

Resumo e Listas: Monte Carlo e Cadeias de Markov

Amanda Ferreira de Azevedo - afazevedo@cos.ufrj.br

PESC/UFRJ — 30 de julho de 2021

Sumário

1	Revisão de Probabilidade	3
1.1	Espaço amostral	3
1.2	Probabilidade	3
1.3	Probabilidade Condicional	5
1.4	Lei da Probabilidade Total	5
1.5	Regra de Bayes	6
1.6	Variável Aleatória	6
1.7	Variável Aleatória Indicadora	7
1.7.1	Função de probabilidade	7
1.8	Distribuição de Bernoulli	8
1.8.1	Distribuição Binomial	8
1.8.2	Distribuição Geométrica	9
1.8.3	Distribuição Zeta	9
1.9	Valor esperado	10
1.10	Variância	10
1.11	Propriedades Importantes	11
1.12	Distribuição conjunta	11
1.13	Independência de v.a.	11
1.14	Valor esperado condicional	12
1.15	Espaço amostral contínuo	12
1.15.1	Função densidade	12
2	Lista 1 - Revisão de Probabilidade	12
3	Limitantes para probabilidade	24
3.1	Cauda e Cabeça	24
3.2	Desigualdade de Markov	24
3.3	Desigualdade de Chebyshev	25
3.4	Desigualdade de Chernoff	25
3.5	With High Probability (whp)	26
3.6	Limitante da União	26
4	Lei dos grandes números	27
4.1	Média Amostral	27
4.2	Lei Fraca dos grandes números	28
4.3	Lei forte dos grandes números	28
5	Erro e Confiança	29
5.1	Margem de erro	30
5.2	Falácia do apostador	30

6	Método de Monte Carlo	30
6.1	Vantagens e Desvantagens	32
6.2	Calculando erro	32
6.3	Estimando π	32
6.3.1	Erro de π	33
6.4	Integração de Monte Carlo	33
6.5	Monte Carlo Ray Tracing	34
7	Gerando amostras de variáveis aleatórias discretas	34
7.1	Gerando Amostras Uniformes	34
7.2	Gerando outras distribuições	34
7.3	Gerando Geométrica	36
7.4	Método da transformada inversa	38
7.5	Gerando Binomial	38
7.6	Gerando permutações	39
8	Rejection Sampling (Amostragem por rejeição)	40
8.1	Exemplo 1	42
8.2	Exemplo 2	42
8.3	Cenário Problemático	43
8.4	Importance Sampling	43
8.5	Generalização	44
9	Lista 2 - Algoritmos de Monte Carlo	45
10	Cadeias de Markov	56
10.1	Definição e Exemplos	57
10.2	Sem memória	57
10.3	Distribuição no tempo	58
10.4	Comunicação entre estados	58
10.5	Irredutibilidade	58
10.6	Periodicidade	59
11	Convergência da CM	59
11.1	Tempo de Chegada	59
11.2	Encontrando a distribuição estacionária	60
11.3	Reversibilidade	60
12	Autovalores e Autovetores	60
12.1	Tempo para Convergência	61
12.2	Spectral Gap	62
12.3	Spectral Gap e Tempo de Mistura	62
13	Caminho amostral	62
14	Teorema Ergódico	63
14.1	Estimando a distribuição estacionária	63
14.2	Simular uma Cadeia de Markov	63
14.3	Gerando amostras de X_t ?	64
14.4	Gerando Amostras Estacionárias	64
14.5	Simulando Cadeias de Markov Grandes	64
14.6	Amostrando Espaços Complicados	65
15	Lista 3 - Cadeias de Markov	65

16 Monte Carlo e Cadeias de Markov	71
16.1 Metropolis-Hastings	71
16.1.1 Caso simétrico	71
16.1.2 Caso Geral	72
16.2 Gibbs Sampling	72
17 Otimização	73
17.1 Exemplo do Caxeiro Viajante	73
18 Simulated Annealing	73
18.1 Estratégias de Resfriamento	74
18.2 Voltando ao Caixeiro	74
19 Lista 4 - Cadeias de Markov e Algoritmos de Monte Carlo	75

1 Revisão de Probabilidade

1.1 Espaço amostral

Um conjunto de objetos que você pode contar, como por exemplo um conjunto de letras.

S é um conjunto:

- Ex: $S = \{a, b, c, \dots, z\}$
- Ex: $|S| = 26$ quantidade de elementos

1.2 Probabilidade

Função que associa cada elemento de S um valor entre 0 e 1.

$$p : S \rightarrow [0, 1]$$

Função que mapeia os elementos de S em um número real entre 0 e 1.

Para p ser uma função de *probabilidade* esta deve atender a seguinte restrição:

- A soma de todos os elementos da imagem tem que ser igual a 1.

Cada função terá sua regra de mapeamento. Tomando o espaço amostral das letras, algumas regras:

- $p_x = \frac{1}{26}$ para qualquer letra x .
- $p_x = \frac{1}{10}$ para qualquer vogal e $\frac{1}{42}$ para consoantes.

Um **evento** é um subconjunto do espaço amostral. Utilizando o espaço amostral das letras, alguns exemplos de eventos:

- $A = \{a, b, c, d\}$
- $B =$ todas as consoantes
- $C =$ todas as letras depois de Q

Definimos como a probabilidade de um evento como a soma das probabilidades de cada elemento dele, isto é:

$$P[A] = \sum_{e \in A} p_e$$

Leia-se: Probabilidade associada ao evento A é a soma das probabilidades associadas a cada elemento e do conjunto A. Exemplo:

- $P[A] = \frac{1}{26} + \frac{1}{26} + \frac{1}{26} + \frac{1}{26} = \frac{4}{26} = \frac{2}{13}$
- $P[B] = \sum_1^{21} \frac{1}{26} = \frac{21}{26}$
- $P[C] = \sum_1^9 \frac{1}{26} = \frac{9}{26}$

Como eventos são *conjuntos*, existem algumas operações básicas para manipular os eventos, como **união**, **interseção** e **complemento**. Exemplo:

- $A \cup B$ = todas as consoantes mais a letra a .
- $A \cap B$ = todas as consoantes b, c, d .
- B^c = todas as vogais.

A probabilidade de eventos resultantes destas operações seguem a mesma lógica, se por exemplo tomarmos o evento $Y = A \cap B$ a $P[Y] = \sum_{e \in Y} p_e$.

Notação:

- $A \cup B = A + B = A \vee B$
- $A \cap B = A.B = A \wedge B$
- $A \cap B = \bar{B} = \neg B$

Definição 1: $P[A \cap B] = P[A \wedge B] = P[A].P[B] \Leftrightarrow$ os eventos A e B são *independentes*.

Exercício 1: $A = \{a, b, c, d\}$ e B = todas as consoantes. São independentes?

- $P[A] = \frac{2}{13}$
- $P[B] = \frac{21}{26}$
- $P[Y = A \cap B] = \frac{1}{26} + \frac{1}{26} + \frac{1}{26} = \frac{3}{26}$

$$P[Y = A \cap B] = \frac{3}{26} \Leftrightarrow P[A].P[B] = \frac{2}{13} \cdot \frac{21}{26} = \frac{42}{338}?$$

Falso. Logo, A e B não são independentes.

Exercício 2: $A = \{a, z\}$ e B = todas as letras antes de N . São independentes?

- $P[A] = \frac{1}{26} + \frac{1}{26} = \frac{2}{26} = \frac{1}{13}$
- $P[B] = \frac{13}{26}$
- $P[Y = A \cap B] = P[Y = \{a\}] = \frac{1}{26}$

$$P[Y] = \frac{1}{26} \Leftrightarrow P[A].P[B] = \frac{1}{13} \cdot \frac{13}{26} = \frac{13}{338} = \frac{1}{26}?$$

Verdade. Logo, A e B são independentes.

Definição 2: Dois eventos A e B são ditos mutualmente exclusivos se $P[A \cup B] = P[A \vee B] = P[A] + P[B]$.

Exercício 3: $A = \{a, b, c, d\}$ e B = todas as consoantes. A e B são mutualmente exclusivos?

- $P[A] = \frac{4}{26}$
- $P[B] = \frac{21}{26}$

- $P[Y = A \cup B] = \frac{22}{26}$

$$P[Y] = \frac{22}{26} \Leftrightarrow P[A] + P[B] = \frac{4}{26} + \frac{21}{26} = \frac{25}{26}?$$

Falso. Logo, A e B não são mutuamente exclusivos.

Exercício 2: $A = \{x, y, z\}$ e $B =$ todas as letras antes de N . São mutuamente exclusivos?

- $P[A] = \frac{1}{26} + \frac{1}{26} + \frac{1}{26} = \frac{3}{26}$
- $P[B] = \frac{13}{26}$
- $P[Y = A \cup B] = \frac{16}{26}$

$$P[Y] = \frac{16}{26} \Leftrightarrow P[A] + P[B] = \frac{3}{26} + \frac{13}{26} = \frac{16}{26}?$$

Correto. Logo, A e B são mutuamente exclusivos.

1.3 Probabilidade Condicional

É a probabilidade de um evento A dada uma ocorrência de um evento B . Reduzir o espaço amostral para o evento que ocorreu.

$$P[A|B] = \frac{P[A \cap B]}{P[B]}$$

Exercício 5: Seja $A = \{a, b, c, d\}$, $B =$ todas as consoantes e $C = \{x, y, z\}$ qual $P[A|B]$? $P[B|A]$? $P[C|B]$?

- $P[A] = \frac{4}{26}$
- $P[B] = \frac{19}{26}$
- $P[C] = \frac{3}{26}$
- $P[A \cap B] = P[B \cap A] = P[\{b, c, d\}] = \frac{3}{26}$
- $P[B \cap C] = P[C] = \frac{3}{26}$

$$P[A|B] = \frac{\frac{3}{26}}{\frac{19}{26}} \Leftrightarrow P[A|B] = \frac{3}{19}$$

$$P[B|A] = \frac{\frac{3}{26}}{\frac{4}{26}} \Leftrightarrow P[B|A] = \frac{3}{4}$$

$$P[C|B] = \frac{\frac{3}{26}}{\frac{19}{26}} \Leftrightarrow P[C|B] = \frac{3}{19}$$

1.4 Lei da Probabilidade Total

Particionamos um espaço amostral S em subconjuntos tal que todo elemento deste conjunto aparece em uma única parte desta partição.

Exemplo: S o espaço amostral das letras. Definimos subconjuntos O_1 : todas as vogais e O_2 : todas as consoantes.

A lei da probabilidade total relaciona a probabilidade de um evento e a probabilidade condicional com eventos de uma partição.

Ou seja, dado um evento qualquer, devemos observar que parte do evento está em que partição. Desta forma, se particionarmos o espaço amostral em $|O|$ partições, a probabilidade de um evento A pode ser dada por:

$$P[A] = \sum_i P[A \cap O_i]$$

Note que: Se houver apenas uma partição própria, isto é, $O_1 = S$, a interseção de A com o espaço amostral é o próprio A .

Dessa forma, podemos desmembrar as probabilidades olhando para a interseção do evento A com cada partição.

Exemplo: Seja S o espaço amostral das letras do alfabeto. Considere as partições O_1 e O_2 . Tome o evento $A = \{a, b, c, d\}$.

- $P[A] = \frac{4}{26}$
- $P[A \cap O_1] = \frac{1}{26}$
- $P[A \cap O_2] = \frac{3}{26}$

$$P[A] = \frac{4}{26} \Leftrightarrow P[A \cap O_1] + P[A \cap O_2] = \frac{1}{26} + \frac{3}{26} = \frac{4}{26}$$

A partir da probabilidade condicional, podemos reescrever a equação da seguinte maneira:

$$\sum_i P[A \cap O_i] = \sum_i P[A|O_i] \cdot P[O_i]$$

1.5 Regra de Bayes

Relação entre probabilidades condicionais. É muito utilizado quando é mais fácil de mensurar a probabilidade condicional contrária.

$$P[A|B] = \frac{P[B|A]P[A]}{P[B]}$$

Ou seja, dados dois eventos A e B , quando é mais fácil calcular $P[B|A]$ do que $P[A|B]$.

1.6 Variável Aleatória

Uma variável aleatória X é uma função que mapeia o espaço amostral S nos inteiros. Assim,

$$X : S \rightarrow \mathbb{Z}$$

Dessa forma, seremos capazes de trabalhar com números inteiros. Podemos usar X para definir eventos em função de seus valores. Exemplo:

$$A = \{X > 5\} = \{e \in S : X(e) > 5\}$$

Ou seja, podemos usar uma variável aleatória para modelar um evento qualquer, neste caso, o evento A são todos os elementos de S tal que a imagem $X(e) > 5$.

Exemplo: Usamos o espaço amostral das letras do alfabeto, onde $|S| = 26$ e a probabilidade de uma dada letra x é $p_x = \frac{1}{26}$, que é *uniforme*.

Definimos X uma variável aleatória que mapeia cada letra e um número. Por exemplo, $X(a) = 1, X(b) = 2, \dots, X(z) = 26$. Outra variável aleatória que podemos considerar é a Y , tal que $Y(\text{vogal}) = 1$ e $Y(\text{consoante}) = 2$. Outra que podemos considerar é uma variável aleatória Z onde $Z(\text{primeiras10letras}) = 1, Z(\text{ultimas10letras}) = 2$ e $Z(\text{oquesobrou}) = 3$.

Definindo eventos a partir de condicionais:

- $A = \{X < 5\} = \{a, b, c, d\}$
- $B = \{Y = 1\} = \{a, e, i, o, u\}$
- $Z = \{Z = 3\} = \{k, l, m, n, o, p\}$

A vantagem de usar variáveis aleatórias para definir eventos é poder calcular sua probabilidade de forma mais fácil.

Exercício 7: $P[X < 5] = ?$

Como mapeamos os elementos em números inteiros, podemos desmembrar da seguinte forma:

$$P[X < 5] = P[\{a, b, c, d\}] = P[X = 1] + P[X = 2] + P[X = 3] + P[X = 4] = \frac{1}{26} + \frac{1}{26} + \frac{1}{26} + \frac{1}{26} = \frac{4}{26}$$

Exercício 8: $P[Y = 1] = ?$

$$P[Y = 1] = P[\{a, e, i, o, u\}] = \frac{5}{26}$$

Exercício 9: $P[Y = 1 \text{ e } Z = 3] = ?$

$$P[Z = 3] = P[\{k, l, m, n, o, p\}] = \frac{6}{26}$$

Se $A = \{a, e, i, o, u\}$ e $B = \{k, l, m, n, o, p\}$, queremos calcular $P[A \cap B]$. Os eventos são independentes?

$$P[A \cap B] = \frac{11}{26} \Leftrightarrow P[A] \cdot P[B] = \frac{5}{26} \cdot \frac{6}{26} = \frac{36}{676}$$

Não são independentes.

Podemos também manipular variáveis aleatórias. Podemos fazer as seguintes operações:

- Multiplicação por escalar
- Novo mapeamento para um espaço amostral: $Y = 2X + 1$, $Z = X + Y$
- Simplificação de eventos: $2X > 4 = X > 2$

1.7 Variável Aleatória Indicadora

É uma das variáveis aleatórias mais importantes e utilizadas, assume apenas dois valores, (1 ou 0) e indica a ocorrência de um evento de interesse.

Exemplo: S = todos os grafos com n vértices

- $X = 1$: todos os grafos conexos, $X = 0$ caso contrário.
- $Y_K = 1$: todos os grafos com diâmetro k , $Y_k = 0$ caso contrário.

1.7.1 Função de probabilidade

Até agora, quando utilizamos uma variável aleatória para definir eventos nós olhamos, a partir do mapeamento, a probabilidade de cada evento e somamos. Porém, existe uma maneira de olhar para a probabilidade diretamente da variável aleatória. Faremos isso utilizando uma *função de probabilidade* que nos ajudará a calcular a probabilidade de uma variável aleatória tomar certo valor.

Exemplo: $P[X = 1] = p_1$. Definiremos uma função de probabilidade que nos ajude a calcular p_1 , por exemplo.

Definição: Seja X uma variável aleatória e x um dos seus possíveis valores. Associaremos um número $f_X(x) = P[X = x]$, denominado probabilidade de x . Seja O_X o domínio da variável aleatória X , uma função de probabilidade $f_X(x)$ deve satisfazer as seguintes condições:

- $0 \leq f_X(x) \leq 1$, para todo $x \in O_X$
- $\sum_{x \in O_X} f_X(x) = 1$

Em especial, chamamos de *função cumulativa de probabilidade* quando queremos calcular a probabilidade de uma variável aleatória estar limitada por certos valores.

- **Função de probabilidade:** $f_X(x) = P[X = x]$
- **Função cumulativa:** $F_X(x) = P[X \leq x]$

É possível calcular a função cumulativa através das funções de probabilidade, através da soma por exclusão mútua dos elementos de X . Ou seja:

$$F_X(x) = \sum_{y \leq x} f_X(y)$$

Considere uma *sequência* de n variáveis aleatórias X_1, X_2, \dots, X_n .

Diremos que uma sequência é *i.i.d.* (independente e identicamente distribuída) se as variáveis aleatórias:

- são independentes;
- possuem a mesma função distribuição;
- são distintas.

Exemplo: Jogar um mesmo dado n vezes: X_i é o valor observado na i -ésima jogada do dado.

1.8 Distribuição de Bernoulli

Uma variável aleatória possui distribuição de *Bernoulli* assume apenas dois valores. Ou seja:

- $f_X(1) = P[X = 1] = p$
- $f_X(0) = P[X = 0] = 1 - p$

Essa distribuição ajuda a dar valores para variáveis aleatórias **indicadoras**!

Essa distribuição possui apenas um parâmetro, p . A única restrição é que $0 < p < 1$. Dizemos que uma variável X possui distribuição de *Bernoulli* de parâmetro p por $X \sim \text{Bernoulli}(p)$.

Comumente utilizada em:

- Cara ou coroa, sim ou não, verdade ou falso

O parâmetro p na distribuição de bernoulli, porquê não pode assumir 0 ou 1? Não seria menor ou igual a 0 e maior igual a 1?

1.8.1 Distribuição Binomial

Considere uma sequência *iid* de n variáveis aleatórias de Bernoulli X_1, X_2, \dots, X_n

- $X_i \sim \text{Bernoulli}(p)$ para $i = 1, 2, \dots, n$

Seja $Z = \sum_{i=1}^n X_i$ a soma destas variáveis aleatórias.

Essa distribuição possui dois parâmetros: n o número de variáveis aleatórias de Bernoulli e p o parâmetro da Bernoulli.

Observando o somatório, Z pode assumir valores entre 0 e n , pois cada variável aleatória de Bernoulli pode assumir, no máximo, 1.

A probabilidade de uma distribuição binomial segue a seguinte ideia:

$$f_Z(i) = P[Z = i] = \binom{n}{i} p^i (1-p)^{n-i}$$

Notação: $Z \sim \text{Bin}(n, p)$

Exemplos:

- Número de caras ao jogar uma moeda 20 vezes;
- Número de resultados pares ao jogar um dado 10 vezes.

1.8.2 Distribuição Geométrica

Considere uma sequência iid de n variáveis aleatórias de Bernoulli X_1, X_2, \dots, X_n

- $X_i \sim \text{Bernoulli}(p)$ para $i = 1, 2, \dots, n$

Seja Z o menor valor tal que $X_z = 1$, isto é, a variável aleatória nesta distribuição identifica o primeiro índice onde a variável aleatória de Bernoulli toma o valor 1. Em outras palavras, $Z = \min\{i | X_i = 1\}$.

Z possui distribuição Geométrica com parâmetro p . Esta variável aleatória pode assumir valores entre 0 e n também, uma vez que representa os possíveis índices de n Bernoullis.

A probabilidade de uma distribuição geométrica segue a seguinte ideia:

$$f_z(i) = P[Z = i] = (1-p)^{i-1} \cdot p \quad i = 1, 2, \dots$$

Notação: $Z \sim \text{Geo}(p)$

Exemplos:

- Número de vezes que uma moeda é jogada até a primeira cara;
- Número de elementos inseridos em tabela *hash* até a colisão com elemento em uma posição fixa.

1.8.3 Distribuição Zeta

Seja Z uma variável aleatória com distribuição Zeta com parâmetro $s > 1$. A sua distribuição de probabilidade é a seguinte:

$$f_Z(i) = P[Z = i] = \frac{C(s)}{i^s}, \quad i = 1, 2, \dots$$

Onde $C(s)$ é a constante de normalização, definida pela função zeta de Riemann.

Zeta é uma distribuição que possui a característica *cauda pesada* onde a probabilidade decai bem mais devagar do que uma exponencial.

Exemplos:

- Número de vezes que uma palavra ocorre na Wikipedia;
- Número de seguidores de um perfil no Twitter.

1.9 Valor esperado

A função de probabilidade caracteriza por completo um comportamento de uma variável aleatória, mas muitas vezes desconhecemos ela. Nesse caso, é melhor falar de um *resumo* dessa função, ou seja, a média que será representada pelo *valor esperado*.

$$\mu_X = E[X] = \sum_{i \in O_X} i f_X(i)$$

Valores Esperados das Distribuições:

- $X \sim \text{Bernoulli}(p)$: Aplicando a definição $E[X] = 0(1-p) + 1p = p$
- $X \sim \text{Bin}(n, p)$: np
- $X \sim \text{Geo}(p)$: $\frac{1}{p}$
- $X \sim \text{Zeta}(s)$: $\frac{C(s-1)}{C(s)}$ $s > 2$

Seja g uma função qualquer tal que: $g : \mathbb{Z} \rightarrow \mathbb{R}$. Podemos aplicar g a X .

$$\sum_{i \in O_X} g(i) f_X(i)$$

1.10 Variância

A variância é outro resumo de um comportamento de uma variável aleatória. É uma medida de *dispersão* ao redor da média.

