizada; o adjetivo "aleatória" é usado para enfatizar que o seu valor é de certo modo incerto. Formalmente, contudo, uma variável aleatória não é nem "aleatória" nem é uma variável.

Definição 4.1.1: Seja $(\Omega, \mathcal{A}, P)$ um espaço de probabilidade. Uma função $X : \Omega \to R$ é chamada de variável aleatória se para todo evento Boreliano B, $X^{-1}(B) \in \mathcal{A}$.

Por definição, temos que $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ é o conjunto de elementos do espaço amostral cuja imagem segundo X está em B. Nós recordamos que um evento Boreliano é qualquer evento pertencente à σ-álgebra de Borel, onde a σ-álgebra de Borel é a menor σ-álgebra contendo todos os intervalos.

Dada uma variável aleatória X, pode-se definir uma probabilidade induzida P_X no espaço mensurável $(R, \mathcal{B})$ da seguinte maneira: para todo $A \in \mathcal{B}$, definimos $P_X(A) = P(X^{-1}(A))$. Por definição de variável aleatória, tem-se que $X^{-1}(A) \in \mathcal{A}$, então P_X está bem definida. Resta provar que P_X satisfaz os axiomas K1, K2, e K4 de probabilidade:

K1. $P_X(A) = P(X^{-1}(A)) \geq 0$.

K2. $P_X(R) = P(X^{-1}(R)) = P(\Omega) = 1$.

K4. Suponha que $A_1, A_2, \ldots$ são eventos Borelianos disjuntos. Então,

$$P_X(\cup_i A_i) = P(X^{-1}(\cup_i A_i)) = P(\cup_i X^{-1}(A_i)) = \sum_i P(X^{-1}(A_i)) = \sum_i P_X(A_i).$$

Vale a pena salientar que em muitos problemas, já teremos a informação sobre a distribuição induzida P_X definida em $(R, \mathcal{B})$. Nestes casos, estaremos "esquecendo" a natureza

funcional de X e nos preocupando apenas com os valores assumidos por X. Estes casos podem ser pensados como se o experimento aleatório fosse descrito por $(R, \mathcal{B}, P_X)$ e $X(w) = w, \forall w \in R$, ou seja, os resultados dos experimento aleatório já são numéricos e descrevem a característica de interesse que queremos analisar.

É importante enfatizar que é usual se referir a variáveis aleatórias por letras maiúsculas $X, Y, Z, \ldots$ e aos valores que tais variáveis podem assumir por letras minúsculas $x, y, z, \ldots$.

4.2 Função de Distribuição Acumulada

Para uma variável aleatória X, uma maneira simples e básica de descrever a probabilidade induzida P_X é utilizando sua *função de distribuição acumulada.*

Definição 4.2.1: A função de distribuição acumulada de uma variável aleatória X, representada por F_X, é definida por

$$F_X(x) = P_X((-\infty, x]), \forall x \in R.$$

A função de distribuição acumulada F_X satisfaz as seguintes propriedades:

F1. Se $x \leq y$, então $F_X(x) \leq F_X(y)$.

$$x \leq y \Rightarrow (-\infty, x] \subseteq (-\infty, y] \Rightarrow P_X((-\infty, x]) \leq P_X((-\infty, y]) \Rightarrow F_X(x) \leq F_X(y).$$

F2. Se $x_n \downarrow x$, então $F_X(x_n) \downarrow F_X(x)$.

Se $x_n \downarrow x$, então os eventos $(-\infty, x_n]$ são decrescentes e $\cap_n(-\infty, x_n] = (-\infty, x]$. Logo, pela continuidade da medida de probabilidade, tem-se que $P_X((-\infty, x_n]) \downarrow P((-\infty, x])$, ou seja, $F_X(x_n) \downarrow F_X(x)$.

F3. Se $x_n \downarrow -\infty$, então $F_X(x_n) \downarrow 0$, e se $x_n \uparrow \infty$, então $F_X(x_n) \uparrow 1$.

Se $x_n \downarrow -\infty$, então os eventos $(-\infty, x_n]$ são decrescentes e $\cap_n(-\infty, x_n] = \emptyset$. Logo, pela continuidade da medida de probabilidade, tem-se que $P_X((-\infty, x_n]) \downarrow P(\emptyset)$, ou seja, $F_X(x_n) \downarrow 0$. Similarmente, se $x_n \uparrow \infty$, então os eventos $(-\infty, x_n]$ são crescentes e $\cup_n(-\infty, x_n] = I\!R$. Logo, pela continuidade da medida de probabilidade, tem-se que $P_X((-\infty, x_n]) \uparrow P(\Omega)$, ou seja, $F_X(x_n) \uparrow 1$.

Teorema 4.2.2: *Uma função real G satisfaz F1–F3 se e somente se G é uma distribuição de probabilidade acumulada.*

Prova: A prova de que se G for uma distribuição de probabilidade acumulada, então G satisfaz F1-F3 foi dada acima. A prova de que toda função real que satisfaz F1-F3 é uma função de probabilidade acumulada é complexa envolvendo o Teorema da Extensão de Carathéodory, e está fora do escopo deste curso. ∎

Condição F2 significa que toda função distribuição de probabilidade acumulada F_X é continua à direita. Ainda mais, como F_X é não-decrescente e possui valores entre 0 e 1, pode-se provar que ela tem um número enumerável de descontinuidades do tipo salto. Pela continuidade à direita , o salto no ponto x é igual a

$$\begin{aligned} &F_X(x) - F_X(x^-) = F_X(x) - \lim_{n\to\infty} F(x - \frac{1}{n}) \\ &= P_X((-\infty, x]) - \lim_{n\to\infty} P_X((-\infty, x - \frac{1}{n}]) \\ &= \lim_{n\to\infty} P_X((x - \frac{1}{n}, x]). \end{aligned}$$

Como a seqüência de eventos $(x - \frac{1}{n}, x]$ é decrescente e $\cap_n (x - \frac{1}{n}, x] = \{x\}$. Temos que $\{x\}$ é Boreliano e

$$P_X(x) = F_X(x) - F_X(x^-).$$

Ou seja, a probabilidade da variável aleatória X assumir o valor x é igual ao salto da função de distribuição acumulada F_X no ponto x.

Exemplo 4.2.3: Determine quais das seguintes funções são funções de distribuição acumuladas, especificando a propriedade que não for satisfeita caso a função não seja uma distribuição acumulada.

(a) $\frac{e^x}{1+e^x}$

(b) $I_{[0,infty)}(x) + [1 - I_{[0,infty)}(x)](1 + e^x)/2$

(c) $e^{-|x|}$

(d) $I_{[0,infty)}(x)$

(e) $I_{(0,infty)}(x)$

Exemplo 4.2.4: Seja K o número de íons emitidos por uma fonte em um tempo T. Se $F_K(1) - F_K(1/2) = 0,1$, qual o valor de $P(K = 1)$?

Exemplo 4.2.5: Uma seqüência de 10 bytes independentes foi recebida. É sábido que a probabilidade é igual a 0,3 que o primeiro símbolo de um byte seja igual a 0. Seja K o número de bytes recebidos tendo 0 como primeiro símbolo.

(a) Calcule $P(K = 2)$

(b) Calcule $F_K(1)$

4.3 Tipos de Variável Aleatória

Definição 4.3.1: Existem três tipos de variáveis aleatórias:

- **Discreta.** Uma variável aleatória X é *discreta* se assume um número enumerável de valores, ou seja, se existe um conjunto enumerável $\{x_1, x_2, \ldots\} \subseteq R$ tal que $X(w) \in \{x_1, x_2, \ldots\}, \forall w \in \Omega$. A função $p(x_i)$ definida por $p(x_i) = P_X(\{x_i\}), i = 1, 2, \ldots$ e $p(x) = 0$ para $x \notin \{x_1, x_2, \ldots\}$, é chamada de *função probabilidade* de X.

- **Contínua.** Uma variável aleatória X é *contínua* se existe uma função $f_X(x) \geq 0$ tal que
$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \forall x \in R.$$
Neste caso, a função f_X é chamada de *função densidade de probabilidade* de X.

- **Singular.** Uma variável aleatória X é *singular* se F_X é uma função contínua cujos pontos de crescimento formam um conjunto de comprimento (medida de Lebesgue) nulo.

Pode-se provar que toda função de distribuição de probabilidade acumulada F_X pode ser decomposta na soma de no máximo três funções de distribuição de probabilidade acumuladas, sendo uma discreta, uma contínua e outra singular.

Na grande maioria dos problemas práticos, não se encontram variáveis aleatórias singulares. Portanto, iremos nos restringir ao estudo de variáveis aleatórias discretas e contínuas. Na próxima seção analisaremos as variáveis aleatórias discretas.

4.4 Variável Aleatória Discreta

Vamos considerar agora o caso das variáveis aleatórias discretas. Nós vimos na seção anterior que se uma variável aleatória é discreta, então nós podemos definir uma função de probabilidade p de modo que $p(x_i) = P_X(\{x_i\}), i = 1, 2, \ldots$, onde $X \subseteq \{x_1, x_2, \ldots\}$ e $p(x) = 0$ para $x \notin \{x_1, x_2, \ldots\}$. Note que toda função de probabilidade é uma função dos reais R e assume valores entre 0 e 1, sendo positiva para um número enumerável de pontos e satisfaz a seguinte propriedade $\sum_i p(x_i) = 1$.

Por outro lado, dada uma função $p : R \to [0, 1]$, onde p é positiva para um número enumerável de pontos $\{x_1, x_2, \ldots\}$ e satisfaz $\sum_i p(x_i) = 1$, uma função P definida nos eventos Borelianos de modo que $P(A) = \sum_{x_i \in A} p(x_i), \forall A \in \mathcal{B}$ é uma medida de probabilidade em $(R, \mathcal{B})$ (é fácil verificar que P satisfaz os axiomas de Kolmogorov e portanto é uma medida de probabilidade). Logo, a distribuição de uma variável aleatória discreta X pode ser determinada tanto pela função de distribuição acumulada F_X ou pela sua função de probabilidade p.

Exemplo 4.4.1: Assuma que X é uma variável aleatória discreta que assume os valores 2, 5, e 7 com probabilidades 1/2, 1/3, e 1/6, então sua função de distribuição acumulada é:

$$F_X(x) = \begin{cases} 0 & \text{se } x < 2, \\ 1/2 & \text{se } 2 \leq x < 5, \\ 5/6 & \text{se } 5 \leq x < 7, \\ 1 & \text{se } x \geq 7. \end{cases}$$

A função de distribuição de uma variável discreta é sempre uma função degrau que tem saltos nos pontos que a variável assume com probabilidade positiva, e o valor do salto em um ponto x_i, como vimos é igual a probabilidade da variável assumir este valor.

4.5 Variável Aleatória Contínua

Vamos considerar agora o caso das variáveis aleatórias contínuas. Nós vimos na seção anterior que se uma variável aleatória é (absolutamente) contínua, então existe uma função $f_X(x) \geq 0$ tal que $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$. Deste modo, F_X é contínua e $f_X(x) = F'_X(x)$, exceto num conjunto de medida de Lebesgue nula. Uma função $f(x) \geq 0$ é densidade de alguma variável aleatória se e somente se, $\int_{-\infty}^{\infty} f(x)dx = 1$, já que neste caso é fácil provar que a função F definida por $\int_{-\infty}^{x} f(t)dt$ satisfaz as condições F1, F2, e F3. Portanto, pelo Teorema **??** F é uma função de distribuição acumulada. Logo, a distribuição de uma variável aleatória contínua X pode ser determinada tanto pela função de distribuição acumulada F_X ou pela sua função de densidade f_X.

Formalmente, uma variável aleatória X tem densidade se F_X é a integral de sua derivada; sendo neste caso a derivada de F_X uma função densidade para X. Além disso, em quase todos os casos encontrados na prática, uma variável aleatória X tem densidade se F_X é (i) contínua e (ii) derivável por partes, ou seja, se F_X é derivável no interior de um número finito ou enumerável de intervalos fechados cuja união é a reta R.

Por exemplo, considere

$$F_X(x) = \begin{cases} 0 & \text{se } x < 0, \\ x & \text{se } 0 \leq x < 1, \\ 1 & \text{se } x \geq 1. \end{cases}$$

Então X tem densidade pois F_X é contínua e derivável em todos os pontos da reta exceto em $\{0, 1\}$.

4.6 Alguns Exemplos de Distribuições de Probabilidade

Vamos agora explorar alguns exemplos importantes de variáveis aleatórias.

4.6.1 Aleatória ou Uniforme Discreta.

Dizemos que X tem uma distribuição *aleatória* com parâmetro n, onde n é um número inteiro positivo, se $X(w) \in \{x_1, x_2, \ldots, x_n\}$ e $p(x_i) = \frac{1}{n}$, para $i \in \{1, \ldots, n\}$.

A função de probabilidade aleatória pode ser utilizada sempre que os possíveis valores da variável aleatória forem eqüiprováveis, como é o caso de modelar mecanismos de jogos (por exemplo, dados e moedas balanceados, cartas bem embaralhadas). Utilizando a propriedade de aditividade da probabilidade, é fácil ver que para qualquer evento $A \subseteq \{x_1, x_2, \ldots, x_n\}$, temos que $P(X \in A) = \frac{||A||}{n}$.

4.6.2 Bernoulli.

Dizemos que X tem uma distribuição *Bernoulli* com parâmetro p, onde $0 \leq p \leq 1$, se $X(w) \in \{x_0, x_1\}$ e $p(x_1) = p = 1 - p(x_0)$.

A função de probabilidade Bernoulli pode ser utilizada para modelar a probabilidade de sucesso em uma única realização de um experimento. Em geral, qualquer variável aleatória dicotômica, ou seja que assume somente dois valores, pode ser modelada por uma distribuição Bernoulli. Denomina-se de *ensaio de Bernoulli*, qualquer experimento que tem uma resposta dicotômica. Um exemplo clássico de um ensaio Bernoulli é o lançamento de uma moeda não necessariamente balanceada.

4.6.3 Binomial.

Dizemos que X tem uma distribuição *Binomial* com parâmetros n e p, onde n é um número inteiro e $0 \leq p \leq 1$, se $X(w) \in \{0, 1, \ldots, n\}$ e $p(k) = \binom{n}{k} p^k (1-p)^{1-k}$, para $k \in \{0, 1, \ldots, n\}$.

Note que utilizando o Teorema Binomial, temos que

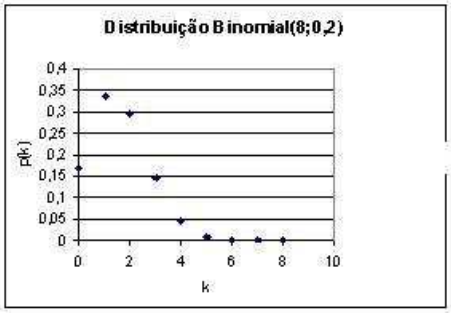
$$\sum_{k=0}^{n} p(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + 1 - p)^n = 1.$$

Logo, esta é uma legítima função probabilidade de massa.

Uma distribuição binomial pode ser obtida quando se considera n repetições independentes de ensaios Bernoulli, e estamos interessados no total de vezes que nesses ensaios obtivemos valor x_1 para a variável. A função de probabilidade binomial pode ser utilizada para modelar a quantidade de erros em um texto de n símbolos quando os erros entre símbolos são assumidos independentes e a probabilidade de erro em um símbolo do texto é igual a p. Também pode ser utilizada para modelar o número de caras em n lançamentos de uma moeda que possui probabilidade p de cair cara em cada lançamento. Se $p = 1/2$, temos um modelo para o número de 1's em uma seqüência binária de comprimento n escolhida aleatoriamente ou o número de caras em n lançamentos de uma moeda justa. A Figura 4.6.3 nos mostra a função probabilidade de massa da Binomial(8; 0,2).

Podemos examinar a função probabilidade de massa binomial analiticamente para encontrarmos seu valor mais provável. Note que a razão entre as probabilidades de dois valores consecutivos da binomial

$$\frac{p(k)}{p(k-1)} = \frac{\frac{n!}{(k)!(n-k)!} p^k (1-p)^{n-k}}{\frac{n!}{(k-1)!(n-k+1)!} p^{k-1} (1-p)^{n-k+1}} = \frac{n-k+1}{k} \frac{p}{1-p}$$



é estritamente decrescente em k. Portanto, se

$$\frac{p(1)}{p(0)} = \frac{np}{1-p} < 1,$$

então as probabilidades são sempre decrescentes em k, e o valor mais provável é 0. No outro extremo, se

$$\frac{p(n)}{p(n-1)} = \frac{p}{n(1-p)} > 1,$$

então as probabilidades são estritamente crescentes em k, e o valor mais provável é n. Se $\frac{1}{n} < \frac{p}{1-p} < n$, então a função começa crescendo em k, enquanto $\frac{n-k+1}{k}\frac{p}{1-p} > 1$, e depois decresce em k. Portanto, se $\frac{p(k)}{p(k-1)} = \frac{n-k+1}{k}\frac{p}{1-p} = 1$ para algum valor de k, temos que k e $k-1$ são os valores mais prováveis. Caso contrário, o valor mais provável será o maior valor de k para o qual $\frac{p(k)}{p(k-1)} = \frac{n-k+1}{k}\frac{p}{1-p} > 1$, isto é, o valor mais provável será o maior valor de k tal que $k < (n+1)p$. No exemplo da Figura 4.6.3, observe que o valor mais provável é para $k = 1$, pois $(n+1)p = 1{,}8$.

Exemplo 4.6.1: Uma moeda com probabilidade 0,4 de cair cara é jogada 5 vezes, qual a probabilidade de se obter exatamente 2 coroas?

Solução: Seja X o número de caras obtidos. Como jogamos a moeda 5 vezes, o evento obter exatamente 2 coroas é igual ao evento obter exatamente 3 caras. Portanto, $P(X = 3) = \binom{5}{3}(0,4)^3(0,6)^2$.

Exemplo 4.6.2: A taxa de sucesso de um bit em uma transmissão digital é 90%. Se 20 bits forem transmitidos, qual a probabilidade de que exatamente 15 deles tenha sido transmitidos com sucesso? Qual a probabilidade de que no máximo 18 deles tenham sido transmitidos com sucesso?

Exemplo 4.6.3: Suponha que para uma dada moeda viciada a probabilidade de que ocorram 3 caras seja igual a probabilidade que ocorram 4 caras se esta moeda for jogada 8 vezes de forma independente. Determine a probabilidade de ocorrerem 3 caras em 8 lançamentos independentes desta moeda.

4.6.4 Uniforme.

Dizemos que X tem uma distribuição *uniforme* com parâmetros a e b, onde a e b são números reais e $a < b$, se a função densidade de X é igual a

$$f_X(x) = \frac{1}{b-a} U(x-a)U(b-x).$$

Este modelo é freqüentemente usado impropriamente para representar "completa ignorância" sobre valores de um parâmetro aleatório sobre o qual apenas sabe-se estar no intervalo finito $[a, b]$. Esta distribuição também é freqüentemente utilizada para, modelar a fase de osciladores e fase de sinais recebidos em comunicações incoerentes. Ela também serve para modelar a escolha de um número aleatório entre a e b.

Neste caso, a função de distribuição acumulada é dada por:

$$F_X(x) = \int_a^x \frac{1}{b-a} dt = \begin{cases} 0 & \text{se } x < a, \\ \frac{x-a}{b-a} & \text{se } a \leq x < b, \\ 1 & \text{se } x \geq b. \end{cases}$$

Exemplo 4.6.4: Sabe-se que é igualmente provável que um dado cliente possa requisitar um serviço no tempo disponível de serviço $[t_0, t_1]$. Se o tempo necessário para executar este serviço é igual a $\tau < t_1 - t_0$, qual a probabilidade que o serviço será executado antes do término do intervalo de tempo disponível de serviço?

Solução: Para que o serviço seja executado em tempo hábil, é necessário que o cliente o requisite antes do tempo $t_1 - \tau$. Logo, $P(X \leq t_1 - \tau) = \frac{1}{t_1 - t_0} \int_{t_0}^{t_1-\tau} dt = \frac{t_1-\tau-t_0}{t_1-t_0}$.

4.7 Variáveis Aleatórias Mistas

Até agora nos restringimos ao estudo de variáveis discretas ou contínuas. No entanto, existem situações práticas, onde a variável aleatória pode tanto assumir valores discretos $x_1, x_2, \ldots$ com probabilidade positiva, como também assumir todos os valores em um determinado intervalo. Tais variáveis são conhecidas como variáveis aleatórias do tipo *misto*. A função de distribuição de uma variável deste tipo é igual a soma de uma função de distribuição de uma variável discreta e de uma função de distribuição de uma variável contínua. Isto é neste caso temos:

$$F_X(x) = \int_{-\infty}^{x} f(x)dx + \sum_{x_i \leq x} p(x_i),$$

onde $p(x_i) \geq 0$, $\sum_{x_i} p(x_i) = p < 1$, $f(x) \geq 0$, e $\int_{-\infty}^{\infty} f(x)dx = 1 - p$.

Um exemplo prático de uma situação onde deve-se usar uma variável aleatória do tipo misto é o caso do tempo de funcionamento de um determinado equipamento. Podem surgir situações em que existe uma probabilidade positiva que o equipamento nunca funcione, isto é, $P(X = 0) = p(0) > 0$ e $P(X > 0) = 1 - p(0)$ e teríamos uma função densidade de probabilidade para descrever a distribuição para valores estritamente positivos de X. Por exemplo, se $f(x) = 0$ quando $x \leq 0$ e $f(x) = (1 - p(0))e^{-x}$, teríamos

$$F_X(x) = \begin{cases} 0 & \text{se } x < 0, \\ p(0) + (1 - p(0)) \int_0^x e^{-t} dt & \text{se } x \geq 0. \end{cases}$$

Note que esta função de distribuição não é nem contínua nem é uma função degrau.

4.8 Variáveis Aleatórias Multidimensionais

Muitas vezes estamos interessados na descrição probabilística de mais de um característico numérico de um experimento aleatório. Por exemplo, podemos estar interessados na distribuição de alturas e pesos de indivíduos de uma certa classe. Para tanto precisamos estender a definição de variável aleatória para o caso multidimensional.

Definição 4.8.1: Seja $(\Omega, \mathcal{A}, P)$ um espaço de probabilidade. Uma função $\vec{X} : \Omega \to R^n$ é chamada de um vetor aleatório se para todo evento B Boreliano de $\mathbb{R}^n$, $\vec{X}^{-1}(B) \in \mathcal{A}$.

Onde um evento é Boreliano em $\mathbb{R}^n$ se pertence a menor σ-álgebra que contem todas regiões da seguinte forma: $C_{\vec{a}} = \{(X_1, X_2, \ldots, X_n) : X_i \leq a_i, 1 \leq i \leq n\}$.

Dado um vetor aleatório $\vec{X}$, pode-se definir uma probabilidade induzida $P_{\vec{X}}$ no espaço mensurável $(\mathbb{R}^n, \mathcal{B}^n)$ da seguinte maneira: para todo $A \in \mathcal{B}^n$, definimos $P_{\vec{X}}(A) = P(\vec{X}^{-1}(A))$. Por definição de vetor aleatório, tem-se que $\vec{X}^{-1}(A) \in \mathcal{A}$, então $P_{\vec{X}}$ está bem definida.

4.8.1 Função de Distribuição Acumulada Conjunta

Para um vetor aleatório $\vec{X}$, uma maneira simples e básica de descrever a probabilidade induzida $P_{\vec{X}}$ é utilizando sua *função de distribuição acumulada conjunta.*

Definição 4.8.2: A função de distribuição acumulada conjunta de um vetor aleatório $\vec{X}$, representada por $F_{\vec{X}}$ ou simplesmente por F, é definida por

$$F_{\vec{X}}(\vec{x}) = P(C_{\vec{x}}) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n), \forall \vec{x} \in \mathbb{R}^n.$$

A função de distribuição acumulada $F_{\vec{X}}$ satisfaz as seguintes propriedades:

F1. Se $x_i \leq y_i$, $\forall i \leq n$, então $F_{\vec{X}}(\vec{x}) \leq F_{\vec{X}}(\vec{y})$.

$$x_i \leq y_i \forall i \leq n \Rightarrow C_{\vec{x}} \subseteq C_{\vec{y}} \Rightarrow P(C_{\vec{x}}) \leq P(C_{\vec{y}}) \Rightarrow F_{\vec{X}}(\vec{x}) \leq F_{\vec{X}}(\vec{y}).$$

F2. Se para algum $i \leq n$ $x_i \to -\infty$, então $C_{\vec{x}}$ decresce monotonicamente para o conjunto vazio $\emptyset$. Logo, pela continuidade monotônica de probabilidade, temos que

$$\lim_{x_i \to -\infty} F_{\vec{X}}(\vec{x}) = 0.$$

F3. Se $x_i \to \infty$, então $C_{\vec{x}}$ cresce monotonicamente para o conjunto $\{X_1 \leq x_1, \ldots X_{i-1} \leq x_{i-1}, X_{i+1} \leq x_{i+1}, \ldots, X_n \leq x_n\}$, ou seja a restrição em X_i é removida. Então, podemos escrever

$$\lim_{x_i \to \infty} F_{\vec{X}}(\vec{x}) = F_{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

Portanto, a função de distribuição acumulada conjunta de $X_1, \ldots, X_{n-1}$ pode ser facilmente determinada da função de distribuição acumulada conjunta de $X_1, \ldots, X_n$ fazendo $x_n \to \infty$. Observe que *funções de distribuição acumuladas conjuntas de ordem maiores determinam as de ordem menores, mas o contrário não é verdadeiro.* Em particular, temos que

$$\lim_{\vec{x} \to \infty} F_{\vec{X}}(\vec{x}) = 1.$$

A função de distribuição acumulada de X_i que se obtém a partir da função acumulada conjunta de $X_1, \ldots, X_n$ fazendo $x_j \to \infty$ para $j \neq i$ é conhecida como *função de distribuição marginal* de X_i.

O próximo exemplo mostra que para $n \geq 2$ as propriedades F1, F2, e F3 não são suficientes para que F seja uma função de distribuição.

Exemplo 4.8.3: Seja $F_0 : I\!R^2 \to I\!R$ uma função definida no plano tal que $F_0(x, y) = 1$ se $x \geq 0$, $y \geq 0$, e $x + y \geq 1$, e $F_0(x, y) = 0$, caso contrário. É claro que F1, F2, e F3 são satisfeitas, mas F_0 não é função de distribuição de nenhum vetor aleatório (X, Y). Se fosse, teríamos uma contradição

$$\begin{aligned} &0 \leq P(0 < X \leq 1, 0 < Y \leq 1) \\ &= F_0(1,1) - F_0(1,0) - F_0(0,1) + F_0(0,0) = 1 - 1 - 1 + 0 = -1 \end{aligned}$$

Os tipos discretos e contínuos de variáveis aleatórias têm os seguintes análogos no caso multivariado. (a) Se $\vec{X}$ for um vetor aleatório discreto, ou seja assumir um número enumerável de valores $\{\vec{x}_1, \vec{x}_2 \ldots, \}$, podemos definir uma função de probabilidade de massa conjunta, p tal que

- $p(\vec{x}_i) \geq 0$.
- $\sum_{i=1}^{\infty} p(\vec{x}_i) = 1$.