Podemos definir uma função de uma variável aleatória g da seguinte maneira:

$$g(X) = (x - \mu)^2$$

É uma função que calcula a diferença entre a variável aleatória e seu valor esperado e eleva ao quadrado. Com isso, podemos finalmente calcular a variância, que será o valor esperado da função acima:

$$\sigma_X^2 = \text{Var}[X] = E[g(X)] = E[(X - \mu)^2]$$

A partir disso, podemos definir também o *desvio padrão*, que será a raiz quadrada da variância:

$$\sigma_X = \sqrt{\text{Var}[X]} = \sqrt{\sigma_X^2}$$

Variância das Distribuições:

- $X \sim \text{Bernoulli}(p)$: Aplicando a definição $E[g(X)] = g(0)(1-p) + g(1)p = (0-p)^2(1-p) + (1-p)^2p = p(1-p)$
- $X \sim \text{Bin}(n, p)$: $np(1-p)$
- $X \sim \text{Geo}(p)$: $\frac{1-p}{p^2}$
- $X \sim \text{Zeta}(s)$: $\frac{C(s-2)}{C(s-1)}$ $s > 3$

1.11 Propriedades Importantes

Linearidade da esperança:

$$E[X + Y] = E[X] + E[Y]$$

Sejam X e Y suas variáveis aleatórias independentes:

$$E[XY] = E[X]E[Y]$$

$$Var[X + Y] = Var[X] + Var[Y]$$

Seja $Y = aX + b$, para constantes a, b :

$$E[Y] = aE[X] + b$$

$$Var[Y] = a^2 Var[X]$$

1.12 Distribuição conjunta

Eventos sobre mais de uma variável aleatória. Sejam X e Y duas variáveis aleatórias definidas sobre um mesmo espaço amostral, definimos uma *distribuição conjunta* de X e Y como:

$$f_{XY}(i, j) = P[X = i \cap Y = j]$$

Podemos calcular a distribuição simples de cada variável aleatória a partir da distribuição conjunta, chamadas **distribuição marginal**:

- $f_X(i) = \sum_{j \in O_Y} f_{XY}(i, j)$
- $f_Y(j) = \sum_{i \in O_X} f_{XY}(i, j)$

1.13 Independência de v.a.

X e Y são variáveis aleatórias independentes se e somente se:

$$f_{XY}(i, j) = P[X = i \cap Y = j] = P[X = i]P[Y = j] = f_X(i)f_Y(j)$$

Exemplos:

Dois dados honestos com k faces. Jogar uma moeda enviesada com probabilidade p para escolher o dado. X = face observada e Y = dado escolhido.

$$f_{XY}(i, j) = \frac{p}{k}, \text{ se } j = 1$$

$$f_{XY}(i, j) = \frac{1-p}{k}, \text{ se } j = 2$$

Exercício: São independentes?

- $f_X = P[X = i] = \text{probabilidade da face } i \text{ sair no dado} = \frac{1}{k}$
- $f_Y = P[Y = j] = \text{probabilidade do dado escolhido ser } j, \text{ a partir da moeda} = p \text{ se } j=1 \text{ e } 1-p \text{ se } j=2.$

$$f_X(i)f_Y(j) = P[X = i]P[Y = j] = \frac{1}{k}p = f_{XY}(i, j) \quad j = 1$$

$$f_X(i)f_Y(j) = P[X = i]P[Y = j] = \frac{1}{k}(1-p) = f_{XY}(i, j) \quad j = 2$$

1.14 Valor esperado condicional

Para isso, precisamos definir a **distribuição condicional** entre variáveis aleatórias. Sejam X e Y duas variáveis aleatórias, a distribuição condicional de X dado Y é:

$$f_{X|Y}(i, j) = P[X = i | Y = j]$$

Então podemos definir o valor esperado condicional calculando o valor esperado restrito a um subconjunto do espaço amostral (definido pelo valor da outra variável aleatória). Então o valor esperado de X dado que $Y = j$ é:

$$E[X|Y = j] = \sum_{i \in O_X} i f_{X|Y}(i, j)$$

Torre da esperança: $E[X] = E[E[X|Y]]$

1.15 Espaço amostral contínuo

Como associar a probabilidade a cada ponto do espaço amostral no espaço contínuo? A solução é dar probabilidade a *subconjuntos* do espaço amostral. Os eventos irão representar esses pedaços.

1.15.1 Função densidade

Seja A um evento qualquer de um espaço amostral qualquer S . Dizemos que $f(x)$ é uma função densidade se e somente se:

$$P[A] = \int_A f(x) dx$$

A probabilidade de um evento A no espaço contínuo é igual a área da função densidade dentro do espaço definido por A .

Para essa ser uma função densidade é necessário ter as seguintes propriedades:

- $0 \leq \int_A f(x) dx \leq 1$
- $\int_S f(x) dx = 1$

Exemplo: $S = [a, b]$ intervalo para a, b constantes. Seja $f(x) = \frac{1}{b-a}$, $x \in [a, b]$, e $A = [a', b']$, com $a' \geq a$ e $b' \leq b$.

$$P[A] = \int_{a'}^{b'} \frac{1}{b-a} dx = \frac{b' - a'}{b - a}$$

2 Lista 1 - Revisão de Probabilidade

1 Questão: Filhos e filhas

Considere um casal que tem dois descendentes e que as chances de cada um deles ser filho ou filha são iguais. Responda às perguntas abaixo:

1. Calcule a probabilidade dos descendentes formar um casal (ou seja, um filho e uma filha).
2. Calcule a probabilidade de ao menos um dos descendentes ser filho.
3. Calcule a probabilidade das duas serem filhas dado que uma é filha.
4. Calcule a probabilidade dos descendentes nascerem no mesmo dia (assuma que a chance de nascer em um determinado dia é igual a qualquer outro)

Resposta: Calcule a probabilidade dos descendentes formar um casal (ou seja, um filho e uma filha).

Considere, sem perda de generalidade, o evento A = ter uma filha menina. A probabilidade $P[A] = \frac{1}{2}$ e a probabilidade de ter um menino é $P[B = \neg A] = \frac{1}{2}$.

O item 1 nos pede a probabilidade dos descendentes formar um casal, isto é, a probabilidade de ter um filho e ter uma filha ou ter uma filha e um filho. Em outras palavras:

$$P[A \cap B] \cup P[B \cap A]$$

A $P[A \cap B]$ será a probabilidade de ter uma menina e um menino. Olhando para o espaço amostral, existe apenas uma única possibilidade em quatro. Logo, $P[A \cap B] = \frac{1}{4}$. De forma análoga, $P[B \cap A] = \frac{1}{4}$. Dessa forma, juntando essas duas possibilidades, temos que a probabilidade dos descendentes formar um casal é de $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.

Resposta: Calcule a probabilidade de ao menos um dos descendentes ser filho.

Para calcular o item 2, calcularemos a probabilidade de nasceram duas filhas. Ou seja, $P[A \cap A] = P[A] \cdot P[A] = \frac{1}{4}$. A probabilidade de ao menos um dos descendentes ser menino é exatamente o complementar de $P[A \cap A]$, ou seja: $1 - \frac{1}{4} = \frac{3}{4}$.

Resposta: Calcule a probabilidade das duas serem filhas dado que uma é filha.

Neste caso, utilizamos o conceito de probabilidade condicional. Dado que existe uma filha, nosso espaço amostral se restringiu a apenas casos que existem ao menos uma menina. Antes, tínhamos quatro casos possíveis. Agora, retirando o caso em que se tem dois meninos, nosso espaço amostral possui três casos possíveis. Desta forma, existe apenas um caso que podem nascer duas meninas, logo a probabilidade é $\frac{1}{3}$.

Resposta: Calcule a probabilidade dos descendentes nascerem no mesmo dia (assuma que a chance de nascer em um determinado dia é igual a qualquer outro)

Dado que o primeiro filho nasceu, a probabilidade de um segundo filho nascer no mesmo dia é independente do nascimento do primeiro filho. Sendo assim, como a probabilidade é uniforme (desconsiderando anos bissexto), esta probabilidade é $\frac{1}{365}$.

2 Questão: Dado

Considere um icosaedro (um sólido Platônico de 20 faces) tal que a chance de sair a face $i = 1, \dots, 20$ seja linearmente proporcional a i . Ou seja, $P[X = i] = c \cdot i$ para alguma constante c , onde X é uma variável aleatória que denota a face do dado. Responda às perguntas abaixo:

1. Determine o valor de c .
2. Calcule o valor esperado de X (obtenha também o valor numérico).
3. Calcule a probabilidade de X ser maior do que seu valor esperado.
4. Calcule a variância de X (obtenha também o valor numérico).
5. Repita os ultimo três itens para o caso do dado ser uniforme, ou seja, $P[X = i] = \frac{1}{20}$, $i = 1, \dots, 20$. Qual dado possui maior variância? Compare os resultados.

i

Resposta: Determine o valor de c .

Para calcular c usaremos a função de probabilidade da variável aleatória X . Uma de suas propriedades nos diz que:

$$1 = \sum_{i=1}^{20} P[X = i] = \sum_{i=1}^{20} c \cdot i = c \sum_{i=1}^{20} i = c \cdot 210$$

Portanto $c = \frac{1}{210}$.

i

Resposta: Calcule o valor esperado de X (obtenha também o valor numérico).

O valor esperado de X é igual a:

$$\mu_X = E[X] = \sum_{i \in O_X} i f_X(i)$$

Ou seja,

$$\sum_{i=1}^{20} i \cdot c \cdot i = c \sum_{i=1}^{20} i^2$$

Calculando numericamente, podemos utilizar a expressão:

$$\sum_{i=1}^{20} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Logo,

$$c \sum_{i=1}^{20} i^2 = \frac{1}{210} \cdot 2870 = \frac{41}{3}$$

i

Resposta: Calcule a probabilidade de X ser maior do que seu valor esperado.

A probabilidade de X ser maior que seu valor esperado é igual a:

$$P[X \geq \frac{41}{3}] = P[X \geq 14] = P[X = 14] + P[X = 15] + P[X = 16] + P[X = 17] + P[X = 18] + P[X = 19] + P[X = 20]$$

Que é igual a:

$$\frac{1}{210} \cdot [14 + 15 + 16 + 17 + 18 + 19 + 20] = \frac{119}{210} = \frac{17}{30}$$

i

Resposta: Calcule a variância de X (obtenha também o valor numérico).

Seja $g(X) = (X - \mu)^2$

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{20} g(i) f_X(i) = \sum_{i=1}^{20} (i - \mu)^2 \cdot \frac{i}{210} = \sum_{i=1}^{20} (i - \frac{41}{3})^2 \cdot \frac{i}{210} \\ \sum_{i=1}^{20} (i - \frac{41}{3})^2 \cdot \frac{i}{210} &= \sum_{i=1}^{20} (i^2 - 2\frac{41}{3}i + \frac{41^2}{3^2}) \cdot \frac{i}{210} = \frac{1}{210} \sum_{i=1}^{20} (i^3 - 2\frac{41}{3}i^2 + \frac{41^2}{3^2}i) \\ \frac{1}{210} \sum_{i=1}^{20} (i^3 - 2\frac{41}{3}i^2 + \frac{41^2}{3^2}i) &= \frac{1}{210} [\sum_{i=1}^{20} i^3 - \frac{82}{3} \sum_{i=1}^{20} i^2 + \frac{41^2}{3^2} \cdot \sum_{i=1}^{20} i = 210] \\ &= \frac{1}{210} \cdot \frac{20^2 \cdot 21^2}{4} - \frac{82}{3} \cdot \frac{41}{3} + \frac{41^2}{3^2} \\ &= 210 - \frac{3362}{9} + \frac{1681}{9} = 210 - \frac{1681}{9} = \frac{209}{9} \end{aligned}$$

i

Resposta: Repita os ultimo três itens para o caso do dado ser uniforme, ou seja, $P[X = i] = \frac{1}{20}$, $i = 1, \dots, 20$. Qual dado possui maior variância? Compare os resultados.

O valor esperado de X , neste caso, pode ser calculado como:

$$\begin{aligned} \mu_X = E[X] &= \sum_{i \in O_X} i f_X(i) = \sum_{i=1}^{20} i \cdot \frac{1}{20} = \frac{1}{20} \sum_{i=1}^{20} i = \frac{1}{20} \cdot 210 = \frac{21}{2} \\ P[X \geq 11] &= \sum_{i=11}^{20} \frac{1}{20} = \frac{10}{20} = \frac{1}{2} \end{aligned}$$

Seja $g(X) = (X - \mu)^2$

$$\begin{aligned} \sigma^2 &= \sum_{i=1}^{20} g(i) f_X(i) = \sum_{i=1}^{20} (i - \frac{21}{2})^2 \cdot \frac{1}{20} = \frac{1}{20} \sum_{i=1}^{20} (i^2 - 21i + \frac{21^2}{2^2}) \\ \frac{1}{20} [\sum_{i=1}^{20} i^2 - 21 \sum_{i=1}^{20} i + \frac{21^2}{2^2} \sum_{i=1}^{20} 1] &= \frac{1}{20} [2870 - 21 \cdot 210 + \frac{21^2}{2^2} \cdot 20] \\ \frac{2870}{20} - \frac{4410}{20} + \frac{21^2}{2^2} &= \frac{441}{4} - \frac{1540}{20} = \frac{2205}{20} - \frac{1540}{20} = \frac{665}{20} = \frac{133}{4} \end{aligned}$$

A variança da questão anterior deu, aproximadamente 23,32. Já esta, dá 33,25. Logo, esta variância é maior.

3 Questão: Dado em ação

Considere a versão uniforme do dado acima, ou seja, $P[X = i] = \frac{1}{20}, i = 1, \dots, 20$. Seja Y uma variável aleatória indicadora da primalidade da face do dado. Ou seja, $Y = 1$ quando o X é um número primo e $Y = 0$ caso contrário. Responda às perguntas abaixo:

1. Determine $P[Y = 1]$.
2. Considere que o dado será jogado n vezes. Seja Y_i a indicadora da primalidade da i -ésima rodada, para $i = 1, \dots, n$ e defina $Z = \sum_{i=1}^n Y_i$. Repare que Z é uma variável aleatória que denota o número de vezes que o resultado é primo. Determine a distribuição de Z , ou seja, $P[Z = k]$, para $k = 0, \dots, n$. Que distribuição é esta?
3. Considere que o dado será jogado até que um número primo seja obtido. Seja Y_i a indicadora da primalidade da i -ésima rodada, para $i = 1, \dots$ e defina $Z = \min\{i | Y_i = 1\}$. Repare que Z denota o número de vezes que o dado é jogado até que o resultado seja um número primo. Determine a distribuição de Z , ou seja, $P[Z = k]$, para $k = 1, \dots$. Que distribuição é esta?

i

Resposta: Determine $P[Y = 1]$.

Existem 8 números primos, sendo eles, 2,3,5,7,11,13,17,19. De 20 possíveis faces, temos 8 faces que são números primos. Dessa forma, $P[Y = 1] = \frac{8}{20} = \frac{2}{5}$.

i

Resposta: Considere que o dado será jogado n vezes. Seja Y_i a indicadora da primalidade da i -ésima rodada, para $i = 1, \dots, n$ e defina $Z = \sum_{i=1}^n Y_i$. Repare que Z é uma variável aleatória que denota o número de vezes que o resultado é primo. Determine a distribuição de Z , ou seja, $P[Z = k]$, para $k = 0, \dots, n$. Que distribuição é esta?

A variável indicadora Y_i possui uma distribuição de Bernoulli, de parâmetro $p = \frac{2}{5}$. Ou seja, $Y_i \sim \text{Bernoulli}(\frac{2}{5})$. Como Z é uma variável aleatória que é a soma de n variáveis aleatórias de Bernoulli, diremos que Z possui uma distribuição Binomial de parâmetros n e $p = \frac{2}{5}$. Ou seja, $Z \sim \text{Bin}(n, \frac{2}{5})$. A função de probabilidade associada a esta variável é:

$$P[Z = k] = \binom{n}{k} \left(\frac{2}{5}\right)^k \cdot \left(\frac{3}{5}\right)^{n-k} \quad k \in 0, \dots, n$$

i

Resposta: Considere que o dado será jogado até que um número primo seja obtido. Seja Y_i a indicadora da primalidade da i -ésima rodada, para $i = 1, \dots$ e defina $Z = \min\{i | Y_i = 1\}$. Repare que Z denota o número de vezes que o dado é jogado até que o resultado seja um número primo. Determine a distribuição de Z , ou seja, $P[Z = k]$, para $k = 1, \dots$. Que distribuição é esta?

A variável indicadora Y_i possui uma distribuição de Bernoulli, de parâmetro $p = \frac{2}{5}$, como na questão anterior. Porém, Z , neste caso, é uma variável aleatória que denota o número de vezes que o dado é jogado até que o número seja primo. Em outras palavras, Z identifica o primeiro índice i onde a variável Y_i de Bernoulli assume valor 1. Diremos que Z possui uma distribuição Geométrica de parâmetro $p = \frac{2}{5}$. Ou seja, $Z \sim \text{Geo}(\frac{2}{5})$. A função de probabilidade associada a esta variável é:

$$P[Z = k] = \left(\frac{3}{5}\right)^{k-1} \cdot \frac{2}{5} \quad k = 1, 2, \dots$$

4 Questão: Cobra

Considere três imagens tiradas em uma floresta, I_1 , I_2 e I_3 . Em apenas uma das imagens existe uma pequena cobra. Um algoritmo de detecção de cobras em imagens detecta a cobra na imagem i com probabilidade α_i . Suponha que o algoritmo não encontrou a cobra na imagem I_1 . Defina o espaço amostral e os eventos apropriados e use regra de Bayes para determinar:

1. A probabilidade da cobra estar na imagem I_1 .
2. A probabilidade da cobra estar na imagem I_2 .

Algumas premissas:

- Probabilidade de existir uma pequena cobra em uma das três imagens é uniforme. Isto é, vale $\frac{1}{3}$ para qualquer imagem I_i .
- Probabilidade do algoritmo detectar cobras em uma imagem i , dependerá da probabilidade de cima. Ou seja, se existir uma cobra, então o algoritmo tem chance de detectar. Assim, a probabilidade do algoritmo acertar dado que existe a cobra naquela imagem i , que é dada por α_i .
- C_i : variável aleatória que mapeia a existência de uma cobra em uma imagem i .
- A_i : variável aleatória associada a detecção de uma cobra na imagem i pelo algoritmo.

Com isso, podemos definir:

- A probabilidade de existir uma cobra na imagem i é $P[C_i] = \frac{1}{3}$ para qualquer i .
- A probabilidade do algoritmo encontrar uma cobra dado que existe uma cobra na imagem i é $P[A_i|C_i] = \alpha_i$.
- Podemos também definir que a $P[A_i|C_j]$ para $j \neq i$ é 0, uma vez que só pode existir cobra em uma única imagem.
- A probabilidade do algoritmo encontrar uma cobra em uma imagem que não existe cobra também será 0, ou seja, $P[A_i|\bar{C}_i] = 0$.
- A probabilidade do algoritmo não encontrar uma cobra em uma imagem que existe a cobra é então $1 - \alpha_i$. Que também a mesma probabilidade do algoritmo não encontrar dado que não tem cobra na imagem. Ou seja, $P[\bar{A}_i|C_i] = P[\bar{A}_i|\bar{C}_i] = 1 - \alpha_i$.

Falta definir a $P[A_i]$. Para isso, usaremos a *lei da probabilidade total*. Definiremos o espaço amostral em duas partições: O_i^1 : existe cobra na imagem i e O_i^2 : não existe cobra na imagem i .

$$P[A_i] = P[A_i|O_i^1]P[O_i^1] + P[A_i|O_i^2]P[O_i^2]$$

Note que o evento O_i^1 é exatamente igual ao evento C_i e o evento $O_i^2 = \bar{C}_i$. Assim, podemos reescrever:

$$P[A_i] = P[A_i|C_i]P[C_i] + P[A_i|\bar{C}_i]P[\bar{C}_i] \Leftrightarrow P[A_i] = \alpha_i \cdot \frac{1}{3} + 0 = \frac{\alpha_i}{3}$$

1

Resposta: A probabilidade da cobra estar na imagem I_1 .

Pelo enunciado do problema, o evento \bar{A}_1 aconteceu. Então dado que este evento aconteceu, a probabilidade da cobra estar na imagem I_1 é $P[C_1|\bar{A}_1]$. Neste caso, usaremos a *Regra de Bayes* pois temos calculado $P[\bar{A}_i|C_i]$ para qualquer i .

$$P[C_1|\bar{A}_1] = \frac{P[\bar{A}_1|C_1] \cdot P[C_1]}{P[\bar{A}_1]}$$

Tomaremos $P[\bar{A}_1] = 1 - P[A_1] = 1 - \frac{\alpha_1}{3}$

$$P[C_1|\bar{A}_1] = \frac{(1 - \alpha_1) \cdot \frac{1}{3}}{1 - \frac{\alpha_1}{3}} = \frac{\frac{1}{3} - \frac{\alpha_1}{3}}{1 - \frac{\alpha_1}{3}} = \frac{1 - \alpha_1}{3 - \alpha_1}$$

Resposta: A probabilidade da cobra estar na imagem I_2 .

Novamente, dado que o evento \bar{A}_1 aconteceu, queremos calcular $P[C_2|\bar{A}_1]$. Como temos calculado $P[\bar{A}_i|C_i]$ para qualquer i , usaremos Bayes novamente.

$$P[C_2|\bar{A}_1] = \frac{P[\bar{A}_1|P[C_2]] \cdot C_2}{P[\bar{A}_1]}$$

Neste caso, tomaremos o complementar de $P[\bar{A}_1|P[C_2]] = 1 - P[A_1|P[C_2]]$ que sabemos que $P[A_1|P[C_2]] = 0$. No denominador, como antes, tomaremos $P[\bar{A}_1] = 1 - P[A_1] = 1 - \frac{\alpha_1}{3}$.

$$P[C_2|\bar{A}_1] = \frac{(1-0) \cdot \frac{1}{3}}{1 - \frac{\alpha_1}{3}} = \frac{1}{3 - \alpha_1}$$

5 Questão: Sem Memória

Seja $X \sim Geo(p)$ uma variável aleatória geométrica com parâmetro p . Mostre que a distribuição geométrica não tem memória. Ou seja dado que $X > k$, o número de rodadas adicionais até que o evento de interesse ocorra possui a mesma distribuição (dica: formalize esta afirmação).

Resposta:

Usando um exemplo para entender melhor, se X for a espera em dias para a ocorrência de um certo evento, a probabilidade condicional de espera de mais de $k+d$ dias, sabendo que o evento não ocorreu antes de d dias é a mesma de esperar mais que k dias.

Dessa forma, queremos provar que:

$$P(X > k + d | X \geq d) = P(X > k)$$

Por definição de probabilidade condicional:

$$P[X > k + d | X \geq d] = \frac{P[X > k + d \cap X \geq d]}{P[X \geq d]}$$

Assumindo $k \geq 0$, a interseção de $[X > k + d \cap X \geq d]$ é $X > k + d$, portanto:

$$\Leftrightarrow \frac{P[X > k + d]}{P[X \geq d]}$$

Aplicando a definição de distribuição geométrica e tomando o complementar, temos:

$$\Leftrightarrow \frac{\sum_{i=k+d+1}^{\infty} p(1-p)^i}{\sum_{j=d}^{\infty} p(1-p)^j} = \frac{1-p \sum_{i=0}^{k+d} (1-p)^i}{1-p \sum_{j=0}^{d-1} (1-p)^j} = \frac{1-P[X \leq k+d]}{1-P[X < d]}$$

Neste caso, podemos usar a definição da soma de termos de uma progressão geométrica finita. A razão dessa progressão é $q = (1-p)$, então a soma de d termos fica:

$$\sum_{j=0}^{d-1} (1-p)^j = \frac{1 \cdot ((1-p)^d - 1)}{(1-p) - 1} = \frac{(1-p)^d - 1}{p}$$

Substituindo e fazendo o análogo para o numerador:

$$\Leftrightarrow \frac{1 - p \sum_{i=0}^{k+d} (1-p)^i}{1 - p \sum_{j=0}^{d-1} (1-p)^j} = \frac{1 - p \left(\frac{(1-p)^{k+d+1} - 1}{p} \right)}{1 - p \left(\frac{(1-p)^d - 1}{p} \right)} = \frac{(1-p)^{k+d+1}}{(1-p)^d} = (1-p)^{k+1} = 1 - p \sum_{t=0}^k (1-p)^t = P(X > k)$$

6 Questão: Ônibus

Considere que o processo de chegada do ônibus 485 no ponto do CT seja bem representado por um processo de Poisson. Ou seja, $X \sim Poi(\lambda, t)$ denota o número (aleatório) de ônibus que chegam ao ponto em um intervalo de tempo t com taxa média de chegada igual a λ . Assuma que $\lambda = 10$ ônibus por hora.

1. Determine a probabilidade de não chegar nenhum ônibus em um intervalo de 30 minutos. (inclusive numericamente).
2. Determine a probabilidade da média ocorrer, ou seja, de chegarem exatamente 10 ônibus em uma hora (inclusive numericamente).
3. Determine a taxa λ tal que a probabilidade de chegar ao menos um ônibus em um intervalo de 5 minutos seja maior que 90% (inclusive numericamente).

i

Resposta: Determine a probabilidade de não chegar nenhum ônibus em um intervalo de 30 minutos. (inclusive numericamente).

A função de probabilidade associada ao processo de Poisson é:

$$P[X = k] = \frac{e^{-\lambda t} (\lambda t)^k}{k!}$$

Nesse caso, queremos a probabilidade de não chegar nenhum ônibus no intervalo de 30 minutos. Sabemos que no intervalo de 60 minutos, a taxa média de chegada é 10 ônibus. Em 30 minutos, a taxa $\lambda t = 5$ ônibus, proporcionalmente. O parâmetro $t = 30$ minutos. A probabilidade de não chegar nenhum ônibus é igual a $P[X = 0]$, já que X representa o número de ônibus que chegam no ponto. Logo,

$$P[X = 0] = \frac{e^{-5} 5^0}{0!} \Leftrightarrow P[X = 0] = e^{-5} = \frac{1}{e^5} = 0.0067$$

i

Resposta: Determine a probabilidade da média ocorrer, ou seja, de chegarem exatamente 10 ônibus em uma hora (inclusive numericamente).

Nesse caso, queremos encontrar $P[X = 10]$ com parâmetro $\lambda t = 10$ e $t = 60min$.

$$P[X = 10] = \frac{e^{-10} \cdot 10^{10}}{10!} = 0.125$$

Resposta: Determine a taxa λ tal que a probabilidade de chegar ao menos um ônibus em um intervalo de 5 minutos seja maior que 90% (inclusive numericamente).

Chegar ao menos um ônibus em um intervalo $t = 5$ minutos, é equivalente a encontrar $P[X \geq 1]$ com uma certa taxa λt . Sem perda de generalidade, para facilitar faremos $\lambda = \lambda t$. Queremos determinar λ de tal forma que $P[X \geq 1] > 0.9$. É mais fácil trabalhar com o complementar, logo $1 - P[X = 0] > 0.9$.

$$1 - P[X = 0] > 0.9 \Leftrightarrow -\frac{e^{-\lambda} \cdot \lambda^0}{0!} > -0.1 \Leftrightarrow e^{-\lambda} \leq 0.1$$

Aplicando logaritmo natural de ambos os lados, temos

$$\Leftrightarrow -\lambda \leq \ln(0.1) \Leftrightarrow -\lambda \leq -2.302 \Leftrightarrow \lambda > 2.302$$

7 Questão: Propriedades

Seja X e Y duas variáveis aleatórias discretas. Mostre as seguintes equivalências usando as definições:

1. $E[X] = E[E[X|Y]]$, conhecida como regra da torre da esperança.
2. $Var[X] = E[X^2] - E[X]^2$

Resposta: $E[X] = E[E[X|Y]]$, conhecida como regra da torre da esperança.

Considere j qualquer um valor para condicionarmos a variável aleatória Y . Por definição de esperança, temos que:

$$E[E[X|Y = j]] \Leftrightarrow \sum_{j \in O_Y} E[X|Y = j] \cdot P[Y = j]$$

Usando a definição de esperança condicional, temos que:

$$\Leftrightarrow \sum_{j \in O_Y} \sum_{i \in O_X} (i \cdot P[X = i|Y = j]) \cdot P[Y = j]$$

Aplicando a definição de probabilidade condicional, temos que:

$$\Leftrightarrow \sum_{j \in O_Y} \sum_{i \in O_X} (i \cdot \frac{P[X = i \cap Y = j]}{P[Y = j]}) \cdot P[Y = j] \Leftrightarrow \sum_{i \in O_X} i \sum_{j \in O_Y} P[X = i \cap Y = j]$$

Note que $P[X = i \cap Y = j]$ é exatamente $P[X = i]$, uma vez que a interseção entre esses elementos j e i é exatamente i , para todo j .

$$\Leftrightarrow \sum_{i \in O_X} iP[X = i] = E[X]$$

i

Resposta: $Var[X] = E[X^2] - E[X]^2$

Por definição de variância, temos que:

$$Var[X] = E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2]$$

Aplicando as propriedades de esperança e assumindo independência:

$$\Leftrightarrow Var[X] = E[X^2] - 2E[X]E[X] + (E[X])^2 = E[X^2] - 2(E[X])^2 + (E[X])^2 = E[X^2] - (E[X])^2$$

8 Questão: Paradoxo do Aniversário

Considere um grupo com n pessoas e assuma que a data de nascimento de cada uma é uniforme dentre os 365 dias do ano. Vamos calcular a chance de duas ou mais pessoas fazerem aniversário no mesmo dia.

- Seja $c(n)$ a probabilidade de duas ou mais pessoas fazerem aniversário no mesmo dia. Determine explicitamente $c(n)$.
- Usando a aproximação $e^x \approx 1 + x$, determine o valor aproximado de $c(n)$.
- Usando o valor aproximado de $c(n)$, determine o menor valor de n tal que a chance de colisão na data de aniversário seja maior do que $\frac{1}{2}$. Você considera este número alto ou baixo?

i

Resposta: Seja $c(n)$ a probabilidade de duas ou mais pessoas fazerem aniversário no mesmo dia. Determine explicitamente $c(n)$.

Neste caso, é mais simples calcular $\bar{C}(n)$, pela independência dos eventos. $P[\bar{C}(n)]$ representa, por sua vez, a probabilidade de **não** ter ninguém fazendo aniversário no mesmo dia em uma sala com n pessoas.

A probabilidade $\bar{C}(n)$ tem a seguinte construção: Se existem apenas duas pessoas na sala, a chance deles não terem o mesmo aniversário é $1 \cdot \frac{364}{365}$. Assim, se houverem três pessoas na sala a chance de todos eles não terem o mesmo aniversário é:

$$\bar{C}(3) = \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365}$$

Agora, se temos n pessoas na sala, a probabilidade $C(n)'$ será:

$$\bar{C}(n) = \frac{365 - 0}{365} \cdot \frac{365 - 1}{365} \cdot \frac{365 - 2}{365} \cdots \frac{365 - (n - 1)}{365} \Leftrightarrow \bar{C}(n) = \frac{365!}{365^n (365 - n)!}$$

Como queremos calcular a probabilidade de $C(n)$, faremos $C(n) = 1 - \bar{C}(n)$.

$$C(n) = 1 - \frac{365!}{365^n (365 - n)!}$$

i

Resposta: Usando a aproximação $e^x \approx 1 + x$, determine o valor aproximado de $C(n)$.

Podemos reescrever a seguinte equação:

$$\frac{365-0}{365} \cdot \frac{365-1}{365} \cdot \frac{365-2}{365} \cdots \frac{365-(n-1)}{365}$$

Como:

$$1 \cdot \left(1 - \frac{1}{365}\right) \cdot \left(1 - \frac{2}{365}\right) \cdot \left(1 - \frac{3}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right)$$

Olhando para os termos, podemos usar expansão e Taylor de primeira ordem. Por hipótese, usaremos a aproximação $e^x \approx 1 + x$. Então, tomando por exemplo $x = -\frac{1}{365}$

$$1 - \frac{1}{365} \approx e^{-\frac{1}{365}}$$

Então, podemos reescrever da seguinte forma:

$$C(n) \approx 1 - 1 \cdot e^{-\frac{1}{365}} \cdot e^{-\frac{2}{365}} \cdot e^{-\frac{3}{365}} \cdots e^{-\frac{(n-1)}{365}}$$

Dai

$$C(n) \approx 1 - e^{-\frac{1+2+3+\cdots+(n-1)}{365}} \Leftrightarrow C(n) \approx 1 - e^{-\frac{1+2+3+\cdots+(n-1)}{365}}$$

Adição de números de 1 até $n-1$ é igual a $\frac{n(n-1)}{2}$

$$C(n) \approx 1 - e^{-\frac{n(n-1)}{2 \cdot 365}} \approx 1 - e^{-\frac{n^2}{730}}$$



Resposta: Usando o valor aproximado de $c(n)$, determine o menor valor de n tal que a chance de colisão na data de aniversário seja maior do que $\frac{1}{2}$. Você considera este número alto ou baixo?

Queremos encontrar

$$C(n) \geq \frac{1}{2} \Leftrightarrow \bar{C}(n) < \frac{1}{2} \Leftrightarrow e^{-\frac{n^2}{730}} < \frac{1}{2}$$

Aplicando logaritmo natural em ambos os lados, temos:

$$\Leftrightarrow -\frac{n^2}{730} < \ln(1) - \ln(2) = -\ln(2) \Leftrightarrow -n^2 < 730 \cdot (-0,693) = -505,83 \Leftrightarrow n > 22,49$$

Ou seja, a partir de $n = 23$ esta probabilidade ultrapassa $\frac{1}{2}$.

9 Questão: Caras em sequência

Considere uma moeda enviesada, tal que a probabilidade do resultado ser cara é p (e a coroa é $1-p$). Considere o número de vezes que a moeda precisa ser jogada para obtermos k caras consecutivas. Por exemplo, na sequência "COCOCCOOCOCCC" a moeda teve que ser jogada 13 vezes até o aparecimento de $k = 3$ caras consecutivas, onde $C = cara$ e $O = coroa$. Seja N_k a variável aleatória que denota esta quantidade. Qual o número médio de vezes que a moeda precisa ser jogada para obtermos k caras consecutivas, ou seja, qual o valor esperado de N_k ? Dica: monte uma recursão e use a regra da torre da esperança.

Resposta:

Seja N_k a variável aleatória que denota o número de vezes que a moeda precisa ser jogada para obter k caras.

Definiremos uma variável aleatória X que denota o índice da primeira coroa obtida no lançamento de moedas. Essa variável aleatória terá distribuição geométrica $X \sim Geo(1 - p)$.

Para um i qualquer, $E[N_k|X = i]$ pode ser definida de duas maneiras. Como k representa o número de caras consecutivas, temos que:

1. Se a primeira coroa i for encontrada depois das k caras consecutivas.
2. Se a primeira coroa i for encontrada antes das k caras consecutivas.

Sabemos que o valor esperado de jogadas para o caso 1. será exatamente k , uma vez que é necessário que as k primeiras jogadas sejam exatamente cara para que a primeira coroa seja encontrada depois.

O valor esperado para o caso 2., no entanto, é um pouco menos claro. Se a primeira coroa i for encontrada antes das k caras, teremos a seguinte situação:

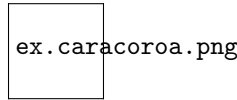


Figura 1: Exemplo com $k = 3$, $N_3 = 8$ e a primeira coroa no índice $i = 3$

Observe que toda vez que temos uma coroa antes das k caras consecutivas, reiniciamos a contagem delas. Note que, como não saiu nas i primeiras jogadas, calculamos o valor esperado do número de jogadas dado que a segunda coroa j aconteceu antes das k caras consecutivas. Dessa forma, nesse caso, $E[N_k|X = i] = i + E[N_k|X \geq i + j]$ para $j \neq i$ e $j > i$. Como a distribuição geométrica é sem memória, podemos definir a seguinte recursão: $E[N_k|X = i] = i + E[N_k|X \geq i] = E[N_k]$.

$$E[N_k] = E[E[N_k|X]] = \sum_{i \in O_X} E[N_k|X = i] \cdot P[X = i]$$

Dividindo nos dois casos:

$$\Leftrightarrow E[N_k] = \sum_{i \in O_X} k \cdot P[X = i] + \sum_{i \in O_X} (i + E[N_k]) \cdot P[X = i]$$

Como $X \sim Geo(1 - p)$, $P[X = i] = p^{i-1} \cdot (1 - p)$ $i = 1, 2, \dots$

$$\Leftrightarrow E[N_k] = \sum_{i \in O_X} k \cdot p^{i-1} \cdot (1 - p) + \sum_{i \in O_X} (i + E[N_k]) \cdot p^{i-1} \cdot (1 - p)$$

$$\Leftrightarrow E[N_k] = k \sum_{i \in O_X} p^{i-1}(1 - p) + \sum_{i \in O_X} ip^{i-1}(1 - p) + E[N_k](1 - p) \sum_{i \in O_X} p^{i-1}$$

$$\Leftrightarrow E[N_k] - \left(E[N_k](1 - p) \sum_{i \in O_X} p^{i-1} \right) = k \sum_{i \in O_X} p^{i-1}(1 - p) + \sum_{i \in O_X} ip^{i-1}(1 - p)$$

$$\Leftrightarrow E[N_k] \left(1 - (1 - p) \sum_{i \in O_X} p^{i-1} \right) = k \sum_{i \in O_X} p^{i-1}(1 - p) + \sum_{i \in O_X} ip^{i-1}(1 - p)$$

$$\begin{aligned}
\Leftrightarrow E[N_k] \left(1 - (1-p) \sum_{i \in O_X} p^{i-1} \right) &= k(1-p) \sum_{i \in O_X} p^{i-1} + (1-p) \sum_{i \in O_X} ip^{i-1} \\
\Leftrightarrow E[N_k] &= \frac{k(1-p) \sum_{i \in O_X} p^{i-1} + (1-p) \sum_{i \in O_X} ip^{i-1}}{1 - (1-p) \sum_{i \in O_X} p^{i-1}} \\
\Leftrightarrow E[N_k] &= \frac{k \sum_{i \in O_X} p^{i-1} + \sum_{i \in O_X} ip^{i-1}}{1 - \sum_{i \in O_X} p^{i-1}}
\end{aligned}$$

3 Limitantes para probabilidade

Calcular a probabilidade de um evento pode ser bem difícil, então pode ser mais fácil calcular limitantes para esse valor.

- U_A é um limitante superior $\Leftrightarrow P[A] \leq U_A$
- L_A é um limitante inferior $\Leftrightarrow P[A] \geq L_A$

3.1 Cauda e Cabeça

Seja X uma variável aleatória com $\mu = E[X]$. Diremos que:

- **Cauda:** valores de X bem maiores que μ
- **Cabeça:** valores de X bem menores que μ

Exemplo: Jogar 50 vezes um dado honesto com 10 faces. Seja N o número de vezes que o resultado foi primo e X_i o resultado na i -ésima rodada.

Então diremos que N pode ser a soma das variáveis indicadoras de um número primo.

$$N = \sum_i I(X_i)$$

Qual $P[N \geq 40]$?

- $P[X_i] = \frac{2}{5}$
- $N \sim \text{Bin}(50, \frac{2}{5})$

$$P[N \geq 40] = \sum_{i=40}^{50} \binom{50}{i} \left(\frac{2}{5}\right)^i \left(\frac{3}{5}\right)^{50-i}$$

Pode ser difícil calcular esses coeficientes de Newton.

3.2 Desigualdade de Markov

Importante limitante superior para probabilidade de um evento, pois relaciona o valor esperado e a probabilidade do evento. Para qualquer variável aleatória X não-negativa e uma constante a não-negativa, temos:

$$P[X \geq a] \leq \frac{E[X]}{a}$$

Note que: só faz sentido para $a > E[X]$, em outras palavras, a tem que estar na cauda.

Voltando ao exemplo, podemos tentar aplicar a desigualdade de Markov. O valor esperado de N é exatamente $50 \cdot \frac{2}{5} = 20$, pois $N \sim \text{Bin}(50, \frac{2}{5})$. Logo,

$$P[N \geq 40] \leq \frac{20}{40} = \frac{1}{2}$$

Ou seja, a chance de ver 40 ou mais primos é menor do que $\frac{1}{2}$.

3.3 Desigualdade de Chebyshev

Outro importante limitante superior para probabilidade de um evento, pois relaciona o valor esperado, a variância e a probabilidade. Em geral, ela será mais precisa que a desigualdade de Markov.

Para qualquer variável aleatória X com valor esperado μ e variância σ^2 e para qualquer k não-negativo, temos

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Quanto maior for k , melhor será o limitante superior. A desigualdade mede a "chance de X está k desvios padrão longe da média".

Caso interessante: Tome $k = \sqrt{2}$

$$P[|X - \mu| \geq 1.41\sigma] \leq \frac{1}{2}$$

A probabilidade de X estar fora do intervalo $[\mu - 1.41\sigma, \mu + 1.41\sigma]$ é menor que $\frac{1}{2}$. Isso vale para qualquer distribuição.

Voltando ao exemplo, podemos tentar aplicar a desigualdade de Chebyshev. A variância de uma distribuição binomial é $\sigma^2 = 50 \cdot \frac{2}{5} \cdot \frac{3}{5} = 12$. Agora falta escolher k . Repare que, como queremos que $P[N \geq 40]$, fazemos:

$$P[N \geq 40] = P[N - \mu \geq 40 - 20]$$

Logo, para aplicar a desigualdade, $k\sigma = 20 \Rightarrow k = \frac{10}{\sqrt{3}}$

$$P[N \geq 40] \leq P[|N - 20| \geq 20] \leq \frac{1}{\frac{100}{3}} = 0.03$$

Resultado bem melhor do que markov!

3.4 Desigualdade de Chernoff

Um limitante superior para a soma de variáveis aleatórias independentes. Seja $Y_i \sim \text{Bern}(p)$, $\mu = E[X] = np$ e qualquer $\delta > 0$, temos

$$P[X \geq (1 + \delta)\mu] \leq \left(\frac{e^\delta}{(1 + \delta)^{1 + \delta}} \right)^\mu$$

Repare que essa é a probabilidade da cauda, isto é, probabilidade de X está depois da média. Também podemos definir a probabilidade da cabeça:

$$P[X \leq (1 - \delta)\mu] \leq \left(\frac{e^{-\delta}}{(1 - \delta)^{1 - \delta}} \right)^\mu$$

Voltando ao exemplo, podemos tentar aplicar as desigualdades de Chernoff. Note que queremos fazer $(1 + \delta) \cdot 20 = 40 \Rightarrow \delta = 1$

$$P[X \geq 40] \leq \left(\frac{e^1}{(1 + 1)^{1 + 1}} \right)^{20} = \frac{e^{20}}{2^{40}} = 0.00044$$

Resultado ainda melhor por chebyshev! Mas observe que nesse caso, é necessário que esteja trabalhando com a soma de variáveis aleatórias.

3.5 With High Probability (whp)

Seja n um parâmetro de um modelo probabilístico. (Ex. Número de rodada de um dado). Seja $A(n)$ um evento no respectivo espaço amostral. Dizemos que $A(n)$ ocorre com alta probabilidade (whp) quando

$$P[A(n)] \geq 1 - \frac{1}{n^\alpha}$$

Para algum $\alpha \geq 1$ constante. Repare que $\lim_{n \rightarrow \infty} P[A(n)] = 1$. Justamente por isso que é chamado dessa maneira, conforme n cresce essa probabilidade associada a ele cada vez mais se aproxima de 1.

De forma análoga, a probabilidade complementar de $A(n)$ decresce

$$P[\bar{A}(n)] = 1 - P[A(n)] \leq \frac{1}{n^\alpha}$$

Exemplo: Considere jogar uma moeda n vezes e Y_i modela se o resultado é cara ao jogar a i -ésima moeda. Dessa forma, definimos X uma variável aleatória que contabiliza o número de caras em n jogadas, ou seja, $X = \sum_{i=1}^n Y_i$. Como a moeda é honesta, $\mu = \frac{n}{2}$.

Onde é que tá o ponto onde a cauda da distribuição tem probabilidade que vai a zero com n ? Ou seja, qual o valor de λ tal que $P[X > \mu + \lambda] < \frac{1}{n}$

3.6 Limitante da União

Chamado de *Union Bound*. É bastante útil para lidar com muitos eventos que não são necessariamente mutuamente exclusivos ou independentes. Sejam A e B dois eventos em um espaço amostral. Temos que:

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

Pois a interseção, no melhor caso, tem valor zero, que é quando A e B são mutuamente exclusivos.

Seja A_i uma sequência de eventos de um espaço amostral com $i = 1, \dots, n$. Temos que

$$P[\bigcup_i A_i] = P[\sum_i A_i] \leq \sum_i P[A_i]$$

Se A_i for uniforme, isto é, tiverem a mesma probabilidade

$$P[\bigcup_i A_i] = P[\sum_i A_i] \leq \sum_i P[A_i] = nP[A_i]$$

Caso contrário, ainda podemos majorar

$$P[\bigcup_i A_i] = P[\sum_i A_i] \leq \sum_i P[A_i] \leq n \cdot \max_i P[A_i]$$

Exemplo: Vou jogar um dado honesto três vezes. Qual a probabilidade de sair 6 ao menos uma vez?