Neste caso, pode-se definir a *função probabilidade de massa marginal* de X_i como sendo

$$p_{X_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

(b) Seja $\vec{X} = (X_1, \ldots, X_n)$ um vetor aleatório e F sua função de distribuição. Se existe uma função $f(x_1, \ldots, x_n) \geq 0$ tal que

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \cdots \int_{-\infty}^{x_1} f(t_1, \ldots, t_n) dt_1 \ldots dt_n, \forall (x_1, \ldots, x_n) \in I\!R^n,$$

então f é chamada de densidade conjunta das variáveis aleatórias $X_1, \ldots, X_n$, e neste caso, dizemos que $\vec{X}$ é (absolutamente) contínuo. Neste caso, define-se a *densidade marginal* de X_i como sendo

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n.$$

Exemplo 4.8.4: Duas linhas de produção fabricam um certo tipo de peça. Suponha que a capacidade em qualquer dia seja 4 peças na linha 1 e 3 peças na linha 2. Admita que o número de peças realmente produzida em uma dada linha em um dado dia seja uma variável aleatória. Sejam X e Y o número de peças produzido pela linha 1 e 2 em um dado dia, respectivamente. A tabela a seguir dá a distribuição conjunta de (X, Y):

		Y			
		0	1	2	3
	0	0	0	0,1	0,2
	1	0,2	0	0	0,1
X	2	0	0,1	0,1	0
	3	0	0,1	0	0
	4	0	0	0,1	0

(a) Determine a probabilidade que mais peças sejam produzidas pela linha 2.

(b) Determine as funções probabilidade de massa marginais de X e Y.

Exemplo 4.8.5: Suponha que um vetor aleatório bidimensional (X, Y) tenha densidade conjunta dada por:

$$f_{X,Y}(x, y) = \begin{cases} x^2 + \frac{xy}{3} & \text{se } 0 \leq x \leq 1 \text{ e } 0 \leq y \leq 2, \\ 0 & \text{, caso contrário.} \end{cases}$$

(a) Determine a probabilidade que $Y - X > 0$.

(b) Determine as densidades marginais de X e Y.

Solução: Para a parte (a), note que

$$\begin{aligned} P(Y - X > 0) &= \int_0^1 \int_x^2 x^2 + \frac{xy}{3} dy dx \qquad (4.1) \\ &= \int_0^1 x^2(2 - x) + \frac{x}{3}(\frac{2^2}{2} - \frac{x^2}{2}) dx = \int_0^1 (\frac{-7x^3}{6} + 2x^2 + \frac{2x}{3}) dx \\ &= (\frac{-7x^4}{24} + 2\frac{x^3}{3} + \frac{x^2}{3})|_0^1 = \frac{17}{24}. \end{aligned}$$

Para a parte (b), temos que a densidade marginal de X é:

$$f_X(x) = \int_0^2 x^2 + \frac{xy}{3} dy = 2x^2 + \frac{2x}{3},$$

para $0 \leq x \leq 1$, $f_X(x) = 0$, caso contrário. E a densidade marginal de Y é

$$f_Y(y) = \int_0^1 x^2 + \frac{xy}{3} dx = \frac{1}{3} + \frac{y}{6},$$

para $0 \leq y \leq 2$, $f_Y(y) = 0$, caso contrário.

4.8.2 Distribuição condicional de X dada Y discreta

Seja X uma variável aleatória no espaço de probabilidade $(\Omega, \mathcal{A}, P)$, e seja A um evento aleatório tal que $P(A) > 0$. Usando o conceito de probabilidade condicional, podemos definir a distribuição condicional de X dado o evento A por

$$P(X \in B|A) = \frac{P([X \in B] \cap A)}{P(A)},$$

para B boreliano. Pode-se verificar facilmente que isto define uma probabilidade nos borelianos verificando-se os axiomas de Kolmogorov. Podemos interpretar a distribuição condicional de X dado A como a nova distribuição que se atribui a X quando sabe-se da ocorrência do evento A. A função de distribuição associada à distribuição condicional é chamada função distribuição condicional de X dado A:

$$F_X(x|A) = P(X \leq x|A).$$

Agora suponhamos que os eventos aleatórios $A_1, A_2, \ldots$ formem uma partição (finita ou enumerável) de Ω. Pelo Teorema da Probabilidade Total, temos

$$P(X \in B) = \sum_n P(A_n)P(X \in B|A_n), \forall B \in \mathcal{B},$$

e

$$F_X(x) = P(X \leq x) = \sum_n P(A_n)P(X \leq x|A_n)$$
$$= \sum_n P(A_n)F_X(x|A_n), \forall x.$$

Em outras palavras, a distribuição de X (resp., função de distribuição) é uma média ponderada da distribuição condicional (resp., função de distribuição condicional) de X dado A_n, onde os pesos são as probabilidades dos membros A_n da partição.

Consideremos agora o caso em que a partição do espaço amostral é gerada por uma variável aleatória discreta. Para tanto, seja Y uma variável aleatória discreta em $(\Omega, \mathcal{A}, P)$, tomando somente os valores $y_1, y_2, \ldots$. Então, os eventos $A_n = [Y = y_n]$ formam uma partição de Ω. Neste caso, a distribuição

$$P(X \in B|Y = y_n) = P(X \in B|A_n),$$

para B boreliano, é chamada de distribuição condicional de X dado que $Y = y_n$, e valem as fórmulas

$$P(X \in B) = \sum_n P(Y = y_n)P(X \in B|Y = y_n),\ B \text{ boreliano}$$
$$F_X(x) = \sum_n P(Y = y_n)F_X(x|Y = y_n).$$

Exemplo 4.8.6: A tabela abaixo dá a distribuição conjunta de X e Y.

		Y		
		0	1	2
	0	0,1	0,1	0,1
X	1	0,2	0	0,3
	2	0	0,1	0,1

(a) Determinar as distribuições marginais de X e Y.

(b) Calcule $P(X = 0|Y = 1)$ e $P(Y = 3|X = 2)$.

(c) Calcule $P(X \leq 2)$ e $P(X \leq 1, Y = 2)$.

4.8.3 Distribuição condicional de X dada Y contínua

Quando temos variáveis aleatórias contínuas, freqüentemente estamos em uma situação onde queremos condicionar em um evento que tem probabilidade zero. Por exemplo, poderemos estar interessados em saber qual a probabilidade de que a altura de um dado indivíduo seja menor ou igual a h sabendo que seu peso é igual a k, ou seja, queremos determinar $P(H \leq h|P = k)$. O problema é que a probabilidade do evento condicionante $[P = k]$ é nula. A definição formal de como calcular estas probabilidades no caso geral envolve conceitos complexos de Teoria da Medida. Mas na maioria dos casos práticos, podemos prosseguir com o seguinte procedimento. Se um evento B tem probabilidade zero, então aproximamos B por uma coleção de eventos $\{B_\delta, \delta > 0\}$, $P(B_\delta) > 0$, $B \subset B_\delta$, e $\cap_{\delta>0} B_\delta = B$. Deste modo, pode-se definir $P(A|B)$ por $\lim_{\delta\to 0} P(A|B_\delta)$, desde que este limite exista e independa da coleção de eventos $\{B_\delta\}$.

Por exemplo, suponha que (X, Y) é um vetor aleatório com densidade conjunta dada por $f_{X,Y}(x, y)$. Suponha que estejamos interessados em obter o valor de $P(X \leq x|Y = y)$. Então, utilizando nosso procedimento descrito acima podemos aproximar esta probabilidade por $\lim_{\delta\to 0} P(X \leq x|y - \delta < Y \leq y + \delta)$. Mas,

$$P(X \leq x|y - \delta < Y \leq y + \delta) = \frac{P(X \leq x, y - \delta < Y \leq y + \delta)}{P(y - \delta < Y \leq y + \delta)} \tag{4.2}$$
$$P(X \leq x, y - \delta < Y \leq y + \delta) = \int_{-\infty}^{x}\int_{y-\delta}^{y+\delta} f_{X,Y}(s, t)dtds.$$

Para δ pequeno o suficiente, assumindo que $f_{X,Y}$ é contínua no ponto (x, y), então é aproximadamente constante no intervalo $(y - \delta, y + \delta)$

$$P(X \leq x, y - \delta < Y \leq y + \delta) \approx 2\delta \int_{-\infty}^{x} f_{X,Y}(s, y)ds.$$

Similarmente,

$$P(y - \delta < Y \leq y + \delta) \approx 2\delta f_Y(y).$$

Portanto,

$$P(X \leq x|Y = y) = \int_{-\infty}^{x} \frac{f_{X,Y}(s, y)}{f_Y(y)}ds,$$

desde que $f_Y(y) > 0$.

Define-se a função de distribuição acumulada condicional de X dado Y, $F_{X|Y}(x|y)$, por $P(X \leq x|Y = y)$. Deste modo, o resultado acima nos permite afirmar que neste caso a densidade condicional, $f_{X|Y}(x|y)$, de X dado Y é dada por $\frac{f_{X,Y}(x,y)}{f_Y(y)}$, desde que $f_Y(y) > 0$ e y não seja ponto de descontinuidade de f_Y. Nos casos em que y é um zero ou um ponto de descontinuidade de $f_Y(y)$, adotaremos a convenção que a densidade condicional é igual a zero nestes pontos.

Exemplo 4.8.7: Considere novamente o vetor aleatório bidimensional (X, Y) que tem densidade conjunta dada por:

$$f_{X,Y}(x, y) = \begin{cases} x^2 + \frac{xy}{3} & \text{se } 0 \leq x \leq 1 \text{ e } 0 \leq y \leq 2, \\ 0 & \text{, caso contrário.} \end{cases}$$

(a) Determine as densidades condicionais $g(x|y)$ e $h(y|x)$.

Solução: Já vimos que $f_X(x) = \int_0^2 x^2 + \frac{xy}{3}dy = 2x^2 + \frac{2x}{3}$, para $0 \leq x \leq 1$, $f_X(x) = 0$, caso contrário; e que $f_Y(y) = \int_0^1 x^2 + \frac{xy}{3}dx = \frac{1}{3} + \frac{y}{6}$, para $0 \leq y \leq 2$, $f_Y(y) = 0$, caso contrário. Portanto, aplicando nosso resultado anterior, temos para $0 \leq y \leq 2$:

$$g(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \begin{cases} \frac{x^2 + \frac{xy}{3}}{\frac{1}{3} + \frac{y}{6}} & \text{se } 0 \leq x \leq 1, \\ 0 & \text{, caso contrário.} \end{cases}$$

Similarmente, para $0 \leq x \leq 1$:

$$h(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \begin{cases} \frac{x^2 + \frac{xy}{3}}{2x^2 + \frac{2x}{3}} & \text{se } 0 \leq y \leq 2, \\ 0 & \text{, caso contrário.} \end{cases}$$

4.8.4 Independência entre Variáveis Aleatórias.

Sejam $X_1, X_2, \ldots, X_n$ variáveis aleatórias definidas no mesmo espaço de probabilidade $(\Omega, \mathcal{A}, P)$. Informalmente, as variáveis aleatórias X_i's são independentes se, e somente se, quaisquer eventos determinados por qualquer grupo de variáveis aleatórias distintas são independentes. Por exemplo, $[X_1 < 5]$, $[X_2 > 9]$, e $0 < X_5 \leq 3$ são independentes. Formalmente,

Definição 4.8.8: Dizemos que um conjunto de variáveis aleatórias $\{X_1, \ldots, X_n\}$ é mutuamente independente se, e somente se, para quaisquer eventos borelianos $A_1, \ldots, A_n$,

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i).$$

O próximo teorema estabelece três critérios para provar que um conjunto de variáveis aleatórias é mutuamente independente.

Teorema 4.8.9: *As seguintes condições são necessárias e suficientes para testar se um conjunto $\{X_1, \ldots, X_n\}$ de variáveis aleatórias é mutuamente independente:*

(a) $F_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} F_{X_i}(x_i)$.

(b) Se $\vec{X}$ for um vetor aleatório discreto,

$$p_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

(c) Se $\vec{X}$ for um vetor aleatório contínuo,

$$f_{\vec{X}}(\vec{x}) = \prod_{i=1}^{n} f_{X_i}(x_i), \forall (x_1, \ldots, x_n) \in I\!R^n.$$

Prova: Omitida, pois está fora do escopo deste curso. ∎

É fácil observar que utilizando, a definição de probabilidade condicional que se X e Y são independentes, então para todo A e B boreliano tal que $P(Y \in B) > 0$:

$$P(X \in A | Y \in B) = P(X \in A),$$

ou seja, se X e Y são independentes o conhecimento do valor de Y não altera a descrição probabilística de X.

Exemplo 4.8.10: Verifique se as variáveis aleatórias X e Y do Exemplo 4.8.6 são independentes.

4.9 Funções de Variáveis Aleatórias

Muitas vezes sabemos a distribuição de probabilidade que descreve o comportamento de uma variável aleatória X definida no espaço mensurável $(\Omega, \mathcal{A})$, mas estamos interessados na descrição de uma função $Y = H(X)$. Por exemplo, X pode ser uma mensagem enviada em um canal de telecomunicações e Y ser a mensagem recebida. Nosso problema é determinar $P(Y \in A)$, onde A é um evento Boreliano, dado P_X. Para determinarmos esta

probabilidade, estaremos interessados na imagem inversas a função H, ou seja, a probabilidade do evento $\{Y \in A\}$ será por definição igual a probabilidade do evento $\{X \in H^{-1}(A)\}$, onde $H^{-1}(A) = \{x \in I\!R : H(x) \in A\}$. Para que esta probabilidade esteja bem definida, precisamos restringir H tal que $H^{-1}(A)$ seja um evento boreliano para todo A boreliano, caso contrário não poderemos determinar $P(\{X \in H^{-1}(A)\})$; uma função que satisfaz esta condição é conhecida como *mensurável com respeito a* $\mathcal{A}$ *e* $\mathcal{B}$. Note que Y também pode ser vista como uma função do espaço amostral Ω, $Y(\omega) = H(X(\omega))$ para todo $\omega \in \Omega$. Visto dessa maneira Y é uma variável aleatória definida em $(\Omega, \mathcal{A})$, pois para todo boreliano A $Y^{-1}(A) = X^{-1}(H^{-1}(A))$ e como por suposição $H^{-1}(A)$ é boreliano e X é uma variável aleatória, temos que $X^{-1}(H^{-1}(A)) \in \mathcal{A}$ e portanto satisfaz a definição de uma variável aleatória. *Nesses problemas é sempre útil fazer um esboço do gráfico da transformação* H *para determinarmos quais são as regiões inversas* $H^{-1}(A)$.

Vamos primeiro tratar este problema no caso de variáveis aleatórias discretas. Neste caso para qualquer função H, temos que $Y = H(X)$ é uma variável aleatória discreta.

Suponha que X assuma os valores $x_1, x_2, \ldots$ e seja H uma função real tal que $Y = H(X)$ assuma os valores $y_1, y_2, \ldots$. Vamos agrupar os valores que X assume de acordo os valores de suas imagens quando se aplica a função H, ou seja, denotemos por $x_{i1}, x_{i2}, x_{i3}, \ldots$ os valores de X tal que $H(x_{ij}) = y_i$ para todo j. Então, temos que

$$P(Y = y_i) = P(X \in \{x_{i1}, x_{i2}, x_{i3}, \ldots\}) = \sum_{j=1}^{\infty} P(X = x_{ij}) = \sum_{j=1}^{\infty} p_X(x_{ij}),$$

ou seja, para calcular a probabilidade do evento $\{Y = y_i\}$, acha-se o evento equivalente em termos de X, isto é, todos os valores x_{ij} de X tal que $H(x_{ij}) = y_i$ e somam-se as probabilidades de X assumir cada um desses valores.

Exemplo 4.9.1: Admita-se que X tenha os valores possíveis $1, 2, 3, \ldots$ e suponha que $P(X = n) = (1/2)^n$. Seja $Y = 1$ se X for par e $Y = -1$ se X for ímpar. Então, temos que

$$P(Y = 1) = \sum_{n=1}^{\infty} (1/2)^{2n} = \sum_{n=1}^{\infty} (1/4)^n = \frac{1/4}{1 - 1/4} = 1/3.$$

Conseqüentemente,

$$P(Y = -1) = 1 - P(Y = 1) = 2/3.$$

Podemos estender este resultado para uma função de um vetor aleatório $\vec{X}$ de forma análoga. Neste caso se $\vec{Y} = H(\vec{X})$, denotemos por $\vec{x}_{i1}, \vec{x}_{i2}, \vec{x}_{i3}, \ldots$ os valores de $\vec{X}$ tal que $H(\vec{x}_{ij}) = \vec{y}_i$ para todo j. Então, temos que

$$P(\vec{Y} = \vec{y}_i) = P(\vec{X} \in \{\vec{x}_{i1}, \vec{x}_{i2}, \vec{x}_{i3}, \ldots\}) = \sum_{j=1}^{\infty} P(\vec{X} = \vec{x}_{ij}) = \sum_{j=1}^{\infty} p_{\vec{X}}(\vec{x}_{ij}),$$

ou seja, para calcular a probabilidade do evento $\{\vec{Y} = \vec{y}_i\}$, acha-se o evento equivalente em termos de $\vec{X}$, isto é, todos os valores $\vec{x}_{ij}$ de $\vec{X}$ tal que $H(\vec{x}_{ij}) = \vec{y}_i$ e somam-se as probabilidades de $\vec{X}$ assumir cada um desses valores.

Vamos ver agora um exemplo no caso em que $\vec{X}$ é contínuo.

Exemplo 4.9.2: Se $X \sim U[0,1]$, qual a distribuição de $Y = -\log(X)$? Como

$$0 < Y < \infty \Leftrightarrow 0 < X < 1$$

e $P(0 < X < 1) = 1$, temos $F_Y(y) = 0, y \leq 0$. Se $y > 0$, então

$$P(Y \leq y) = P(-\log(X) \leq y) = P(X \geq e^{-y}) = 1 - e^{-y},$$

ou seja, $Y \sim Exp(1)$.

Capítulo 5

Esperança e Momentos

5.1 O Conceito de Esperança

O conceito de Esperança ou Valor Esperado de uma variável aleatória X, ou a "média" é tão antigo quanto o próprio conceito de probabilidade. Na verdade, é até possível definir probabilidade em termos de esperança, mas esta não é uma maneira comum de se apresentar a teoria. Existem quatro tipos de interpretações da Esperança:

1. Parâmetro m de uma medida de probabilidade, função de distribuição, ou função probabilidade de massa, também conhecido como média.

2. Um operador linear em um conjunto de variáveis aleatórias que retorna um valor típico da variável aleatória interpretado como uma medida de localização da variável aleatória.

3. média do resultado de repetidos experimentos independentes no longo prazo.

4. preço justo de um jogo com pagamentos descritos por X.

5.1.1 Definição da Esperança - Caso Discreto

Vamos motivar a definição de esperança considerando o cálculo do resultado médio de 1000 lançamentos de um dado. Uma maneira de calcular este resultado médio seria somar todos os resultados e dividir por 1000. Uma maneira alternativa seria calcular a fração $p(k)$ de todos os lançamentos que tiveram resultado igual a k e calcular o resultado médio através da soma ponderada:

$$1p(1) + 2p(2) + 3p(3) + 4p(4) + 5p(5) + 6p(6).$$

Quando o número de lançamentos se torna grande as frações de ocorrência dos resultados tendem a probabilidade de cada resultado. Portanto, em geral definimos a esperança de uma variável discreta como uma soma ponderada onde as probabilidades são os pesos de ponderação.

Definição 5.1.1: Se X é uma variável aleatória discreta assumindo valores $\{x_1, x_2, x_3, \ldots\}$ com probabilidade $\{p_1, p_2, p_3, \ldots\}$, respectivamente, então sua esperança é dada pela fórmula

$$EX = \sum_{i:x_i<0} x_i p_i + \sum_{i:x_i\geq 0} x_i p_i,$$

desde que pelo menos um dos somatórios seja finito. Em caso os dois somatórios não sejam finitos, a esperança não existe. Caso EX seja finita, diz-se que X é integrável.

Exemplo 5.1.2: Considere uma variável aleatória X tal que: $P(X = -1) = 0.25$, $P(X = 0) = 0.5$ e $P(X = 2) = 0.25$. Então,

$$EX = -1(0.25) + 0(0.5) + 2(0.25) = 0.25.$$

Exemplo 5.1.3: Considere uma variável aleatória X tal que: $P(X = -a) = P(X = a) = 1/2$. Então,

$$EX = -a(0.5) + a(0.5) = 0.$$

Note então que muitas variáveis aleatórias diferentes podem ter o mesmo valor esperado ou esperança. (É só variar o valor de a no exemplo anterior.)

Exemplo 5.1.4: Aleatória. Se $X \in \{1, 2, \ldots, n\}$ for uma variável aleatória com distribuição de probabilidade aleatória com parâmetro n, temos que sua esperança é dada por:

$$EX = \sum_{k=1}^{n} kp(k) = \sum_{k}^{n} k\frac{1}{n} = \frac{1}{n}\sum_{k}^{n} k = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}.$$

Onde utilizamos a fórmula da soma dos primeiros n termos de uma progressão aritmética. Em geral, se X for uma variável aleatória com distribuição de probabilidade aleatória assumindo os valores $\{x_1, x_2, \ldots, x_n\}$, então:

$$EX = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Exemplo 5.1.5: Bernoulli. Se $X \in \{0, 1\}$ for uma variável aleatória com distribuição de probabilidade Bernoulli com parâmetro p, temos que sua esperança é dada por:

$$EX = 0(1-p) + 1(p) = p.$$

Exemplo 5.1.6: Binomial. Se X for uma variável aleatória com distribuição de probabilidade Binomial com parâmetros n e p, temos que sua esperança é dada por:

$$EX = \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=1}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$$\sum_{k=1}^{n} n\frac{(n-1)!}{(k-1)!(n-k)!}p^k(1-p)^{n-k} = np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} = np.$$

Onde utilizamos o Teorema Binomial na última igualdade.

5.1.2 Definição da Esperança - Caso Contínuo

Definição 5.1.7: Se X é uma variável aleatória contínua com função densidade de probabilidade f, então sua esperança é dada pela fórmula

$$EX = \int_{-\infty}^{0} xf(x)dx + \int_{0}^{\infty} xf(x)dx,$$

desde que pelo menos uma das integrais seja finita. Em caso as duas integrais não sejam finitas, a esperança não existe. Caso EX seja finita, diz-se que X é integrável.

Deve-se observar a analogia entre o valor esperado de uma variável aleatória e o conceito de centro de gravidade em Mecânica. Se um objeto tem massa distribuída sobre a reta, em pontos discretos, $x_1, x_2, \ldots$, e se $p(x_i)$ for a massa do ponto x_i, então vemos que $\sum_{i=1}^{\infty} x_i p(x_i)$ representa o centro de gravidade do objeto em relação a origem. Similarmente, se um objeto tem massa distribuída continuamente sobre uma reta, e se $f(x)$ representar a densidade de massa em x, então $\int_{-\infty}^{\infty} xf(x)dx$ determina o centro de gravidade deste objeto. Então, podemos interpretar a esperança de uma variável aleatória X como sendo o centro da distribuição de probabilidade de X.

Considere o seguinte exemplo:

Exemplo 5.1.8: Uniforme. Se $X \sim U(a, b)$, então X possui densidade igual a $f(x) = \frac{1}{b-a}$ se $x \in (a, b)$, e $f(x) = 0$, caso contrário. Logo, temos que sua esperança é dada por:

$$EX = \int_{a}^{b} \frac{x}{b-a}dx = \frac{a+b}{2}.$$

5.2 Esperança de Funções de Variáveis Aleatórias

Vamos iniciar considerando o caso discreto.

5.2.1 Caso Discreto

Como vimos anteriormente, se X for uma variável aleatória discreta e se $Y = H(X)$, então Y também será uma variável aleatória discreta. Conseqüentemente, pode-se calcular EY. Existem duas maneiras de calcular EY que são equivalentes.

Definição 5.2.1: Seja X uma variável aleatória discreta e seja $Y = H(X)$. Se Y assumir os seguintes valores $y_1, y_2, \ldots$ e se $p(y_i) = P(Y = y_i)$, definimos:

$$EY = \sum_{i=1}^{\infty} y_i p(y_i).$$

Conforme vimos no capítulo anterior podemos determinar as probabilidades $p(y_i)$ dado que sabemos a distribuição de X. No entanto, podemos encontrar EY sem preliminarmente encontrarmos a distribuição de probabilidade de Y, partindo-se apenas do conhecimento da distribuição de probabilidade de X, conforme mostra o seguinte teorema.

Teorema 5.2.2: *Seja X uma variável aleatória discreta assumindo os valores $x_1, x_2, \ldots$ e seja $Y = H(X)$. Se $p(x_i) = P(X = x_i)$, temos*

$$EY = E(H(X)) = \sum_{i=1}^{\infty} H(x_i)p(x_i).$$

Prova: Vamos re-ordenar o somatório $\sum_{i=1}^{\infty} H(x_i)p(x_i)$, agrupando os termos onde x_i tem a mesma imagem de acordo com a função H, ou seja, sejam $x_{i1}, x_{i2}, \ldots$, todos os valores x_i tal que $H(x_{ij}) = y_i$ para $j \geq 1$, onde $y_1, y_2, \ldots$ são os possíveis valores de Y. Desse modo podemos reescrever

$$\sum_{i=1}^{\infty} H(x_i)p(x_i) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} H(x_{ij})p(x_{ij}) = \sum_{i=1}^{\infty} y_i \sum_{j=1}^{\infty} p(x_{ij}) = \sum_{i=1}^{\infty} y_i p(y_i) = EY.$$

∎

Exemplo 5.2.3: Suponha que X é uma variável aleatória tal que $P(K = k) = e^{-\lambda}\frac{\lambda^k}{k!}$, para $k = 0, 1, 2, \ldots$. (Veremos adiante que esta é uma distribuição de Poisson com parâmetro λ.) Seja $Y = X^2$, vamos calcular EY. Utilizando o Teorema 5.2.2, temos

$$EY = \sum_{k=0}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k^2 e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} + \sum_{k=1}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!}$$
$$= \lambda^2 \sum_{k=2}^{\infty} e^{-\lambda}\frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 + \lambda.$$

Também podemos estender este resultado para o caso de uma função real de um vetor aleatório. Neste caso, se $Y = H(\vec{X})$, temos que $EY = \sum_i H(\vec{x_i})p_{\vec{X}}(\vec{x_i})$, onde os $\vec{x_i}$ são os valores assumidos pelo vetor aleatório $\vec{X}$.

5.2.2 Caso Contínuo

No caso de uma variável aleatória contínua X também podemos calcular a esperança de uma função $Y = \varphi(X)$ de maneira análoga.

Teorema 5.2.4: *Seja X uma variável aleatória contínua, $Y = \varphi(X)$ uma outra variável aleatória, então*

$$EY = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_{-\infty}^{\infty} \varphi(x) f_X(x)dx,$$

desde que estas integrais existam.

Prova: Omitida. ∎

Uma fórmula análoga também é válida quando consideramos funções de vetores aleatórios.