- Seja X_i o resultado da i -ésima rodada.

$$P[X_1 = 6 \cup X_2 = 6 \cup X_3 = 6] \leq P[X_1 = 6] + P[X_2 = 6] + P[X_3 = 6] = \frac{1}{2}$$

Mas qual a probabilidade exata? Podemos usar o complemento dessa probabilidade, isto é, de não sair 6 em nenhuma rodada. Conseguimos essa exatidão pois garantimos que são eventos independentes.

$$1 - P[X_1 \neq 6 \cap X_2 \neq 6 \cap X_3 \neq 6] = 1 - (P[X_1 \neq 6] \cdot P[X_2 \neq 6] \cdot P[X_3 \neq 6]) = 1 - \left(\frac{5}{6}\right)^3 = 0.42$$

Note que o limitante deu um bom resultado!

OBSERVAÇÃO: É bom utilizar o limitante da união quando existem poucas parcelas ou quando a probabilidade das parcelas são pequenas. Caso contrário, o resultado não é muito útil.

Exemplo 2: Considere kn bolas jogadas aleatoriamente sobre n urnas para $k \geq 1$. Qual a probabilidade de termos ao menos uma urna vazia (p_0)?

- Definimos uma indicadora X_i que indica se a urna i está vazia.
- $P[X_i] = (1 - \frac{1}{n})^{kn}$ (Se a chance de acertar a urna é $\frac{1}{n}$ a chance de não acertar é $1 - \frac{1}{n}$. Então a chance de não ter nenhuma bola na urna i é multiplicar kn vezes.

Podemos aplicar o limitante da união! Pois queremos que ou uma urna esteja vazia, ou outra esteja e assim por diante. Assim

$$p_0 = P[\bigcup X_i] = P[\sum_{i=1}^n X_i \leq n \left(1 - \frac{1}{n}\right)^{kn}]$$

Se tomarmos $n = 10$ e $k = 3$, $p_0 \leq 0.42$. No entanto, se tomarmos $n = 100$ e $k = 2$, $p_0 \leq 13.4$ que não é nada útil.

4 Lei dos grandes números

Começaremos com um exemplo. Suponha que desejamos jogar um dado honesto n vezes. Seja X_i a variável aleatória que dá o resultado da i -ésima jogada. Definiremos uma indicadora de $X_i = 1$, $Y = I[X_i = 1]$, que definirá se a i -ésima jogada é ou não o número 1. Seja $N_1(n)$ o número de vezes que o resultado dá 1 em n jogadas, logo $N_1(n) = \sum_{i=1}^n Y_i$. Definiremos uma **fração relativa de resultados** $F_1(n)$ da seguinte maneira:

$$F_1(n) = \frac{N_1(n)}{n}$$

Observe que $F_1(n)$ é uma variável aleatória, uma vez que teremos, muito provavelmente, resultados diferentes para um mesmo n .

Mas quando $n \rightarrow \infty$? A frequência relativa converge para a probabilidade do evento, nesse caso, converge para a probabilidade do dado sair 1.

Essa lei é fundamental, pois é a conexão entre teoria e prática. Atribui um significado físico a um conceito abstrato (probabilidade). Então a fração de resultados aleatórios quando você repete muitas vezes o experimento converge para um número.

4.1 Média Amostral

Seja X_i uma sequência de variáveis aleatórias iid, tal que $\mu = E[X_i]$ e $\sigma^2 = Var[X_i]$.

Seja $M_n = \frac{1}{n} \sum_{i=1}^n X_i$ a média amostral, em outras palavras, a média aritmética dos n valores da amostra. Repare que M_n é uma variável aleatória. Qual o valor esperado e a variância de M_n ?

$$E[M_n] = \frac{1}{n} E\left[\sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n E[X_i] = \frac{1}{n} \cdot n \cdot \mu = \mu$$

Logo $E[M_n] = E[X_i]$.

$$Var[M_n] = \frac{1}{n^2} Var\left[\sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n Var[X_i] = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

Ou seja, a variância depende do número de amostras, diferentemente do valor esperado.

4.2 Lei Fraca dos grandes números

Se μ for finito, para qualquer $\epsilon > 0$, temos

$$\lim_{n \rightarrow \infty} P[|M_n - \mu| \leq \epsilon] = 1$$

É chamado de convergência em probabilidade, pois estamos tomando um limite de uma probabilidade que converge para um número.

Repare que esse limite analisa a diferença entre o valor da média amostral com o valor esperado. A probabilidade dessa diferença ser tão pequena quanto se queira (tomando ϵ ao seu favor), converge para 1 quando temos um número grande de amostras.

Em outras palavras, a nossa média amostral converge para o valor esperado dela, nessas condições.

Para provar essa lei, podemos usar a *desigualdade de Chebyshev*.

$$P[|X - \mu| \geq k\sigma_{M_n}] \leq \frac{1}{k^2}$$

Tomamos $k\sigma_{M_n} = \epsilon \Rightarrow k = \frac{\epsilon\sqrt{n}}{\sigma}$.

Como em *Chebyshev* a probabilidade está indo para zero, tomaremos o complementar, já que a Lei dos grandes números vai para 1. Ou seja, $P[|M_n - \mu| < \epsilon] = 1 - P[|M_n - \mu| \geq \epsilon]$

Usando *Chebyshev*

$$P[|X - \mu| \geq k\sigma_{M_n} = \epsilon] \leq \frac{1}{k^2} = \frac{\sigma^2}{\epsilon^2 n}$$

$$\Leftrightarrow P[|M_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{\epsilon^2 n} \Rightarrow -P[|M_n - \mu| \geq \epsilon] \geq -\frac{\sigma^2}{\epsilon^2 n} \Rightarrow 1 - P[|M_n - \mu| \geq \epsilon] \geq 1 - \frac{\sigma^2}{\epsilon^2 n}$$

Logo, como $P[|M_n - \mu| < \epsilon] = 1 - P[|M_n - \mu| \geq \epsilon]$

$$P[|M_n - \mu| < \epsilon] \geq 1 - \frac{\sigma^2}{\epsilon^2 n}$$

Fazendo o limite disso, vemos que $1 - \frac{\sigma^2}{\epsilon^2 n}$ converge para 1 quando $n \rightarrow \infty$. O que prova a Lei.

4.3 Lei forte dos grandes números

M_n possui o mesmo valor esperado que X_i e a variância vai a zero com n . Para μ, σ^2 finitos, a lei diz o seguinte:

$$P\left[\lim_{n \rightarrow \infty} M_n = \mu\right] = 1$$

Repare que o limite está dentro da probabilidade agora. É um resultado bem mais forte que a anterior, pois não depende de ϵ . Essa é chamada de *convergência quase certamente*.



- Moeda honesta, fração de caras

$$E[M_n] = \frac{1}{2}$$

$$\text{Var}[M_n] = \frac{1}{4n}$$

- Conforme n aumenta, M_n fica mais centrada!

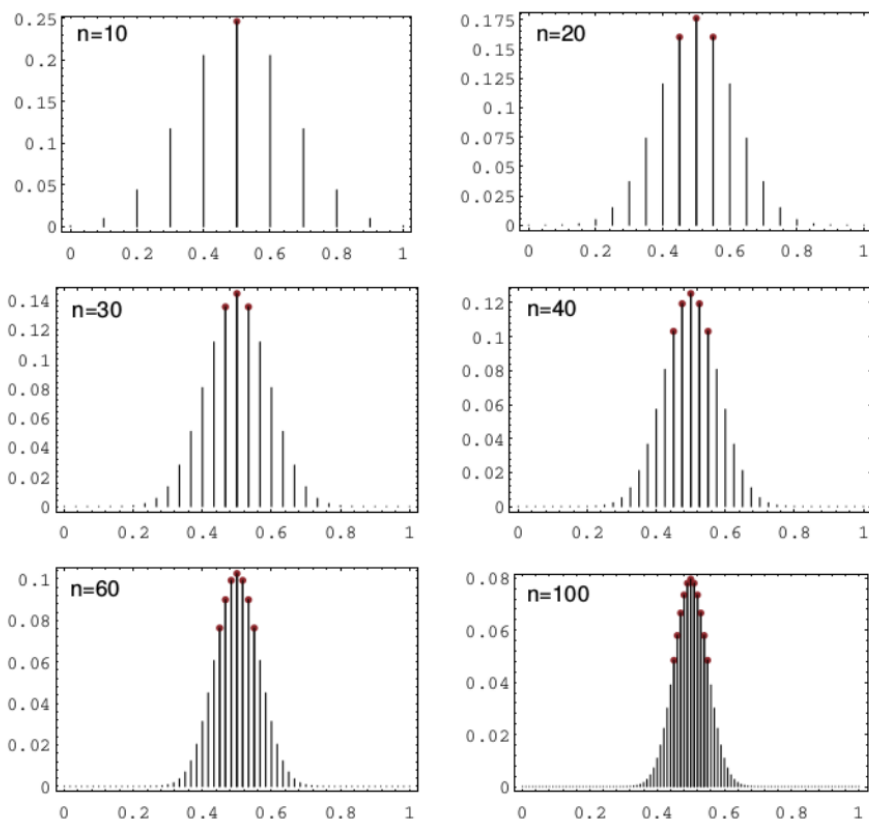


Figura 2: Lei dos grandes números em ação!

5 Erro e Confiança

Podemos usar *Chebyshev* para calcular precisão e confiança de lei dos grandes números. Seja a precisão ϵ e a confiança β . A ideia é calcular o valor de n a partir de seu erro e confiança, que dão o quão perto a média amostral está da média. Em outras palavras, temos que:

$$P[M_n \in [\mu - \epsilon, \mu + \epsilon]] > \beta$$

Em tese, queremos determinar n para garantir que a média amostral vai estar ϵ perto do valor esperado e que sua probabilidade tem que ser maior que β .

Voltando a demonstração da Lei fraca dos grandes números, temos que:

$$P[|M_n - \mu| < \epsilon] \geq 1 - \frac{\sigma^2}{\epsilon^2 n}$$

Ora, por definição $|M_n - \mu| < \epsilon \iff M_n \in [\mu - \epsilon, \mu + \epsilon]$. Assim, basta definirmos $1 - \frac{\sigma^2}{\epsilon^2 n} = \beta$. Isolando n , temos:

$$n = \frac{\sigma^2}{\epsilon^2(1 - \beta)}$$

- A *confiança* vai influenciar n linearmente e inversamente.
- A *precisão* vai influenciar n de forma quadrática e inversamente.
- Aumentar a precisão (reduzir ϵ) demanda mais rodadas do que aumentar a confiança (aumentar β).

Exemplo: Suponha uma moeda enviesada, com probabilidade de sair cara igual a 45%. Queremos testar se a moeda é mesmo enviesada. Quantas vezes preciso lançar a moeda?

- Tome $\epsilon = 0.01$
- Tome $\beta = 0.95$
- Temos que $\mu = 0.45$ pois a probabilidade de sair cara é 0.45
- Temos que $\sigma^2 = 0.45 \cdot 0.55$, pois é uma Bernoulli.

$$P[|M_n - 0.45| < 0.01] = P[M_n \in [0.44, 0.46]] > 1 - \frac{0.45 \cdot 0.55}{(0.01)^2 n} = 0.95$$

Logo, $n = 49500$.

5.1 Margem de erro

Suponha que é feita uma pesquisa com 1000 cariocas que revela que entre praia e cachoeira, 70% preferem praia. Qual a margem de erro da pesquisa com 90% da confiança?

- Cada pessoa pode ser uma *Bernoulli*, $X_i = 1$ se a pessoa prefere praia.
- Supor que a probabilidade $p = \mu = 0.7$ (resultado da pesquisa). Logo, $\sigma^2 = 0.7 \cdot 0.3 = 0.21$

Neste caso, queremos encontrar ϵ tal que $n = 1000$ e $\beta = 0.9$

$$P[|M_n - \mu| < \epsilon] \geq 1 - \frac{\sigma^2}{\epsilon^2 n} = 0.9$$

$$1 - \frac{\sigma^2}{\epsilon^2 n} = \beta \Rightarrow -\frac{\sigma^2}{\epsilon^2 n} = \beta - 1 \Rightarrow \sigma^2 = \epsilon^2 n(1 - \beta) \Rightarrow \epsilon^2 = \frac{\sigma^2}{n(1 - \beta)}$$

Assim, substituindo, $\epsilon = 0.046 = 4.6\%$. Então podemos dizer que esta é a **margem de erro** desse experimento com confiança 90%.

5.2 Falácia do apostador

A ideia dessa falácia é a sensação de que se um evento ocorre mais frequentemente que o esperado no passado então ele vai ter menos chance de ocorrer no futuro. **Falácia!**. Passado não influencia na aleatoriedade do futuro, pois já foi observado e é independente.

Porém, a Lei dos grandes números não torna a falácia verdadeira? Uma vez que a Lei garante e converge a uma certa probabilidade quando um experimento é feito diversas vezes.

Resposta: Não! São afirmações diferentes.

- A Lei diz que a fração relativa das observações converge para sua respectiva probabilidade.
- A Falácia diz que observações passadas influenciam a probabilidade de observações futuras.

6 Método de Monte Carlo

É uma classe de algoritmos baseados em amostragem aleatória repetida, no intuito de obter soluções aproximadas para problemas determinísticos. A ideia central é que tomar um grande número de amostras repetidas acabam revelando a solução. Em especial, quando o número de amostras tende ao infinito, no limite as amostras dão a solução. A base teórica para esta ideia é dada pela *Lei dos grandes números*.

Exemplo: Calcular o valor de um somatório com N parcelas muito grandes.

$$G_N = \sum_{i=1}^N g(i)$$

Como usar aleatoriedade para aproximar esse somatório? Usando o valor esperado! Seja X uma variável aleatória uniforme onde $P[X = i] = \frac{1}{N}$ para $i = 1, \dots, N$.

$$E[g(X)] = \sum_{i=1}^N g(i) \cdot P[X = i] = \frac{1}{N} \sum_{i=1}^N g(i) = \frac{G_N}{N}$$

Isolando G_N , temos:

$$E[g(X)] = \frac{G_N}{N} \Leftrightarrow G_N = N \cdot E[g(X)]$$

Podemos agora tentar estimar $E[g(X)]$, como? **Gerando amostras e fazendo a média!** Seja X_i a sequência iid de variáveis aleatórias uniformes $[1, N]$, escolhemos n como o número de amostras. A média será:

$$M_n = \frac{1}{n} \sum_{i=1}^n g(X_i)$$

Pela Lei dos grandes números, sabemos que quando $n \rightarrow \infty$ $M_n \rightarrow E[g(X)]$. Logo, M_n é um estimador de $E[g(X)]$. Como

$$G_N = N \cdot E[g(X)] \Rightarrow G_N = N \cdot M_n$$

Logo $N \cdot M_n$ é um estimador para G_N .

Exemplo 2: Quantas arestas tem a rede de amizade do Facebook?

- Função indicadora de aresta: $g(i, j) = 1$ se existe aresta entre o perfil i e j
- Número de perfis $n_p = 5 \cdot 10^9$

O número total de ligações será:

$$T = \sum_{i=1}^{n_p} \sum_{j=i+1}^{n_p} g(i, j)$$

Note que, as ligações **não são direcionadas**. Por isso contaremos apenas uma vez a aresta (i, j) .

Não tem nada de aleatório sobre essa soma. O problema é que o somatório tem aprox. 10^{19} termos. Como construir um método de Monte Carlo para obter um valor aproximado para T ?

Faremos o mesmo procedimento acima. Definimos como $N = \frac{n_p(n_p-1)}{2}$ o número total de pares, tal que $P[Z = i] = \frac{1}{N}$. Definimos $Z = (i, j)$ uma variável aleatória uniforme iid no conjunto de pares entre os n_p perfis. Pelo resultado anterior, temos:

$$E[g(Z)] = \frac{T}{N} \Rightarrow T = N \cdot E[g(Z)]$$

De forma semelhante, definimos uma sequência iid de variáveis aleatórias uniformes Z_i sobre pares.

$$M_n = \frac{1}{n} \sum_{i=1}^n g(Z_i)$$

Logo, T pode ser aproximado por $M_n \cdot N$.

6.1 Vantagens e Desvantagens

O que estamos fazendo aqui, no fundo, é trocar $G_N = \sum_{i=1}^N$ por $M_n = \frac{1}{n} \sum_{i=1}^n g(X_i)$.

- Se N for pequeno, essa ideia **não** é boa. Nesse caso, é mais fácil computar o somatório de forma determinística.
- Se calcular $g()$ é muito caro, no entanto, talvez não seja fácil calcular deterministicamente.
- Se $g(i)$ existir um valor, por exemplo, que é maior que todo o resto da soma (impactando diretamente na média). O estimador pode ser muito ruim se n não for grande o suficiente.

Logo, a qualidade da aproximação depende de $N, n, g()$.

6.2 Calculando erro

Podemos usar *Chebyshev* para calcular n com precisão ϵ e confiança β . Relembrando:

$$P[|M_n - \mu| < \epsilon] \geq 1 - \frac{\sigma^2}{\epsilon^2 n} = \beta$$

- $\mu = E[g(X)]$ e $\sigma^2 = Var[g(X)]$
- $g()$ é geralmente uma função indicadora (variável binária) e X é uma variável aleatória uniforme nos valores que g pode assumir.
- **Problema:** muitas vezes não sabemos σ^2 ! Temos que estimar.

Exemplo: Jogo da Paciência. Cálculo do erro. Seja $F_A = 0.1$ a fração de vezes que o algoritmo A vence. Ou seja, em 1 de 10 jogadas o algoritmo vence.

- S é uma variável aleatória uniforme das permutações possíveis de cartas.
- $\mu = E[f_A(S)] = F_A = 0.1$ e $\sigma^2 = Var[f_A(S)] = 0.1 \cdot 0.9$ por S ser uma indicadora.
- Tomamos $\epsilon = 10^{-4}$ e $\beta = 0.99$.

Qual a margem de erro?

$$P[|M_n - 0.1| < 10^{-4}] \geq 1 - \frac{0.1 \cdot 0.9}{10^{-8} n} = 0.99 \Leftrightarrow n = 9 \cdot 10^8$$

Logo, o número de vezes que é necessário jogar é igual a $n \approx 10^9$ e é muito menor que $52! \approx 10^{68}$ que é a permutação de 52 cartas.

6.3 Estimando π

Como estimar o valor de π ? Ou qualquer valor que tenha alguma relação com a **geometria**. A ideia é escrever π como uma relação entre áreas e usar monte carlo para estimar essa relação.

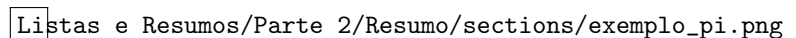
 `Listas e Resumos/Parte 2/Resumo/sections/exemplo_pi.png`

Figura 3: Círculo circunscrito em um quadrado de lado 1

- A_q = Área do quadrado que é igual a 1.
- A_c = Área do círculo que é igual a πr^2 . Como $r = \frac{1}{2}$ então $A_c = \frac{\pi}{4}$
- Isolando π , $\pi = 4A_c$. Como $A_q = 1$, podemos escrever $\pi = 4 \frac{A_c}{A_q}$

O problema é estimar a relação entre essas áreas. Para isso, geramos n pontos uniformes dentro do quadrado. Vamos medir a fração dos pontos que estão dentro do círculo. Seja X e Y duas variáveis aleatórias uniformes contínuas em $[0, 1]$. Seja $g(x, y)$ a indicadora se o ponto está dentro do círculo ou não.

- $g(x, y) = 1$ se $(x - 0.5)^2 + (y - 0.5)^2 \leq 1$

$$E[g(X, Y)] = \sum_{x=1}^n \sum_{y=1}^n g(x, y) \cdot \frac{1}{N} = \frac{\pi}{4}$$

Seja X_i e Y_i a sequência iid uniforme em $[0, 1]$. A fração de pontos que estão dentro do círculo é:

$$M_n = \frac{1}{n} \sum_{i=1}^n g(X_i, Y_i)$$

M_n converge para $\frac{\pi}{4}$ pela lei dos grande números e então π pode ser estimado por $4 \cdot M_n$.

6.3.1 Erro de π

Podemos usar *Chebyshev* para calcular quantos pontos são necessários para que essa estimativa esteja correta com precisão ϵ e confiança β .

$$P[|M_n - \mu| < \epsilon] \geq 1 - \frac{\sigma^2}{\epsilon^2 n} = \beta$$

- $\mu = E[g(X, Y)] = P[g(X, Y) = 1] = \frac{\pi}{4}$
- $\sigma^2 = Var[g(X, Y)] = \frac{\pi}{4} \cdot (1 - \frac{\pi}{4})$

Resolvendo a equação com $\epsilon = 10^{-4}$ e $\beta = 0.99$, temos $n = 2.5 \cdot 10^9$.

MSE e SEM

Mean Squared Error (MSE) ou erro quadrático médio é uma medida clássica para erro de preditores ou estimadores onde $MSE(\phi') = E_{\phi'}[(\phi' - \phi)^2]$ onde ϕ' é o estimador e ϕ o valor a ser estimado.

- $MSE(M_n) = Var[M_n] = \frac{\sigma^2}{n}$ pela definição.

Standart Error of the Mean (SEM) ou medida de erro relativo, pois normaliza o MSE. Isto é, tira raiz quadrada do MSE.

$$SEM(M_n) = \frac{\sigma}{\sqrt{n}} = \sqrt{MSE(M_n)}$$

6.4 Integração de Monte Carlo

Para problemas onde não podemos definir analiticamente o valor da integral. Suponha que a função $0 \leq h(x) \leq 1$.

$$I = \int_{x=0}^{x=1} h(x) dx$$

- Definir uma função indicadora para ponto de baixo da curva: $g(x, y) = 1$ se $h(x) \leq y$, e zero c.c.

Seja X e Y variável aleatória contínuas uniformes em $[0, 1]$.

$$E[g(X, Y)] = \int_{y=0}^{y=1} \int_{x=0}^{x=1} f_{xy}(x, y) \cdot g(x, y) dx dy$$

- $f_{xy}(x, y) = f_x(x)f_y(y) = 1$ (densidade conjunta de duas variáveis aleatórias contínuas e independentes em $[0, 1]$).