Teorema 5.2.5: *Seja $\vec{X} = (X_1, X_2, \ldots, X_n)$ um vetor aleatório contínuo e $Y = \varphi(\vec{X})$ uma variável aleatória. Então,*

$$EY = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \varphi(\vec{x}) f_{\vec{X}}(\vec{x}) dx_1 \cdots dx_n.$$

Exemplo 5.2.6: Suponha que X seja uma variável aleatória com densidade dada por:

$$f(x) = \begin{cases} \frac{e^x}{2} & \text{se } x \leq 0, \\ \frac{e^{-x}}{2} & \text{se } x > 0. \end{cases}$$

Seja $Y = |X|$, vamos determinar EY. Usando o Teorema 5.2.4, teremos

$$\begin{aligned} EY &= \int_{-\infty}^{\infty} |x| f(x) dx \qquad (5.1) \\ &= \frac{1}{2}\left(\int_{-\infty}^{0} -xe^x dx + \int_{0}^{\infty} xe^{-x} dx\right) \\ &= \frac{1}{2}\left(-xe^x|_{-\infty}^{0} + \int_{-\infty}^{0} e^x dx - xe^{-x}|_{0}^{\infty} + \int_{0}^{\infty} e^{-x} dx\right) \\ &= \frac{1}{2}(0 + 1 + 0 + 1) = 1. \end{aligned}$$

Exemplo 5.2.7: Podemos utilizar o valor esperado de uma variável aleatória a fim de tomar uma decisão ótima. Por exemplo, suponha que um fabricante produza certo tipo de equipamento e seja X o número de tais equipamentos que são vendidos por dia. Suponha que X seja uma variável aleatória uniformemente distribuída em $[3, 6]$. Suponha que cada equipamento vendido dê um lucro de R$200,00, enquanto se um dado equipamento não for vendido no mesmo dia como o fabricante não tem onde armazená-lo ele será destruído dando um prejuízo de R$50,00. Suponha que o fabricante deve decidir no dia anterior quantos equipamentos deverá produzir no dia seguinte. Queremos saber quantos equipamentos o fabricante deverá produzir de forma a maximizar o seu lucro esperado. Suponha que o fabricante decida produzir k unidades, então seu lucro será igual a

$$Z = L(X) = \begin{cases} 200k & \text{se } X \geq k, \\ 200X - 50(k - X) & \text{se } X < k. \end{cases}$$

Podemos então calcular o lucro esperado EZ da seguinte forma:

$$EZ = \int_{-\infty}^{\infty} L(x) f_X(x) dx = \frac{1}{3}\int_{3}^{6} L(x) dx.$$

Obviamente, o lucro $L(x)$ depende do valor de k. Se $k \leq 3$, então $L(x) = 200k$, para todo x entre 3 e 6, logo $EZ = \frac{1}{3}\int_3^6 200k dx = 200k$. Se $k \geq 6$, então $L(x) = 250x - 50k$, para todo x entre 3 e 6, logo

$$EZ = \frac{1}{3}\int_{3}^{6} (250x - 50k) dx = \frac{1}{3}(125x^2 - 50kx)|_3^6 = \frac{1}{3}(3375 - 150k).$$

Por fim, se $3 < k < 6$, temos

$$EZ = \frac{1}{3}(\int_3^k (250x - 50k)dx + \int_k^6 (200k)dx) = \frac{1}{3}(125(k^2 - 9) - 50k(k-3) + 200k(6-k))$$
$$= \frac{1}{3}(-125k^2 + 1350k - 1125)$$

Resumindo, temos:

$$EZ = \begin{cases} 200k & \text{se } k \leq 3, \\ \frac{1}{3}(-125k^2 + 1350k - 1125) & \text{se } 3 < k < 6, \\ \frac{1}{3}(3375 - 150k) & \text{se } k \geq 6. \end{cases}$$

Note que para $k \leq 3$, o lucro esperado é crescente em k, e que para $k \geq 6$ o lucro esperado decresce com k. Na região, $3 < k < 6$, temos que o lucro esperado é uma parábola que atinge o máximo para $k = 5{,}4$. Como o número de equipamentos fabricados tem que ser inteiro então o máximo deve ocorrer ou em $k = 5$ ou em $k = 6$, comparando estes valores temos que, se $k = 5$, então $EZ = 833{,}33$; e se $k = 6$, então $EZ = 825$. Portanto, o fabricante deve produzir 5 equipamentos por dia para maximizar seu lucro esperado.

5.3 Propriedades da Esperança

As seguintes propriedades são aplicações imediatas da definição de esperança:

1. $P(X = c) = 1 \Rightarrow EX = c$.

2. $P(X \geq 0) = 1 \Rightarrow EX \geq 0$.

3. $E(aX) = aEX$, onde a um número real qualquer. Esta propriedade segue facilmente da expressão da esperança de uma função de variável aleatória.

4. $E(X + Y) = EX + EY$. Para provar esta propriedade, note que

$$E(X+Y) = \sum_i \sum_j (x_i + y_j)p(x_i, y_j) = \sum_i x_i \sum_j p(x_i, y_j) + \sum_i \sum_j y_j p(x_i, y_j)$$
$$= \sum_i x_i p(x_i) + \sum_j y_j \sum_i p(x_i, y_j) = EX + \sum_j y_j p(y_j) = EX + EY.$$

5. $P(X \geq Y) = 1 \Rightarrow EX \geq EY$. Propriedade 5 segue das propriedades 2, 3, e 4, pois

$$P(X \geq Y) = P(X - Y \geq 0),$$

o que, pela propriedade 2, implica que $E(X - Y) \geq 0$. Pela propriedade 4, temos que $E(X - Y) = EX + E(-Y)$. Finalmente, pela propriedade 3, temos que $E(X - Y) = EX - EY$, ou seja podemos concluir que $EX - EY \geq 0$.

6. Se $\{X_1, \ldots, X_n\}$ são variáveis aleatórias mutuamente independentes, então $E(\prod_{i=1}^n X_i) = \prod_{i=1}^n EX_i$. Para provar esta propriedade note que

$$E(\prod_{i=1}^{n} X_i) = \sum_{i_1} \cdots \sum_{i_n} x_{i_1} \cdots x_{i_n} p(x_{i_1}, \ldots, x_{i_n})$$

$$= \sum_{i_1} \cdots \sum_{i_n} x_{i_1} \cdots x_{i_n} \prod_{j=1}^{n} p(x_{i_j}) = \sum_{i_1} x_{i_1} p(X_{i_1}) \cdots \sum_{i_n} x_{i_n} p(x_{i_n}) = \prod_{i=1}^{n} EX_i.$$

7. Se X tem uma distribuição simétrica em torno de a, ou seja, $P(X-a \geq x) = P(X-a \leq -x)$, e se a esperança de X tiver bem definida, então $EX = a$. Para provar esta expressão, primeiro note que se X é simétrica em relação a a então $Y = X - a$ é simétrica em relação a zero. Se provarmos que $EY = 0$, então segue da linearidade da esperança que $EX = a$. No caso discreto, como Y é simétrica em torno de 0, temos que $p(x_i) = p(-x_i)$ para todo x_i, portanto segue que $EY = \sum_i x_i p(x_i) = 0$. No caso contínuo, como Y é simétrica em torno de 0, temos que $P(Y \geq y) = P(Y \leq -y)$, o que implica que $1 - F_Y(y) = F_Y(-y)$. Finalmente, derivando obtemos $f_Y(y) = f_Y(-y)$, ou seja, Y possui densidade par, logo $EY = \int_{-\infty}^{\infty} y f_Y(y) dy = 0$, pois é a integral de uma função ímpar $y f_Y(y)$ em torno de um intervalo simétrico em torno de zero.[1]

Pode-se definir outras medidas de posição de uma variável aleatória, tais como: *mediana* e *moda*. A *mediana de uma v.a.* X é qualquer número m tal que $P(X \geq m) \geq 0{,}5$ e $P(X \leq m) \geq 0{,}5$. Por exemplo, se X assume os valores $-1, 0, 1$ com probabilidades $1/4, 1/4, 1/2$, respectivamente, então qualquer número no intervalo fechado de 0 a 1. A *moda* de uma variável aleatória discreta é o seu valor mais provável. Como para uma variável aleatória contínua todos os valores tem probabilidade zero, define-se como moda neste caso, o valor que maximiza a função densidade de probabilidade. A moda não é necessariamente única, pois a função probabilidade de massa (caso discreto) ou a função densidade de probabilidade (caso contínuo) pode atingir seu máximo em vários valores $x_1, x_2, \ldots$.

Quando uma função densidade de probabilidade tem múltiplos máximos locais, é comum se referir a todos os máximos locais como modas da distribuição (mesmo que a definição formal implique que apenas o máximo global é uma moda da distribuição). Tais distribuições contínuas são chamadas de multimodais (em oposição a unimodal).

Em distribuições unimodais simétricas, isto é, distribuições tal que existe um número m tal que $P(X - m \geq x) = P(X - m \leq -x)$ para todo $x \in \mathbb{R}$, a esperança (se bem definida), mediana, e moda coincidem e são iguais a m.

5.4 Momentos

Momentos dão informações parciais sobre a medida de probabilidade P, a função de distribuição acumulada, ou a função probabilidade de massa de uma variável aleatória discreta

[1] Como assumimos que EX é bem definida, segue da linearidade que $EY = EX - a$ também é bem definida, donde concluímos que pelo menos uma das integrais $\int_{-\infty}^{0} y f_Y(y) dy$ ou $\int_{0}^{\infty} y f_Y(y) dy$ é finita, e como $f_Y(y)$ é par estas integrais tem mesmo módulo mas sinais contrários, o que nos permite afirmar que a integral sobre toda reta é nula.

X. Momentos de X são esperanças de potências de X.

Definição 5.4.1: Para qualquer inteiro não-negativo n, o n-ésimo momento da variável aleatória X é EX^n, se esta esperança existe.

Na seção anterior, vimos que o segundo momento de uma variável aleatória Poisson com parâmetro λ é dado por: $\lambda^2 + \lambda$. Vamos agora calcular o segundo momento de uma variável aleatória X Binomial com parâmetros n e p:

$$EX^2 = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k^2 \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} =$$
$$\sum_{k=1}^{n} k(k-1) \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} + \sum_{k=1}^{n} k \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$
$$n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2} (1-p)^{n-k} + np$$
$$= n(n-1)p^2 \sum_{j=0}^{m} \frac{(m)!}{(j)!(m-j)!} p^j (1-p)^{m-j} + np = n(n-1)p^2 + np.$$

Pode-se provar que *momentos de ordem superiores finitos implicam momentos de ordem inferiores finitos.*

5.4.1 Momentos Centrais

Definição 5.4.2: Se X é uma variável aleatória seu n-ésimo momento central é: $E(X - EX)^n$, se esta esperança existir.

Note que o primeiro momento central é zero, pois $E(X - EX) = EX - EEX = EX - EX = 0$. O segundo momento central é conhecido como *variância* e denota-se por $VarX$. A variância pode ser também calculada por:

$$VarX = E(X - EX)^2 = E(X^2 - 2XEX + (EX)^2) = EX^2 - 2E(XEX) + E((EX)^2)$$
$$= EX^2 - 2(EX)^2 + (EX)^2 = EX^2 - (EX)^2. \quad (5.2)$$

Do Teorema Binomial e da linearidade da esperança, temos

$$E(X - EX)^n = \sum_{k=0}^{n} \binom{n}{k} (-EX)^{n-k} EX^k$$

e

$$EX^n = E(X - EX + EX)^n = \sum_{k=0}^{n} \binom{n}{k} (EX)^{n-k} E(X - EX)^k.$$

Como um corolário, temos que o n-ésimo momento central existe se, e somente se, o n-ésimo momento existe.

Exemplo 5.4.3: Considere uma variável aleatória X tal que

$$P(X = m - a) = P(X = m + a) = \frac{1}{2} \Rightarrow EX^k = \frac{1}{2}[(m-a)^k + (m+a)^k].$$

$$EX = m, EX^2 = \frac{1}{2}[2m^2 + 2a^2] = m^2 + a^2, VarX = a^2.$$

Este exemplo, mostra que podemos encontrar uma variável aleatória bem simples possuindo qualquer esperança e variância predeterminadas.

Exemplo 5.4.4: (**Aleatória ou Uniforme Discreta.**) Se X tem uma distribuição uniforme discreta assumindo os valores $\{x_1, x_2, \ldots, x_n\}$ com mesma probabilidade, então:

$$VarX = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n^2}(\sum_{i=1}^{n} x_i)^2.$$

Exemplo 5.4.5: (**Binomial.**) Já demonstramos que se X tem uma distribuição binomial, então $EX = np$ e $E(X^2) = n(n-1)p^2 + np$. Portanto, $VarX = n(n-1)p^2 + np - n^2p^2 = np(1-p)$.

Exemplo 5.4.6: (**Uniforme Contínua.**) Se X tem uma distribuição uniforme em [a,b], então

$$EX^2 = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^3 - a^3}{3(b-a)}$$

e

$$(EX)^2 = (\frac{a+b}{2})^2.$$

Portanto, $VarX = \frac{b^3-a^3}{3(b-a)} - (\frac{a+b}{2})^2 = \frac{(b-a)^2}{12}$.

O *desvio-padrão* σ de uma variável aleatória X é definido como a raiz quadrada da variância, $\sigma(X) = \sqrt{VarX}$.

Propriedades da Variância

As seguintes propriedades da variância são conseqüências imediatas de sua definição.

1. $VarX \geq 0$.
2. Se $X = c$, $Var(X) = 0$.

   Prova: Temos que $EX = c$, logo $Var(X) = E(X-c)^2 = E(0) = 0$. ∎

3. $Var(X + a) = VarX$, onde a é uma constante real.

   Prova:

   $$Var(X+a) = E(X+a)^2 - (E(X+a))^2$$
   $$= EX^2 + 2aEX + a^2 - (EX)^2 - 2aEX - a^2 = EX^2 - (EX)^2 = VarX.$$

   ∎

4. $Var(aX) = a^2 VarX$

   Prova:

   $$Var(aX) = E(aX)^2 - (E(aX))^2 = a^2 EX^2 - a^2 (EX)^2 = a^2 VarX.$$

   ∎

5. Se X e Y forem variáveis aleatórias mutuamente independentes, então $Vax(X+Y) = VarX + VarY$.

   Prova:

   $$\begin{aligned} &Var(X+Y) = E(X+Y)^2 - [E(X+Y)]^2 \\ &= E(X^2 + 2XY + Y^2) - (EX)^2 - 2EXEY - (EY)^2 \\ &= EX^2 + EY^2 - (EX)^2 - (EY)^2 + 2E(XY) - 2EXEY = VarX + VarY \end{aligned}$$

   ∎

6. Se $X_1, \ldots, X_n$ são variáveis aleatórias independentes, então $Var(X_1 + \cdots X_n) = VarX_1 + \cdots + VarX_n$. Esta propriedade segue da propriedade anterior e de uma aplicação de indução matemática.

7. **Desigualdade de Chebyshev Generalizada.** Dado um conjunto A e uma função $g(x)$ tal que $\forall x\ g(x) \geq I_A(x)$, tem-se que $P(X \in A) \leq \min(1, Eg(X))$.

   Prova: Pela monotonicidade da Esperança, temos que $Eg(X) \geq EI_A(X) = P(X \in A)$. Mas, como a cota superior pode exceder 1, temos que $\min(1, Eg(X)) \geq P(X \in A)$.
   ∎

Corolário 5.4.7: *Seja X uma variável aleatória, então para todo $\epsilon > 0$, $P(|X| \geq \epsilon) \leq \frac{E|X|}{\epsilon}$.*

Prova: Escolha $A = \{x : |x| \geq \epsilon\}$ e $g(x) = \frac{|x|}{\epsilon}$. Note que $g(x) \geq I_A(x)$, então $P(|X| \geq \epsilon) \leq \frac{E|X|}{\epsilon}$. ∎

Corolário 5.4.8: *Se $Z \geq 0$ e $EZ = 0$, então $P(Z = 0) = 1$.*

Prova: $P(Z \geq \frac{1}{n}) \leq nEZ = 0$. Como $[Z > 0] = \cup_n [Z \geq \frac{1}{n}]$, temos que

$$P(Z > 0) = P(\cup_n [Z \geq \frac{1}{n}]) \leq \sum_n P(Z \geq \frac{1}{n}) = 0.$$

Portanto, $P(Z = 0) = 1 - P(Z > 0) = 1$. ∎

Note que este último corolário implica que, quando $Var(X) = 0$, ou seja $E(X - EX)^2 = 0$, temos que $P(X = EX) = 1$, ou seja X é constante com probabilidade 1.

Corolário 5.4.9: Desigualdade (Original) de Chebyshev. *Seja X uma variável aleatória, então $P(|X - EX| \geq \epsilon) \leq \frac{VarX}{\epsilon^2}$.*

Prova: Escolha $A = \{x : |x| \geq \epsilon\}$ e $g(x) = \frac{x^2}{\epsilon^2}$. Note que $g(x) \geq I_A(x)$, então pelo teorema anterior, $P(X \in A) = P(|X| \geq \epsilon) \leq \frac{EX^2}{\epsilon^2}$. Substituindo X por $X - EX$, temos $P(|X - EX| \geq \epsilon) \leq \frac{VarX}{\epsilon^2}$. ∎

Note que a desigualdade de Chebyshev converte conhecimento sobre um momento de segunda ordem ou uma variância numa cota superior para a probabilidade da cauda de uma variável aleatória.

8. $VarX = E(X - \mu)^2 = \min_{c \in \mathbb{R}} E(X - c)^2$.

   Prova:

   $$(X - c)^2 = (X - \mu + \mu - c)^2 = (X - \mu)^2 + 2(\mu - c)(X - \mu) + (\mu - c)^2,$$

   logo

   $$\begin{aligned} E(X - c)^2 &= E(X - \mu)^2 + 2(\mu - c)(EX - \mu) + (\mu - c)^2 \\ &= VarX + (\mu - c)^2. \end{aligned}$$

   Portanto, $E(X - c)^2 \geq E(X - \mu)^2, \forall c \in \mathbb{R}$. ∎

5.5 Correlação, Covariância, e Desigualdade de Schwarz

Correlação e covariância são quantidades parecidas com momentos que são medidas do grau de dependência linear entre duas variáveis.

Definição 5.5.1: A correlação entre duas variáveis aleatórias X e Y é dada por EXY se esta esperança existe. A covariância entre elas é dada por $Cov(X, Y) = E[(X - EX)(Y - EY)] = EXY - (EX)(EY)$.

Note que $Cov(X, X) = VarX$. Pela prova da propriedade 5 de variância, vemos que a seguinte relação é válida:

$$Var(X + Y) = VarX + VarY + 2Cov(X, Y).$$

Diz-se que duas varáveis são *não-correlacionadas* se $Cov(X, Y) = 0$. Como já provamos que se X e Y são independentes, então $EXY = EXEY$. Temos que se X e Y são independentes, elas necessariamente são não-correlacionadas. O contrário nem sempre é verdadeiro como o próximo exemplo ilustra.

Exemplo 5.5.2: Se X é uma variável aleatória tal que $P(X = -a) = P(X = a) = 1/2$ e $Y = X^2$, temos que $EXY = -a^3(1/2) + a^3(1/2) = 0$ e $EX = -a(1/2) + a(1/2) = 0$. Logo, $EXY = EXEY = 0$, ou seja, $Cov(X, Y) = 0$. Porém, X e Y não são independentes, pois Y é uma função de X.

O próximo teorema trata de uma importante desigualdade em teoria da probabilidade:

Teorema 5.5.3: $(E(XY))^2 \leq EX^2EY^2$ *e* $(Cov(X,Y))^2 \leq VarXVarY$.

Prova: $(aX+Y)^2 \geq 0 \Rightarrow E(aX+Y)^2 \geq 0 \Rightarrow a^2EX^2 + 2aEXY + EY^2 \geq 0$. Observe que está equação do segundo grau em a não pode ter duas raízes reais diferentes, pois caso contrário essa expressão seria negativa para os valores entre as raízes. Então, utilizando a regra do discriminante, temos que

$$4(EXY)^2 - 4EX^2EY^2 \leq 0,$$

e temos a primeira desigualdade. A segunda desigualdade segue da primeira trocando X por $X - EX$ e Y por $Y - EY$ na expressão da primeira desigualdade. ∎

O *coeficiente de correlação* entre duas variáveis aleatórias X e Y é dado por

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

O teorema anterior provou que $|\rho(X,Y)| \leq 1$.

5.6 Esperança Condicional

Nesta seção nos apresentamos o conceito de esperança condicional $E(Y|X)$ de uma variável aleatória Y dado outra variável aleatória X através do uso de funções probabilidade de massa no caso discreto e funções densidade de massa no caso contínuo. A interpretação é que $E(Y|X=x)$ é a média da variável aleatória Y sabendo que a variável aleatória X é igual a x. Por exemplo, podemos estar interessados na média do peso de indivíduos que têm $1{,}70m$ de altura.

Definição 5.6.1:

(a) Se (X,Y) for uma vetor aleatório contínuo bidimensional, define-se o valor esperado condicional de Y dado que $X = x$, como sendo

$$E(Y|X=x) = \int_{-\infty}^{0} y f_{Y|X}(y|x)dy + \int_{0}^{\infty} y f_{Y|X}(y|x)dy,$$

desde que pelo menos uma das integrais seja finita.

(b) Se (X,Y) for uma vetor aleatório discreto bidimensional, define-se o valor esperado condicional de Y dado que $X = x_i$, como sendo

$$E(Y|X=x_i) = \sum_{j:y_j \leq 0} y_j p_{Y|X}(y_j|x_i) + \sum_{j:y_j > 0} y_j p_{Y|X}(y_j|x_i),$$

desde que pelo menos uma das séries seja convergente.

Exemplo 5.6.2: Considere um vetor aleatório com densidade conjunta dada por

$$f_{X,Y}(x,y) = e^{-2|y-x^2|-x} I_{[0,\infty]}(x).$$

Determine $E(Y|X=x)$.

Solução: Vamos primeiro obter a densidade marginal de X

$$f_X(x) = \int_{-\infty}^{\infty} e^{-2|y-x^2|-x} I_{[0,\infty]}(x) dy$$

$$= e^{-x} I_{[0,\infty]}(x)(e^{-2x^2} \int_{-\infty}^{x^2} e^{2y} dy + e^{2x^2} \int_{x^2}^{\infty} e^{-2y} dy)$$

$$= e^{-x} I_{[0,\infty]}(x)(\frac{1}{2} + \frac{1}{2}) = e^{-x} I_{[0,\infty]}(x).$$

Logo, a densidade condicional de Y dado X é igual a

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = e^{-2|y-x^2|}.$$

Portanto,

$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy = \int_{-\infty}^{\infty} y e^{-2|y-x^2|} dy$$

$$= \int_{-\infty}^{x^2} y e^{-2(x^2-y)} dy + \int_{x^2}^{-\infty} y e^{-2(y-x^2)} dy$$

$$= (\frac{y}{2} e^{-2(x^2-y)}|_{-\infty}^{x^2} - \int_{-\infty}^{x^2} \frac{e^{-2(x^2-y)}}{2} dy) + (\frac{-y}{2} e^{-2(y-x^2)}|_{x^2}^{\infty} - \int_{x^2}^{\infty} \frac{-e^{-2(y-x^2)}}{2} dy)$$

$$= (\frac{x^2}{2} - \frac{1}{2}) + (\frac{x^2}{2} + \frac{1}{2}) = x^2.$$

Observe que $E(Y|X=x)$ é uma função de x, chamemos esta função de h(x). Então, temos que $E(Y|X) = h(X)$ é uma função da variável aleatória X e portanto é uma variável aleatória. Por outro lado, $E(Y|X)$ é uma média da variável Y. A seguir listamos algumas propriedades da esperança condicional:

1. $E(aY_1 + bY_2 + c|X) = aE(Y_1|X) + bE(Y_2|X) + c$.
2. $E(g(Y,X)|X=x) = E(g(Y,x)|X=x)$.
3. Se X e Y são independentes, então $E(Y|X) = E(Y)$.
4. $EY = E[E(Y|X)]$.

   Prova:

$$EY = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_{-\infty}^{\infty} y (\int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dx) dy$$

$$= \int_{-\infty}^{\infty} f_X(x) (\int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy) dx = \int_{-\infty}^{\infty} E(Y|X=x) f_X(x) dx$$

$$= E[E(Y|X)].$$

▮

Exemplo 5.6.3: Podemos utilizar este último resultado para calcular a esperança (incondicional) de Y no exemplo anterior.

$$\begin{aligned} EY = E[E(Y|X)] = E(X^2) = \int_0^\infty x^2 e^{-x} \\ = -x^2 e^{-x}|_0^\infty + 2\int_0^\infty x e^{-x} dx \\ = 0 + 2(-xe^{-x}|_0^\infty + \int_0^\infty e^{-x} dx) \\ = 2(0+1) = 2. \end{aligned}$$

Capítulo 6

Principais Variáveis Aleatórias Discretas

6.1 Introdução

Neste capítulo descreveremos um pouco sobre os principais modelos de variáveis aleatórias discretas.

6.2 Geométrica.

Dizemos que X tem uma distribuição *Geométrica* com parâmetro β, onde $0 \leq \beta < 1$, se $X(w) \in \{1, 2, 3, \ldots\}$ e $p(k) = (1-\beta)\beta^{k-1}$, para $k \in \{1, 2, 3, \ldots\}$.

Utilizando o resultado de uma soma infinita de uma Progressão Geométrica, temos que

$$\sum_{k=1}^{\infty} p(k) = \sum_{k=1}^{\infty} (1-\beta)\beta^{k-1} = (1-\beta)\sum_{k=1}^{\infty} \beta^{k-1} = 1.$$

Logo, esta é uma legítima função probabilidade de massa.

A função de probabilidade Geométrica pode ser utilizada para modelar o número de repetições do lançamento de uma moeda até a primeira ocorrência de cara, tempo de espera medido em unidades de tempo inteira até a chegada do próximo consumidor em uma fila, ou até a próxima emissão de um fóton.

Exemplo 6.2.1: Suponha que joga-se uma moeda com probabilidade de cara igual a $0 < p < 1$ independentemente até que uma coroa ocorra. Seja X o número de repetições necessárias até que coroa apareça nesta seqüência, de modo que se o primeiro lançamento for coroa temos que $X = 1$. Qual a probabilidade do evento $X = k$ para $k \in \{1, 2, 3, \ldots\}$? Note que para que $X = k$ é necessário que os primeiros $k-1$ lançamentos sejam caras e o k-ésimo lançamento seja coroa, logo pela independência dos lançamentos, temos que $P(X = k) = p^{k-1}(1-p)$. Ou seja X é uma variável geométrica com parâmetro p.

Se X for uma variável aleatória com distribuição de probabilidade Geométrica com pa-

râmetro β, temos que sua esperança é dada por:

$$EX = \sum_{k=1}^{\infty} k(1-\beta)\beta^{k-1} = \sum_{k=1}^{\infty}\sum_{j=1}^{k}(1-\beta)\beta^{k-1}$$
$$= (1-\beta)\sum_{j=1}^{\infty}\sum_{k=j}^{\infty}\beta^{k-1} = \sum_{j=1}^{\infty}\beta^{j-1} = \frac{1}{1-\beta}$$

Onde utilizamos a fórmula da soma infinita de uma progressão geométrica com razão β. Com um cálculo similar, porém mais longo, pode-se provar que $VarX = \frac{\beta}{(1-\beta)^2}$.

Exemplo 6.2.2: Suponha que X tenha uma distribuição geométrica com parâmetro β. Mostre que para quaisquer dois inteiros positivos s e t,

$$P(X > s+t|X > s) = P(X > t).$$

Solução: Note que

$$P(X > s+t|X > s) = \frac{P(X > s+t, X > s)}{P(X > s)} = \frac{P(X > s+t)}{P(X > s)}.$$

Mas

$$P(X > s+t) = \sum_{k=s+t+1}^{\infty}(1-\beta)\beta^{k-1} = \beta^{s+t}.$$

Similarmente, temos que $P(X > s) = \beta^s$. Portanto,

$$P(X > s+t|X > s) = \beta^t = P(X > t).$$

Esta propriedade da distribuição geométrica é conhecida como *falta de memória*.

6.3 Binomial Negativa ou Pascal.

Esta distribuição é uma generalização óbvia da distribuição geométrica. Suponha que ao invés de estarmos interessados no número de repetições de um experimento até a primeira ocorrência de um evento, estejamos interessados em calcular o número de repetições até a r-ésima ocorrência de um evento. Seja Y o número de repetições necessário a fim de que um evento A possa ocorrer exatamente r vezes. Temos que $Y = k$ se, e somente se, A ocorrer na k-ésima repetição e A tiver ocorrido $r-1$ vezes nas $(k-1)$ repetições anteriores. Assumindo independência entre os experimentos, esta probabilidade é igual $p\binom{k-1}{r-1}p^{r-1}(1-p)^{k-r}$. Portanto,

$$P(Y = k) = \binom{k-1}{r-1}p^r(1-p)^{k-r}, \text{ onde } k \geq r.$$

Note que se $r = 1$, temos que Y tem uma distribuição geométrica com parâmetro $\beta = 1-p$. No caso geral, dizemos que Y tem uma distribuição *Binomial Negativa ou Pascal*.

Para calcular EY e $VarY$ podemos proceder da seguinte maneira. Seja $Z_1, Z_2, \ldots$ uma seqüência de variáveis aleatórias tal que Z_1 é o número de repetições necessárias até a primeira ocorrência de um evento A, e Z_i é o número de repetições necessárias entre a $(i-1)$-ésima e a i-ésima ocorrência de A, para $i = 2, 3, \ldots, r$. Então, as variáveis Z_i são independentes e cada uma delas tem uma distribuição geométrica com parâmetro $\beta = 1 - p$, e temos que $Y = Z_1 + Z_2 + \cdots + Z_r$. Logo, usando propriedades da esperança e da variância, temos que $EY = rEZ_1 = \frac{r}{p}$ e $VarY = rVarZ_1 = \frac{r(1-p)}{p^2}$.

6.3.1 Relação entre as Distribuições Binomial e Binomial Negativa.

Suponhamos que X tenha distribuição binomial com parâmetros n e p, ou seja, X é igual ao número de sucessos em n ensaios repetidos de Bernoulli com probabilidade de sucesso p. Suponhamos que Y tenha uma distribuição Binomial Negativa com parâmetros r e p, ou seja, Y é o número de ensaios de Bernoulli necessários para se obter r sucessos com probabilidade de sucesso p. Então, temos que $\{X \geq r\} = \{Y \leq n\}$, ou seja, o número de sucessos em n ensaios é maior ou igual a r se, e somente se, o número de ensaios Bernoulli até a ocorrência do r-ésimo sucesso for menor ou igual a n. Portanto,

$$P(X \geq r) = P(Y \leq n).$$

Observe que estas duas distribuições tratam de ensaios de Bernoulli repetidos. A distribuição binomial surge quando lidamos com um número fixo de ensaios e estamos interessados no número de sucessos que venham a ocorrer. A distribuição binomial negativa é encontrada quando fixamos o número de sucessos e então registramos o número de ensaios necessário.

6.4 Poisson.

Dizemos que X tem uma distribuição *Poisson* com parâmetro λ, onde $\lambda \geq 0$, se $X(w) \in \{0, 1, \ldots\}$ e $p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$, para $k \in \{0, 1, \ldots\}$.

Usando o resultado da expansão em série de Taylor da função exponencial, temos que para todo x real,

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

Utilizando este fato, temos que

$$\sum_{k=0}^{\infty} p(k) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1.$$

Logo, esta é uma legítima função probabilidade de massa.

A função de probabilidade Poisson é utilizada para modelar a contagem do número de ocorrências de eventos aleatórios em um certo tempo T: número de fótons emitidos por uma fonte de luz de intensidade I fótons/seg em T segundos ($\lambda = IT$), número de clientes chegando em uma fila no tempo T ($\lambda = CT$), número de ocorrências de eventos raros no tempo T ($\lambda = CT$).

Exemplo 6.4.1: Se a probabilidade de 0 fótons serem emitidos no tempo T é igual a 0,1, então qual a probabilidade de pelo menos 2 fótons serem emitidos no tempo T?

Se X for uma variável aleatória com distribuição de probabilidade Poisson com parâmetros λ, temos que sua esperança é dada por:

$$EX = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda\sum_{k=1}^{\infty}\frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda.$$

Já vimos que o segundo momento de uma variável aleatória com distribuição Poisson(λ) é igual a $\lambda^2+\lambda$. Portanto, $VarX = \lambda^2+\lambda-(\lambda)^2 = \lambda$.

Podemos analisar o valor mais provável de uma distribuição de Poisson, através da razão de dois valores sucessivos da função probabilidade de massa:

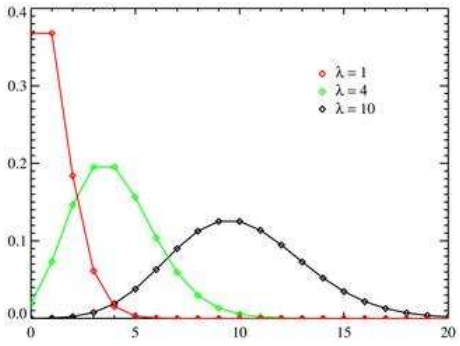
$$\frac{p_{k+1}}{p_k} = \frac{\lambda}{k+1}.$$

Note que esta razão é estritamente decrescente em k. Logo, $\{p_k\}$ é sempre decrescente se $\lambda < 1$, decresce após $p_0 = p_1$ se $\lambda = 1$, e cresce inicialmente se $\lambda > 1$ e eventualmente decresce qualquer que seja o valor de λ. Formalmente, um valor mais provável de uma distribuição de Poisson é definido como k^* se $p_{k^*+1} \leq p_{k^*}$ e $p_{k^*-1} \leq p_{k^*}$. (Note que podem existir valores adjacentes que possuam o mesmo valor.) Mas esta condição é equivalente a,

$$k^* \leq \lambda \leq k^*+1, \text{ ou}$$

$$\lambda - 1 \leq k^* \leq \lambda.$$

Note que se tomarmos k^* como sendo o maior inteiro menor ou igual a λ esta restrição é satisfeita, e portanto este é um valor mais provável desta distribuição. A Figura 6.4 nos mostra a função probabilidade de massa da Poisson para 3 valores de parâmetros 1, 4, e 10.



Exemplo 6.4.2: Suponha que o número de clientes que chegam em um banco segue uma distribuição de Poisson. Se a probabilidade de chegarem 3 clientes for o triplo da de chegarem 4 clientes em um dado período de 10 minutos. Determine:

(a) Qual o número esperado de clientes que chegam em um período de 1 hora neste banco?

(b) Qual o número mais provável de clientes que chegam em um período de 1 hora neste banco?

6.5 Hipergeométrica.

A distribuição hipergeométrica descreve o número de sucessos em uma seqüência de n amostras de uma população finita sem reposição.

Por exemplo, considere que tem-se uma carga com N objetos dos quais D têm defeito. A distribuição hipergeométrica descreve a probabilidade de que em uma amostra de n objetos distintos escolhidos da carga aleatoriamente exatamente k objetos sejam defeituosos.

Em geral, se uma variável aleatória X segue uma distribuição hipergeométrica com parâmetros N, D, e n, então a probabilidade de termos exatamente k sucessos é dada por

$$p(k) = \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}.$$

Esta probabilidade é positiva se: $N - D \geq n - k$, ou seja $k \geq \max(0, D + n - N)$, e $k \leq \min(n, D)$.

Esta fórmula pode ser entendida assim: existem $\binom{N}{n}$ possíveis amostras sem reposição. Existem $\binom{D}{k}$ maneiras de escolher k objetos defeituosos e existem $\binom{N-D}{n-k}$ maneiras de preencher o resto da amostra com objetos sem defeito.

Quando a população é grande quando comparada ao tamanho da amostra (ou seja, N for muito maior que n) a distribuição hipergeométrica é aproximada razoavelmente bem por uma distribuição binomial com parâmetros n (tamanho da amostra) e $p = D/N$ (probabilidade de sucesso em um único ensaio).

Se X for uma variável aleatória com distribuição de probabilidade Hipergeométrica com parâmetro N, D, n, temos que sua esperança é dada por:

$$EX = \sum_{k=0}^{n} k\frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}} = \sum_{k=1}^{n} \frac{D!(N-D)!(N-n)!n!}{k!(D-k)!(n-k)!(N-D-n+k)!N!}$$
$$= \frac{nD}{N}\sum_{k=1}^{n} \frac{(D-1)!(N-D)!(N-n)!(n-1)!}{(k-1)!(D-k)!(n-k)!(N-D-n+k)!(N-1)!} = \frac{nD}{N}\sum_{k=1}^{n} \frac{\binom{D-1}{k-1}\binom{N-D}{n-k}}{\binom{N-1}{n-1}}$$

Substituindo no somatório $D^* = D - 1, k^* = k - 1, n^* = n - 1$ e $N^* = N - 1$, temos

$$EX = \frac{nD}{N}\sum_{k^*=0}^{n^*} \frac{\binom{D^*}{k^*}\binom{N^*-D^*}{n^*-k^*}}{\binom{N^*}{n^*}} = \frac{nD}{N}.$$

Onde utilizamos o fato que o somatório é igual soma da função probabilidade de massa de uma variável aleatória Hipergeométrica para todos os valores que tem probabilidade positiva, e portanto, é igual a 1. Com um cálculo similar, porém mais longo, pode-se provar que $VarX = \frac{nD}{N}\frac{(N-D)(N-n)}{N(N-1)}$.

Exemplo 6.5.1: Suponha que uma urna contém 20 bolas brancas e 10 bolas pretas. Se 4 bolas são retiradas da urna. Determine:

(a) A probabilidade de pelo menos uma bola ser branca, se as bolas são retiradas com reposição.

(b) A probabilidade de pelo menos uma bola ser branca, se as bolas são retiradas sem reposição.

Exemplo 6.5.2: Por engano 3 peças defeituosas foram misturadas com boas formando um lote com 12 peças no total. Escolhendo ao acaso 4 dessas peças, determine a probabilidade de encontrar:

(a) Pelo menos 2 defeituosas.

(b) No máximo 1 defeituosa.

(c) No mínimo 1 boa.

6.6 Poisson como um Limite de Eventos Raros de Binomial

Suponhamos que chamadas telefônicas cheguem em uma grande central, e que em um período particular de três horas (180 minutos), um total de 270 chamadas tenham sido recebidas, ou seja, 1,5 chamadas por minuto. Suponhamos que queiramos calcular a probabilidade de serem recebidas k chamadas durante os próximos três minutos.

Ao considerar o fenômeno da chegada de chamadas, poderemos chegar à conclusão de que, a qualquer instante, uma chamada telefônica é tão provável de ocorrer como em qualquer outro instante. Como em qualquer intervalo de tempo, temos um número infinito de pontos, vamos fazer uma série de aproximações para este cálculo.

Para começar, pode-se dividir o intervalo de 3 minutos em nove intervalos de 20 segundos cada um. Poderemos então tratar cada um desses nove intervalos como um ensaio de Bernoulli, durante o qual observaremos uma chamada (sucesso) ou nenhuma chamada (falha), com probabilidade de sucesso igual a $p = 1,5 \times \frac{20}{60} = 0,5$. Desse modo, poderemos ser tentados a afirmar que a probabilidade de 2 chamadas é igual a $\binom{9}{2}(0,5)^9 = \frac{9}{128}$. Porém, este cálculo ignora a possibilidade de que mais de uma chamada possa ocorrer em um único intervalo. Então, queremos aumentar o número n de subintervalos de tempo de modo que cada subintervalo corresponde a $\frac{180}{n}$ segundos e então a probabilidade de ocorrência de uma chamada em um subintervalo é igual a $p = 1,5 \times \frac{180}{60n}$. Desta maneira temos que $np = 4,5$ permanece constante ao crescermos o número de subintervalos. Utilizando novamente o modelo binomial, temos que a probabilidade de ocorrerem k chamadas é dada por: $\binom{n}{k}(\frac{4,5}{n})^k(1-\frac{4,5}{n})^{n-k}$. Queremos saber então o que acontece com esta probabilidade quando $n \to \infty$. A resposta como veremos a seguir é que esta distribuição tende a distribuição de Poisson e este resultado é conhecido como *limite de eventos raros*.

Consideremos a expressão geral da probabilidade binomial,

$$p(k) = \binom{n}{k}p^k(1-p)^{n-k} = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!}p^k(1-p)^{n-k}.$$

Como queremos estudar o caso em que np é constante, façamos $np = \alpha$, ou seja, $p = \alpha/n$ e $1 - p = \frac{n-\alpha}{n}$. Então,

$$p(k) = \frac{n(n-1)\cdots(n-k+1)}{k!}(\frac{\alpha}{n})^k(\frac{n-\alpha}{n})^{n-k}$$
$$= \frac{\alpha^k}{k!}[(1)(1-\frac{1}{n})\cdots(1-\frac{k-1}{n})][1-\frac{\alpha}{n}]^{n-k}$$

Fazendo $n \to \infty$, temos que os termos da forma $(1-\frac{j}{n})$, para $1 \leq j \leq k-1$, tendem para 1 e como existe um número fixo $k-1$ deles, o seu produto também tende a 1. O mesmo ocorre com $(1-\frac{\alpha}{n})^{-k}$. Finalmente, por definição do número e, temos que $(1-\frac{\alpha}{n})^n \to e^{-\alpha}$ quando $n \to \infty$. Portanto,

$$\lim_n p(k) = e^{-\alpha}\frac{\alpha^k}{k!},$$

ou seja obtemos a expressão de Poisson.

Então, provamos o seguinte teorema:

Teorema 6.6.1: *Se* $\lim_{n\to\infty} np_n = \alpha > 0$, *então*

$$\lim_{n\to\infty}\binom{n}{k}p_n^k(1-p_n)^{n-k} = e^{-\alpha}\frac{\alpha^k}{k!}.$$

Exemplo 6.6.2: Ao formar números binários com n dígitos, a probabilidade de que um dígito incorreto possa aparecer é 0,002. Se os erros forem independentes, qual é a probabilidade de encontrar k dígitos incorretos em um número binário de 25 dígitos? Se um computador forma 10^6 desses números de 25 dígitos por segundo, qual é a probabilidade de que pelo menos um número incorreto seja formado durante qualquer período de 1 segundo?

Solução: A probabilidade de que k dígitos sejam incorretos em um número binários de 25 dígitos é igual a $\binom{25}{k}(0{,}002)^k(0{,}998)^{25-k}$. Em particular, a probabilidade de que pelo menos um dígito seja incorreto é igual a $1-(0{,}998)^{25} \approx 0{,}049$. Se tivéssemos usado a aproximação pela Poisson então teríamos uma Poisson com parâmetro $25 \times 0{,}002 = 0{,}05$, logo a probabilidade de pelos menos um dígito incorreto neste número de 25 dígitos seria $1 - e^{-0{,}05} \approx 0{,}049$.

A probabilidade de que pelo menos um número incorreto seja formado durante o período de 1 segundo é igual a $1-(0{,}049)^{10^6} \approx 1 - e^{-49000} \approx 1$.

6.7 A Distribuição Multinomial

Vamos dar o exemplo de uma distribuição conjunta de variáveis aleatórias discretas, que pode ser considerada como uma generalização da distribuição binomial. Considere um experimento aleatório qualquer e suponha que o espaço amostral deste experimento é particionado em k eventos $\{A_1, A_2, \ldots, A_k\}$, onde o evento A_i tem probabilidade p_i. Suponha que se repita este experimento n vezes de maneira independente e seja X_i o número de vezes que o evento A_i ocorreu nestas n repetições. Então,

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1!n_2!\cdots n_k!}p_1^{n_1}p_2^{n_2}\cdots p_k^{n_k},$$

onde $\sum_{i=1}^{k} n_i = n$. (Relembre que o número de maneiras de arranjar n objetos, n_1 dos quais é de uma espécie, n_2 dos quais é de uma segunda espécie, ..., n_k dos quais são de uma k-ésima espécie é dado pelo coeficiente multinomial $\frac{n!}{n_1!n_2!\cdots n_k!}$.)

Capítulo 7

Principais Variáveis Aleatórias Contínuas

7.1 Introdução

Neste capítulo, vamos explorar alguns exemplos importantes de variáveis aleatórias contínuas.

7.2 Normal ou Gaussiana

Dizemos que X tem uma distribuição *Normal (ou Gaussiana)* com parâmetros μ e σ, onde μ e $\sigma > 0$ são números reais, se a função densidade de X é igual a

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

Vamos verificar que esta realmente é uma função densidade de probabilidade. Fazendo a substituição de variáveis $t = \frac{x-\mu}{\sigma}$, obtemos

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt = I.$$

Vamos agora utilizar um artifício de calcular I^2. Temos

$$I^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{\frac{-t^2}{2}} dt \int_{-\infty}^{\infty} e^{\frac{-s^2}{2}} ds = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{\frac{-(t^2+s^2)}{2}} dt ds.$$

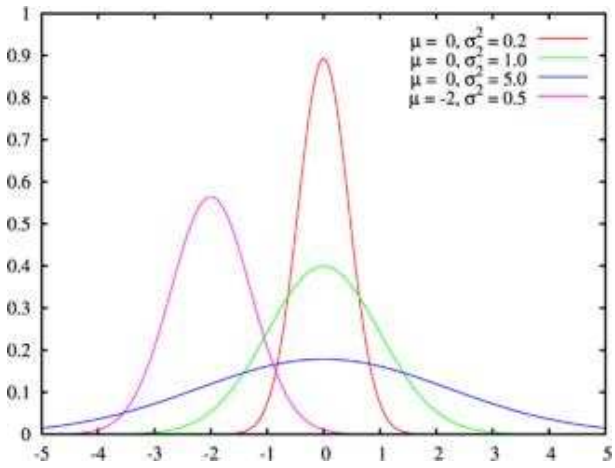
Fazendo a seguinte mudança de variável: $t = r\cos\theta$ e $s = r\text{sen}\,\theta$, temos

$$\begin{aligned} I^2 &= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} r e^{\frac{-r^2}{2}} dr d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} -e^{\frac{-r^2}{2}} |_0^{\infty} d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} 1 d\theta = 1. \end{aligned}$$

Portanto, temos que $I = 1$.

Historicamente, esta distribuição foi chamada de "normal" porque ela era amplamente aplicada em fenômenos biológicos e sociais que era sempre tida como a distribuição antecipada ou normal. Aplicações da distribuição normal incluem ruído térmico em resistores e em outros sistemas físicos que possuem um componente dissipativo; ruídos de baixa-freqüência como os em encontrados em amplificadores de baixa freqüência; e variabilidade em parâmetros de componentes manufaturados e de organismos biológicos (por exemplo, altura, peso, inteligência). (Pode parecer estranho, modelar quantidades que só assumem valores positivos por uma distribuição normal onde valores negativos aparecem. Nestes casos o que ocorre é que os parâmetros μ e σ^2 devem ser escolhidos de modo que a probabilidade da variável assumir um valor negativo seja aproximadamente nula de modo que a representação seja válida.)

A Figura 7.2 nos mostra a função probabilidade de massa da Normal para 4 pares de parâmetros. Observe que a densidade é simétrica em torno do parâmetro μ, e quanto menor o parâmetro σ mais concentrada é a densidade em torno deste parâmetro μ. Pode-se provar que os pontos $\mu - \sigma$ e $\mu + \sigma$ são os pontos de inflexão do gráfico de f_X. Veremos adiante que μ e σ^2 são iguais a esperança e a variância da distribuição normal, respectivamente. Se $\mu = 0$ e $\sigma^2 = 1$ chamamos esta densidade de *normal padrão* ou *normal reduzida*.



Se $X \sim \mathcal{N}(\mu, \sigma)$, temos que sua esperança é dada por:

$$EX = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx.$$

Fazendo a mudança de variável $y = \frac{x-\mu}{\sigma}$, temos

$$EX = \int_{-\infty}^{\infty} \frac{\sigma y + \mu}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} dy = \int_{-\infty}^{\infty} \frac{\sigma y}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} dy + \int_{-\infty}^{\infty} \frac{\mu}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} dy = 0 + \mu = \mu.$$

Para o cálculo do segundo momento, vamos também realizar a mudança de variável

$y = \frac{x-\mu}{\sigma}$, logo

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma y + \mu)^2 e^{\frac{-z^2}{2}} dz$$
$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{\frac{-z^2}{2}} dz + 2\mu\sigma \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{\frac{-z^2}{2}}$$
$$+\mu^2 \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz.$$

A segunda parcela, pela resultado da esperança da normal padrão é igual a zero. A última parcela pelo resultado da integral da densidade da normal, temos que é igual a μ^2. Para calcular a primeira parcela, vamos usar integral por partes onde $u = z$ e $dv = ze^{\frac{-z^2}{2}}$. Assim obtemos

$$E(X^2) = \frac{\sigma^2}{\sqrt{2\pi}}(-ze^{\frac{-z^2}{2}}|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{\frac{-z^2}{2}} dz) + \mu^2$$
$$= \sigma^2 + \mu^2.$$

O seguinte teorema afirma que transformações lineares de variáveis aleatórias com distribuição normal também são distribuídas normalmente.

Teorema 7.2.1: *Se $X \sim N(\mu, \sigma^2)$ e se $Y = aX + b$, onde $a > 0$ e $b \in \mathbb{R}$, então Y terá distribuição $N(a\mu + b, a^2\sigma^2)$.*

Prova: Note que

$$F_Y(y) = P(Y \leq y) = P(X \leq \frac{y-b}{a}) = F_X(\frac{y-b}{a}).$$

Derivando a expressão acima em relação a y, temos

$$f_Y(y) = \frac{1}{a} f_X(\frac{y-b}{a}) = \frac{1}{\sqrt{2\pi}a\sigma} e^{\frac{-(\frac{y-b}{a}-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}a\sigma} e^{\frac{(y-(b+a\mu))^2}{2a^2\sigma^2}},$$

ou seja, $Y \sim N(a\mu + b, a^2\sigma^2)$. ∎

Corolário 7.2.2: *Se $X \sim N(\mu, \sigma^2)$, então $Y = \frac{X-\mu}{\sigma}$ tem distribuição normal padrão.*

Pode-se provar que se $X_i \sim N(\mu_i, \sigma_i^2)$ são independentes, e $a_i \in \mathbb{R}$, para $i = 1, 2, 3, \ldots$, então $Y = c + \sum_{i=1}^{n} a_i X_i$ também tem distribuição normal com média $EY = c + \sum_{i=1}^{n} a_i \mu_i$ e variância $VarY = \sum_{i=1}^{n} (a_i \sigma_i)^2$.

7.2.1 Tabulação da Distribuição Normal

Se $X \sim N(0,1)$, então

$$P(a < X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx.$$

Esta integral não pode ser resolvida analiticamente, contudo métodos de integração numérica podem ser empregados para calcular integrais da forma acima e de fato valores de $P(X \leq s)$ existem em várias tabelas. A função de distribuição acumulada de uma normal padrão é usualmente denotada por Φ. Portanto, temos que $\Phi(s) = \int_{-\infty}^{s} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx$. Então, consultando valores de Φ em uma tabela, podemos determinar que $P(a < X \leq b) = \Phi(b) - \Phi(a)$.

Utilizando o resultado do Corolário 7.2.2 e valores de Φ, podemos obter para qualquer $X \sim N(\mu, \sigma^2)$, o valor de $P(a < X \leq b)$:

$$P(a < X \leq b) = P(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma})$$
$$\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

Em especial podemos estar interessados em calcular $P(\mu - k\sigma \leq X \leq \mu + k\sigma)$, usando o resultado acima temos que esta probabilidade é igual a $\Phi(k) - \Phi(-k)$.

Da simetria em torno de zero da normal padrão, temos que $\Phi(s) = P(X \leq s) = P(X \geq -s) = 1 - \Phi(-s)$ para qualquer valor de s. Esta relação pode ser útil, pois freqüentemente tabelas da distribuição normal só possuem os valores positivos de s.

Exemplo 7.2.3: Suponha que X tenha uma distribuição $N(2; 0{,}16)$. Empregando a tábua de distribuição normal, calcule as seguintes probabilidades:

(a) $P(X \geq 2{,}3)$.

(b) $P(1{,}8 \leq X \leq 2{,}1)$.

Solução: Parte (a),

$$P(X \geq 2{,}3) = 1 - P(\leq 2{,}3) = 1 - \Phi(\frac{2{,}3-2}{0{,}4}) = 1 - \Phi(0{,}75) = 1 - 0{,}7734 = 0{,}2266.$$

Parte (b),

$$P(1{,}8 \leq X \leq 2{,}1) = \Phi(\frac{2{,}1-2}{0{,}4}) - \Phi(\frac{1{,}8-2}{0{,}4}) = \Phi(0{,}25) - \Phi(-0{,}5) = 0{,}5987 - 0{,}3085 = 0{,}2902.$$

Exemplo 7.2.4: Um equipamento com dois terminais com uma resistência equivalente de 1 Megohm opera em uma sala com temperatura de 300K. A voltagem térmica V que ele gera é observada na banda de 1,5GHz até 2,5GHz. Qual é a probabilidade que a magnitude da voltagem exceda 8 milivolts? Assuma que $V \sim N(0, \sigma^2)$, onde $\sigma^2 = 4\kappa TRB$, κ é a constante de Boltzman que é igual a $1{,}38 \times 10^{-23}$, V é medido em volts, T é medido em graus Kelvin, R medido em ohms, e B medido em Hertz.

Solução: Das informações podemos calcular que $\sigma^2 = 4(1{,}38 \times 10^{-23})(300)(10^6)(10^9) = 16{,}5 \times 10^{-6}$. Logo, $\sigma \approx 0{,}004$. Portanto,

$$P(|V| > 0{,}008) = P(V > 0{,}008) + P(V < -0{,}008) = (1 - \Phi(\frac{0{,}008 - 0}{0{,}004})) + \Phi(\frac{-0{,}008 - 0}{0{,}004})$$
$$= 1 - \Phi(2) + \Phi(-2) = 2(1 - \Phi(2)) = 2(1 - 0{,}9772) = 0{,}456.$$

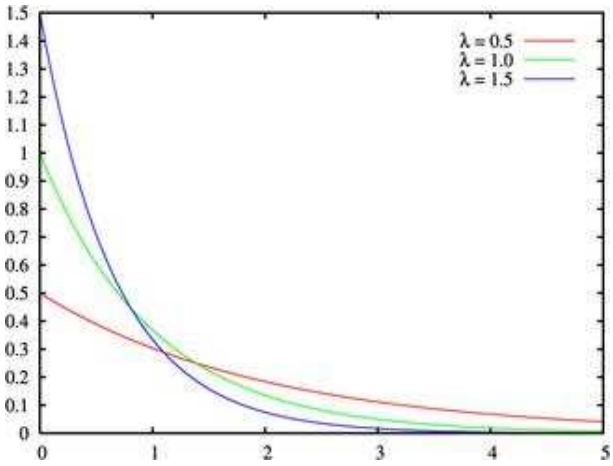
7.3 Exponencial

Dizemos que X tem uma distribuição *Exponencial* com parâmetro λ, onde $\lambda > 0$ é um número real, se a função densidade de X é igual a

$$f_X(x) = \lambda e^{-\lambda x} U(x),$$

onde $U(x) = I_{[0,\infty)}(x)$ é conhecida como *função degrau*.

A Figura 7.3 mostra a função densidade exponencial para $\lambda = 0{,}5$, $\lambda = 1$, e $\lambda = 1{,}5$.



A densidade exponencial pode ser utilizada para modelar os seguintes fenômenos: tempo de vida de componentes que falham sem efeito de idade; tempo de espera entre sucessivas chegadas de fótons, emissões de elétrons de um cátodo, ou chegadas de consumidores; e duração de chamadas telefônicas.

Se $X \sim Exp(\lambda)$, então X possui densidade igual a $f_X(x) = \lambda e^{-\lambda x} U(x)$. Logo, temos que sua esperança é dada por:

$$EX = \int_0^\infty x\lambda e^{-\lambda x} dx = -xe^{-\lambda x}|_0^\infty + \int_0^\infty e^{-\lambda x} dx = \frac{-e^{-\lambda x}}{\lambda}|_0^\infty = \frac{1}{\lambda}.$$

Para o cálculo da variância, vamos calcular o segundo momento:

$$EX^2 = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = -x^2 e^{-\lambda x}|_0^\infty + 2\int_0^\infty xe^{-\lambda x} dx = \frac{2}{\lambda^2}.$$

Portanto,

$$VarX = EX^2 - (EX)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

A distribuição exponencial também possui a propriedade de falta de memória, ou seja, para quaisquer $s \geq 0$ e $t \geq 0$, temos

$$P(X > s+t|X > s) = P(X > t).$$

Para verificar este fato, note que

$$P(X > s+t|X > s) = \frac{P(X > s+t, X > s)}{P(X > s)} = \frac{P(X > s+t)}{P(X > s)}.$$

Mas

$$P(X > s+t) = \int_{s+t}^{\infty} \lambda e^{-\lambda x} dx = [-e^{-\lambda x}]_{s+t}^{\infty} = e^{-\lambda(s+t)}.$$

Similarmente, temos que $P(X > s) = e^{-\lambda s}$. Portanto,

$$P(X > s+t|X > s) = e^{-\lambda t} = P(X > t).$$

Exemplo 7.3.1: Observa-se que um tipo particular de *chip* é igualmente provável durar menos que 5.000 horas ou mais que 5.000 horas. Determine:

(a) Determine o tempo de duração médio de um *chip* deste tipo.