Logo, ficamos apenas com a integral dupla de $h(x)$.

6.5 Monte Carlo Ray Tracing

Integração numérica aparece em computação gráfica. **Problema:** Determinar a cor (intensidade de luz) em um pixel em uma cena construída por computador.

Resolução: Integrar por todos os caminhos que a luz pode percorrer.

7 Gerando amostras de variáveis aleatórias discretas

A motivação é a realização ou a observação de um fenômeno aleatório. Usamos o computador para poder gerar amostras aleatórias. Fazemos isso para implementar métodos de Monte Carlo, simular sistemas aleatórios, construir jogos, projetar algoritmos, etc.

7.1 Gerando Amostras Uniformes

Seja U uma variável aleatória contínua com distribuição uniforme em $(0, 1]$. O computador consegue gerar amostras de U ?

- O número gerado não é contínuo: representação discretas de números no computador.
- Gerador não é aleatório: computador é determinístico, e um algoritmo vai gerar o número aleatório.

Usaremos então um **gerador pseudo aleatório**.

- O problema 1 é contornado com maior precisão: Divide o intervalo $[0, 1]$ em 2^{64} pedaços.
- O problema 2 é contornado usando algoritmos que misturam bits com manipulações algébricas. (Mersenne-Twister).

7.2 Gerando outras distribuições

Como gerar variáveis aleatórias com outras distribuições? Existem muitas técnicas e algoritmos com diferentes complexidades e aplicabilidade. A melhor abordagem em geral depende da distribuição a ser gerada.

Exemplo: Considere uma moeda enviesada. Seja $B \sim \text{Bernoulli}(p)$ uma variável aleatória que representa a moeda. $P[B = 1] = p$. Como gerar a face da moeda (0 ou 1)?

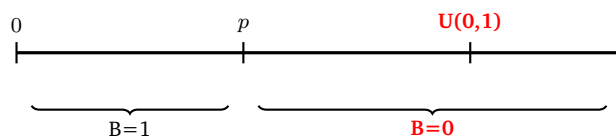


Figura 4: Divisão do intervalo $[0, 1]$

- Dividimos o intervalo $[0, 1]$ em p e $p - 1$.
- Geramos U uniforme
- Se $U \leq p \Rightarrow 1$, caso contrário, retorna 0.

Exemplo: Considere um dado honesto com k faces. Seja D uma variável aleatória que denota o valor da face. $P[D = i] = \frac{1}{k}$. Como gerar o valor da face?

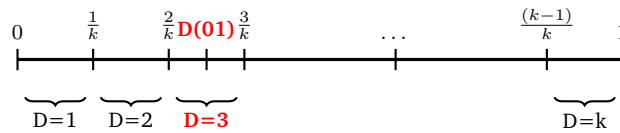


Figura 5: Diviso do intervalo de um dado honesto de k faces

- Dividimos o intervalo $[0, 1]$ em k faixas.
- Geramos D uniforme
- Encontra a faixa que possui o valor gerado de D

Algorithm 1 Gerar Dado Aleatrio: Distribuio Uniforme

```

1: procedure gerar_dado( $k$ )
2:    $D \leftarrow \text{unif}(0, 1)$ 
3:    $i \leftarrow 1$ 
4:   while  $\frac{i-1}{k} \geq D > \frac{i}{k}$  do
5:      $i \leftarrow i + 1$ 
6:   end while
7:   return  $i$ 
8: end procedure

```

Note que D pode estar entre $\frac{i-1}{k}$ e $\frac{i}{k}$.

- Complexidade de melhor caso: 1 (passa uma vez)
- Complexidade de pior caso: k (passa por todos)
- Caso mdio: $\frac{k}{2}$

D pra ser mais eficiente?

Algorithm 2 Gerar Dado Aleatrio: Distribuio Uniforme

```

1: procedure gerar_dado( $k$ )
2:    $D \leftarrow \text{unif}(0, 1)$ 
3:    $i \leftarrow \text{int}(k \cdot D) + 1$ 
4:   return  $i$ 
5: end procedure

```

▷ Pega o teto do nmero

Complexidade: tempo constante, independe de k .

Exemplo 3: Lanando outro dado. Considere agora um dado com k faces que no  honesto. Seja D a varivel aleatria que denota o valor da face. Probabilidade da face i  proporcional ao peso w_i . Isto :

$$P[D = i] = \frac{w_i}{W} \quad W = \sum_{i=1}^k w_i$$

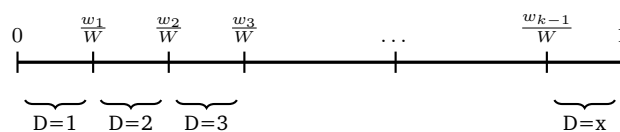


Figura 6: Diviso do intervalo de um dado enviesado k faces

Seguindo a mesma ideia do algoritmo 1, sabemos que

$$\frac{w_{i-1}}{W} < D \leq \frac{w_i}{W}$$

Algorithm 3 Gerar Dado Aleatório

```

1: procedure gerar_dado( $k$ )
2:    $D \leftarrow \text{unif}(0, 1)$ 
3:    $i \leftarrow 1$ 
4:   while  $W \cdot D \neq w_{i-1} \wedge W \cdot D \neq w_i$  do
5:      $i \leftarrow i + 1$ 
6:   end while
7:   return  $i$ 
8: end procedure

```

A complexidade aqui também vai ser proporcional ao número de faces.

Ideia: pré-processamento

- Alocar um vetor de tamanho W
- Preencher w_i posições com valor i
- Escolher um inteiro uniforme em $[1, W]$.
- Acessar o vetor para retomar a face.

1	1	2	2	2	3	4	4
---	---	---	---	---	---	---	---

Figura 7: Exemplo: $W = 8$, $k = 4$, $w_1 = 2$, $w_2 = 3$ e $w_3 = 1$ e $w_4 = 2$

Algorithm 4 Gerar Dado Aleatório: Distribuição Uniforme

```

1: procedure gerar_dado( $k$ )
2:    $W \leftarrow []$ 
3:   for  $i \in 1 : k$  do
4:     for  $j \in 1 : \text{len}(w_i)$  do
5:        $W.append(i)$ 
6:     end for
7:   end for
8:    $D \leftarrow \text{unif}(1, \text{len}(W))$ 
9:    $i \leftarrow W[D]$ 
10:  return  $i$ 
11: end procedure

```

Complexidade dependendo de W .

Se w_i não for inteiro, o método anterior não funciona. Esse método é chamado *Alias Method*.

7.3 Gerando Geométrica

Seja X uma variável aleatória com distribuição geométrica p onde

$$p_i = P[X = i] = (1 - p)^{i-1}p, \quad i = 1, 2, \dots$$

Ideia: Normalizar p_i por $P = \sum_{i=1}^{\infty} p_i$. Ou seja, $P[X = i] = \frac{p_i}{P}$. Se dividirmos o espaço sobre P , teríamos:

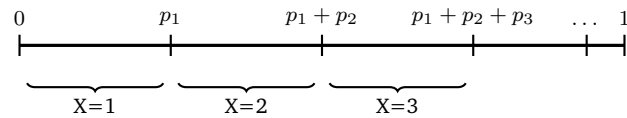


Figura 8: Divisão do intervalo P

Como melhorar? Podemos encarar a Geométrica como uma sequência de Bernoullis até a ocorrência de positivo. Geramos uma sequência de Bernoullis(p) até observar um evento positivo.

Algorithm 5 Sequências de Bernoullis até ocorrência de positivo

```

1: procedure gerar_dado( $k$ )
2:    $c \leftarrow 1$ 
3:   while True do
4:     if  $\text{unif}(0, 1) \leq p$  then                                     ▷ Ocorreu
5:       return  $c$ 
6:     else
7:        $c \leftarrow c + 1$ 
8:     end if
9:   end while
10: end procedure

```

- **Vantagem:** Não precisa calcular o valor p_i da geométrica.
- **Desvantagem:** Gera muitas amostras uniformes.

Uma outra abordagem é usar a inversa da função de *distribuição cumulativa* $F_X(x)$.

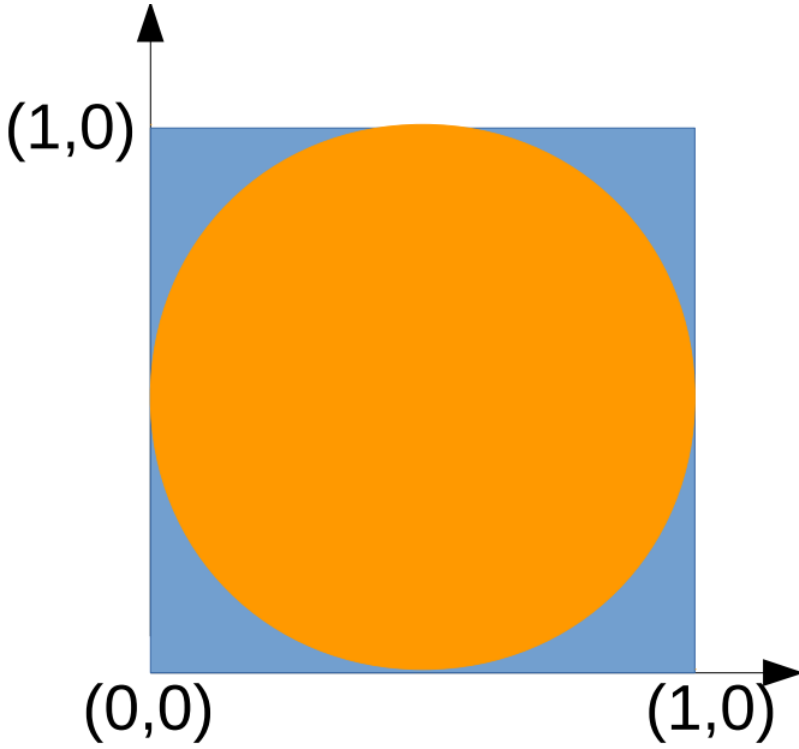


Figura 9: Distribuição cumulativa da Geométrica: No eixo y , $F_X(x)$ e no eixo x os valores do domínio de F_X .

Dessa forma, para saber que x corresponde a $U(0,1)$ basta tomar $F_X^{-1}(unif(0,1))$. Se você tem a inversa, a complexidade é constante. Esse método é chamado *Método da transformada inversa*. Falaremos a seguir.

7.4 Método da transformada inversa

Seja $U \sim unif(0,1)$ e X uma variável aleatória com função cumulativa $F_X(x) = P[X \leq x]$. Logo $X = F_X^{-1}(U)$, onde:

- F_X^{-1} é a inversa de F_X (o valor de x para qual $F_X(x) = u$).
- Podemos usar a inversa para gerar amostras!

Prova: $X = F_X^{-1}(U)$

$$P[X \leq x] = P[F_X^{-1}(U) \leq x] = P[F_X(F_X^{-1}(U)) \leq F_X(x)] = P[U \leq F_X(x)] = F_X(x)$$

A distribuição cumulativa da geométrica com parâmetro p é:

$$F_X(i) = P[X \leq i] = 1 - (1-p)^i \quad i = 1, 2, \dots$$

Pegando a função inversa:

$$1 - (1-p)^i = u \Rightarrow F_X^{-1} = \text{int} \left(\frac{\log(u)}{\log(1-p)} \right) + 1$$

7.5 Gerando Binomial

A cumulativa da binomial é analiticamente muito difícil de inverter. Porém, a binomial também é feita de sequências de N bernoullis com parâmetro p .

Algorithm 6 Sequências de Bernoullis positivas

```
1: procedure binomial( $n$ )
2:    $c \leftarrow 1$ 
3:    $k \leftarrow 0$ 
4:   while  $k \leq n$  do
5:     if  $\text{unif}(0, 1) \leq p$  then                                     ▷ Ocorreu
6:        $c \leftarrow c + 1$ 
7:     end if
8:      $k \rightarrow k + 1$ 
9:   end while
10:  return  $c$ 
11: end procedure
```

Problema: p muito baixo e N muito grande, teremos muitos zeros.

Segunda ideia: usar uma geometria para acelerar a convergência.

Podemos encarar uma sequência de n jogadas como várias geométricas.

001	0001	01	0001	001	01	1	00100
-----	------	----	------	-----	----	---	-------

Figura 10: 8 distribuições geométricas para cada acerto em 23 jogadas

Em outras palavras, X vai ser computada pelo menor valor que faz com que o somatório ultrapasse N jogadas:

$$X = k | \min_k \sum_{i=1}^{\infty} G_i \geq N$$

Onde $G \sim \text{Geom}(p)$.

Algoritmo: Gerar sequência iid $\text{Geom}(p)$ até a soma ser $\geq N$.

Complexidade: Na média o número de geométrica é Np .

7.6 Gerando permutações

Gerar uma permutação das cartas de um baralho com probabilidade uniforme. Suponha que o baralho tem N cartas.

Ideia 1:

- Gerar i uniforme entre 1 e $N!$.
- Retornar a i -ésima permutação.

Ideia 2:

- Escolher um elemento por vez de forma uniforme. (sem repetição)
- Fazer n escolhas sucessivas.

Nenhuma dessas ideias é efetiva.

Algoritmo de Knuth Suffle: Usar um vetor para alocar elementos e as escolhas realizadas.

Exemplo: $N = 14$.

Construímos um vetor permuta:

1	2	3	4	5	6	7	8	9	10	11	12	13	14
---	---	---	---	---	---	---	---	---	----	----	----	----	----

Figura 11: Vetor permuta

- Escolher uniforme entre 1 e 14: Saiu 9. Permutar 9 com 14.
- Permutar com o último elemento.

1	2	3	4	5	6	7	8	14	10	11	12	13	9
---	---	---	---	---	---	---	---	----	----	----	----	----	---

Figura 12: Vetor permuta

- Escolher uniforme entre 1 e 13: Saiu 5. Permutar 9 com 14.
- Permutar com o último elemento.

1	2	3	4	13	6	7	8	14	10	11	12	9	5
---	---	---	---	----	---	---	---	----	----	----	----	---	---

Figura 13: Vetor permuta

Ao final, o vetor contém uma permutação uniforme de complexidade $O(n)!$.
O pseudo código é descrito a seguir:

Algorithm 7 Knuth Shuffle

```

1: procedure KS(Vetor_permuta)
2:   for  $i = 0$  até  $N - 1$  do
3:      $j \leftarrow \text{unif}(1, N - i)$ 
4:      $\text{tmp} \leftarrow \text{permuta}[j]$ 
5:      $\text{permuta}[j] \leftarrow \text{permuta}[N - i]$ 
6:      $\text{permuta}[N - i] \leftarrow \text{tmp}$ 
7:   end for
8: end procedure

```

8 Rejection Sampling (Amostragem por rejeição)

Técnica fundamental para geração de amostras aleatórias. A ideia é usar distribuição simples para gerar amostras de outra distribuição mais complicada. Sejam X e Y duas variáveis aleatórias com distribuições p_x e q_x , definidas no mesmo domínio.

- $p_x = P[X = x]$, $q_x = P[Y = x]$

Assuma que $q_x \leq cp_x$ para alguma constante c e todo x , ou seja, cp_x é um função envelope para q_x . Vamos assumir também que sabemos gerar amostras para X . Mas como gerar amostras para Y ?

Algorithm 8 Rejection Sampling

```
1: procedure  $RS$ 
2:   while True do
3:     Gerar valor para  $i$  a partir de  $p_x$ 
4:     Gerar  $u$  uniforme  $(0, cp_i)$  contínua, usando o  $i$  gerado
5:     if  $u < q_i$  then
6:       return  $i$ 
7:     end if
8:   end while
9: end procedure
```

- Algoritmo pode rejeitar amostra de X várias vezes
- Aceita com probabilidade dada pela razão entre q_x e cp_x

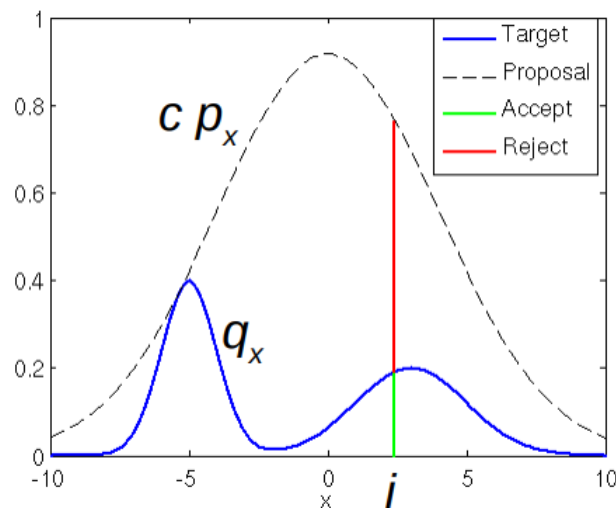


Figura 14: Algoritmo representado graficamente

OBS: Achar o valor de c para que a probabilidade p seja um envelope para uma probabilidade difícil q .

Demonstração: O algoritmo gera amostra i com probabilidade q_i

$$q_i = P[Y = i] = P[X = i | \text{aceitar}]$$

Considere um valor i e o evento aceitar

$$P[X = i \cap \text{aceitar}] = P[X = i]P[\text{aceitar} | X = i] = \frac{p_i q_i}{c p_i} = \frac{q_i}{c}$$

Probabilidade de aceitar (por prob. Total)

$$P[\text{aceitar}] = \sum_i P[X = i, \text{aceitar}] = \sum_i \frac{q_i}{c} = \frac{1}{c}$$

Ao final de cada rodada, algoritmo aceita com probabilidade $\frac{1}{c}$. Temos então

$$P[X = i | \text{aceitar}] = \frac{P[X = i \cap \text{aceitar}]}{P[\text{aceitar}]} = q_i = P[Y = i]$$

- A cada rodada, algoritmo aceita com probabilidade $\frac{1}{c}$.
- Número de rodadas é aleatório.
- A distribuição é geométrica com parâmetro $\frac{1}{c}$.

- O valor esperado é, então, c . Complexidade de caso médio para cada amostra.
- Escolha do valor para c é muito importante. A escolha tem que ser o menor valor tal que $q_x \leq cp_x$ para todo x .
- Valor depende da *distância* entre q_x e p_x . Se muito diferentes, pode demandar c muito grande.
- Esta técnica também funciona com variáveis aleatórias contínuas.

8.1 Exemplo 1

Seja uma moeda enviesada em que a probabilidade de ter cara maior que $\frac{1}{2}$. Como gerar moeda sem viés?

- Encontrar constante c tal que $q_x \leq cp_x$
- $c(1 - p) = \frac{1}{2} \Rightarrow c = \frac{1}{2(1-p)}$

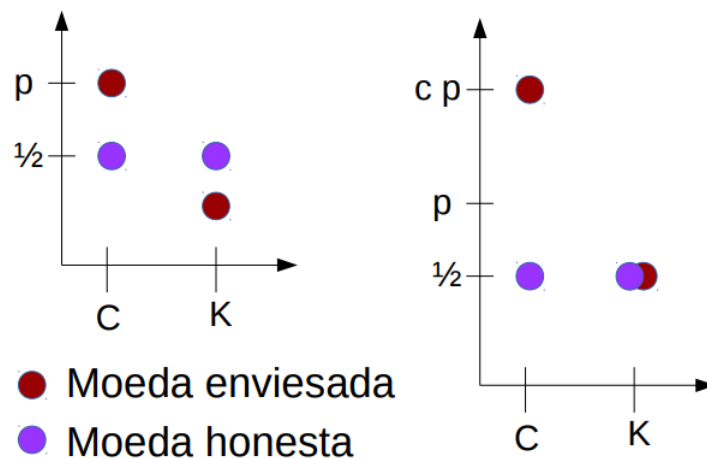


Figura 15: Geração de moeda

Aplicação do algoritmo:

1. Gerar valor para $i = \{C, K\}$ a partir de p_x (moeda enviesada).
2. Gerar u uniforme $(0, cp_i)$ contínua, usando o i gerado.
3. Se $u < q_i$, retorna i , caso contrário volta para o passo 1.

8.2 Exemplo 2

Gerar amostras da variável aleatória Y contínua com densidade $f_y(x) = 20x(1-x)^3, 0 \leq x \leq 1$. Usando o método da rejeição, qual a distribuição proponente?

- Uniforme em $[0, 1]$, $g_X(x) = 1, 0 \leq x \leq 1$.
- Determinar c tal que $f(x) \leq cg(x)$ para $0 \leq x \leq 1$.
- A ideia é encontrar o máximo de $\frac{f(x)}{g(x)}$.
- O máximo é no ponto $x = \frac{1}{4}$, logo $c = \frac{135}{64}$

Aplicação do algoritmo:

1. Gerar valor para x a partir de $g(x)$ (uniforme $[0,1]$).
2. Gerar u uniforme $(0, cg(x))$ contínua, usando o x gerado.
3. Se $u < f(x)$, retorna x . Caso contrário, volta ao passo 1.

8.3 Cenário Problemático

Algoritmo de Monte Carlo é um estimador universal e a média amostral converge para o valor esperado. O problema é quando a variância do estimador é alta! Com isso, muitas amostras seriam necessárias. Exemplo:

$$G_N = \sum_{i=1}^N g(i) \quad M_n = \frac{1}{n} \sum_{i=1}^n g(X_i), X \sim Unif(1, N)$$

- $g(i)$ é desprezível (ou zero) para muitos valores de i , e grande para poucos valores.
- Teremos que gerar muitas amostras para "acertar" os valores importantes de $g(i)$.

8.4 Importance Sampling

Amostrar com probabilidade diferente da original. Buscando dar mais probabilidade a região mais importante do problema em questão. Pra isso, é necessário compensar de alguma forma o aumento dessa probabilidade. Então a ideia é usar o método de Monte Carlo em um problema reformulado, com novas funções de distribuição, no intuito de reduzir a variância do estimador.

Seja X uma variável aleatória uniforme $(1, N)$.

$$E[g(X)] = \frac{1}{N} \sum_{i=1}^N g(i) = \frac{G_N}{N}$$

Seja Y uma variável aleatória com distribuição dada por $h(i) > 0$, para todo i .

$$E_h\left[\frac{g(Y)}{h(Y)}\right] = \sum_{i=1}^N \frac{g(i)}{h(i)} \cdot h(i) = \sum_{i=1}^N g(i) = G_N$$

Podemos estimar G_N estimando o valor esperado através da média amostral nesse caso. Usar h para reduzir a variância.