(b) Qual a probabilidade que o *chip* durará menos de 1.000 horas ou mais de 10.000 horas?

Solução: Seja X o tempo de duração de um *chip* deste tipo. Tempos que X tem uma distribuição exponencial, devemos agora determinar seu parâmetro. Sabe-se que $P(X < 5000) = P(X > 5000)$, e como $P(X < 5000) + P(X > 5000) = 1$, temos que $P(X < 5000) = 0{,}5$. Portanto, $1 - e^{-\lambda(5000)} = 0{,}5$, ou seja, $\lambda = \frac{\log 2}{5000}$. Então, o tempo de duração médio deste tipo de *chip* é $\frac{5000}{\log 2}$ horas.

Para calcular a probabilidade desejada temos que

$$P([X < 1000] \cup [X > 10000]) = P(X < 1000) + P(X > 10000) = 1 - e^{-\frac{\log 2}{5}} + e^{-2\log 2}$$
$$= 1 - (2)^{-\frac{1}{5}} + (2)^{-2} = 1 - 0{,}8706 + 0{,}25 = 0{,}3794.$$

7.4 Cauchy

Dizemos que X tem uma distribuição *Cauchy* com parâmetro x_0 e $\gamma > 0$, se a função densidade de X é igual a

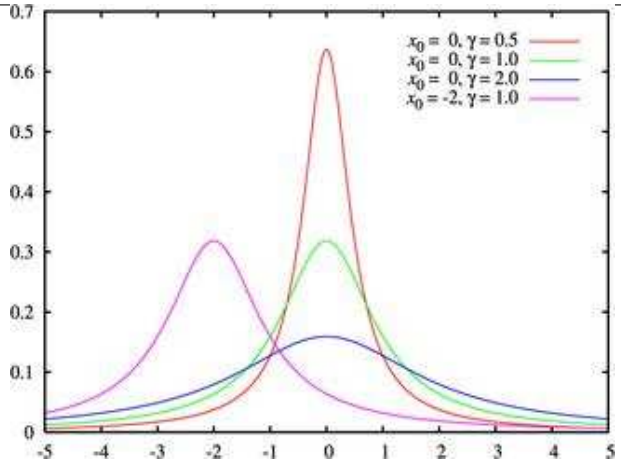
$$f_X(x) = \frac{1}{\pi} \cdot \frac{\gamma}{\gamma^2 + (x - x_0)^2}.$$

A Figura 7.4 mostra a função densidade Cauchy para alguns pares de parâmetros.

Pode-se provar que a razão entre duas variáveis aleatórias com distribuição Normal padrão independentes tem uma distribuição Cauchy com parâmetros $x_0 = 0$ e $\gamma = 1$.

Se $X \sim Cauchy(x_0, \gamma)$, então X não é integrável, ou seja EX não está definida, pois:

$$\int_{-\infty}^{0} \frac{x}{\pi} \cdot \frac{\gamma}{\gamma^2 + (x - x_0)^2} dx = -\infty, \text{ e}$$
$$\int_{0}^{\infty} \frac{x}{\pi} \cdot \frac{\gamma}{\gamma^2 + (x - x_0)^2} dx = \infty.$$



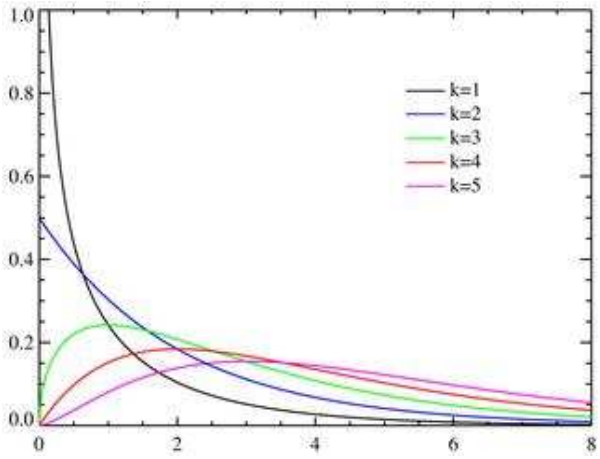
7.5 Qui-quadrado

Dizemos que X tem uma distribuição *Qui-quadrado* com parâmetro n, onde n é número natural, se a função densidade de X é igual a

$$f_X(x) = \frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)}U(x),$$

onde $\Gamma(p) = \int_0^\infty x^{p-1}e^{-x}dx$ para $p > 0$ é a *função gama*. n é conhecido como número de *graus de liberdade* da distribuição Qui-quadrado.

A Figura 7.5 mostra a função densidade Qui-quadrado para 1, 2, 3, 4, e 5 graus de liberdade.



Pode-se provar que se $X_1, X_2, X_3, \ldots, X_n$ são n variáveis aleatórias independentes com densidade normal padrão, então $X = X_1^2 + X_2^2 + \cdots + X_n^2$ tem densidade Qui-quadrado com n graus de liberdade. A distribuição Qui-quadrado tem inúmeras aplicações em inferência estatística. Por exemplo, na estimação de variâncias. Pode-se provar que $EX = n$ e $VarX = 2n$.

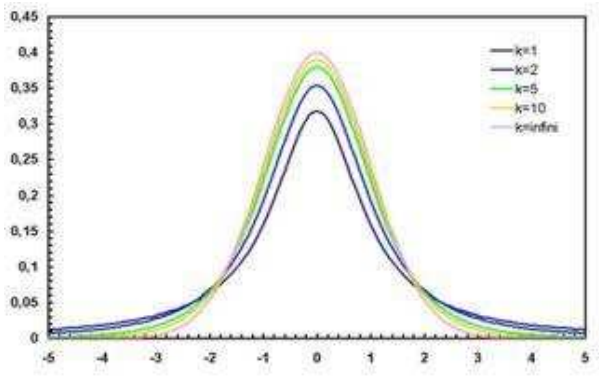
7.6 t de Student

Dizemos que X tem uma distribuição *t de Student* com parâmetro n, onde n é número natural, se a função densidade de X é igual a

$$f_X(x) = \frac{\Gamma[(n+1)/2]}{\Gamma[n/2]\sqrt{\pi n}}(1+\frac{x^2}{n})^{\frac{-(n+1)}{2}},$$

onde n é conhecido como número de *graus de liberdade* da distribuição t de Student.

A Figura 7.6 mostra a função densidade t de Student para 1, 2, 5, 10 e infinitos graus de liberdade.



Note que se $n = 1$, temos que a distribuição t de Student é igual a distribuição Cauchy(0,1). Se $n \to \infty$, a distribuição t de Student converge para a distribuição normal padrão. Pode-se provar que se Z é uma distribuição normal padrão independente de V que tem distribuição Qui-quadrado com n graus de liberdade, então $X = \frac{Z}{\sqrt{\frac{V}{n}}}$ tem uma distribuição t de Student com n graus de liberdade. A distribuição t de Student é bastante utilizada em inferência estatística. Por exemplo, pode-se utilizá-la para calcular intervalos de confiança para a média de uma amostra quando a variância da população não é conhecida. Pode-se provar que se $n > 1$, então $EX = 0$; que se $n > 2$, então $VarX = \frac{n}{n-2}$.

7.7 A Distribuição Normal Bivariada

Vamos agora dar o exemplo de uma distribuição conjunta de variáveis aleatórias contínuas. Dizemos que o vetor aleatório (X, Y) possui distribuição normal bivariada quando tem densidade dada por

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_1}{\sigma_1})^2 - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2}) + (\frac{y-\mu_2}{\sigma_2})^2]\},$$

onde $\sigma_1 > 0, \sigma_2 > 0, -1 < \rho < 1, \mu_1 \in I\!R, \mu_2 \in I\!R$.

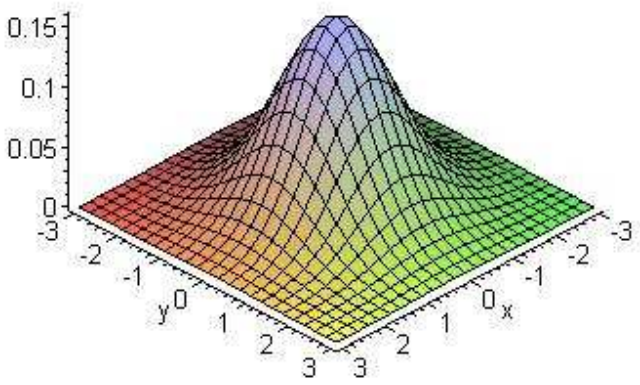
Se $\rho = 0$, esta densidade fatora e temos que X e Y são independentes. Se $\rho \neq 0$, esta densidade não fatora e X e Y não são independentes. Além disso, a distribuição normal bivariada satisfaz as seguintes propriedades:

1. As distribuições marginais de X e de Y são $N(\mu_1, \sigma_1^2)$ e $N(\mu_2, \sigma_2^2)$, respectivamente.

2. O parâmetro ρ é igual ao coeficiente de correlação entre X e Y.

3. As distribuições condicionais de X dado que $Y = y$ e de Y dado que $X = x$ são, respectivamente, $N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2))$ e $N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y - \mu_1), \sigma_2^2(1 - \rho^2))$.

A Figura 7.7 nos mostra a função densidade da normal bivariada, onde $\rho = \mu_1 = \mu_2 = 0$ e $\sigma_1 = \sigma_2 = 1$.



Capítulo 8

Análise Exploratória de Dados

8.1 Resumo de Dados

8.1.1 Tipos de Variáveis

Quando se faz um experimento científico, em geral queremos observar os resultados referentes à alguma característica de interesse. Tais características de interesse são denominadas *variáveis*. Por exemplo, podemos estar interessados no tempo de vida útil de um dado equipamento eletrônico. As variáveis podem ser classificadas como *qualitativas* quando descrevem possíveis atributos de um dado experimento ou *quantitativas* quando descrevem possíveis números resultantes de um processo de contagem ou mensuração. Por exemplo, a marca e o modelo de um equipamento eletrônico são variáveis qualitativas, porém o temo de vida útil é uma variável quantitativa.

As variáveis qualitativas podem ser classificadas como *nominais* ou *ordinais* dependendo se existe ou não uma ordem natural em seus possíveis resultados. No exemplo anterior tanto a marca como o modelo são variáveis nominais. Para um exemplo de uma variável ordinal, considere o grau de escolaridade de um individuo em uma dada pesquisa.

As variáveis quantitativas podem ser classificadas como *discretas* ou *contínuas* dependendo se o conjunto de possíveis resultados é um conjunto enumerável ou não enumerável. O tempo de vida útil de um equipamento pode ser considerado como uma variável contínua. Já o número de fótons emitidos por uma fonte radioativa é uma variável discreta.

Em algumas situações podem se atribuir valores numéricos aos diversos atributos ou classes de uma variável qualitativa para que se possa efetuar uma análise como se esta fosse quantitativa, desde que haja alguma possível interpretação desta atribuição. Um caso bastante útil é no caso de uma variável dicotômica, ou seja, que assume apenas dois possíveis valores. Por exemplo, o sexo de um indivíduo em uma dada observação. Pode-se neste caso associar-se o valor zero a um sexo e o valor 1 ao outro.

Um outro possível critério para classificar variáveis é em função da escala de medida adotada para se analisar o resultado do experimento. As escalas de medidas podem ser: *nominais*, *ordinais*, *intervalares*, e de *razão*.

- Uma escala nominal é utilizada para classificar os resultados de um experimento, por exemplo, se dado equipamento falhou ou não durante o período de estudo, a marca e

o modelo do equipamento em questão.

- Uma escala ordinal além de classificar os resultados também pode ser utilizada para estabelecer uma ordem entre as diferentes classes de possíveis resultados, por exemplo, grau de escolaridade de um indivíduo, classe socio-econômica de um indivíduo, posição que um dado indivíduo conclui uma certa corrida. Transformações que preservam a ordem não alteram a estrutura de uma classe ordinal.

- Uma escala intervalar pode ser utilizada para além de classificar e ordenar os resultados também quantificar a diferença entre as classes. Nesta escala necessitamos estabelecer uma origem arbitrária e uma unidade de medida nesta escala. Por exemplo, a temperatura de um dado equipamento em funcionamento medida em graus centígrados constitui uma medida numa escala intervalar. Considere o caso em que temos três equipamentos E1, E2, e E3, operando em temperaturas de 40, 45 e 80 graus centígrados, respectivamente. É válido afirmar que a diferença de temperatura entre E3 e E2 é 7 vezes maior que a diferença de temperatura entre E2 e E1. Contudo, neste escala não faz sentido afirmar que E3 tem uma temperatura 2 vezes maior que E1, pois lembre que a origem e a unidade de graus centígrados escolhidas são arbitrárias, se estivéssemos medindo a temperatura em graus Fahrenheits não se observaria esta relação.

- Uma escala de razão podem ser utilizada para além de classificar e ordenar os resultados também estabelecer quão maior é um resultado que outro. A diferença com a escala intervalar é que agora existe um zero bem definido neste escala. A altura de um indivíduo, o tempo até ocorrência de um dado evento, o número de ocorrências de um dado evento em um dado intervalo de tempo são exemplos de medidas que utilizam uma escala de razão. Observe que se no caso em que temos dois equipamentos E1 e E2 com tempo de vida útil de 100h e 200h, respectivamente. É válido afirmar que o tempo de vida útil de E2 é o dobro do tempo de vida útil de E1.

8.1.2 Distribuições de Freqüências

Considere a seguinte tabela que contém informações sobre alguns empregados de uma companhia.

No.	Est. Civil	Grau de Instrução	No. de Filhos	Salário	Idade	Sexo
1	S	Médio	0	3	34	F
2	C	Superior	2	5	25	M
3	C	Fundamental	1	4	46	M
4	C	Fundamental	3	5,5	32	M
5	S	Médio	1	7,3	23	F
6	C	Médio	2	3,5	39	M
7	S	Superior	3	10	50	M
8	C	Médio	4	6	47	M
9	C	Médio	0	2	21	F
10	S	Médio	1	3,7	33	M

Uma maneira útil de se descrever os resultados das variáveis é através das medidas de freqüência, freqüência relativa (proporção), e porcentagem. Por exemplo, vamos considerar a variável Grau de Instrução na tabela anterior anterior. A freqüência de uma dada classe nada mais é do que o número de vezes que determinada classe ocorreu nos resultados do experimento. A freqüência relativa nada mais é que a proporção de vezes que dada classe ocorreu em relação ao número total de indivíduos que participaram do experimento. A porcentagem é igual a 100 vezes a freqüência relativa. A tabela abaixo é conhecida como tabela de freqüência para a variável Grau de Instrução.

Grau de Instrução	Freqüência (n_i)	Freqüência Relativa (f_i)	Porcentagem $100f_i$
Fundamental	2	0,2	20
Médio	6	0,6	60
Superior	2	0,2	20
Total	10	1	100

Quando desejamos comparar esta variável grau de instrução entre diferentes empresas, deve-se usar ou a freqüência relativa ou a porcentagem, pois possuem o mesmo total para qualquer empresa, enquanto o número total de empregados varia de empresa para empresa.

Em geral, quando construímos uma tabela de freqüência estamos interessados em resumir os resultados no que diz respeito a uma dada classe. No caso de uma variável quantitativa, às vezes se faz necessário que dividamos em intervalos os possíveis resultados do experimento para esta variável, pois caso contrário pode ocorrer que cada resultado ocorra somente um número pequeno de vezes e não se possa resumir a informação a respeito da dada variável. Esta situação ocorre freqüentemente no caso de variáveis que assumem valores reais. No nosso exemplo anterior, suponha que queiramos construir uma tabela de freqüência para a variável Salário. Neste caso, podemos considerar intervalos de tamanho 3 para construir a seguinte tabela:

Salário	Freqüência (n_i)	Freqüência Relativa (f_i)	Porcentagem $100f_i$
$[0, 3)$	1	0,1	10
$[3, 6)$	6	0,6	60
$[6, 9)$	2	0,2	20
$[9, 12)$	1	0,1	10
Total	10	1	100

A escolha dos intervalos acima em geral é arbitrária, dependendo do contexto cada profissional pode escolher um conjunto diferente de intervalos. A única restrição que tal escolha deve satisfazer é que estes intervalo sejam disjuntos e que cubram todos os valores que foram obtidos pela variável no experimento. Se escolhermos poucos intervalos, perdemos informação, pois note que a tabela só afirma que 6 pessoas têm salário entre 3 e 6 salários mínimos sem especificar qual o salário exato deles. Por outro lado, se escolhermos muitos intervalos, então nossa intenção de resumir os resultados do experimento não é cumprida. Em geral, recomenda-se o uso de 5 a 15 intervalos de comprimentos iguais.

8.1.3 Representação Gráfica

Variáveis Qualitativas

Existem vários tipos de gráficos para representar a distribuição dos dados de uma variável qualitativa. Os dois mais utilizados são: o *gráfico de barras* e o *gráfico de setores ou pizza*.

O gráfico de barras consiste em construir retângulos ou barras, uma para cada classe, em que uma das dimensões é proporcional à freqüência de ocorrência desta classe, e a outra dimensão é arbitrária porém igual para todas as barras. As barras são dispostas paralelamente umas às outras, horizontal ou verticalmente.

O gráfico de setores destina-se a representar a composição, usualmente em porcentagem, de partes de um todo. Consiste de um círculo de raio arbitrário, representando o todo, dividido em setores, sendo que cada setor corresponde a uma classe e tem área proporcional a freqüência relativa de ocorrência desta classe.

Variáveis Quantitativas

Para uma variável quantitativa discreta podemos também utilizar um gráfico de barras como no caso de variáveis quantitativas, onde agora temos uma barra para cada possível valor que a variável pode assumir. Também podemos considerar um gráfico de dispersão unidimensional onde desenhamos apenas pontos no plano cartesiano da forma (x_i, n_i), isto é, onde a abscissa do ponto é um possível valor da variável e a ordenada é a freqüência de ocorrência deste valor.

Uma outra alternativa de gráfico para variável quantitativa que é muito útil no caso de variáveis contínuas é conhecida como *histograma*.

Para a construção de um histograma, o primeiro passo é definir os intervalos contíguos e disjuntos que cubram todos os resultados observados. Uma vez definidos os intervalos, um histograma nada mais é do que um gráfico de barras contíguas, onde a base é proporcional ao comprimento do intervalo e a área da barra é proporcional a freqüência relativa de ocorrência de intervalos neste dado intervalo. Logo, se o i-ésimo intervalo tem comprimento Δ_i e a freqüência relativa de ocorrência de resultados neste intervalo é f_i, então a altura da barra deve ser proporcional a f_i/Δ_i, que é chamada de *densidade de freqüência* da i-ésima classe. Com essa convenção a área total do histograma deve ser proporcional a 1.

Exemplo 8.1.1: Duzentas baterias automotivas de uma dada marca foram testadas quanto a sua vida útil. Os resultados do teste em meses são reportados na tabela abaixo:

Durabilidade	Freq. Relativa
$0 \vdash 3$	0,02
$3 \vdash 6$	0,05
$6 \vdash 9$	0,15
$9 \vdash 12$	0,25
$12 \vdash 15$	0,30
$15 \vdash 18$	0,23

(a) Construa um histograma referente a tabela acima.

(b) Quantas baterias, em 1000 fabricadas, serão repostas pelo fabricante se a garantia for de 6 meses?

8.1.4 Medidas de Posição

Vimos como um resumo dos resultados (dados) através de uma tabela de freqüência pode ser útil para a descrição dos mesmos. Muitas vezes porém estaremos interessados em apenas um ou alguns valores que possam *representar* todos os resultados de uma dada variável. As medidas de posição mais utilizadas são: média (aritmética), mediana, ou moda.

A *moda* de uma variável é definida como sendo o seu resultado mais freqüente durante o experimento. Por exemplo, na tabela anterior, considere a variável número de filhos, vemos que 1 é a realização mais freqüente tendo ocorrido 3 vezes. A moda de uma variável não necessariamente é única, se houver empate entre a freqüência de ocorrência de mais de dos possíveis resultados, então todos estes serão moda da variável em questão. A moda não necessariamente é numérica, por exemplo, a moda da variável Grau de Instrução é médio, pois este é o grau de instrução mais freqüente entre os funcionários da companhia.

A média (aritmética) de uma variável é a soma dos seus resultados divididos pelo número total de resultados obtidos. Portanto, a média aritmética da variável número de filhos é: 17/10. Note que apenas faz sentido calcular média de variáveis quantitativas.

A *mediana* é o resultado que ocupa a posição central da série de observações, quando estes estão ordenados em ordem crescente. Quando o número de observações for par e a variável for quantitativa, usa-se a média aritmética das duas observações centrais como sendo a mediana. Por exemplo, considere a variável salário as observações desta variável foram: 2, 3, 3,5, 3,7, 4, 5, 5,5, 6, 7,3, 10. As duas observações centrais são 4 e 5, logo a mediana desta variável é 4,5. Quando o número de observações for par e a variável for ordinal, define-se ambas as classes das duas observações centrais e todas as outras classes entre elas como sendo medianas. Portanto, podemos definir a mediana para qualquer variável ordinal ou quantitativa.

Note que a presença de valores extremos ou muito pequenos ou muito grandes em comparação com os demais valores de uma variável, alteram significativamente sua média. Por outro lado, o valor da mediana não se altera muito com a presença destes valores extremos e por isso às vezes é mais recomendado o uso da mediana para representar a posição de uma variável. Por exemplo, se o indivíduo que ganha 10 salários mínimos passasse a ganhar 100, a média dos salários passaria de 5 para 14, enquanto a mediana permaneceria igual a 4,5.

A determinação de medidas de posição para uma variável quantitativa contínua, através de sua distribuição de freqüências, exige aproximações, pois perdemos as informações dos valores das observações. Para o cálculo da média, uma aproximação razoável é supor que todos os valores dentro de uma classe tenham seus valores iguais ao ponto médio desta classe, o que nos deixa na mesma situação para utilizarmos o procedimento anterior. A moda é definida como sendo o ponto médio da classe de maior freqüência. No caso da mediana, podemos ainda obter uma estimativa mais aproximada considerando que as ocorrências em cada classe são uniformemente distribuídas, e deste modo calculando a mediana através de uma proporção, como no exemplo a seguir.

Exemplo 8.1.2: O número de divórcios na cidade, de acordo com a duração do casamento,

está representado na tabela abaixo:

Anos de Casamento	No. de Divórcios
$0 \vdash 6$	2.800
$6 \vdash 12$	1.400
$12 \vdash 18$	600
$18 \vdash 24$	150
$24 \vdash 30$	50

(a) Construa o histograma da distribuição.

(b) Qual a duração média dos casamentos? E a mediana? E a moda?

Solução: (a) Para a construção do histograma note que a freqüência relativa das 5 classes são, respectivamente: 14/25, 7/25, 3/25, 3/100, e 1/100. Como cada classe tem comprimento 6, a altura de cada barra no histograma deve ser 1/6 da freqüência relativa da sua classe correspondente.

(b) Para o cálculo da média, devemos utilizar a aproximação de cada classe pelo seu ponto médio deste modo:

$$\overline{x} = (3 \times (14/25)) + (9 \times (7/25)) + (15 \times (3/25)) + (21 \times (3/100)) + (27 \times (1/100)) = 6{,}9.$$

Para o cálculo da mediana observe que a primeira classe já contém mais de 50% das observações, como a mediana deve ser o valor para o qual 50% dos valores são menores que a mediana, então podemos determinar a mediana através de uma proporção: o intervalo de 0 a 6 contém 56% das observações, o intervalo de 0 até a mediana deve conter 50%, então

$$\frac{md}{50} = \frac{6}{56},$$

portanto, $md = 5{,}36$. Como a classe de maior freqüência é a primeira, temos que a moda é igual a 3. ∎

8.1.5 Medidas de Dispersão

As medidas de posição que vimos na seção anterior, nos dão informação sobre a posição central dos resultados mas não nos fornecem nenhuma informação sobre a variabilidade dos resultados. Para tanto, precisamos de medidas de dispersão. Por exemplo, considere dois grupos de resultados de uma certa variável: Grupo 1 - 3,4,5,6,7; e Grupo 2 - 1,3,5,7,9. Ambos grupos possuem a mesma média e mediana que é igual a 5, porém os resultados do Grupo 1 estão mais aglutinados ao redor deste valor. Medidas de dispersão são utilizadas para mensurar esta variabilidade. As duas medidas de dispersão mais utilizadas são: desvio médio e variância. Estas medidas analisam quão distante da média estão os resultados.

Seja $\overline{x}$ a média dos resultados do experimento. Para cada valor x_i do resultado podemos definir a distância entre x_i e $\overline{x}$ de diversas maneiras. O desvio médio é calculado considerando distância como sendo o valor absoluto da diferença entre x_i e $\overline{x}$. Formalmente, temos

$$dm(X) = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}.$$

Logo, para o Grupo 1, o desvio médio é $\frac{2+1+0+1+2}{5} = 1{,}2$. Para o Grupo 2, o desvio médio é $\frac{4+2+0+2+4}{5} = 2,4$.

A variância, por sua vez, é calculada considerando como distância o quadrado da diferença entre x_i e $\overline{x}$. Logo, temos que

$$var(X) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2)}{n} \\ = \frac{\sum_{i=1}^{n} x_i^2}{n} - 2\overline{x}\frac{\sum_{i=1}^{n} x_i}{n} + \overline{x}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2. \tag{8.1}$$

Logo, para o Grupo 1, a variância é $\frac{4+1+0+1+4}{5} = 2$. Para o Grupo 2, a variância é $\frac{16+4+0+4+16}{5} = 8$. Como a variância é uma medida de dimensão igual ao quadrado da dimensão dos resultados, é freqüente usar a medida do *desvio padrão*, que é igual a raiz quadrada da variância, como medida de dispersão. Assim como a média, as medidas de dispersão são afetadas de forma excessiva por valores extremos.

No caso de variáveis contínuas descritas através de sua distribuição de freqüências, para o cálculo de medidas de dispersão também devemos aproximar cada classe pelo seu ponto médio e proceder como anteriormente.

Exemplo 8.1.3: Determine a variância do número de divórcios do Exemplo 8.1.2.

Solução: Devemos aproximar cada classe pelo valor do seu ponto médio, então:

$varX = (9\times(14/25))+(81\times(7/25))+(225\times(3/25))+(441\times(3/100))+(729\times(1/100))-(6{,}9^2) = 27{,}63.$

∎

8.1.6 Quantis

Apenas a informação da medida de posição e de dispersão não nos dão informação a respeito da simetria ou assimetria da distribuição dos resultados. Os quantis são medidas que servem para informar a este respeito. Vimos que a mediana é uma medida tal que metade dos resultados são menores e a outra metade é maior que a mediana. Analogamente, podemos definir um *quantil de ordem p* ou *p-quantil*, indicado por $q(p)$, onde p é uma proporção qualquer, $0 < p < 1$, tal que $100p\%$ dos resultados sejam menores que $q(p)$. Existem alguns quantis que são usados mais freqüentemente e recebem nomes particulares: $q(0{,}25)$ é o 1o. quartil ou 25o. percentil; $q(0{,}5)$ é a mediana, 5o. decil, ou 50o. percentil; $q(0{,}75)$ é o terceiro quartil ou 75o. percentil; e $q(0{,}95)$ é o 95o. percentil.

Por exemplo, se temos uma coleção de n resultados, como deveríamos definir $q(1/n)$? Seja $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ uma reordenação dos resultados em ordem crescente, conhecida como *estatística de ordem* dos resultados. Então, em analogia com a definição da mediana, o quantil $q(1/n)$ é definido como sendo a média aritmética entre $x_{(1)}$ e $x_{(2)}$, de modo que exatamente $100/n\%$ dos resultados são menores que $q(1/n)$. Similarmente, o quantil $q(2/n)$ é definido como sendo a média aritmética entre $x_{(2)}$ e $x_{(3)}$. Mas neste caso como $q(1/n) \leq x_{(2)} \leq q(2/n)$, o resultado $x_{(2)}$ deve corresponder a um quantil $q(p)$, onde $\frac{1}{n} < p < \frac{2}{n}$. Para a definição formal dos quantis assume-se linearidade entre os quantis da forma $q(m/n)$, para $m \leq n$.

Então, como $x_{(2)} = \frac{q(1/n)+q(2/n)}{2}$, $x_{(2)}$ é igual ao quantil $q(\frac{\frac{1}{n}+\frac{2}{n}}{2}) = q(\frac{3}{2n})$. Em geral, seguindo o mesmo argumento, $x_{(i)}$ é igual ao quantil $q(\frac{\frac{i-1}{n}+\frac{i}{n}}{2}) = q(\frac{2i-1}{2n}) = q(\frac{i-0,5}{n})$, para $i = 1, 2, \ldots, n$.

Contudo, dependendo do valor de p, precisamos ter cuidado ao definir o quantil. Se $p < \frac{1}{2n}$, como $x_{(1)}$ é o menor valor observado dos resultados e é igual ao quantil $q(\frac{1}{2n})$, define-se $q(p)$ como sendo igual a $x_{(1)}$. Similarmente, se $p > \frac{2n-1}{2n}$, como $x_{(n)}$ é o maior valor observado dos resultados e é igual ao quantil $q(\frac{n-0,5}{n})$, define-se $q(p)$ como sendo igual a $x_{(n)}$. Finalmente, se $p = \alpha\frac{2(i-1)-1}{2n} + (1-\alpha)\frac{2i-1}{2n}$, onde $0 < \alpha < 1$, então define-se $q(p)$ como sendo igual a $\alpha x_{(i-1)} + (1-\alpha)x_{(i)}$.

Resumindo, temos que

$$q(p) = \begin{cases} x_{(1)} & \text{, se } p < \frac{1}{2n}, \\ x_{(n)} & \text{, se } p > \frac{2n-1}{2n}, \\ x_{(i)} & \text{, se } p = \frac{2i-1}{2n}, \\ \alpha x_{(i-1)} + (1-\alpha)x_{(i)} & \text{, se } p = \alpha\frac{2(i-1)-1}{2n} + (1-\alpha)\frac{2i-1}{2n}, \text{ onde } 0 < \alpha < 1. \end{cases}$$

Exemplo 8.1.4: Considere os resultados de um teste foram: 3,4,5,6, e 7. Vamos determinar (a) $q(0{,}05)$, (b) $q(0{,}25)$, e (c) $q(0{,}75)$.

Solução: Para (a), como $0,05 < \frac{1}{10}$, temos que $q(0{,}05) = 3$. Para (b), note que $0{,}25 = \alpha(0{,}1)+(1-\alpha)0{,}3$, se $\alpha = 1/4$. Portanto, $q(0{,}25) = (1/4)3+(3/4)4 = 15/4$. Finalmente, para (c), note que $0{,}75 = \alpha(0{,}7) + (1-\alpha)0{,}9$, se $\alpha = 3/4$. Portanto, $q(0{,}75) = (3/4)6 + (1/4)7 = 25/4$. ∎

Uma medida de dispersão alternativa é a *distância interquartil*, d_q, definida como sendo a diferença entre o terceiro e o primeiro quartil, isto é, $d_q = q(0{,}75) - q(0{,}25)$.

Os cinco valores $x_{(1)}$, $q(0{,}25)$, $q(0{,}5)$, $q(0{,}75)$, e $x_{(n)}$ são importantes para se ter uma idéia a respeito da assimetria da distribuição dos dados. Para se ter uma distribuição aproximadamente simétrica, precisamos ter:

(a) $q(0{,}5) - x_{(1)} \simeq x_{(n)} - q(0{,}5)$;

(b) $q(0{,}5) - q(0{,}25) \simeq q(0{,}75) - q(0{,}5)$; e

(c) $q(0{,}25) - x_{(1)} \simeq x_{(n)} - q(0{,}75)$.

Exemplo 8.1.5: O serviço de atendimento ao consumidor de uma concessionária de veículos recebe as reclamações dos clientes. Tendo em vista a melhoria na qualidade do atendimento foram anotados o número de reclamações diárias nos últimos 30 dias: 4, 5, 3, 4, 2, 6, 4, 1, 6, 5, 3, 4, 4, 5, 2, 3, 6, 5, 4, 2, 2, 3, 4, 3, 3, 2, 1, 1, 5, e 2.

(a) Faça uma tabela de freqüências desses dados.

(b) Determine o valor da média, moda, mediana, desvio padrão, e do 1o. e 3o. quartis desta distribuição de dados.

(c) Com base nos valores obtidos na letra (b), você diria que esta é uma distribuição simétrica de dados?

Solução: A tabela de freqüência dos dados é dada por:

No. de Reclamações	Freq. Relativa
1	3/30
2	6/30
3	6/30
4	7/30
5	5/30
6	3/30

A média dos dados é dada por:

$$\overline{x} = (1 \times 3/30) + (2 \times 6/30) + (3 \times 6/30) + (4 \times 7/30) + (5 \times 5/30) + (6 \times 3/30) \simeq 3{,}47.$$

A moda é igual a 4. A mediana é dada por 3,5. A variância é dada por:

$$\sigma^2 = (1\times 3/30)+(4\times 6/30)+(9\times 6/30)+(16\times 7/30)+(25\times 5/30)+(36\times 3/30)-3{,}47^2 \simeq 2{,}16.$$

Logo, o desvio padrão é igual aproximadamente a 1,47. O primeiro quartil é dado por $x_{(8)} = 2$, e o terceiro quartil é dado por $x_{(23)} = 5$. Com estes resultados podemos observar que

(a) $q(0{,}5) - x_{(1)} = 2{,}5 = x_{(n)} - q(0{,}5)$;

(b) $q(0{,}5) - q(0{,}25) = 1{,}5 = q(0{,}75) - q(0{,}5)$; e

(c) $q(0{,}25) - x_{(1)}1 = x_{(n)} - q(0{,}75)$.

Logo, podemos concluir que estes dados formam uma distribuição simétrica. ∎

No caso de variáveis contínuas descritas através de sua distribuição de freqüências, para o cálculo dos quantis utilizamos uma metodologia similar a do cálculo da mediana, sendo que agora $q(p)$, $0 < p < 1$, é calculado através de uma proporção de forma que $p\%$ da área do histograma esteja antes de $q(p)$ e $(1-p)\%$ esteja após $q(p)$, como no seguinte exemplo.

Exemplo 8.1.6: O número de divórcios na cidade, de acordo com a duração do casamento, está representado na tabela abaixo:

Anos de Casamento	No. de Divórcios
$0 \vdash 6$	2.800
$6 \vdash 12$	1.400
$12 \vdash 18$	600
$18 \vdash 24$	150
$24 \vdash 30$	50

(a) Encontre o 1o. e o 9o. decis.

(b) Qual o intervalo interquartil?

Solução: (a) Podemos encontrar o primeiro decil através de uma proporção, pois a primeira classe contém 56% das observações, então

$$\frac{q(0{,}1)}{10} = \frac{6}{56},$$

logo, $q(0{,}1) = 1{,}07$. Para o nono decil note que as duas primeiras classes contém 84% das observações, e as três primeiras contém 96% das observações, então o nono decil deve estar na terceira classe e podemos determiná-lo também por uma proporção:

$$\frac{q(0{,}9) - 12}{6} = \frac{6}{12},$$

logo $q(0{,}9) = 15$.

Para obtermos o intervalo interquartil, precisamos encontrar o primeiro e o terceiro quartil que podemos obter de maneira similar a parte (a).

$$\frac{q(0{,}25)}{25} = \frac{6}{56},$$

logo, $q(0{,}25) = 2{,}68$. O terceiro quartil deve estar na segunda classe, então como a primeira claase já contém 56% das observações:

$$\frac{q(0{,}75) - 6}{19} = \frac{6}{28},$$

logo, $q(0{,}75) = 10{,}07$. Portanto, o intervalo interquartil é $[2{,}68; 10{,}07]$. ∎

Capítulo 9

Distribuições Amostrais

9.1 Introdução

Quando vamos aplicar modelos probabilísticos em algum problema prático, precisamos ter uma informação a respeito da distribuição de probabilidade da variável aleatória de interesse. Existem dois processos clássicos para a obtenção da distribuição de uma variável aleatória: eduzir uma distribuição a priori de um especialista da área, ou inferir a distribuição a partir de uma análise de dados. Neste curso, não trataremos de métodos de edução, mas nos concentraremos em métodos de inferência.

9.2 População e Amostra

Suponha que estivéssemos interessados na distribuição do consumo mensal de energia elétrica de todos os domicílios brasileiros. Caso tivéssemos meios de obter os valores para todos os domicílios, poderíamos obter sua distribuição exata e daí calcular parâmetros de posição e dispersão, por exemplo. Nesse caso, não necessitaríamos de inferência estatística, pois teríamos acesso a todos os valores de interesse.

Porém, é raro a situação em que se consegue obter a distribuição exata de alguma variável, ou porque os custos são muito elevados, ou o tempo para a coleta de tais dados é muito longo, ou porque às vezes o experimento aleatório que se realiza consiste de um processo destrutivo. Por exemplo, poderíamos estar interessados em medir a tensão máxima de entrada que um determinado tipo de estabilizador suporta. Nosso experimento poderia começar com tensão de 0 volts e ir aumentado gradativamente e definiríamos a tensão máxima como a tensão a partir da qual o estabilizador queimou. Deste modo se fôssemos testar todos os estabilizadores, não restaria nenhum para ser vendido. Assim a solução é selecionar parte dos estabilizadores (amostra), analisá-la e inferir propriedades para todos os estabilizadores (população). Esta questão dentre outras é objeto de estudo da área de *inferência estatística*.

Definição 9.2.1: *População* é o conjunto de todos os elementos ou resultados sob investigação. *Amostra* é um subconjunto formado por elementos selecionados da população.

Freqüentemente, usa-se uma distribuição de probabilidades como um modelo para uma população. Por exemplo, um engenheiro de estruturas pode considerar como normalmente

distribuída, com média μ e variância σ^2 desconhecidas, a população de resistências a tração de um elemento estrutural de um chassi. Usualmente se refere a isso como uma população normal ou uma população distribuída normalmente.

Para outro exemplo, suponha que estejamos interessados em investigar se uma dada moeda é honesta e para isso nós lançamos a moeda 50 vezes. Neste caso, a população pode ser considerada como sendo a distribuição de uma variável aleatória X que assume o valor 1, se ocorrer cara, e 0 em caso contrário, e tem distribuição Bernoulli com parâmetro p desconhecido. A amostra será a seqüencia binária de comprimento 50.

Observe que neste dois últimos caso a população foi especificada como sendo uma distribuição de uma variável aleatória X que modela a característica de interesse. Este artifício exige a proposta de um modelo para a variável X. Neste caso, é comum usar expressões "a população $f(x)$" ou "a população das resistências $X \sim N(\mu, \sigma^2)$".

9.3 Seleção de uma Amostra

A fim de obtermos inferências realmente informativas a respeito de uma dada população, precisa-se de cuidado com os métodos de seleção de uma amostra; é necessário que a amostra seja *representativa* da população. Por exemplo, ao se fazer uma pesquisa de opinião pública a respeito de um dado governo, se escolhêssemos só pessoas que vivem em uma dada região beneficiada por esse governo, a amostra pode não ser representativa de toda a população, pois esta contém pessoas que não necessariamente foram diretamente beneficiadas pelo governo, neste caso diz-se que a amostra é *viesada*. Neste curso, iremos apenas analisar o caso de *amostragem aleatória simples*.

9.3.1 Amostra Aleatória Simples

Este procedimento é o método mais simples de selecionarmos uma amostra aleatória de uma população e serve de base para outros métodos de amostragem mais complexos. No caso de uma população finita, poderemos implementar este método numerando os elementos da população e em seguida escolher um número ou olhando uma tabela de números aleatórios ou gerando números aleatórios em um computador. Neste caso, todos os elementos da população têm a mesma probabilidade de ser selecionados. Repete-se o processo até que n elementos sejam selecionadas. Teremos uma amostragem *com reposição*, se for permitido que uma unidade possa ser sorteada mais de uma vez, e *sem reposição*, se o elemento for removido da população. Do ponto de vista da quantidade de informação contida na amostra, amostrar sem reposição é mais adequado. Contudo, a amostragem com reposição, implica que tenhamos independência entre os elementos selecionados e isto facilita o desenvolvimento de propriedades de estimadores, conforme veremos adiante. Portanto, nos restringiremos ao caso com reposição. Em geral, temos a seguinte definição de amostra aleatória simples:

Definição 9.3.1: Uma amostra aleatória simples de tamanho n de uma população modelada por uma variável aleatória X, com uma dada distribuição, é um conjunto de n variáveis aleatórias independentes $X_1, X_2, \ldots, X_n$, cada uma com a mesma distribuição de X.

Intuitivamente, X_i representa a observação do i-ésimo elemento sorteado. Portanto, no caso de uma população X contínua, com função densidade de probabilidade f, a função densidade de probabilidade conjunta da amostra $(X_1, X_2, \ldots, X_n)$, será dada por:

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

Quando geramos números números aleatórios em um programa, sabemos forma da distribuição da variável aleatória que estamos simulando, e os parâmetros que estamos simulando. Por exemplo, ao gerarmos 50 números de uma distribuição normal padrão, estamos obtendo uma amostra aleatória simples de tamanho 50 desta população normal. Se outra pessoa observa apenas estes 50 números gerados, ela não conhecerá nada a respeito da distribuição que se gerou nem dos parâmetros dessa distribuição que foram utilizados. O objetivo da inferência estatística é fornecer critérios para que se possa descobrir a forma da distribuição e/ou os parâmetros da população que gerou a amostra que se observa.

9.4 Estatísticas e Parâmetros

Uma vez obtida a amostra de uma dada população, muitas vezes estaremos interessados em calcular alguma função desta amostra. Por exemplo, a média da amostra $(X_1, X_2, \ldots, X_n)$ é dada por

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Como $\overline{X}$ é uma função contínua de variáveis aleatórias, ela também é uma variável aleatória.

Definição 9.4.1: Uma estatística T é uma função de uma amostra $X_1, X_2, \ldots, X_n$.

As estatísticas mais comuns são:

1. Média da amostra: $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$;
2. Variância da amostra: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$;
3. O menor valor da amostra: $X_{(1)} = \min(X_1, X_2, \ldots, X_n)$;
4. O maior valor da amostra: $X_{(n)} = \max(X_1, X_2, \ldots, X_n)$;
5. Amplitude amostral: $W = X_{(n)} - X_{(1)}$;
6. A i-ésima maior observação da amostra: $X_{(i)}$.[1]

Para diferenciar características da amostra com características da população, chama-se de *parâmetro* uma medida usada para descrever uma característica da população. Assim se uma população for modelada por uma variável aleatória X, a esperança e a variância EX e $VarX$, respectivamente, seriam parâmetros.

[1] Os elementos da amostra ordenados, isto é, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, são conhecidos como estatísticas de ordem da amostra.

9.5 Distribuições Amostrais

Suponha que estejamos interessados em algum parâmetro θ da população. Suponha que decidimos usar uma estatística T de uma amostra aleatória simples $X_1, X_2, \ldots, X_n$ da população. Uma vez que a amostragem é realizada, pode-se calcular que $T = t_0$ e é baseado neste valor que faremos uma afirmação sobre θ. Como vimos T sendo uma função de variáveis aleatórias também é uma variável aleatória e, portanto, possui uma dada distribuição. Esta distribuição é conhecida como *distribuição amostral da estatística* T.

Exemplo 9.5.1: Suponha que retiramos com reposição todas as amostras de tamanho 2 da população $\{1, 3, 5, 5, 7\}$. A distribuição conjunta da amostra (X_1, X_2) é dada por:

	1	3	5	7
1	1/25	1/25	2/25	1/25
3	1/25	1/25	2/25	1/25
5	2/25	2/25	4/25	2/25
7	1/25	1/25	2/25	1/25

Vamos calcular então a distribuição da média amostral. Por exemplo, $P(\overline{X} = 3) = P(X_1 = 1, X_2 = 5) + P(X_1 = X_2 = 3) + P(X_1 = 5, X_2 = 1) = 5/25$. Similarmente, os demais valores podem ser obtidos conforme tabela a seguir:

$\overline{x}$	1	2	3	4	5	6	7
$P(\overline{X} = \overline{x})$	1/25	2/25	5/25	6/25	6/25	4/25	1/25

Exemplo 9.5.2: No caso do lançamento de uma moeda 50 vezes, usando como estatística X o número de caras obtidas, a distribuição amostral desta estatística é uma binomial com parâmetros $n = 50$ e p, onde p é a probabilidade de cara em um lançamento qualquer desta moeda. Se estivermos interessados em saber se esta moeda é honesta, ou seja, em checar se $p = 0{,}5$, e soubermos que em 50 lançamentos ocorreram 36 caras podemos calcular que $P_{0,5}(X \geq 36) = 0,0013$, ou seja, se a moeda for honesta, então a probabilidade de se obterem 36 ou mais caras é igual a 0,0013, então existe evidência que p deve ser diferente de 0,5. Por outro lado, se obtivermos 29 caras, obtemos que $P_{0,5}(X \geq 29) = 0,1611$, então se a moeda for honesta aproximadamente 1/6 das vezes observa-se um valor maior ou igual a 29, então não temos dados suficientes para descartar a hipótese que a moeda seja honesta neste caso.

Exemplo 9.5.3: Uma população consiste de quatro números 1, 3, 5, e 7. Considere todas as possíveis amostras de tamanho 2 de elementos que podem ser selecionadas com reposição desta população. Determine.

(a) A média e variância populacional.

(b) A distribuição da média e variância amostrais.

Solução: A média populacional é dada por: $\mu = \frac{1+3+5+7}{4} = 4$, e a variância populacional é dada por $\sigma^2 = \frac{1^2+3^2+5^2+7^2}{4} - 4^2 = 5$. Para determinarmos a média a variância amostrais, considere a seguinte tabela onde todos as possíveis amostras estão enumeradas:

x_1	x_2	$\overline{x}$	s^2
1	1	1	0
1	3	2	2
1	5	3	8
1	7	4	18
3	1	2	2
3	3	3	0
3	5	4	2
3	7	5	8
5	1	3	8
5	3	4	2
5	5	5	0
5	7	6	2
7	1	4	18
7	3	5	8
7	5	6	2
7	7	7	0

Como cada uma das possíveis amostrais tem probabilidade 1/16, temos que a distribuição da média amostral e da variância amostral são respectivamente descritas pelas tabelas a seguir:

$\overline{x}$	1	2	3	4	5	6	7
$P(\overline{X} = \overline{x})$	1/16	2/16	3/16	4/16	3/16	2/16	1/16

e

s^2	0	2	8	18
$P(S^2 = s^2)$	4/16	6/16	4/16	2/16

Para algumas estatísticas não conseguiremos obter analiticamente sua distribuição amostral, então podemos simular um número grande de amostras diferentes e calcular a estatísticas de cada uma dessas amostras para obter uma distribuição amostral empírica da estatística de interesse. Por exemplo, para obter a mediana das alturas de amostras de 5 mulheres retiradas da população $X \sim N(167, 25)$, podemos gerar, via qualquer software, 200 amostras de tamanho 5 desta população, determinar a mediana de cada uma dessas amostras e calcular medidas de posição e dispersão dos valores das medianas obtidos com essas amostras, bem como representação gráficas destes valores.

9.5.1 Distribuição Amostral da Média Amostral

Vamos agora estudar a distribuição amostral da média amostral $\overline{X}$. Antes de obtermos informações sobre a forma desta distribuição, podemos determinar a esperança e a variância da distribuição amostral de $\overline{X}$.

Teorema 9.5.4: *Seja X uma variável aleatória com média μ e variância σ^2, e seja $(X_1, X_2, \ldots, X_n)$ uma amostra aleatória simples de X. Então,*

$$E(\overline{X}) = \mu \text{ e } Var(\overline{X}) = \frac{\sigma^2}{n}.$$

Prova: Pela linearidade da esperança, temos:

$$E(\overline{X}) = \frac{1}{n}(EX_1 + EX_2 + \cdots + EX_n) = \mu.$$

Como $X_1, X_2, \ldots, X_n$ são independentes, temos

$$Var(\overline{X}) = \frac{1}{n^2}(VarX_1 + VarX_2 + \cdots + VarX_n) = \frac{\sigma^2}{n}.$$

∎

Note que conforme n vai aumentando a distribuição de $\overline{X}$ tende a ficar mais concentrada em torno de sua média μ, pois sua variância vai diminuindo. Além disso, o próximo teorema nos dá uma informação mais detalhada para a distribuição amostral da média para valores grandes de n. Este teorema é conhecido como *Teorema do Limite Central.*

Teorema 9.5.5: *Para amostras aleatórias simples $(X_1, X_2, \ldots, X_n)$, retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média $\overline{X}$ aproxima-se, para n grande, de uma distribuição normal, com média μ e variância σ^2/n., ou seja, se $F_{\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}}$ for a função de distribuição acumulada de $\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}$, temos que $\forall x \in \mathbb{R}$*

$$\lim_n F_{\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}}(x) = \Phi(x).$$

Prova: A prova deste teorema está fora do escopo deste curso. ∎

Caso a população já possua uma distribuição normal, então como $\overline{X}$ é uma combinação linear de $X_1, X_2, \ldots, X_n$ que são independentes e possuem distribuição normal, então a distribuição amostral da média amostral será exatamente uma normal para qualquer valor de n, e a média dessa distribuição será igual a média da população e variância será igual a variância da população dividida por n.

Em geral o TLC afirma que para valores grandes de n, $\overline{X}$ terá uma distribuição aproximadamente normal, a velocidade desta convergência depende da distribuição da população. Se esta for próxima da normal, a convergência é mais rápida; se for muito diferente a convergência é mais lenta. Como regra empírica, para amostras de tamanhos de 30 elementos, a aproximação já pode ser considerada boa.

A diferença entre a média amostral e a média da população é conhecida como *erro amostral da média*, isto é, $e = \overline{X} - \mu$. A partir do Teorema do Limite Central, podemos obter que $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1)$, ou seja, $\frac{\sqrt{n}e}{\sigma} \sim N(0,1)$.

Exemplo 9.5.6: Suponha que uma máquina está regulada para produzir lâmpadas com tempo de vida útil médio de 10.000horas. De uma amostra de 50 lâmpadas produzidas por esta máquina, verifica-se o tempo de vida útil de cada uma delas. Determine a probabilidade de que o tempo de vida útil médio seja menor ou igual a 8.000horas.

Solução: Sabe-se que o tempo de vida útil de uma lâmpada é distribuído de acordo com uma Exponencial. Portanto, como o tempo de vida útil médio é de 10.000horas, temos que

a média populacional é 10.000horas e a variância populacional é igual a 10^8horas^2. Além disso, como temos uma amostra maior que 30, podemos utilizar o TCL para afirmar que a média amostral tem uma distribuição $N(10^4, \frac{10^8}{50})$. Portanto,

$$P(\overline{X} \leq 8000) = P(Z \leq \frac{\sqrt{50}(8000 - 10000)}{10000}) = \Phi(-\sqrt{2}) = 0{,}0793.$$

∎

9.5.2 Distribuição Amostral de uma Proporção

Vamos supor que a proporção de indivíduos de uma população que são portadores de uma determinada característica seja igual a p. Logo, pode-se definir uma variável aleatória X que assume o valor um se o indivíduo possui a característica e o valor 0, em caso contrário. Portanto, X tem uma distribuição Bernoulli de parâmetro p. Considere agora que escolhemos uma amostra aleatória simples de tamanho n desta população e seja Y_n o número total de indivíduos na amostra que possuem a característica de interesse. Então, temos que Y_n tem uma distribuição binomial com parâmetros n e p. A proporção de indivíduos portadores da característica é dada por

$$\hat{p} = \frac{Y_n}{n}.$$

Portanto, podemos determinar a distribuição de $\hat{p}$ a partir da distribuição de Y_n, utilizando a seguinte relação: $P(\hat{p} = \frac{k}{n}) = P(Y_n = k)$.

Pelo Teorema Central do Limite se $X_1, X_2, \ldots, X_n$ formam uma amostra aleatória simples desta população, a distribuição amostral de $\overline{X}$ é aproximadamente igual a $N(p, p(1-p)/n)$ para valores grandes de n. Portanto, a distribuição de $Y_n = n\overline{X}$ pode ser aproximada por uma normal $N(np, np(1-p))$. Como $\hat{p} = \overline{X}$, temos que a distribuição da proporção amostral também pode ser aproximada por $N(p, p(1-p)/n)$ para valores grandes de n.

Exemplo 9.5.7: Suponha que uma máquina está regulada para produzir lâmpadas de modo que 10% delas tenham tempo de vida útil menor ou igual a 1.000horas. De uma amostra de 50 lâmpadas produzidas por esta máquina, qual a probabilidades de encontrarmos no máximo 90% com tempo de vida útil maior que 1.000 horas.

Solução: Como temos uma amostra maior que 30, podemos utilizar o TCL para afirmar que a proporção amostral tem uma distribuição $N(0{,}1, \frac{(0{,}1)(0{,}9)}{50})$. Portanto,

$$P(1 - \hat{p} \leq 0{,}9) = P(\hat{p} \geq 0{,}1) = P(Z \geq 0) = 0{,}5.$$

∎

9.6 Determinação do Tamanho de uma Amostra

Em certas situações estamos interessados em determinar o tamanho de uma amostra que selecionaremos de uma população de modo a obter um erro de estimação previamente estipulado, com certo grau de confiança. Por exemplo, suponha que iremos estimar a média

populacional μ através da média amostral $\overline{X}$ de uma amostra de tamanho n. Nosso objetivo é então determinar o menor valor de n tal que

$$P(|\overline{X} - \mu| \leq \epsilon) \geq \gamma,$$

onde γ representa o grau de confiança necessário para que o erro amostral seja no máximo igual a ϵ. Como a distribuição amostral de $\overline{X}$ é $N(\mu, \frac{\sigma^2}{n})$, temos que o tamanho mínimo da amostra n tem que satisfazer

$$P(-\epsilon \leq \overline{X} - \mu \leq \epsilon) = P(\frac{-\sqrt{n}\epsilon}{\sigma} \leq Z \leq \frac{\sqrt{n}\epsilon}{\sigma}) = \gamma,$$

onde Z tem uma distribuição normal padrão. Da distribuição normal padrão, temos que

$$P(-\Phi^{-1}(\frac{\gamma+1}{2}) \leq Z \leq \Phi^{-1}(\frac{\gamma+1}{2})) = P(Z \leq \Phi^{-1}(\frac{\gamma+1}{2})) - P(Z < -\Phi^{-1}(\frac{\gamma+1}{2}))$$
$$= \frac{\gamma+1}{2} - (1 - \frac{\gamma+1}{2}) = \gamma.$$

Portanto,

$$\frac{\sqrt{n}\epsilon}{\sigma} = \Phi^{-1}(\frac{\gamma+1}{2}), \text{ ou seja,}$$
$$n = \frac{\sigma^2(\Phi^{-1}(\frac{\gamma+1}{2}))^2}{\epsilon^2}.$$

Note que o tamanho da amostra depende da variância da população. Como era de se esperar, quanto mais variabilidade tiver a população, mais amostras serão necessárias para que se possa fazer afirmações confiáveis a respeito dos erros dos estimadores. Contudo, em geral o valor da variância da população é desconhecido. Na prática, pode-se fazer um projeto piloto para que se possa estimar o valor desta variância e, em seguida, usá-la para determinar o tamanho de amostra do estudo principal.