Assim, seja Y_i uma sequência iid de variáveis aleatórias com distribuição $h(i)$

$$M_n = \frac{1}{n} \sum_{i=1}^n \frac{g(Y_i)}{h(Y_i)}$$

O valor esperado de M_n é:

$$E_h[M_n] = \frac{1}{n} \sum_{i=1}^n E_h\left[\frac{g(Y_i)}{h(Y_i)}\right] = G_N$$

E a variância:

$$Var_h[M_n] = \frac{\sigma_{g/h}^2}{n}$$

Onde $\sigma_{g/h}^2$ é a variância da variável aleatória $\frac{g(Y)}{h(Y)}$.

M_n converge para $E_h\left[\frac{g(Y)}{h(Y)}\right] = G_N$ e a variância do estimador depende da variância de $\frac{g(Y)}{h(Y)}$.

$\sigma_{g/h}^2$ depende apenas do momento:

$$E_h\left[\left(\frac{g(Y)}{h(Y)}\right)^2\right] = \sum_{i=1}^N \left(\frac{g(i)}{h(i)}\right)^2 \cdot h(i) = \sum_{i=1}^N \frac{g(i)^2}{h(i)}$$

Exemplo: Dado N , calcular $G_N = \sum_{i=1}^N g(i)$ onde $g(i) = i \log i$. Seja Y_i uma sequência de iid de variáveis aleatórias com distribuição $h(i) > 0$ para todo i .

Opção 1:

- $h(i) = \frac{1}{N}$ para todo i , ou seja, $h(i)$ é uniforme em $(1, N)$.

$$\sum_{i=1}^N \frac{g(i)^2}{h(i)} = N \sum_{i=1}^N g(i)^2 = N \sum_{i=1}^N i^2 \log^2 i$$

Como reduzir este segundo momento? A ideia é escolher $h(i)$ proporcionalmente a $g(i)$.

Opção 2:

- $h(i) = \frac{i}{K_2}$, ou seja, linearmente proporcional a i .
- onde $K_2 = 1 + 2 + 3 + 4 + \dots + N = \frac{N(N+1)}{2}$

$$\sum_{i=1}^N \frac{g(i)^2}{h(i)} = K_2 \sum_{i=1}^N \frac{g(i)^2}{i} = K_2 \sum_{i=1}^N i \log^2 i$$

Opção 3:

- $h(i) = \frac{i^3}{K_3}$, ou seja, cúbica em i .
- onde $K_3 = 1^3 + 2^3 + 3^3 + \dots + N^3 = \frac{N^2(N+1)^2}{4}$

$$\sum_{i=1}^N \frac{g(i)^2}{h(i)} = K_3 \sum_{i=1}^N \frac{g(i)^2}{i^3} = K_3 \sum_{i=1}^N \frac{\log^2 i}{i}$$

Qual a melhor opção? A opção que possui menor variância! Ou menor segundo momento!

- **Opção 1:** $1.44 \cdot 10^1 3$
- **Opção 2:** $1.03 \cdot 10^1 3$
- **Opção 3:** $2.75 \cdot 10^1 3$

Logo, o melhor estimador é a **opção 2**. Isso significa que vamos ter menos amostras para um mesmo erro e um erro menor para um mesmo número de amostras. Mas qual seria a melhor $h(i)$ possível? A melhor função é a que induz variância igual a zero.

$$\sigma_{g/h}^2 = E_h \left[\left(\frac{g(Y)}{h(Y)} - \mu_{g/h} \right)^2 \right]$$

Se $\frac{g(i)}{h(i)} = \mu_{g/h}$ para todo i , então a variância é zero!

A ideia é tentar aproximar $\mu_{g/h}$ por heurísticas.

8.5 Generalização

Suponha que queremos calcular um determinado valor esperado, onde X tem distribuição dada por f .

$$\mu_f = E_f[g(X)] = \sum_i g(i)f(i)$$

Podemos aplicar *Importance Sampling* neste problema, ou seja, amostrar mais regiões importantes para g . Seja h outra distribuição para variável aleatória X , tal que $f(i) > 0 \rightarrow h(i) > 0$.

$$\mu_f = \sum_i \frac{g(i)f(i)}{h(i)} h(i) = E_h \left[\frac{g(X)f(X)}{h(X)} \right]$$

Ou seja, $E_f[g(X)] = E_h \left[\frac{g(X)f(X)}{h(X)} \right]$. Esse valor esperado E_h pode ser estimado, X em E_h tem distribuição h . Podemos reduzir a variância escolhendo h .

Algoritmo para estimar μ_f .

Algorithm 9 Monte Carlo

```
1: procedure  $MS$ 
2:    $S \leftarrow 0$ 
3:   for  $i = 1$  até  $n$  do
4:     Gerar amostra  $x$  com distribuição  $h$ 
5:      $S \leftarrow S + \frac{g(x)f(x)}{h(x)}$ 
6:   end for
7:   return  $\frac{S}{n}$ 
8: end procedure
```

Note que, se h for bem escolhida, a variância do estimador pode ser bem menor que estimador usando f .

9 Lista 2 - Algoritmos de Monte Carlo

10 Cauda do dado em ação

Considere um icosaedro (um sólido Platônico de 20 faces) honesto, tal que a probabilidade associada a cada face é $\frac{1}{20}$. Considere que o dado será lançado até que um número primo seja observado e seja Z a variável aleatória que denota o número de vezes que o dado é lançado. Responda às perguntas abaixo:

1. Determine a distribuição Z , ou seja, $P[Z = k], k = 1, 2, \dots$. Que distribuição é esta?
2. Utilize a desigualdade de *Markov* para calcular um limitante para $P[Z \geq 10]$.
3. Utilize a desigualdade de *Chebyshev* para calcular um limitante para $P[Z \geq 10]$.
4. Calcule o valor exato de $P[Z \geq 10]$. (dica: use probabilidade complementar). Compare os valores obtidos.

i

Resposta: Determine a distribuição Z , ou seja, $P[Z = k], k = 1, 2, \dots$. Que distribuição é esta?

Considere uma variável aleatória indicadora Y que denota a primalidade da face do dado. A probabilidade da face ter um número primo é $P[Y = 1] = \frac{2}{5}$. A variável Z , por sua vez, denota o número de vezes que o dado é lançado até o primeiro primo. Em outras palavras:

$$Z = \min\{i | Y_i = 1\}$$

Dessa forma, $Z \sim \text{Geo}(\frac{2}{5})$. Assim:

$$P[Z = k] = \frac{3}{5}^{k-1} \cdot \frac{2}{5}$$

i

Resposta: Utilize a desigualdade de *Markov* para calcular um limitante para $P[Z \geq 10]$.

Para uma v.a. qualquer, desigualdade de Markov é:

$$P[X \geq a] \leq \frac{E[X]}{a}$$

Assim,

$$P[Z \geq 10] \leq \frac{E[Z]}{10}$$

Como $Z \sim Geo(\frac{2}{5})$, então sabemos que a esperança é $\frac{1}{\frac{2}{5}} = 2.5$. Portanto:

$$P[Z \geq 10] \leq \frac{2.5}{10} \Leftrightarrow P[Z \geq 10] \leq \frac{1}{4} = 0.25$$

Resposta: Utilize a desigualdade de *Chebyshev* para calcular um limitante para $P[Z \geq 10]$.

Para qualquer variável aleatória X com valor esperado μ e variância σ^2 e para qualquer k não-negativo, temos

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Para adaptar ao nosso caso, faremos $P[Z \geq 10] = P[Z - \mu \geq 10 - \mu]$. Assim, para aplicar Chebyshev, devemos resolver $k\sigma = 10 - \mu$. Como $Z \sim Geo(\frac{2}{5})$, $\mu = 2.5$ e $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{(1-p)}{p^2}} = \frac{5\sqrt{3}}{2\sqrt{5}}$.

$$k\sigma = 10 - \mu \Leftrightarrow k \cdot \frac{5\sqrt{3}}{2\sqrt{5}} = 7.5 = \frac{15}{2} \Leftrightarrow k = \frac{15}{2} \cdot \frac{2\sqrt{5}}{5\sqrt{3}} = \frac{3\sqrt{5}}{\sqrt{3}} \Leftrightarrow k = \sqrt{15}$$

Assim, aplicando Chebyshev:

$$P[Z \geq 10] \leq \frac{1}{15} = 0.067$$

Resposta: Calcule o valor exato de $P[Z \geq 10]$. (dica: use probabilidade complementar). Compare os valores obtidos.

Como $P[Z \geq 10] = 1 - P[Z < 10]$,

$$1 - P[Z < 10] = 1 - (P[Z = 1] + P[Z = 2] + P[Z = 3] + \dots + P[Z = 9]) = \sum_{i=1}^9 \frac{3^{i-1}}{5} \cdot \frac{2}{5}$$

Esta é uma soma de uma PG de razão $r = \frac{3}{5}$.

$$S = \frac{a_1(r^n - 1)}{r - 1} \Leftrightarrow S = \frac{\frac{2}{5} \left(\frac{3^9}{5} - 1 \right)}{\frac{3}{5} - 1} \Leftrightarrow S = 1 - \frac{3^9}{5^9} \approx 0.9899$$

Assim,

$$P[Z \geq 10] = 1 - 0.9899 = 0.0101$$

11 Moedas

Você tem duas moedas: uma honesta e outra enviesada que produz cara com probabilidade $\frac{3}{4}$. Uma das duas moedas é escolhida aleatoriamente e lançada n vezes. Seja S_n o número total de caras observadas nas n jogadas. Responda às perguntas abaixo:

1. Determine a fração média de caras que será observada.
2. Podemos determinar qual moeda foi escolhida, depois da mesma ser lançada n vezes?
3. Determine o valor de n tal que tenhamos 95% de chance de acertar qual moeda foi escolhida.

Resposta: Determine a fração média de caras que será observada.

Seja X_i uma sequência de variáveis aleatórias iid, tal que $\mu = E[X_i]$ e $\sigma^2 = Var[X_i]$.

Seja $M_n = \frac{1}{n} \sum_{i=1}^n X_i$ a média amostral, em outras palavras, a média aritmética dos n valores da amostra.

Seja A a moeda honesta e B a moeda enviesada. Definimos X uma v.a. indicadora que define se a moeda é cara ou coroa. A $P[X_A = 1 = cara] = \frac{1}{2}$ e $P[X_B = 1 = cara] = \frac{3}{4}$. Para n amostras aleatórias, temos:

$$M_n^A = \frac{1}{n} \sum_{i=1}^n X_A$$

$$M_n^B = \frac{1}{n} \sum_{i=1}^n X_B$$

Resposta: Podemos determinar qual moeda foi escolhida, depois da mesma ser lançada n vezes?

Para uma certa precisão e confiança, quando $n \rightarrow \infty$ pela Lei dos Grandes Números, a probabilidade do número de caras se aproximará do valor esperado. Se essa probabilidade for próximo de M_n^A , diremos que a moeda é honesta. Por outro lado, se aproximar de M_n^B , diremos que é enviesada.

Resposta: Determine o valor de n tal que tenhamos 95% de chance de acertar qual moeda foi escolhida.

Usaremos a desigualdade de *Chebyshev*:

$$P[|M_n - \mu| < \epsilon] \geq 1 - \frac{\sigma^2}{\epsilon^2 n} = \beta$$

Daí,

$$n = \frac{\sigma^2}{\epsilon^2(1 - \beta)}$$

Precisamos calcular a variância. Logo, a variância associada a moeda honesta:

$$Var[X_A] = p_A(1 - p_A) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

E a variância associada a moeda eviesada:

$$Var[X_B] = p_B(1 - p_B) = \frac{3}{4} \cdot \frac{1}{4} = \frac{3}{16}$$

Calculando o número de jogadas para precisão $\epsilon = 0.01$ e confiança $\beta = 0.95$, para a moeda honesta:

$$n_A = \frac{\frac{1}{4}}{(0.01)^2(1 - 0.95)} = \frac{\frac{1}{4}}{(0.0001)(0.05)} = 50.000$$

Para a moeda enviesada:

$$n_B = \frac{\frac{3}{16}}{(0.01)^2(1 - 0.95)} = \frac{\frac{3}{16}}{(0.0001)(0.05)} = 37.500$$

Tomamos o maior valor, devido a Lei dos Grandes Números. Neste caso, vemos que a probabilidade do evento, no infinito, parece se aproximar de M_n^A depois de 50.000 jogadas.

Dessa forma, serão necessárias 50.000 jogadas.

12 Sanduíches

Você convidou 64 pessoas para uma festa e agora precisa preparar sanduíches para os convidados. Você acredita que cada convidado irá comer 0, 1 ou 2 sanduíches com probabilidades $1/4$, $1/2$ e $1/4$, respectivamente. Assuma que o número de sanduíches que cada convidado irá comer é independente de qualquer outro convidado. Quantos sanduíches você deve preparar para ter uma confiança de 95% de que não vai faltar sanduíches para os convidados?

Seja X a variável aleatória que contabiliza a quantidade de sanduíches por pessoa. Modelamos de tal forma que $P[X = 0] = \frac{1}{4}$, $P[X = 1] = \frac{1}{2}$, $P[X = 2] = \frac{1}{4}$.

$$E[X] = \sum_{i=0}^2 X_i P[X = i] = 0 \frac{1}{4} + 1 \frac{1}{2} + 2 \frac{1}{4} = 1$$

$$E[X^2] = \sum_{i=0}^2 X_i^2 P[X = i] = 0^2 \frac{1}{4} + 1^2 \frac{1}{2} + 2^2 \frac{1}{4} = \frac{3}{2}$$

$$Var[X] = E[X^2] - E[X]^2 = \frac{3}{2} - 1 = \frac{1}{2}$$

Seja Y a quantidade total de sanduíches consumidos. Sendo assim:

$$Y = \sum_{i=1}^{64} X_i$$

O valor esperado de Y , por sua vez:

$$E[Y] = \sum_{i=1}^{64} E[X_i] \Leftrightarrow E[Y] = \sum_{i=1}^{64} 1 = 64$$

E sua variancia:

$$Var[Y] = \sum_{i=1}^{64} Var[X_i] \Leftrightarrow Var[Y] = \sum_{i=1}^{64} \frac{1}{2} = 32$$

Agora, aplicaremos a Lei dos Grandes Números (fraca).

$$P[|M_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{\epsilon^2 n} = 1 - \beta$$

Pela desigualdade de Chebyshev:

$$P[|Y - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

Tomaremos como precisão $\epsilon = k\sigma$ e confiança $1 - \beta = \frac{1}{k^2}$. Assim:

$$\frac{1}{k^2} = 1 - 0.95 \Leftrightarrow k \approx 4,47$$

Aplicando:

$$P[Y - 64 \geq 4,47\sqrt{32}] \leq \frac{1}{4,47^2} \Rightarrow P[Y - 64 \geq 25,28] \leq \frac{1}{4,47^2} \Rightarrow P[Y \geq 89,28] \leq \frac{1}{4,47^2}$$

Devem ser preparados 90 sanduíches.

13 Graus improváveis

Considere o modelo de grafo aleatório de Erdos-Renyi (também conhecido por $G(n, p)$), onde cada possível aresta de um grafo rotulado com n vértices ocorre com probabilidade p , independentemente. Responda às perguntas abaixo:

1. Determine a distribuição do vértice 1 (em função de n e p).
2. Determine o valor de γ (em função de n e p) tal que com alta probabilidade $(1 - \frac{1}{n})$ o grau observado no vértice 1 é menor ou igual a γ .

1

Resposta: Determine a distribuição do vértice 1 (em função de n e p).

Seja X_i uma variável aleatória que denota o grau de um vértice i . Segue o seguinte exemplo para $n = 4$

- $P[X_i = 0] = (1 - p)^3$ (a probabilidade de não ter aresta entre os 3 outros vértices)
- $P[X_i = 1] = p(1 - p)^2 \binom{3}{1}$ (probabilidade de ter uma aresta e não ter as outras duas, nas três possíveis permutações)
- $P[X_i = 2] = p^2(1 - p) \binom{3}{1}$ (probabilidade de ter duas arestas e não ter uma, nas três possíveis permutações)
- $P[X_i = 3] = p^3$ (probabilidade de ter todas as arestas possíveis)

Generalizando, temos então:

$$P[X = i] = \binom{n-1}{i} p^i (1-p)^{n-1-i}$$

Uma distribuição Binomial. Logo, $X \sim \text{Bin}(n-1, p)$.

1

Resposta:

Determine o valor de γ (em função de n e p) tal que com alta probabilidade $(1 - \frac{1}{n})$ o grau observado no vértice 1 é menor ou igual a γ .

Definiremos o evento A como sendo o evento do grau ser menor ou igual a γ . Ou seja,

$$P[X_1 \leq \gamma(n, p)] \geq 1 - \frac{1}{n}$$

Usaremos a distribuição cumulativa da Binomial. Assim:

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

Utilizando chernoff, de cauda:

$$P[X \leq (1 - \delta)\mu] \leq e^{-\mu \frac{\delta^2}{2}}$$

Como $\mu = (n-1)p$,

$$(1 - \delta)(n-1)p = \gamma \Rightarrow -(n-1)p\delta = \gamma - (n-1)p \Rightarrow (n-1)p\delta = (n-1)p - \gamma \Rightarrow \delta = \frac{(n-1)p - \gamma}{(n-1)p}$$

$$P[X \leq \gamma] \leq e^{-\frac{(n-1)p \left(\frac{(n-1)p - \gamma}{(n-1)p} \right)^2}{2}}$$

$$P[X \leq \gamma] \leq e^{-\frac{(n-1)p \frac{((n-1)p - \gamma)^2}{(n-1)^2 p^2}}{2}}$$

$$P[X \leq \gamma] \leq e^{-\frac{((n-1)p-\gamma)^2}{2(n-1)p}}$$

Com alta probabilidade, temos:

$$1 - \frac{1}{n} \leq P[X \leq \gamma] \leq e^{-\frac{((n-1)p-\gamma)^2}{2(n-1)p}}$$

Portanto,

$$1 - \frac{1}{n} \leq e^{-\frac{((n-1)p-\gamma)^2}{2(n-1)p}}$$

Utilizando a ideia de que $e^x \approx 1 + x$, podemos aproximar $1 - \frac{1}{n} \approx e^{-\frac{1}{n}}$. Assim,

$$e^{-\frac{1}{n}} \leq e^{-\frac{((n-1)p-\gamma)^2}{2(n-1)p}}$$

Aplicando logaritmo dos dois lados:

$$\begin{aligned} -\frac{1}{n} &\leq -\frac{((n-1)p-\gamma)^2}{2(n-1)p} \Rightarrow \frac{1}{n} \leq \frac{((n-1)p-\gamma)^2}{2(n-1)p} \\ -\frac{1}{n} &\leq -\frac{((n-1)p-\gamma)^2}{2(n-1)p} \Rightarrow \frac{2(n-1)p}{n} \leq ((n-1)p-\gamma)^2 \\ \sqrt{\frac{2(n-1)p}{n}} &\leq (n-1)p-\gamma \end{aligned}$$

Portanto,

$$\gamma \leq (n-1)p - \sqrt{\frac{2(n-1)p}{n}}$$

14 Calculando uma importante constante

Questão feita no notebook: [Google Colab](#)

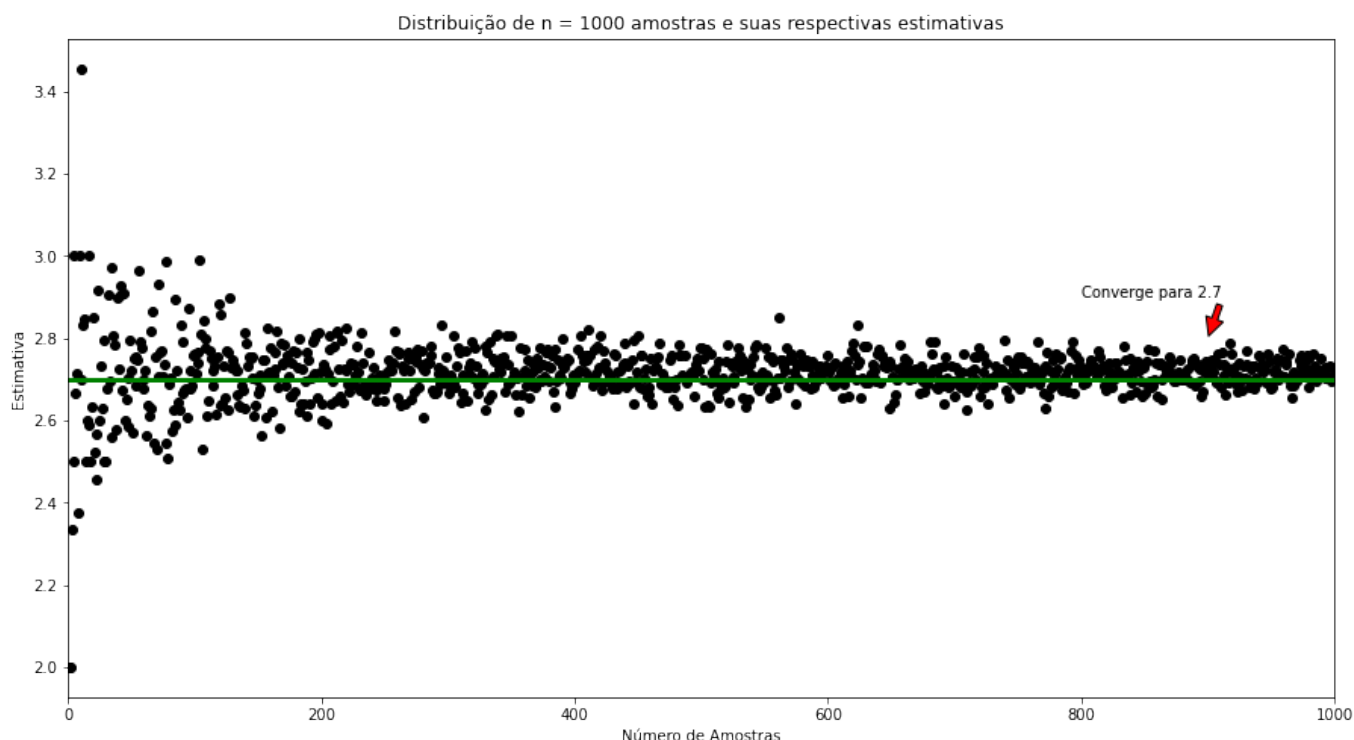


Figura 16: Número de amostras x Estimativa

Está convergindo para, aproximadamente, e .