No caso de proporções, como neste caso, $\sigma = p(1-p)$, temos que

$$n = \frac{(\Phi^{-1}(\frac{\gamma+1}{2}))^2 p(1-p)}{\epsilon^2}.$$

Como na prática, na maioria dos casos não se conhece o verdadeiro valor da proporção populacional p, pode-se usar o fato que $p(1-p) \leq \frac{1}{4}$, para obtermos que

$$n = \frac{(\Phi^{-1}(\frac{\gamma+1}{2}))^2}{4\epsilon^2}.$$

Exemplo 9.6.1: Uma variável aleatória X tem distribuição amostral $N(3, 2^2)$. Qual deve ser o tamanho n de uma amostra aleatória de X para que a média amostral $\overline{X}$ tenha 84,13% dos valores menores que 3,4?

Solução: Queremos que $P(\overline{X} \leq 3{,}4) = 0{,}8413$. Portanto, como $\frac{\sqrt{n}(\overline{X}-3)}{2}$ tem distribuição normal padrão, temos que

$$0{,}8413 = P(\overline{X} \leq 3{,}4) = P(Z \leq \frac{\sqrt{n}(3{,}4-3)}{2}).$$

Logo, $\sqrt{n} = \frac{2\Phi^{-1}(0{,}8413)}{0{,}4} = 5$, ou seja, $n = 25$.

Capítulo 10

Estimação

10.1 Estimativas e Estimadores

Uma aplicação muito importante de estatísticas é a obtenção de *estimativas* dos parâmetros da população, tais como média e variância da população. O objetivo da estimação é selecionar um único número baseado nos dados da amostra, sendo esse número o mais plausível para um parâmetro θ. Em geral, se X for uma variável aleatória com distribuição de probabilidades caracterizada por um parâmetro desconhecido θ, e se $X_1, X_2, \ldots, X_n$ for uma amostra aleatória de tamanho n de X, então a estatística $\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$ é chamada de um estimador de θ. Note que depois da amostra ter sido selecionada, $\hat{\Theta}$ assume um valor $\hat{E}$, chamado *estimativa* de θ. Portanto, uma estimativa pontual de algum parâmetro θ da população é um único valor numérico $\hat{E}$ de uma estatística $\hat{\Theta}$.

Problemas de estimação ocorrem freqüentemente, os parâmetros mais comuns que se desejam estimar são:

- A média de uma única população.
- A variância σ^2 (ou desvio-padrão σ) de uma única população.
- A proporção p de ítens em uma população que pertencem a uma classe de interesse.
- A diferença nas médias de duas populações, $\mu_1 - \mu_2$.
- A diferença nas proporções de duas populações, $p_1 - p_2$.

Estimadores razoáveis desses parâmetros são, respectivamente:

- A média amostral $\overline{X}$.
- A variância amostral $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.
- A proporção amostral $\hat{p}$ de ítens em uma amostra que pertencem a uma classe de interesse.
- A diferença nas médias amostrais $\overline{X}_1 - \overline{X}_2$ de duas amostras aleatórias independentes.

- A diferença nas proporções amostrais $\hat{p}_1 - \hat{p}_2$ de duas amostras aleatórias independentes.

Existem várias possibilidades para a escolha de um estimador de um parâmetro. Por exemplo, poderíamos utilizar o estimador $\frac{(n-1)S^2}{n}$ para estimar a variância populacional. Necessitamos estudar propriedades dos estimadores para podermos desenvolver algum critério para determinar qual melhor estimador para determinado parâmetro.

Exemplo 10.1.1: Suponha que desejassemos comprar um rifle e para tanto podemos testar quatro opções de rifles A, B, C, e D. Para tanto, podemos executar 15 tiros a um alvo com cada um deles. Para chegarmos a conclusão de qual a melhor arma, precisamos de alguns critérios. Quanto a qualidade da arma, poderíamos definir três critérios, o critério da *acurácia* que mede a proximidade de cada observação do valor do alvo que se procura atingir, o critério da *precisão* que mede a proximidade de cada observação da média de todas as observações, e o critério do *viés* que mede a proximidade da média de todas as observações do valor do alvo que se procura atingir.

10.2 Propriedades de Estimadores

Como vimos o problema da estimação é determinar uma função $h(X_1, X_2, \ldots, X_n)$ que seja próxima de θ, segundo algum critério. O primeiro critério é o seguinte:

Definição 10.2.1: O estimador T é *não-viesado* para θ se $ET = \theta$ para todo θ, onde ET é calculada segundo a distribuição amostral de T.

O viés de um estimador T para um parâmetro θ é igual a $ET - \theta$. Logo, um estimador T é não-viesado para θ, se o seu viés for igual a zero para todo θ.

Exemplo 10.2.2: A média amostral $\overline{X}$ é um estimador não-viesado para média populacional μ, pois

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \mu.$$

A proporção amostral $\hat{p}$ é um estimador não-viesado para proporção populacional p que possui uma certa característica, pois chamando de Y_i a variável aleatória que é igual a 1 se o i-ésimo indivíduo da amostra possui a característica de interesse, e igual a zero, em caso contrário, temos que

$$E(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n} EY_i = p.$$

Exemplo 10.2.3: Considere uma população com N elementos, com média populacional $\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$, e variância populacional

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2.$$

Um possível estimador para σ^2, baseado numa amostra aleatória simples de tamanho n dessa população, é

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Vamos mostrar que esse estimador é viesado. Note que

$$\begin{aligned}
&\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - 2\sum_{i=1}^{n}(X_i - \mu)(\overline{X} - \mu) + \sum_{i=1}^{n}(\overline{X} - \mu)^2 \\
&= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2.
\end{aligned}$$

Portanto,

$$\begin{aligned}
E(\hat{\sigma}^2) &= \frac{1}{n}(\sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\overline{X} - \mu)^2) \\
&= \frac{1}{n}(\sum_{i=1}^{n} Var(X_i) - nVar(\overline{X})) \\
&= \frac{1}{n}(n\sigma^2 - n\frac{\sigma^2}{n}) = \frac{n-1}{n}\sigma^2.
\end{aligned}$$

Logo, o viés de $\hat{\sigma}^2$ é igual a $\frac{n-1}{n}\sigma^2 - \sigma^2 = \frac{-\sigma^2}{n}$. Portanto, o estimador $\hat{\sigma}^2$ em geral subestima o verdadeiro parâmetro σ^2. Por outro lado, o viés diminui com n tendendo a zero quando n tende a infinito. É fácil ver que $S^2 = \frac{n}{n-1}\hat{\sigma}^2$ é um estimador não-viesado para σ^2. Portanto, a variância de uma amostra de tamanho n é dada por S^2, onde o denominador é igual a $n-1$, enquanto que a variância de uma população de tamanho N é dada por σ^2, onde o denominador é igual a N.

O segundo critério que iremos analisar é o critério da consistência de um estimador. Intuitivamente, temos que um estimador é consistente se quando aumentamos o tamanho da amostra n, a probabilidade de que este difira do parâmetro por mais que qualquer erro pre-especificado $\epsilon > 0$ tende a zero. Formalmente,

Definição 10.2.4: Uma seqüência $\{T_n\}$ de estimadores de um parâmetro θ é *consistente* se, para todo $\epsilon > 0$,

$$\lim_{n\to\infty} P(|T_n - \theta| > \epsilon) = 0.$$

Exemplo 10.2.5: A seqüência de estimadores $\overline{X}_n$ é consistente, pois como $E(\overline{X}_n) = \mu$ e $Var(\overline{X}_n) = \frac{\sigma^2}{n}$, utilizando a desigualdade de Chebyshev, temos:

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0,$$

quando $n \to \infty$, para qualquer $\epsilon > 0$.

Podemos utilizar o seguinte, teorema para determinar se uma dada seqüência de estimadores é consistente:

Teorema 10.2.6: *Se* $\{T_n\}$ *é uma seqüência de estimadores de* θ *tal que* $\lim_{n\to\infty} E(T_n) = \theta$ *e* $\lim_{n\to\infty} Var(T_n) = 0$*, então* $\{T_n\}$ *é consistente.*

Prova: Note que pela desigualdade triangular, se $|T_n - \theta| > \epsilon$, então $|ET_n - \theta| > \frac{\epsilon}{2}$ ou $|T_n - ET_n| > \frac{\epsilon}{2}$. Portanto,

$$P(|T_n - \theta| > \epsilon) \leq P(|ET_n - \theta| > \frac{\epsilon}{2}) + P(|T_n - ET_n| > \frac{\epsilon}{2}).$$

Logo, pela desigualdade de Chebyshevy

$$P(|T_n - \theta| > \epsilon) \leq P(|ET_n - \theta| > \frac{\epsilon}{2}) + \frac{4Var(T_n)}{\epsilon^2}.$$

Então tomando os limites quando $n \to \infty$, temos que

$$\lim_n P(|T_n - \theta| > \epsilon) \leq \lim_n P(|ET_n - \theta| > \frac{\epsilon}{2}) + \lim_n \frac{4Var(T_n)}{\epsilon^2} = 0.$$

Portanto, $\{T_n\}$ é consistente. ∎

Note que se T_n for um estimador não-viesado, então obviamente $\lim_{n\to\infty} E(T_n) = \theta$, e portanto se a variância do estimador T_n tender a zero, ele é um estimador consistente.

Exemplo 10.2.7: Vimos que S^2 é um estimador não-viesado para σ^2. É possível demonstrar no caso em que a população tem distribuição normal com média μ e variância σ^2 que

$$Var(S^2) = \frac{2\sigma^4}{n-1}.$$

Logo, S^2 é consistente para σ^2.

Exemplo 10.2.8: Como $\hat{\sigma}^2 = \frac{n-1}{n}S^2$, temos que $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \to \sigma^2$ quando $n \to \infty$, e $Var(\hat{\sigma}^2) = (\frac{n-1}{n})^2 \frac{2\sigma^4}{n-1} \to 0$ quando $n \to \infty$. Logo, pelo teorema $\hat{\sigma}^2$ também é consistente para σ^2.

Um outro critério para comparação de estimadores é o seguinte:

Definição 10.2.9: Se T e T' são dois estimadores não-viesados de um mesmo parâmetro θ, e $VarT < VarT'$, então T é mais *eficiente* que T'.

Exemplo 10.2.10: Consideremos uma população normal X, com parâmetros μ e σ^2. Queremos estimar a mediana desta população. Como a distribuição é simétrica temos que a mediana e a média coincidem e são iguais a μ. Definindo $\overline{X}$ e md como a média e a mediana de uma amostra de tamanho n dessa população, qual dos dois estimadores é mais eficiente para estimar a mediana populacional?

Sabemos que $\overline{X} \sim N(\mu, \sigma^2/n)$ e pode-se demonstrar que a distribuição da mediana pode ser aproximada por $N(Md(X), \frac{\pi\sigma^2}{2n})$. Portanto, os dois estimadores são não-viesados, mas $\overline{X}$ é mais eficiente, pois $Var(md) > Var(\overline{X})$. Conclui-se que para estimar a mediana dessa população, é preferível usar a média da amostra como estimador, o que contraria um pouco a nossa intuição.

Finalmente, podemos considerar o critério do *erro quadrático médio* para comparar estimadores. Denomina-se de *erro amostral* de um estimador T para um parâmetro θ a diferença $e = T - \theta$. Note o erro amostral é uma v.a. pois é uma função de T que é uma v.a., além disso note que o viés de T é igual a esperança do erro amostral.

Definição 10.2.11: O *erro quadrático médio* (EQM) do estimador T para o parâmetro θ é igual ao segundo momento do erro amostral com respeito a distribuição amostral do estimador T, ou seja, $EQM(T, \theta) = E(e^2) = E(T - \theta)^2$.

Podemos desenvolver a expressão do EQM para obter:

$$\begin{aligned} &EQM(T,\theta) = E(T - E(T) + E(T) - \theta)^2 \\ &= E(T - E(T))^2 + 2E[(T - E(T))(E(T) - \theta)] + E(E(T) - \theta)^2 \\ &= Var(T) + V^2. \end{aligned}$$

Vemos então que o erro quadrático médio leva em consideração tanto o viés V do estimador como sua variabilidade medida através de $Var(T)$. Segundo este critério o estimador é tão melhor quanto menor for seu erro quadrático médio.

Exemplo 10.2.12: Determine o erro quadrático médio do estimador $\overline{X}$ para μ.

Solução: Neste caso, temos que

$$E(\overline{X} - \mu)^2 = Var(\overline{X}) = \frac{\sigma^2}{n}.$$

∎

10.3 Intervalo de Confiança

Até agora os estimadores apresentados foram pontuais, isto é, especificam um único valor para o estimador. Esse procedimento não permite julgar qual a possível magnitude do erro que estamos cometendo. Daí surge a idéia de construir *intervalos de confiança* que são baseados na distribuição amostral do estimador.

Um intervalo de confiança de um parâmetro desconhecido θ é um intervalo da forma $[L, U]$, em que os pontos extremos do intervalo L e U dependem da amostra, e portanto são, na verdade, estatísticas, isto é variáveis aleatórias. Nosso objetivo ao construir intervalos de confiança é determinar funções da amostra L e U tal que a seguinte afirmação seja verdadeira:

$$P(L \leq \theta \leq U) = \gamma,$$

onde $0 < \gamma < 1$. Assim, existe uma probabilidade γ de selecionarmos uma amostra tal que o intervalo $[L, U]$ contenha o valor de θ. Note que θ não é aleatório, L e U é que são aleatórios. Se a afirmação acima for verdadeira o que estamos afirmando é que se forem construídos vários intervalos de confiança usando as estimativas L e U, em $100\gamma\%$ das vezes θ estará incluso no intervalo $[L, U]$. Tal intervalo é chamado de um intervalo de $100\gamma\%$ de confiança para θ, e γ é conhecido como coeficiente (ou nível) de confiança do intervalo.

Na prática, obtemos somente uma amostra aleatória e calculamos um intervalo de confiança. Calculado este intervalo de confiança, então duas situações podem existir: ele contém ou não o verdadeiro valor de θ. Neste ponto, não existe mais nenhum valor aleatório, portanto não faz sentido associar uma probabilidade ao intervalo conter o verdadeiro valor θ. A afirmação apropriada é: o intervalo observado $[l, u]$ contém o verdadeiro valor θ, com $100\gamma\%$ de confiança. E esta afirmação tem uma interpretação freqüentista, ou seja, não sabemos se a afirmação é ou não verdadeira para esta amostra específica, mas o método usado para obter o intervalo $[l, u]$ resulta em afirmações corretas em $100\gamma\%$ das vezes.

Note que quanto maior o intervalo de confiança, mais confiantes estaremos que ele contenha o verdadeiro valor θ. Por outro lado, quanto maior for o intervalo, menos informação teremos a respeito do verdadeiro valor de θ. Em uma situação ideal, obtemos um intervalo relativamente pequeno com alta confiança.

O intervalo de confiança descrito acima é um intervalo bilateral de confiança, pois especificamos tanto o limite inferior como o limite superior do intervalo. Podemos também obter um intervalo unilateral inferior de confiança para θ com nível de confiança γ, escolhendo um limite inferior L de tal forma que

$$P(L \leq \theta) = \gamma.$$

Analogamente, um intervalo unilateral superior de confiança para θ com nível de confiança γ, pode ser obtido escolhendo um limite superior U tal que

$$P(\theta \leq U) = \gamma.$$

10.3.1 Intervalo de Confiança para Média com Variância Conhecida

Nesta seção estaremos interessados em construir um intervalo de confiança para média populacional μ admitindo-se que a variância populacional σ^2 é conhecida. Recorde que pelo Teorema Central do Limite, a distribuição amostral de $\overline{X}$ é aproximadamente normal com média μ e variância σ^2/n, desde que n seja suficientemente grande (≥ 30). Neste caso,

$$Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

tem uma distribuição normal padrão. Seja $\Phi^{-1}(\alpha)$ o valor tal que $P(Z \leq \Phi^{-1}(\alpha)) = \alpha$. Então, temos que

$$P(-\Phi^{-1}(w) \leq Z \leq \Phi^{-1}(w)) = P(Z \leq \Phi^{-1}(w)) - P(Z \leq -\Phi^{-1}(w)) = w - (1 - w) = 2w - 1.$$

Deste modo,

$$P(-\Phi^{-1}((\gamma + 1)/2) \leq Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \leq \Phi^{-1}((\gamma + 1)/2)) = \gamma.$$

Rearrumando as desigualdades obtemos

$$P(\overline{X} - \Phi^{-1}((\gamma+1)/2)\sigma/\sqrt{n} \leq \mu \leq \overline{X} + \Phi^{-1}((\gamma+1)/2)\sigma/\sqrt{n}) = \gamma.$$

Deste modo temos que $[\overline{X}-\Phi^{-1}((\gamma+1)/2)\sigma/\sqrt{n}, \overline{X}+\Phi^{-1}((\gamma+1)/2)\sigma/\sqrt{n}]$ é um intervalo com $100\gamma\%$ de confiança para μ. Note que a amplitude deste intervalo é $L = 2\Phi^{-1}((\gamma+1)/2)\sigma/\sqrt{n}$, que é uma constante que independe de $\overline{X}$. Note que com esta fórmula, dado uma amplitude desejada L, podemos determinar o tamanho da amostra necessária para atingir um nível de confiança desejado γ em um intervalo com amplitude L.

Para amostras provenientes de uma população normal ou para amostras de tamanho $n \geq 30$, independente da forma da população, o intervalo fornecerá bons resultados. Caso contrário, não podemos esperar que o nível de confiança seja exato.

Podemos também obter intervalos de confiança unilaterais para μ, neste caso sabemos que

$$P(Z = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \geq -\Phi^{-1}(\gamma)) = \gamma.$$

Rearrumando a desigualdade, temos:

$$P(\mu \leq \overline{X} + \Phi^{-1}(\gamma)\sigma/\sqrt{n}) = \gamma.$$

Deste modo, temos que $(-\infty, \overline{X} + \Phi^{-1}(\gamma)\sigma/\sqrt{n}]$ é um intervalo unilateral superior com $100\gamma\%$ de confiança para μ. Analogamente, podemos obter que $[\overline{X} - \Phi^{-1}(\gamma)\sigma/\sqrt{n}, \infty)$ é um intervalo unilateral inferior com $100\gamma\%$ de confiança para μ.

Exemplo 10.3.1: Suponha que temos uma população com distribuição $Bernoulli(p)$. Por exemplo, p pode representar a probabilidade de um determinado tipo de capacitor ser produzido com defeito por uma determinada fábrica. Dada uma amostra aleatória $X_1, X_2, \ldots, X_n$ de tamanho n da produção de capacitores desta fábrica, podemos estimar um intervalo de confiança bilateral para p. Note que a variância da população é dada por $p(1-p)$. Portanto, sendo $\hat{p}$ a proporção de capacitores com defeito na amostra, como $\sigma^2 = p(1-p)$, o resultado anterior nos leva a afirmar que $[\hat{p} - \Phi^{-1}((\gamma+1)/2)\sqrt{\frac{p(1-p)}{n}}, \hat{p} + \Phi^{-1}((\gamma+1)/2)\sqrt{\frac{p(1-p)}{n}}]$ é um intervalo com $100\gamma\%$ de confiança para p. Como não conhecemos p, podemos proceder de duas maneiras: (1) utilizar o fato que $p(1-p) \leq 1/4$, obtendo o intervalo $[\hat{p} - \Phi^{-1}((\gamma+1)/2)\sqrt{\frac{1}{4n}}, \hat{p} + \Phi^{-1}((\gamma+1)/2)\sqrt{\frac{1}{4n}}]$, ou (2) utilizar $\hat{p}$ como estimativa para p, obtendo o intervalo $[\hat{p}-\Phi^{-1}((\gamma+1)/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p}+\Phi^{-1}((\gamma+1)/2)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$. O primeiro método é sempre correto, porém muito conservador pois em geral $p(1-p)$ pode ser bem menor que $1/4$, e então estamos propondo um intervalo com amplitude maior que a necessária. O segundo método é válido desde que np e $n(1-p)$ sejam maiores que 5, pois caso contrário, se for pequeno a distribuição normal não poderá mais ser usada e teremos que utilizar a distribuição binomial.

Exemplo 10.3.2: O comprimento dos eixos produzidos por uma empresa tem aproximadamente uma distribuição normal com desvio padrão de 4cm. Uma amostra com 16 eixos forneceu uma média de 4,52cm.

(a) Determine um intervalo de confiança de 90% para o comprimento médio real dos eixos.

(b) Com que probabilidade podemos afirmar que o comprimento médio desta amostra não difere da média por mais de 0,5cm?

Solução: O intervalo de confiança é dado por:

$$[4{,}52 - \Phi^{-1}(0{,}95)\frac{4}{\sqrt{16}}; 4{,}52 + \Phi^{-1}(0{,}95)\frac{4}{\sqrt{16}}] = [2{,}875; 6{,}165].$$

Para o ítem (b), como $\sigma/\sqrt{n} = 1$, temos $|\overline{X} - \mu|$ tem distribuição normal padrão, logo

$$P(|\overline{X} - \mu| \leq 0{,}5) = P(|Z| \leq 0{,}5) = 0{,}383.$$

Exemplo 10.3.3: Uma amostra de 400 domicílios mostra que 25% deles são de casas alugadas. Qual é o intervalo de confiança para o número de casas alugadas numa cidade supondo que ela tem 20.000 casas? Considere um coeficiente de confiança de 98%.

Solução: Podemos primeiro determinar o intervalo de confiança para a proporção de casas alugadas. Neste caso, então temos $\hat{p} = 0{,}25$, $n = 400$, e $\gamma = 0{,}98$. Utilizando $\hat{p}(1-\hat{p})$ como uma estimativa para a variância $p(1-p)$, temos que o intervalo de confiança para a população é:

$$[0{,}25 - \Phi^{-1}(0{,}99)\sqrt{\frac{0{,}25(0{,}75)}{400}}, 0{,}25 + \Phi^{-1}(0{,}99)\sqrt{\frac{0{,}25(0{,}75)}{400}}]$$

Então, o intervalo de confiança para o número de casas alugadas é dado por:

$$[20.000(0{,}25 - \Phi^{-1}(0{,}99)\sqrt{\frac{0{,}25(0{,}75)}{400}}), 20.000(0{,}25 + \Phi^{-1}(0{,}99)\sqrt{\frac{0{,}25(0{,}75)}{400}})]$$
$$= [5.000 - 1.006{,}75, 5.000 + 1006{,}75] = [3.993{,}25, 6006{,}75].$$

Exemplo 10.3.4: Uma pesquisa sobre renda familiar foi realizada entre as famílias que tem rendimento de até 5 salários mínimos. Sabe-se que o desvio padrão populacional é de 1,2. Uma amostra de 200 famílias foram selecionadas e seus resultados aparecem na tabela abaixo:

Rendimento	Freqüência
1	90
2	50
3	30
4	20
5	10

(a) Estime, com 95% de confiabilidade, o intervalo de confiança para a média de renda familiar desta população.

(b) Estime a proporção real de famílias que tem rendimento de até 2 salários mínimos, com 95% de confiabilidade.

Solução: Primeiro vamos determinar o valor de $\overline{x}$. Temos que

$$\overline{x} = 1(90/200) + 2(50/200) + 3(30/200) + 4(20/200) + 5(10/200) = 2{,}05.$$

Então, o intervalo de confiança de 95% é dado por:

$$[2{,}05 - \Phi^{-1}(0{,}975)\frac{1{,}2}{\sqrt{200}}, 2{,}05 + \Phi^{-1}(0{,}975)\frac{1{,}2}{\sqrt{200}}] = [1{,}884; 2{,}216].$$

Para o ítem (b), temos que $\hat{p} = 140/200 = 0{,}7$. Usando $\hat{p}(1-\hat{p})$ como estimativa para a variância populacional, temos que o intervalo de confiança de 95% para proporção populacional é:

$$[0{,}7 - \Phi^{-1}(0{,}975)\sqrt{\frac{0{,}7(0{,}3)}{200}}, 0{,}7 + \Phi^{-1}(0{,}975)\sqrt{\frac{0{,}7(0{,}3)}{200}}] = [0{,}636; 0{,}764].$$

10.3.2 Intervalo de Confiança para Média com Variância Desconhecida

Quando estamos construindo intervalos de confiança para a média μ de uma população quando σ^2 for desconhecida, devido ao Teorema Central do Limite, podemos continuar usando os procedimentos da seção anterior desde que o tamanho da amostra seja grande ($n \geq 30$), usando s^2 como estimativa para σ^2. Entretanto, quando a amostra for pequena e σ^2 desconhecida, teremos de fazer uma suposição sobre a forma da distribuição em estudo. Assumiremos nesta seção que a população tem uma distribuição normal. Na prática, muitas populações podem ter suas distribuições aproximadas por uma normal, assim esta restrição não é tão restritiva e o método apresentado nesta seção tem larga aplicabilidade.

Pode-se provar que se a população tem uma distribuição normal, então $T = \frac{\sqrt{n}(\overline{X}-\mu)}{S}$ tem uma distribuição t de student com $n-1$ graus de liberdade. Seja $\tau(\gamma, n-1)$ o valor tal que $P(T \leq \tau(\gamma, n-1)) = \gamma$. Então, utilizando o mesmo procedimento da seção anterior, podemos verificar que

1. $[\overline{X} - \tau((\gamma+1)/2, n-1)s/\sqrt{n}, \overline{X} + \tau((\gamma+1)/2, n-1)s/\sqrt{n}]$ é um intervalo bilateral com $100\gamma\%$ de confiança para a média da população μ.

2. $(-\infty, \overline{X} + \tau(\gamma, n-1)s/\sqrt{n}]$ é um intervalo unilateral superior com $100\gamma\%$ de confiança para μ.

3. $[\overline{X} - \tau(\gamma, n-1)s/\sqrt{n}, \infty)$ é um intervalo unilateral inferior com $100\gamma\%$ de confiança para μ.

Capítulo 11

Testes de Hipótese

11.1 Teste de Hipótese

Na seção anterior, estudamos o problema de estimar um parâmetro de uma população através de uma amostra selecionada desta população. Em muitas situações práticas não estamos interessados em estimar o parâmetro, mas ao invés estamos interessados em aceitar ou rejeitar uma afirmação a respeito do parâmetro. Tal afirmação é conhecida como *hipótese*. E o método utilizado para decidirmos aceitar ou rejeitar uma dada hipótese a partir de dados amostrais é conhecido como *Teste de Hipótese*. A idéia central deste procedimento é assumir que a hipótese é verdadeira e verificar se a amostra observada parece "razoável" ou "consistente", dada esta suposição.

Definição 11.1.1: Uma *hipótese estatística* é uma afirmação sobre os parâmetros de uma ou mais populações.

Como usamos distribuições de probabilidade para representar populações, uma hipótese estatística pode também ser pensada como uma afirmação acerca da distribuição de probabilidades de uma variável aleatória.

Por exemplo, suponha que estejamos interessados em verificar a tensão em uma dada tomada. A tensão na tomada é uma variável aleatória que sofre alterações ao longo do dia e pode ser descrita por uma variável aleatória. Suponha que nosso interesse seja no valor esperado desta distribuição, ou seja, estamos interessados em decidir se a tensão é ou não igual a $220v$. Então, $\mu = 220v$ é chamada de *hipótese nula*, representada por H_0. Esta hipótese nula pode ser aceita ou rejeitada, no caso dela ser rejeitada, precisamos de uma outra hipótese que seja aceitável, conhecida como *hipótese alternativa*, representada por H_1. Por exemplo, uma hipótese alternativa seria $\mu \neq 200v$. Neste caso, como a hipótese alternativa especifica valores de μ maiores e menores que o valor especificado por H_0, ela é chamada de *hipótese alternativa bilateral*. Em algumas situações podemos desejar formular uma hipótese alternativa unilateral, como em $H_0 : \mu = 220v$ e $H_1 : \mu < 220v$, $H_0 : \mu = 220v$ e $H_1 : \mu > 220v$, ou $H_0 : \mu = 220v$ e $H_1 : \mu = 240v$.

Então, a hipótese nula é uma afirmação a respeito da população, mais especificamente uma afirmação a respeito de um parâmetro da população. Esta afirmação, pode ter sido

originada de conhecimento experiência *a priori* da população em estudo, de testes ou experimentos anteriores; pode ter sido determinado de alguma teoria ou modelo da população em estudo; ou pode surgir de considerações exógenas, por exemplo, parâmetros que devem obedecer certos critérios de controle de qualidade.

Estabelecidas as hipóteses nulas e alternativas, a informação contida na amostra é analisada para verificar se a hipótese nula é consistente com esta informação. Caso seja, conclui-se que a hipótese nula é verdadeira, caso contrário, conclui-se que a hipótese é falsa, o que implicará na aceitação da hipótese alternativa. Porém, note que para sabermos com certeza se a hipótese nula é ou não verdadeira, precisaríamos analisar toda a população, o que na prática é freqüentemente impossível. Portanto, todo procedimento de testes de hipóteses tem alguma probabilidade de erro associada.

Para ilustrar alguns conceitos, considere o exemplo descrito anteriormente, ou seja, $H_0 : \mu = 220v$ e $H_1 : \mu = 240v$. Suponha que n medidas na tensão da tomada sejam feitas e que a média dos valores obtidos nesta amostra $\overline{x}$ seja observada. Como vimos, $\overline{x}$ é uma estimativa para o valor de μ, logo se obtivermos um valor de $\overline{x}$ próximo a $220v$, temos uma evidência que a hipótese nula é verdadeira. Precisa-se então estabelecer uma região de valores, conhecida como *região de aceitação* tal que se $\overline{x}$ cair nesta região iremos aceitar a hipótese nula, e se $\overline{x}$ cair fora dessa região, ou seja, na região conhecida como *região crítica* (RC), iremos aceitar a hipótese alternativa. Por exemplo, poderíamos considerar a região de aceitação como sendo o intervalo $(-\infty, 230]$. Os limites da região de aceitação são chamados de *valores críticos*.

Esse procedimento de decisão pode acarretar um de dois tipos de erros diferentes. O primeiro, conhecido como *erro tipo I* ocorre quando a tensão média na tomada é realmente $220v$, mas por chance o conjunto de medidas aleatórios que obtivemos nos levou a obter um valor de $\overline{x}$ na região crítica. Ou seja, um erro do tipo 1 ocorre quando rejeitamos a hipótese nula quando na verdade ela é verdadeira. O segundo, conhecido como *erro do tipo II* ocorre quando apesar da hipótese nula ser falsa, a média das medidas de tensão obtidas cai na região de aceitação. Ou seja, um erro do tipo II ocorre sempre que aceitamos a hipótese nula apesar dela ser falsa.

A probabilidade de ocorrência de um erro tipo I é chamada de nível de significância, tamanho do teste, ou ainda, p-valor do teste, e é denotada por α. O *poder de um teste* é igual a probabilidade de rejeitarmos a hipótese nula quando ela realmente é falsa. Note que o poder do teste é igual a 1 menos a probabilidade de ocorrência de um erro do tipo II, que é usualmente denotada por β.

Quando H_0 for verdadeira, isto é, a tensão for realmente de $220v$, sabemos do TCL que $\overline{X} \sim N(220, \frac{\sigma^2}{n})$. Então, podemos determinar o nível de significância do teste:

$$\alpha = P(\text{erro I}) = P(\overline{X} > 230 | \overline{X} \sim N(220, \frac{\sigma^2}{n}))$$
$$= P(\frac{\sqrt{n}(\overline{X} - 220)}{\sigma} > \frac{\sqrt{n}(230 - 220)}{\sigma})$$

Se soubermos que a variância da tensão na tomada é $64v^2$, e tivermos uma amostra de 4 medidas do valor de tensão, podemos obter:

$$\alpha = P(Z > \frac{2(10)}{8}) = P(Z > 2{,}5) = 0{,}0062.$$

De modo análogo, podemos obter a probabilidade do erro tipo II. Neste caso, se H_1 for verdadeira, temos $\overline{X} \sim N(240, 16)$, então:

$$\beta = P(\text{erro II}) = P(\overline{X} \leq 230 | \overline{X} \sim N(240, 16))$$
$$= P(\frac{(\overline{X} - 240)}{4} \leq \frac{(230 - 240)}{4}) = P(Z \leq -2{,}5) = 0{,}0062.$$

Neste caso, α e β foram iguais devido a simetria da região crítica em relação as hipóteses nula e alternativa. Note que se ao invés de termos escolhido o valor crítico 230, aumentássemos esse valor, então α diminuiria e β aumentaria.

Poderíamos também especificar um valor para a probabilidade de erro do tipo I e verificar qual seria a região crítica que satisfaria esta probabilidade de erro pre-especificada. Por exemplo, suponha que queiramos encontrar a região crítica cujo o α seja igual a 0,01. Temos:

$$0{,}01 = \alpha = P(Z > 2{,}325) = P(\frac{2(\overline{X} - 220)}{8} > 2{,}325) = P(\overline{X} > 229{,}3).$$

Para a região crítica $(229{,}3, \infty)$, podemos determinar o valor de β para esta região.

$$\beta = P(\text{erro II}) = P(\overline{X} \leq 229{,}3 | \overline{X} \sim N(240, 16))$$
$$= P(\frac{(\overline{X} - 240)}{4} \leq \frac{(229{,}3 - 240)}{4}) = P(Z \leq -2{,}675) = 0{,}0038.$$

Este segundo tipo de procedimento é bastante utilizado, pois em geral a hipótese alternativa não contém apenas um único valor de parâmetro como no exemplo acima. Muitas vezes, se nossa hipótese nula é $H_0 : \mu = 220v$, nossa hipótese alternativa será $H_1 : \mu \neq 220v$. Como os parâmetros da hipótese alternativa são muitos, a solução é adotar o último procedimento descrito acima, ou seja, pre-estabelecer um valor α, e calcular uma região crítica que satisfaça esta restrição. No caso de uma hipótese alternativa bilateral, em geral toma-se como região de aceitação um intervalo simétrico ao redor da hipótese nula, deste modo se fixarmos $\alpha = 0{,}01$, teremos

$$0{,}01 = \alpha = P(|Z| > 2{,}575) = P(|\frac{2(\overline{X} - 220)}{8}| > 2{,}575) = 1 - P(209{,}7 \leq \overline{X} \leq 230{,}3).$$

Deste modo, determinamos a região de aceitação $[209{,}7; 230{,}3]$ de modo que o nível de significância de 0,01 seja satisfeito. Mesmo determinada esta regra de decisão, não poderemos determinar β, pois não existe um único valor de μ na hipótese alternativa. Neste caso, poderemos considerar uma função $\beta(\mu)$, conhecida como *função característica de operação.*

Definição 11.1.2: A *função característica de operação* (função CO) de um teste de hipótese é definida como:

$$\beta(\mu) = P(\text{aceitar } H_0 | \mu),$$

ou seja, $\beta(\mu)$ é a probabilidade de aceitar H_0 como função de μ. A *função poder do teste*, que é a probabilidade de se rejeitar H_0 como função de μ é dad por $\pi(\mu) = 1 - \beta(\mu)$.

As seguintes propriedades de $\pi(\mu)$ são facilmente verificadas:

i. $\pi(\mu_0) = \alpha$;

ii. No caso de hipótese alternativa bilateral ($H_1 : \mu \neq \mu_0$), $\pi(-\infty) = \pi(+\infty) = 1$ e $\pi(\mu)$ decresce para $\mu < \mu_0$ e cresce para $\mu > \mu_0$;

iii. No caso de hipótese alternativa unilateral superior ($H_1 : \mu > \mu_0$), $\pi(-\infty) = 0$, $\pi(+\infty) = 1$, e $\pi(\mu)$ é sempre crescente;

iv. No caso de hipótese alternativa unilateral inferior ($H_1 : \mu < \mu_0$), $\pi(-\infty) = 1$, $\pi(+\infty) = 0$, e $\pi(\mu)$ é sempre decrescente.

Na construção das hipóteses, sempre estabeleceremos a hipótese nula como uma igualdade, de modo que o analista pode controlar α, ao estabelecer uma região crítica para o teste. Então o analista, pode controlar diretamente a probabilidade de rejeitar erroneamente H_0, então a rejeição da hipótese nula é uma *conclusão forte*. Note que quanto menor o valor de α, ao rejeitarmos a hipótese nula, estaremos cada vez mais seguros da hipótese alternativa, portanto maior será a significância da nossa conclusão. Por isso, α é chamado de nível de significância do teste. Por outro lado, β não é constante, mas depende do verdadeiro valor do parâmetro, por este motivo a aceitação de H_0 é tida como uma *conclusão fraca*, a não ser que saiba-se que β é aceitavelmente pequena. Então, a nomenclatura mais correta seria ao invés de dizermos "aceitamos H_0" deveríamos dizer "a amostra não apresentou evidência suficiente para rejeitarmos H_0". Neste último caso, não necessariamente afirma-se que existe uma alta probabilidade de que H_0 seja verdadeira, isto pode significar apenas que mais dados são necessários para atingirmos uma conclusão forte.

Na determinação de quem é a hipótese nula, devemos adotar como H_0 aquela hipótese, que se rejeitada erroneamente, conduza a um erro mais importante de se evitar, pois esta probabilidade de erro é controlável. Então, por exemplo, se estivermos interessados em saber se um novo medicamento é eficaz no combate a uma doença, a hipótese nula seria que ele é não eficaz, pois os danos causados por usarmos um remédio não eficaz são maiores que se deixássemos de usar um remédio que seria eficaz. Ou ainda, se desejamos saber se certa substância é radioativa, então a hipótese nula seria que ela é radioativa, pois os danos causados pela manipulação radioativa são maiores que se deixássemos de manipular uma substância por acharmos falsamente que ela é radioativa. Como a rejeição da hipótese nula é que é uma conclusão forte, escolhe-se como H_1 a hipótese que se deseja comprovar. Por exemplo, no caso do novo medicamento H_1 será a hipótese que o novo medicamento é melhor que os existentes.

11.2 Procedimento Geral Para Testes de Hipóteses

A seguir daremos uma seqüência de passos que pode ser seguida em qualquer teste de hipóteses:

0. A partir do contexto do problema, identifique o parâmetro de interesse.

1. Fixe qual a hipótese nula H_0 e alternativa H_1.

2. Use teoria estatística e informações disponíveis para decidir que estimador será usado para testar H_0.

3. Obtenha a distribuição do estimador proposto.

4. Determine α.

5. Construa a região crítica para o teste de modo que α seja satisfeita.

6. Use os dados da amostra para determinar o valor do estimador, ou seja, uma estimativa para o parâmetro.

7. Se o valor do estimador pertencer a região crítica, então rejeite H_0. Caso contrário, reporte que não existe evidência suficiente para se rejeitar H_0.

11.3 Teste de Hipótese para a Média de Uma População com Variância Conhecida

Suponhamos que desejamos testar as hipóteses $H_0 : \mu = \mu_0$ e $H_1 : \mu \neq \mu_0$, sendo μ_0 uma constante especificada. Para testar a hipótese nula, usaremos o estimador média amostral de uma amostra aleatória simples de tamanho n. Deste modo, sabemos pelo TCL que $\overline{X} \sim N(\mu_0, \sigma^2/n)$, se a hipótese nula for verdadeira, e então poderemos proceder como anteriormente.

Note que a estatística padronizada $Z_0 = \frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}}$ tem uma distribuição normal padrão, se a hipótese nula for verdadeira. Portanto, temos que para a região de aceitação $[-\Phi^{-1}(1-\alpha/2), \Phi^{-1}(1-\alpha/2)]$, temos que $P(Z_0 \in RC|\mu = \mu_0) = \alpha$.

É mais fácil entender a região crítica e o procedimento do teste quando a estatística de teste é Z_0 e não $\overline{X}$. Entretanto, a mesma região crítica pode ser calculada em termos do valor da estatística $\overline{X}$. Neste caso, a região de aceitação é $[\mu_0 - \Phi^{-1}(1-\alpha/2)\frac{\sigma}{\sqrt{n}}, \mu_0 + \Phi^{-1}(1-\alpha/2)\frac{\sigma}{\sqrt{n}}]$.

De modo similar, podemos obter a região crítica para o caso de um teste de hipótese unilateral $H_0 : \mu = \mu_0$ e $H_1 : \mu > \mu_0$, ou $H_0 : \mu = \mu_0$ e $H_1 : \mu < \mu_0$. No primeiro caso, temos que a região de aceitação para a estatística Z_0 é $(-\infty, \Phi^{-1}(1-\alpha)]$, o que implica que a região de aceitação para a estatística $\overline{X}$ é $(-\infty, \mu_0 + \Phi^{-1}(1-\alpha)\frac{\sigma}{\sqrt{n}}]$. No segundo caso, temos que a região para a estatística Z_0 é $[\Phi^{-1}(\alpha), \infty)$, o que implica que a região de aceitação para a estatística $\overline{X}$ é $[\mu_0 + \Phi^{-1}(\alpha)\frac{\sigma}{\sqrt{n}}, \infty)$.

11.3.1 Teste para Proporção

O caso do teste para proporção é um caso particular do caso do teste para a média com variância conhecida. Neste caso, temos que cada amostra pode ser considerada como uma variável Bernoulli com parâmetro p que representa a proporção de indivíduos da população que possuem uma determinada característica. Já vimos que a média de uma Bernoulli é igual ao seu parâmetro p, e que sua variância é igual a $p(1-p)$. Logo, utilizando a proporção

amostral como estatística e os resultados gerais da seção anterior temos que a região de aceitação para a proporção é

- No caso de hipótese alternativa bilateral: $H_0 : p = p_0$ e $H_1 : p \neq p_0$, a região de aceitação é

$$[p_0 - \Phi^{-1}(1-\alpha/2)\sqrt{\frac{p_0(1-p_0)}{n}}, p_0 + \Phi^{-1}(1-\alpha/2)\sqrt{\frac{p_0(1-p_0)}{n}}].$$

- No caso de hipótese alternativa unilateral superior: $H_0 : p = p_0$ e $H_1 : p > p_0$, a região de aceitação é

$$(-\infty, p_0 + \Phi^{-1}(1-\alpha)\sqrt{\frac{p_0(1-p_0)}{n}}].$$

- No caso de hipótese alternativa unilateral inferior: $H_0 : p = p_0$ e $H_1 : p < p_0$, a região de aceitação é

$$[p_0 + \Phi^{-1}(\alpha)\sqrt{\frac{p_0(1-p_0)}{n}}, \infty).$$

Exemplo 11.3.1: Um relatório afirma que 40% de toda água obtida através de poços artesianos é salobra. Existem controvérsias sobre esta afirmação, alguns dizem que a proporção é maior outros que é menor. Para acabar com a dúvida, sorteou-se 400 poços e observou-se que em 120 deles a água era salobra. Qual devia ser a conclusão ao nível de significância de 3%?

Solução: Neste caso, estamos testando $H_0 : p = 0{,}4$ contra uma hipótese alternativa bilateral $H_1 : p \neq 0{,}4$. Logo, a região de aceitação é dada por:

$$[0{,}4-\Phi^{-1}(0{,}985)\sqrt{\frac{(0{,}4)(0{,}6)}{400}}, 0{,}4+\Phi^{-1}(0{,}985)\sqrt{\frac{(0{,}4)(0{,}6)}{400}}] = [0{,}4-0{,}053; 0{,}4+0{,}053] = [0{,}347, ; 0{,}453].$$

Como $\hat{p} = 120/400 = 0{,}3$, podemos rejeitar a hipótese nula ao nível de confiança de 3%.

Exemplo 11.3.2: O governo afirma que a taxa de desemprego é de no máximo 15% da população economicamente ativa. Uma amostra aleatória de 1500 pessoas revelou que 1300 destas pessoas estão empregadas. Para um nível de significância de 5%, pode-se dizer que a afirmação está correta?

Solução: Neste caso, temos a hipótese nula $H_0 : p = 0{,}15$ contra a hipótese alternativa $H_1 : p < 0{,}15$. Logo, a região de aceitação é dada por:

$$[0{,}15 + \Phi^{-1}(0{,}05)\sqrt{\frac{(0{,}15)(0{,}85)}{1500}}, \infty] = [0{,}135; +\infty).$$

Como $\hat{p} = 200/1500 = 0{,}133$, podemos rejeitar a hipótese nula ao nível de confiança de 5%, e portanto, concluir que a afirmação estava correta.

11.3.2 Testes para Amostras Grandes

Como se $n \geq 30$, a variância da amostra s^2 é próxima de σ^2, temos que s pode ser usado no lugar de σ nos procedimentos acima sem grande prejuízo aos cálculos. Deste modo o teste para a média de uma população com variância conhecida pode ser utilizado, no caso de $n \geq 30$, para testar a média de uma população com variância desconhecida. O tratamento exato no caso em que σ^2 é desconhecida e a amostra é pequena envolve o uso da distribuição t de student e será estudado mais adiante.

11.4 Teste Sobre a Média de Uma População Normal com Variância Desconhecida

Assim como no caso de intervalos de confiança, quando a amostra for pequena e σ^2 desconhecida, teremos de fazer uma suposição sobre a forma da distribuição em estudo. Assumiremos nesta seção que a população tem uma distribuição normal. Já vimos que se a população tem uma distribuição normal, então $T = \frac{\sqrt{n}(\overline{X}-\mu)}{S}$ tem uma distribuição t de student com $n-1$ graus de liberdade. Seja $\tau(\gamma, n-1)$ o valor tal que $P(T \leq \tau(\gamma, n-1)) = \gamma$. Então, utilizando um procedimento similar ao caso de variância conhecida, podemos verificar que se a estatística utilizada for a média amostral $\overline{X}$, então

- no caso de hipótese alternativa bilateral: $H_0 : \mu = \mu_0$ e $H_1 : \mu \neq \mu_0$, a região de aceitação é

$$[\mu_0 - \tau(1-\alpha/2, n-1)\frac{S}{\sqrt{n}}, \mu_0 + \tau(1-\alpha/2, n-1)\frac{S}{\sqrt{n}}];$$

- no caso de hipótese alternativa unilateral superior: $H_0 : \mu = \mu_0$ e $H_1 : \mu > \mu_0$, a região de aceitação é

$$(-\infty, \mu_0 + \tau(1-\alpha, n-1)\frac{S}{\sqrt{n}}];$$

- no caso de hipótese alternativa unilateral inferior: $H_0 : \mu = \mu_0$ e $H_1 : \mu < \mu_0$, a região de aceitação é

$$[\mu_0 + \tau(\alpha, n-1)\frac{S}{\sqrt{n}}, \infty).$$

Exemplo 11.4.1: O McDonald's pretende instalar uma nova lanchonete se no local transitarem no mínimo 200 carros por hora durante certos períodos do dia. Para 20 horas selecionadas aleatoriamente durante tais períodos, o número médio de carros que transitaram pelo lugar foi de 208,5 com desvio padrão de 30,0. O gerente assume a hipótese de que o volume de carro não satisfaz a exigência de 200 ou mais carros por hora. Para um nível de significância de 5% esta hipótese pode ser rejeitada?

Solução: Neste caso, temos que a hipótese nula é dada por $H_0 : \mu = 200$ e a hipótese alternativa é $H_1 : \mu > 200$. Como a amostra é pequena (<30) e a variância da população é

desconhecida, devemos usar o teste t de student unilateral superior. Neste caso a região de aceitação é dada por:

$$(-\infty, 200 + \tau(0{,}95, 19)\frac{30}{\sqrt{20}}] = (-\infty, 200 + 1{,}729\frac{30}{\sqrt{20}}] = (-\infty, 211{,}6].$$

Portanto, a hipótese não pode ser rejeitada a este nível de confiança.

Exemplo 11.4.2: Num estudo sobre resistência de um dado material, com distribuição normal, foi coletada uma amostra de 25 unidades, resultando num valor médio de 230,4Kg e desvio-padrão de 100Kg. O estudo está interessado em saber se essa amostra é suficiente para garantir ao nível de significância de 5% que a resistência média do material seja superior a 200Kg. Qual a sua conclusão?

Solução: O estudo quer realizar o seguinte teste: $H_0 : \mu = 200$ contra $H_1 : \mu > 200$. Como a variância é desconhecida e a amostra é menor que 30, devemos utilizar o teste t de student. Neste caso, a região de aceitação é

$$(-\infty, 200 + \tau(0, 95, 24)\frac{100}{\sqrt{25}}] = (-\infty, 234{,}2].$$

Logo, a amostra não é grande o suficiente para garantirmos que a resistência média é maior que 200 ao nível de significância de 5%.

11.5 Probabilidade de Significância

O procedimento do testes de hipóteses descrito até agora parte de pré-estabelecimento de um valor para α. Deste modo como a escolha de α é arbitrária pode acontecer que para um determinado valor de α a hipótese nula seja rejeitada, porém para um valor menor de α ela não seja rejeitada. Além disso, no procedimento descrito se a estimativa do parâmetro caia na região crítica a hipótese nula era rejeitada e nenhuma informação a respeito de quão próximo essa estimativa estava da região de aceitação. Uma maneira alternativa para evitarmos tais problemas consiste em apresentar a *probabilidade de significância*, *nível descritivo*, ou *p-valor* do teste. Os passos são muito parecidos, sé que ao invés de construirmos a região crítica, apresentamos o valor da probabilidade de ocorrerem valores da estatística mais extremos que o observado quando a hipótese nula é verdadeira. O p-valor também pode ser definido como o menor nível de significância que conduz a rejeição da hipótese nula com os dados observados.

Suponha que estejamos no caso de um teste para a média de uma população com variância conhecida (ou então variância desconhecida, mas amostra grande). Seja $\overline{x}_0$ a média amostral observada na amostra. Então, para um teste bilateral $H_0 : \mu = \mu_0$ e $H_1 : \mu \neq \mu_0$, temos

$$p = P(|\overline{X} - \mu_0| > |\overline{x}_0 - \mu_0|) = P(\frac{\sqrt{n}|\overline{X} - \mu_0|}{\sigma} > \frac{\sqrt{n}|\overline{x}_0 - \mu_0|}{\sigma})$$
$$P(|Z| > \frac{\sqrt{n}|\overline{x}_0 - \mu_0|}{\sigma}) = 2(1 - \Phi(\frac{\sqrt{n}|\overline{x}_0 - \mu_0|}{\sigma})).$$

Similarmente, para um teste unilateral superior $H_0 : \mu = \mu_0$ e $H_1 : \mu > \mu_0$, temos:

$$p = P(\overline{X} > \overline{x}_0) = P(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} > \frac{\sqrt{n}(\overline{x}_0 - \mu_0)}{\sigma})$$
$$P(Z > \frac{\sqrt{n}|\overline{x}_0 - \mu_0|}{\sigma}) = 1 - \Phi(\frac{\sqrt{n}|\overline{x}_0 - \mu_0|}{\sigma}).$$

Finalmente, para um teste unilateral inferior $H_0 : \mu = \mu_0$ e $H_1 : \mu < \mu_0$, temos:

$$p = P(\overline{X} < \overline{x}_0) = P(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} < \frac{\sqrt{n}(\overline{x}_0 - \mu_0)}{\sigma})$$
$$P(Z < \frac{\sqrt{n}|\overline{x}_0 - \mu_0|}{\sigma}) = \Phi(\frac{\sqrt{n}|\overline{x}_0 - \mu_0|}{\sigma}).$$

Exemplo 11.5.1: Suponha novamente a situação anterior onde queremos testar a hipótese nula $H_0 : \mu = 220v$ *versus* $H_1 : \mu \neq 220v$, temos uma amostra de tamanho 4 e sabemos que a variância é igual a $64v^2$. Suponha ainda que a média amostral deu igual a $227v$, podemos então calcular o p-valor:

$$p = 2(1 - \Phi(\frac{\sqrt{4}|227 - 220|}{\sqrt{64}})) = 2(1 - \Phi(\frac{7}{4})) = 2(1 - 0{,}9599) = 0{,}0802.$$

Portanto, a probabilidade de quando a hipótese nula é verdadeira uma amostra selecionada de tamanho 4 tenha média amostral mais distante de $220v$ que $227v$ é igual a 0,0802, ou ainda, a um nível de significância de 10% a hipótese nula seria rejeitada, mas a um nível de significância de 5% a hipótese nula não pode ser rejeitada.

Temos então que rejeitaremos a hipótese H_0 se o p-valor for "bastante pequeno". A tabela a seguir ilustra a escala de evidências de Fisher contra a hipótese H_0:

p-valor	0,1	0,05	0,025	0,001	0,005	0,001
Natureza da Evidência	marginal	moderada	substancial	forte	muito forte	fortíssima

11.6 Significância Estatística *versus* Significância Prática

Quando aplicamos o procedimento de um teste de hipótese na prática precisamos além de considerar a significância estatística medida pelo p-valor, considerar quais diferenças entre valores dos parâmetros tem implicações práticas. Isto é, pode acontecer que o p-valor seja pequeno levando então a rejeição da hipótese H_0, mas que o desvio real entre o valor do parâmetro na hipótese nula e a estimativa do parâmetro obtida na amostra não tenha *significância prática*. Isto pode ocorrer por exemplo, para tamanhos de amostras grandes. Por exemplo, se obtivéssemos uma amostra de 1600 medidas o observássemos a média amostral de $220{,}5v$, então obteríamos o p-valor bilateral de

$$p = 2(1 - \Phi(\frac{\sqrt{1600}(|220{,}5 - 220|)}{\sqrt{64}})) = 2(1 - \Phi(20/8)) = 0{,}0124.$$

Portanto, temos uma evidência estatística substancial para rejeitarmos H_0. Contudo, do ponto de vista prático, se a média for realmente for $220{,}5v$ não haverá nenhum efeito prático observável no desempenho de qualquer equipamento elétrico. log, esta diferença detectada pelo teste de hipótese apesar de ter significância estatística não tem significância prática.

Logo devemos ter cuidado ao interpretar os resultados de um teste de hipótese principalmente quando a amostra tiver tamanho grande, pois qualquer desvio pequeno do valor do parâmetro testado na hipótese nula será detectado como tendo significância estatística pelo teste, contudo em muitos casos esta diferença poderá ter pouca ou nenhuma significância prática.

Referências Bibliográficas

Livros Textos:

1. Meyer, P. (1983), “Probabilidade - Aplicações à Estatística”, 2a. edição, Livros Técnicos e Científicos Editora, Rio de Janeiro.

2. Bussab, W. & Moretin, P. (2002), “Estatística Básica”, 5a. edição, Saraiva, São Paulo.

Livros Suplementares:

1. Davenport Jr., W. (1987), "Probability and Random Processes - an introduction for applied scientists and engineers", McGraw-Hill Book Company Inc.

2. Fine, T. (2006), “Probability and Probabilistic Reasoning for Electrical Engineering”, Prentice Hall.

3. Montgomery, D. & Runger, G. (2003), “Estatística e Aplicada e Probabilidade para Engenheiros”, 2a. edição, LTC, Rio de Janeiro.