15 Transformada Inversa

Mostre como o método da transformada inversa pode ser usado para gerar amostras de uma v.a. contínua X com as seguintes densidades:

1. Distribuição exponencial com parâmetro $\lambda > 0$, cuja função de densidade é dada por $f_X(x) = \lambda e^{-\lambda x}$, para $x \geq 0$.
2. Distribuição de Pareto com parâmetros $x_0 > 0$ e $\alpha > 0$, cuja função densidade é dada por $f_X(x) = \frac{\alpha x_0^\alpha}{x^{\alpha+1}}$, para $x \geq x_0$.

i

Resposta: Distribuição exponencial com parâmetro $\lambda > 0$, cuja função de densidade é dada por $f_X(x) = \lambda e^{-\lambda x}$, para $x \geq 0$.

$$f_X(x) = \lambda e^{-\lambda x} \Rightarrow F(x) = \int_0^x f_X(x) dx = \lambda \int_0^x e^{-\lambda x} dx$$

Fazendo a substituição $u = -\lambda x$ e $du = -\frac{1}{\lambda}$

$$F(x) = \lambda \int_0^{-\lambda x} -\frac{e^u}{\lambda} du \Rightarrow - \int_0^{-\lambda x} e^u du$$

Aplicando os limites de integração

$$F(x) = -[e^u]_0^{-\lambda x} \Rightarrow F(x) = -e^{-\lambda x} + 1$$

Tomando $F^{-1}(x)$ temos:

$$y = -e^{-\lambda x} + 1 \Rightarrow \ln(1 - y) = \lambda x \Rightarrow x = \frac{\ln(1 - y)}{\lambda}$$

Criando uma variável aleatória uniforme $U \sim \text{unif}(0, 1)$, temos pela transformada inversa que $x = F^{-1}(U)$, basta gerar amostras de:

$$x = \frac{\ln(1 - U)}{\lambda} \quad U \sim \text{unif}(0, 1)$$

i

Resposta: Distribuição de Pareto com parâmetros $x_0 > 0$ e $\alpha > 0$, cuja função densidade é dada por $f_X(x) = \frac{\alpha x_0^\alpha}{x^{\alpha+1}}$, para $x \geq x_0$.

$$f_X(x) = \frac{\alpha x_0^\alpha}{x^{\alpha+1}} \Rightarrow F(x) = \int_{x_0}^x f_X(x) dx = \alpha x_0^\alpha \int_{x_0}^x \frac{1}{x^{\alpha+1}} dx$$

Aplicando os limites de integração:

$$F(x) = \alpha x_0^\alpha \int_{x_0}^x x^{-\alpha-1} \Rightarrow \alpha x_0^\alpha \left[-\frac{x^{-\alpha}}{\alpha} \right]_{x_0}^x \Rightarrow F(x) = -x_0^\alpha x^{-\alpha} + 1$$

Tomando $F^{-1}(x)$ temos:

$$y = -x_0^\alpha x^{-\alpha} + 1 \Rightarrow -y + 1 = x_0^\alpha x^{-\alpha} \Rightarrow \frac{1 - y}{x_0^\alpha} = \frac{1}{x^\alpha} \Rightarrow x = \sqrt[\alpha]{\frac{x_0^\alpha}{1 - y}}$$

Criando uma variável aleatória uniforme $U \sim \text{unif}(0, 1)$, temos pela transformada inversa que $x = F^{-1}(U)$, basta gerar amostras de:

$$x = \frac{x_0}{(1 - U)^{\frac{1}{\alpha}}} \quad U \sim \text{unif}(0, 1)$$

16 Contando domínios na web

Questão feita no notebook: [Google Colab](#)

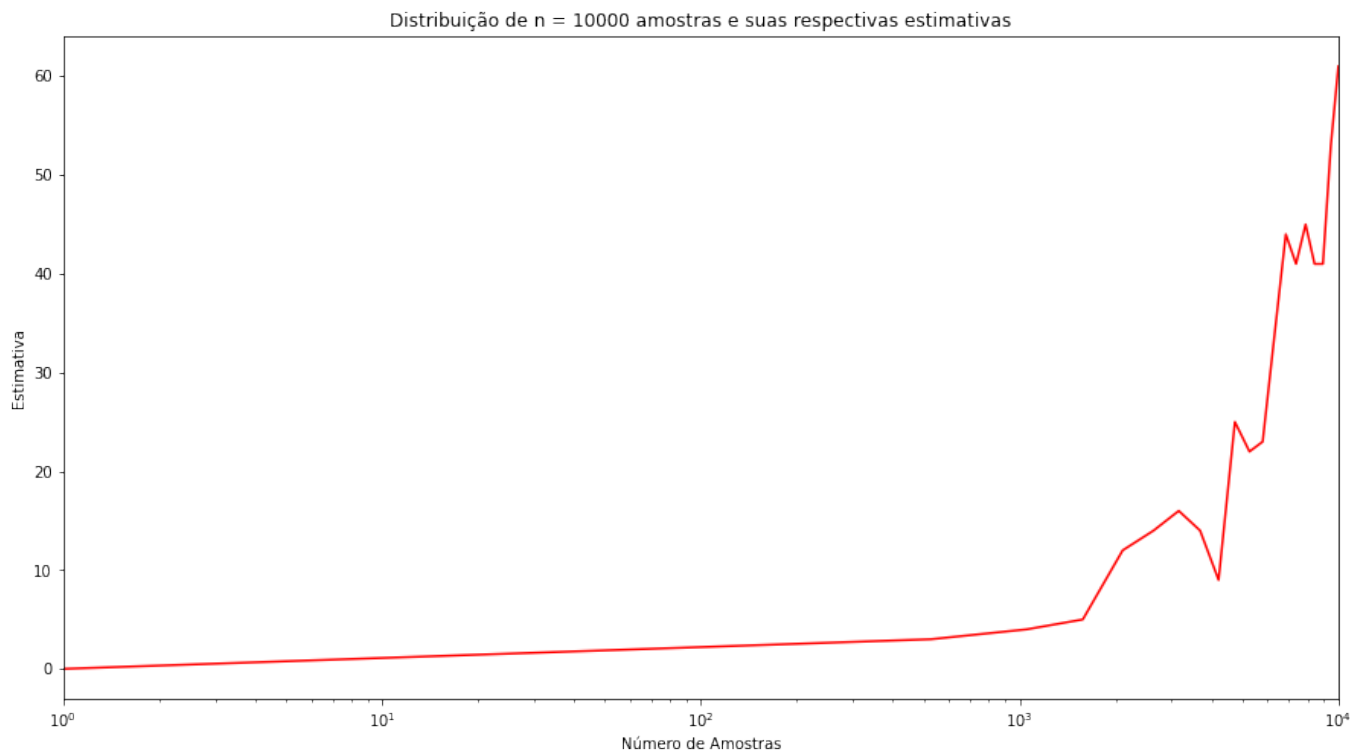


Figura 17: Número de amostras x Estimativa

w_n parece crescer conforme n cresce.

17 Rejection Sampling

Questão feita no notebook: [Google Colab](#)

Usaremos $g(X) = 1$. Logo, $f_x(X) = c * g(X) \Rightarrow c = \max_{f_x(X)}$ A eficiência é de 0.031.

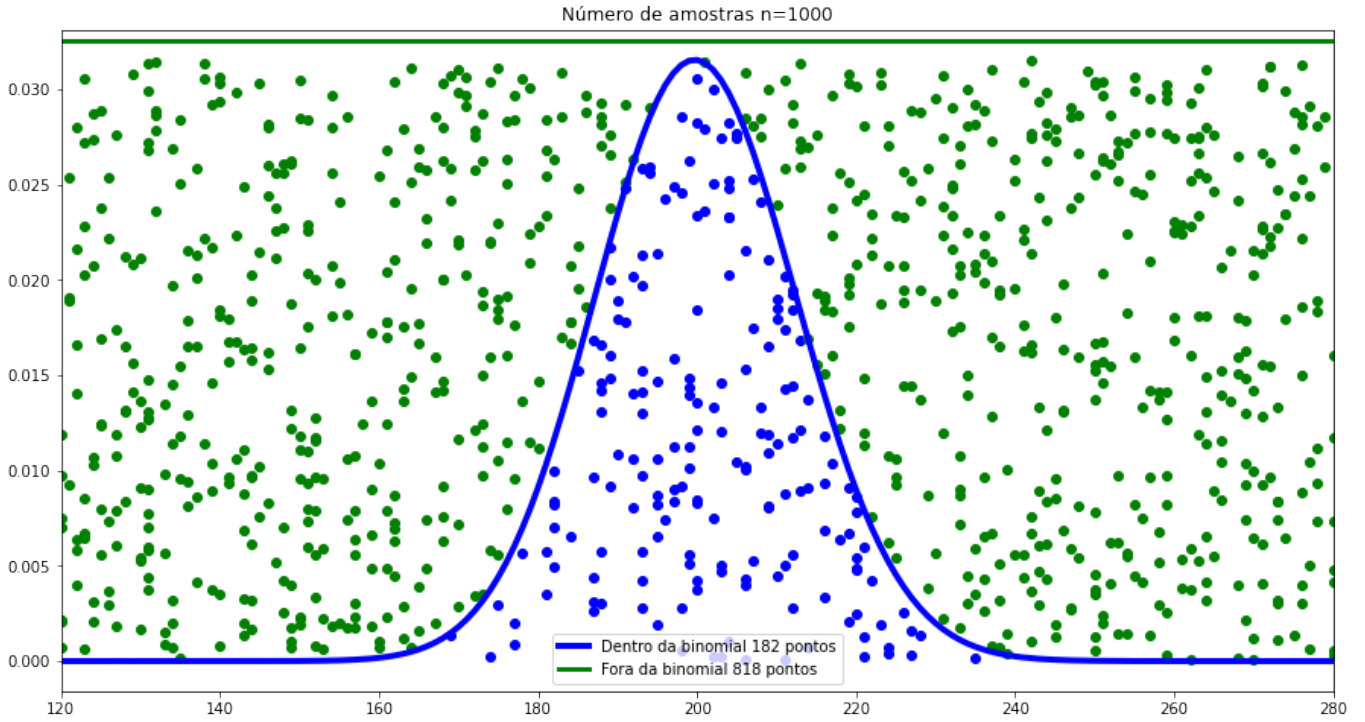


Figura 18: Usando $g(X) = 1$ e $c = 0.034$

Usaremos a distribuição normal para estimar a binomial. Assim, $X \sim \text{Bin}(n, p) \Rightarrow X \sim \text{norm}(np, \sqrt{np(1-p)})$, onde $g_X(x)$ é a função de densidade dessa distribuição.

$$f_X(x) = g_X(x) * c \Rightarrow c = \max\left\{\frac{f_X(x)}{g_X(x)}\right\} \Rightarrow c = 1.0004375953661317$$

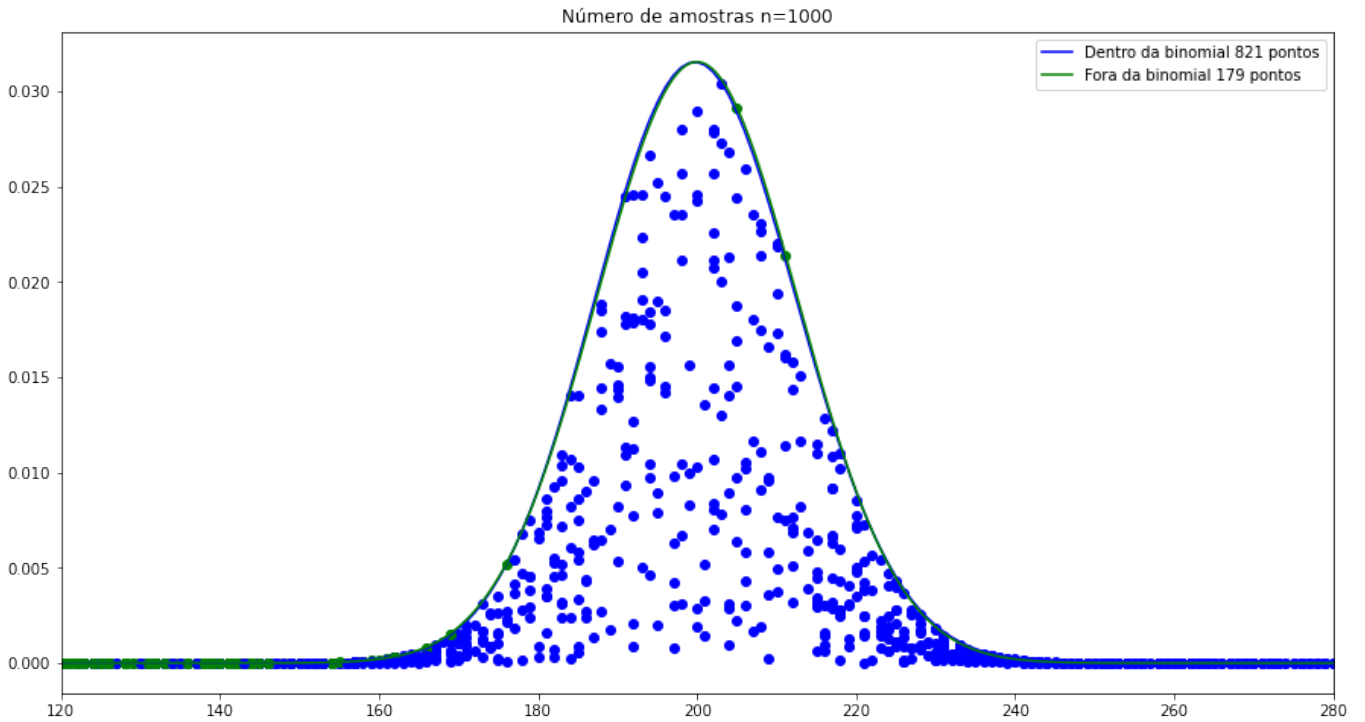


Figura 19: Usando $g \sim \text{norm}(np, \sqrt{np(1-p)})$ e $c = 1.004$

Logo, a eficiência é dada por $c = 1.0004$. Podemos concluir que essa aproximação é muito mais eficiente, uma vez que é bem próximo de 1. Já a outra abordagem, é pouco eficiente, como vimos no gráfico.

18 Integração de Monte Carlo e Importance Sampling

Questão feita no notebook: [Google Colab](#)

Queremos estimar $I = \int_0^{10} e^{-x^2} dx$ que é uma integral não-analítica. Pelo método tradicional de Monte Carlo, estimaremos fazendo:

$$I_{trad} = \int_0^{10} e^{-x^2} dx \approx \frac{10}{N} \sum_{i=1}^N e^{-x_i^2}$$

Onde x_i é uniformemente distribuído entre $[0,10]$ e N é o número de amostras. Queremos que $h(x)$ definida em $[0,10]$ seja tal que:

$$\frac{g(x)}{h(x)} \approx constant$$

Com a propriedade de que:

$$\int_0^{10} h(x) dx = 1 \quad h(x) > 0 \quad \forall x \in [0, 10]$$

A É uma constante de normalização. Assim, tomaremos $f(x) = 1/A$. Para A , vou considerar o valor que

$$\int_0^{10} A \cdot e^{-x} = 1$$

Portanto, $A = \frac{e^{10}}{e^{10}-1}$ e $h(x) = \frac{e^{10}}{e^{10}-1} \cdot e^{-x}$
Com isso, podemos reescrever:

$$I = \int_0^{10} g(x) dx = \int_0^{10} \frac{g(x)}{h(x)} \cdot h(x) dx$$

Agora a integral pode ser calculada gerando números aleatórios pela distribuição $h(x)$ e calculando $\frac{g(x_i)}{h(x_i)}$ nesses pontos. Ou seja

$$\frac{1}{N} \sum_{i=1}^N \frac{g(x_i)}{h(x_i)}$$

Onde, nesse caso, x_i são números aleatórios distribuídos em $h(x)$. Assim, como

$$H(x) = \int_0^x h(x) dx$$

Podemos definir $dH(x) = h(x) dx$. Fazendo uma mudança de variável, $u = H(x)$, onde u é a sequência de números aleatórios uniformemente distribuídos em $[0,1]$.

$$I = \int_0^{10} \frac{g(x)}{h(x)} dH(x) \Rightarrow \int_0^1 \frac{g(H^{-1}(u))}{h(H^{-1}(u))} du \approx \frac{1}{N} \sum_{i=1}^N \frac{g(H^{-1}(u_i))}{h(H^{-1}(u_i))}$$

Valor numérico da integral é: $I = 0.88622$.

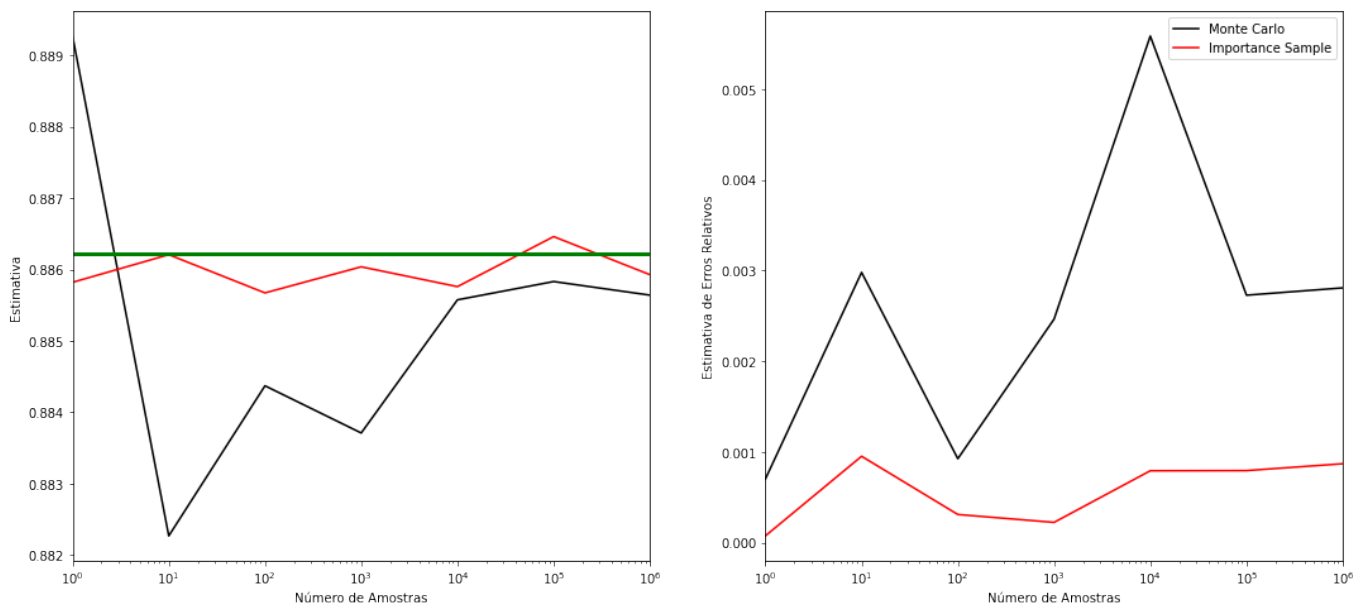


Figura 20: Curva verde $y=I$. Gráfico 1: Número de amostras x Estimativa. Gráfico 2: Número de Amostras x Erros Relativos

10 Cadeias de Markov

É uma **ferramenta de modelagem matemática**, como equações diferenciais. Representam dinâmicas de distemas com aleatoriedade, mais especificamente, uma sequência de variáveis aleatórias dependentes, como uma previsão de tempo.

- **Espaço de estados:** São todos os estados que o sistema pode se encontrar, que são os possíveis valores que a sequência de variáveis aleatórias pode assumir. Pode ser finito ou infinito, contável ou não.
- **Matriz de transição:** Todas as possíveis transições de um estado para o outro que o sistema pode fazer. Essas transições são aleatórias, a partir de uma certa probabilidade, dependendo apenas do estado atual.
- **Tempo discreto:** Faremos uma transição a cada passo de tempo.
- **Estado Inicial:** Estado onde o sistema começa. Pode ser determinística ou aleatória.

É representada por um grafo direcionado com pesos, onde os **vértices** são os estados do sistema, as **arestas** são as possíveis transições e os **pesos** a probabilidade de cada transição, onde a soma total dos pesos é exatamente igual a 1.

Por exemplo, o **modelo on e off**.

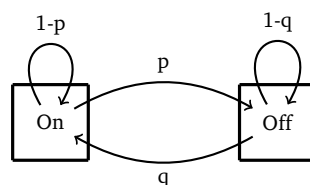


Figura 21: Modelo On-Off

Assim, a **Matriz de transição de estados**:

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix} \quad \leftarrow P(i,j) = \text{prob. de transição do estado } i \text{ para o estado } S.$$

10.1 Definição e Exemplos

Exemplo 1: Sistema Computacional. Os estados são os recursos utilizados pelo processo, como o HD, CPU, Cache, Memória. As transições são definidas empiricamente. Sabemos que a cadeia se comunica linearmente, isto é, $1 \leftrightarrow 2 \leftrightarrow 3 \leftrightarrow 4$. Assim, sua matriz de transição é:

$$P = \begin{bmatrix} p_{11} & p_{12} & 0 & 0 \\ p_{21} & p_{22} & p_{23} & 0 \\ 0 & p_{32} & p_{33} & p_{34} \\ 0 & 0 & p_{43} & p_{44} \end{bmatrix}$$

Note que, a matriz de transição vai ter dimensão no tamanho de estados e cada linha representa uma possível transição, logo a probabilidade de saída de cada linha tem que ser igual a 1. Ou seja, $\sum_{j=1}^4 p_{ij} = 1, \forall i$.

Definição formal:

- S : espaço de estados da cadeia de markov
- P : matriz de transição de estados
- X_t : Variável Aleatória que determina o valor do estado do sistema no instante de tempo $t = 0, 1, 2, \dots$
- Para cada t , X_t possui uma distribuição diferente.
- $P[X_t = s]$ é a probabilidade de no tempo t , você encontrar o sistema no estado $s \in S$.

Exemplo 2: On-Off

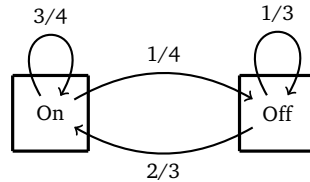


Figura 22: Modelo On-Off

- (Valor inicial) $X_0 = 1$ (On) $\rightarrow P[X_0 = 1] = 1$
- Qual a probabilidade de, no tempo 1, eu encontrar o estado 1 ou 2? $P[X_1 = 1] = 3/4$ e $P[X_1 = 2] = 1/4$.
- $P[X_2 = 1] = 35/48$ e $P[X_2 = 2] = 13/48$.

Neste último caso, se tomarmos como nota $P[X_2 = 1]$, isso é a probabilidade de, no estado anterior, ter ficado on (3/4) e e no estado seguinte continuar no on (3/4) ou probabilidade de ter, no estado anterior, ficado no off (1/4) e no estado seguinte voltar pro on (2/3). Ou seja:

$$\frac{3}{4} \cdot \frac{3}{4} + \frac{1}{4} \cdot \frac{2}{3} = \frac{35}{48}$$

10.2 Sem memória

Cadeia de Markov não possui memória. Ou seja, o próximo estado só depende do estado atual e não de como chegamos a ele. Considere uma trajetória de estados pelo sistema:

$$X_0 = s_0, X_1 = s_1, \dots, X_{t-1} = s_{t-1}, X_t = s_t$$

Onde s_i é um estado qualquer de S . Temos que:

$$P[X_{t+1} = s | X_0 = s_0, X_1 = s_1, \dots, X_t = s_t] = P[X_{t+1} = s | X_t = s_t]$$

Então A DISTRIBUIÇÃO X_{t+1} SÓ DEPENDE DE X_t E NÃO DA TRAJETÓRIA!

10.3 Distribuição no tempo

A cadeia começa em algum estado no tempo $t = 0$. A escolha é de quem constrói a cadeia. O estado inicial pode ser aleatório, e, dessa forma, temos que definir uma probabilidade de iniciar em cada estado.

Então, $P[X_0 = s] = \pi_s(0)$ a probabilidade de começar no estado s no tempo $t = 0$, donde $\sum_{s \in S} \pi_s(0) = 1$.

No exemplo do modelo on off, definimos $\pi(0) = (\pi_1(0), \pi_2(0)) = (1, 0)$. Onde $\pi(0)$ é o nosso **vetor de probabilidade inicial**. Formalizando

- $\pi(t)$: vetor com distribuição de X_t (estado da cadeia)
- $P[X_t = s] = \pi_s(t)$: probabilidade do sistema estar no estado s no tempo t .
- Em forma vetorial, $\pi(t) = (\pi_1(t), \pi_2(t), \dots, \pi_n(t)) = (P[X_t = 1], \dots, P[X_t = n])$

Voltando ao exemplo, formalizando:

- $\pi_1(t) = 3/4 * \pi_1(t-1) + 2/3 * \pi_2(t-1)$
- $\pi_2(t) = 1/4 * \pi_1(t-1) + 1/3 * \pi_2(t-1)$

Assim, usando lei da probabilidade total:

$$\pi_i(t) = P[X_t = i] = \sum_j P[X_t = i | X_{t-1} = j] P[X_{t-1} = j] = \sum_j P_{ji} \pi_j(t-1)$$

De forma matricial, $\pi(t) = \pi(t-1)P = \pi(0)P^t$. Onde P^t é a multiplicação de P t vezes.

Má notícia: multiplicação de matriz é $O(n^3)$.

Veremos algumas propriedades importantes.

10.4 Comunicação entre estados

Considere dois estados s_i e s_j em S . Dizemos que s_i se comunica com s_j se e somente se existe algum $t > 0$ tal que:

$$P[X_{t_0+t} = s_j | X_{t_0} = s_i] > 0$$

Ou seja, $s_i \rightarrow s_j$. Ou seja, existe uma trajetória (uma probabilidade não nula) de sair do estado s_i e chegar no estado s_j . Basta ter um caminho não nulo no grafo não direcionado de s_i para s_j . Se $s_i \rightarrow s_j$ e $s_j \rightarrow s_i$, então dizemos que s_i e s_j se intercomunicam, ou seja $s_i \leftrightarrow s_j$.

10.5 Irredutibilidade

Uma cadeia de markov é dita irredutível se para todo par de estados s_i e s_j em S temos que $s_i \leftrightarrow s_j$. Caso contrário, é dita irredutível. Considerando o grafo direcionado induzido pela matriz de transição P . Se há caminho direcionado entre qualquer par de vértices, então CM é irredutível. Esse grafo é fortemente conexo. Então se o grafo é fortemente conexo então é irredutível.

10.6 Periodicidade

Seja $A = \{a_1, a_2, \dots\}$ um conjunto de inteiros. $\gcd(A) =$ maior divisor comum dentre os inteiros de A .

Seja s_i um estado da CM, e A_i o conjunto dos comprimentos de caminho que iniciam e terminam em s_i . Ou seja a probabilidade de sair e voltar a s_i tem de ser $A_i = \{t : P^t[i, i] > 0\}$. Assim, o período de s_i é dado por $d(s_i) = \gcd(A_i)$

Diremos que uma cadeia de markov é **aperiódica**, isto é, um estado s_i é aperiódico se e somente se $d(s_i) = 1$. Caso contrário, s_i é dito periódico. Uma CM é dita aperiódica se todos os seus estados são aperiódicos. Caso contrário, a CM é dita periódica.

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ .6 & 0 & .4 & 0 \\ 0 & .9 & 0 & .1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

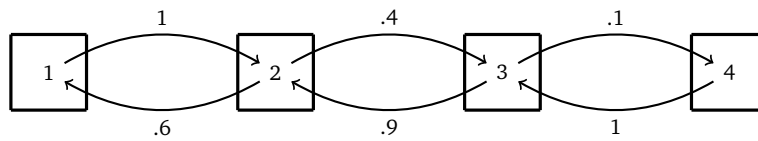


Figura 23: Modelo On-Off

Esta cadeia de Markov não é aperiódica! Pois o período de qualquer estado é 2. Em particular, não é possível voltar ao estado 1 em tempo ímpar. Todos os caminhos são pares.

Considere uma CM irredutível, tal que existe s_i em S tal que $p_{ii} > 0$. CM tem ao menos um estado com aresta em loop. Esta CM é aperiódica ou periódica? Aperiódica!

Lema: Em uma CM irredutível, todos os estados são aperiódicos ou todos são periódicos com o mesmo período.

11 Convergência da CM

Estamos interessados em valores grandes de t tal que

$$\lim_{t \rightarrow \infty} \pi(t) = \lim_{t \rightarrow \infty} \pi(0)P^t$$

Seja π um vetor de distribuição em uma CM com matriz de transição P . Dizemos que π é uma distribuição estacionária se e somente se:

1. $\pi_s \geq 0$ e $\sum_{s \in S} \pi_s = 1$, isto é, é um **vetor de probabilidade**.
2. $\pi P = \pi \Leftrightarrow \pi_i = \sum_j \pi_j P_{ji}$

Ou seja, ao multiplicar o vetor π por P , temos π de volta, "estacionando". Assim, se $\pi(0)$ é uma distribuição estacionária, então $\pi(1) = \pi(0)P = \pi(0)$.

11.1 Tempo de Chegada

Seja T_{ij} o tempo (em transições) necessário para sair de um estado s_i e chegar a outro estado s_j . Isto é, $T_{ij} = \min\{t | X_t = s_j \cap X_0 = s_i\}$. O número de transições T_{ij} é aleatório. Seja τ_{ij} o valor esperado de T_{ij} .

Teorema: Para qualquer CM irredutível e aperiódica e para qualquer dois estados s_i e s_j , temos o seguinte:

- $P[T_{ij} < \infty] = 1$.

- $\tau_{ij} = E[T_{ij}] \leq \infty$

Ou seja, não há chance de sair de s_i e não chegar em s_j e, além disso, o tempo médio de retorno é um valor finito!

Teorema: Para qualquer CM irreduzível e aperiódica, para qualquer estado s_i , temos a seguinte relação:

$$\pi_i = \frac{1}{\tau_{ij}}$$

Na média, a dinâmica visita s_i uma vez a cada τ_{ij} passos. Então conhecer um deles determina o outro.

Teorema: Para qualquer CM irreduzível e aperiódica e para qualquer condição inicial $\pi(0)$, temos que:

$$\lim_{t \rightarrow \infty} d_{TV}(\pi(t), \pi) = 0$$

Onde π é uma distribuição estacionária da CM e d_{TV} é a distância utilizando a métrica de *variação total*. A distância entre dois vetores quaisquer, α e β , é calculada da seguinte forma:

$$d_{TV}(\alpha, \beta) = \frac{1}{2} \sum_k |\alpha_k - \beta_k|$$

Ou seja, independentemente de quem você comece, a cadeia converge para a distribuição estacionária.

11.2 Encontrando a distribuição estacionária

Método 1: Método Iterativo.

$$\pi(t) = \pi(t-1)P = \pi(0)P^t$$

Fazer iteração até critério de convergência.

Método 2: Método Direto.

$$\pi = \pi P$$

Resolver o sistema de equações, adicionando a equação $\sum_i \pi_i = 1$.

Método 3: Monte Carlo.

Usar a própria cadeia para gerar amostras para estimar π_i ou estimar t_{ij} para todo s_i .

11.3 Reversibilidade

Uma CM é dita reversível para a distribuição de probabilidade π se e somente se

$$\pi_i P_{ij} = \pi_j P_{ji}$$

Daí, podemos concluir que **se π , nessas condições, existe então π é estacionário**. A ideia é que a massa de probabilidade fluindo de s_i para s_j seja igual a massa de probabilidade fluindo de s_j para s_i .

12 Autovalores e Autovetores

Dada uma matriz P , v é chamado de autovetor associado ao autovalor λ se

$$Pv = \lambda v$$

u é chamado de autovetor à esquerda se:

$$uP = \lambda u$$

Se u é autovetor a esquerda de P , então existe autovetor v de P' tal que:

$$P'v = \lambda v$$

Nesse caso os autovalores serão os mesmos.

Sabemos que π é uma distribuição estacionária com matriz P , tal que $\pi P = \pi$. Ou seja, π é o autovetor à esquerda de P associado ao autovalor $\lambda = 1$. Em outras palavras, existe um autovetor v tal que:

$$P'v = \pi$$

Para atender a condição que $\sum_{s \in S} \pi_s = 1$, normalizamos o autovetor para garantir que a soma seja 1.

Teorema: Se P é uma matriz de transição de probabilidade (estocástica), temos que os valores de λ em módulo são menores ou iguais a 1 para todo autovalor. Em especial, um dos autovalores **sempre** será igual a 1.

Uma matriz P pode ser escrita através de seus autovetores e autovalores:

$$P = QLQ^{-1}$$

Onde Q é a matriz com autovetores como colunas e L a matriz diagonal com L_i os autovalores associados.

P e L são semelhantes pois possuem os **mesmo autovalores**.

Algumas propriedades de matrizes semelhantes:

- $\det(P) = \det(L)$
- P é inversível se e somente se L é inversível
- P e L tem os mesmos autovalores com a mesma multiplicidade.
- P e L tem o mesmo polinômio característico
- P e L possuem o mesmo *traço*.

Além disso, essa relação é RST (reflexiva, simétrica e transitiva).

Como $\pi(t) = \pi(t-1)P = \pi(0)P^t$, podemos ter a relação:

$$\pi(t) = \pi(0) \cdot QL^tQ^{-1}$$

Para onde vai L^t com t crescente? Para $\lambda = 1$, fica no mesmo lugar. Todos os outros valores vão a zero, pois $|\lambda| < 1$.

Observação: Podemos escrever $\pi(0)$ como uma combinação linear dos autovetores!

Teorema da Convergência: Considere uma CM aperiódica e irreduzível com matriz de probabilidade P com distribuição estacionária π , existem constantes $\alpha \in (0, 1)$ e $C > 0$, tal que:

$$\max_{\pi(0)} d_{TV}(\pi(t), \pi) \leq C \cdot \alpha^t$$

Disso podemos tirar que a distribuição $\pi(t)$ converge exponencialmente rápido em t para distribuição estacionária π , independente de P e $\pi(0)$.

Note que: α é exatamente o segundo maior autovalor em módulo!

12.1 Tempo para Convergência

A ideia é: Quantos passos são necessários para determinar essa convergência? Ou seja, quantos passos t para garantir-mos que $\pi(t)$ está ϵ perto de π ? Na essência, queremos encontrar t tal que $d_{TV}(\pi(t), \pi) = \epsilon$

Com isso, podemos definir **tempo de mistura- ϵ** , τ_ϵ .

$$\tau_\epsilon = \min\{t | \max_{\pi(0)} d_{TV}(\pi(t), \pi) \leq \epsilon\}$$

Teorema: Para qualquer CM aperiódica, irreduzível, temos $\tau_\epsilon \leq \tau_{\frac{1}{4}} \log \frac{1}{\epsilon}$

Com isso, podemos dizer que o tempo de mistura fracamente depende de ϵ . Tomaremos como **constante**.

12.2 Spectral Gap

A convergência depende dos autovalores de P , de grosso modo, **o segundo autovalor (em módulo) domina a convergência**. O **Spectral Gap** (δ): distância entre dois maiores (em módulo) autovalores de P (maior é sempre igual a 1).

Quanto maior for delta, mais rápido é a convergência.

$$\delta = 1 - \max_{k>1} \{|\lambda_k|\}$$

12.3 Spectral Gap e Tempo de Mistura

Considere uma CM irreduzível aperiódica com *spectral gap* δ e $\pi_o = \min_i \pi_i$ (menor valor da distribuição estacionária). Temos a seguinte relação:

$$\underbrace{\left(\frac{1}{\delta} - 1\right) \log\left(\frac{1}{2\epsilon}\right)}_{\text{Limitante inferior}} \leq \tau_\epsilon \leq \underbrace{\frac{\log\left(\frac{1}{\pi_o \epsilon}\right)}{\delta}}_{\text{Limitante superior}}$$

Note que:

- Maior δ , menor τ_ϵ
- Maior π_o , menor τ_ϵ

13 Caminho amostral

Um **caminho amostral** é uma realização de uma sequência de tamanho k de variáveis aleatórias X_t , para $t = 0, 1, \dots, k$. A probabilidade de um caminho amostral é dado pelo vetor $\omega = (\omega_0, \omega_1, \omega_2, \dots, \omega_k)$. Em outras palavras:

$$P[\omega] = P[X_0 = \omega_0, X_1 = \omega_1, \dots, X_k = \omega_k] = \pi_{\omega_0}(0) \cdot T_{\omega_0 \omega_1} \cdot T_{\omega_1 \omega_2} \dots T_{\omega_{k-1} \omega_k}$$

Onde T é a matriz de transição do problema.

Note que, como os valores de T são sempre entre 0 e 1, quando $t \rightarrow \infty$, a sequência de probabilidades vai a zero!

Mas o que acontece com X_t no infinito? Para isso, podemos intuitivamente usar a média sobre os valores da sequência. Podemos analisar de duas formas:

$$S_k = \frac{1}{k} \sum_{t=0}^{k-1} X_t$$

A média amostral dos valores de estados observados.

$$f_k(s) = \frac{1}{k} \sum_{t=0}^{k-1} I(X_t = s)$$

A fração de vezes que um estado s é visitado, contabilizando indicadoras.

Com isso, podemos analisar intuitivamente a convergência dessas médias.

$$S_k = \frac{1}{k} \sum_{t=0}^{k-1} X_t \longrightarrow E_\pi[X] = \sum_s s \pi_s$$

$$f_k(s) = \frac{1}{k} \sum_{t=0}^{k-1} I(X_t = s) \longrightarrow \pi_s$$

14 Teorema Ergódico

Seja f uma função sobre o espaço de estados da CM que mapeia cada estado da CM em um valor real. Se a CM for irredutível e aperiódica, com distribuição estacionária π , temos:

$$P \left[\lim_{k \rightarrow \infty} \frac{1}{k} \sum_{t=0}^{k-1} f(X_t) = E_{\pi}[f(X)] \right] = 1$$

Ou seja, no infinito, essa média amostral converge para o valor esperado da função sobre os estados!

14.1 Estimando a distribuição estacionária

Podemos usar a CM para **gerar um caminho amostral** ω longo e calcular a **fração de visitas a cada estado** nesse caminho.

$$\bar{\pi}_s(k) = \frac{1}{k} \sum_{t=0}^{k-1} I(\omega_t = s)$$

O teorema ergódico garante a convergência de $\bar{\pi}_s(k)$ para π , com certo viés. Porém, a medida que k cresce, esse viés vai a zero.

14.2 Simular uma Cadeia de Markov

Como entrada, temos a matriz T de transição e uma distribuição inicial $\pi(0)$ inicial. Para simular uma CM, precisamos gerar um caminho amostral. Considere a cadeia a seguir:

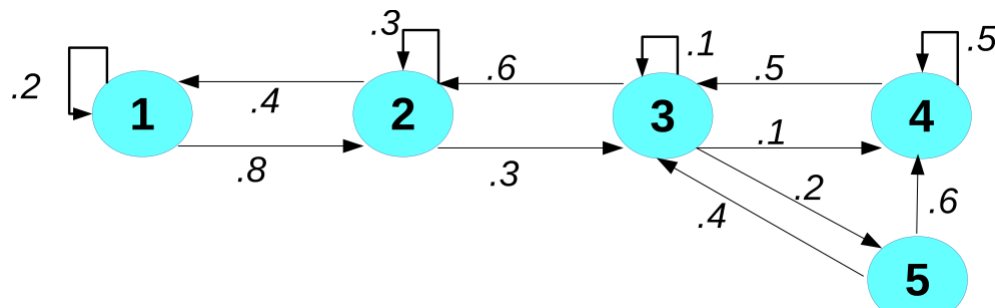


Figura 24: Cadeia de Markov

Passo 1: Representar a matriz T como um **vetor de adjacência**. Ou seja, tomamos apenas as entradas não-nulas da matriz T .

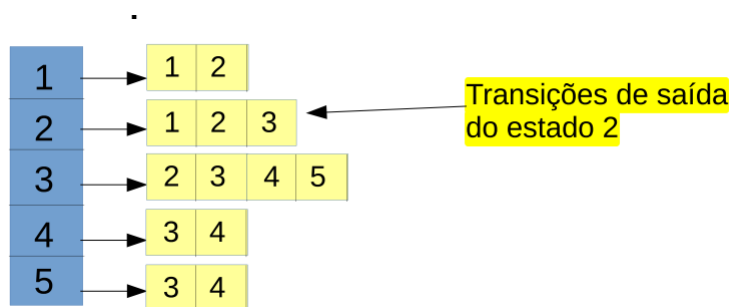


Figura 25: Vetor de Adjacência

Em azul, representamos os possíveis estados $j = 1, 2, 3, 4, 5$. Em amarelo, os vizinhos de cada estado j . Como cada linha da matriz T representa as ligações de um estado j , os valores representados em amarelo são os vizinhos que possuem probabilidade não nula.

Definiremos dois vetores:

- $y_j[i]$ é o vetor de adjacência para cada estado j . O índice i representa a posição do valor associado a chave j .
- $q_j[i]$ é o vetor das probabilidades acumuladas até o valor i . Ou seja, somamos as probabilidades desse vizinho i e dos vizinhos anteriores de i associados ao estado j .

Passo 2: Dividir o espaço de probabilidade para cada estado usando q_j . Cada fatia representa o estado destino y_j . Note que, basta tomarmos uma uniforme entre 0 e 1.

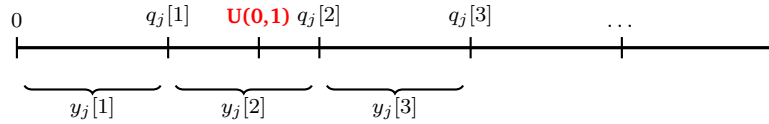


Figura 26: Divisão do intervalo

Nesse exemplo, $y_j[3]$ é o próximo estado.

14.3 Gerando amostras de X_t ?

O caminho amostral não interessa, apenas amostra no tempo t .

Ideia 1:

- Calcular $\pi(t)$ (distribuição no tempo t)
- Gerar amostras desta distribuição

Ideia 2:

- Simular CM até X_t , gerando caminho amostral
- Retornar a amostra gerada para X_t

14.4 Gerando Amostras Estacionárias

Iremos utilizar a mesma abordagem.

Ideia 1:

- Calcular π (distribuição estacionária)
- Gerar amostras desta distribuição

Ideia 2:

- Simular CM até X_t para t suficientemente grande (com tempo de mistura τ_ϵ)
- Retornar a amostra gerada para X_t

14.5 Simulando Cadeias de Markov Grandes

Quando a CM é impossível de representar na memória, podemos gerar apenas *possíveis próximos estados a partir do estado atual*. Ou seja, vamos determinar as transições de saída a partir do estado atual, escolhendo o próximo estado de acordo com uma regra de probabilidade.

14.6 Amostrando Espaços Complicados

Considere todos os percursos por n cidades de comprimento L ou menor. O espaço amostral S cresce exponencialmente com n e não é fácil construir amostra de forma iterativa.

Poderíamos utilizar a técnica *rejection sampling* para rejeitar amostras que não atendem essa restrição, rodando para um espaço "relaxado". Porém, o tempo médio para gerar uma amostra nesse sentido é de $\frac{1}{p}$, onde p é a fração de amostras que atendem a restrição. Se p for muito baixo, isto é, são poucos os aceites, o algoritmo se torna muito ineficiente.

15 Lista 3 - Cadeias de Markov

19 Passeios aleatórios enviesados

Considere um grafo não direcionado $G = (V, E)$ com peso nas arestas, tal que $w_{ij} > 0$ para toda aresta $(i, j) \in E$. Considere um andarilho aleatório que caminha por este grafo em tempo discreto, mas cujos passos são enviesados pelos pesos das arestas. Em particular, a probabilidade do andarilho ir do vértice i para o vértice j é dado por w_{ij}/W_i , onde $W_i = \sum_j w_{ij}$ (W_i é a soma dos pesos das arestas incidentes ao vértice $i \in V$). Temos assim um passeio aleatório enviesado linearmente pelos pesos das arestas.

1. Mostre que este passeio aleatório induz uma cadeia de Markov calculando a matriz de transição de probabilidade.
2. Determine a distribuição estacionária desta cadeia de Markov (dica: use o método da inspeção).
3. Determine se esta cadeia de Markov é reversível no tempo.

Resposta: Mostre que este passeio aleatório induz uma cadeia de Markov calculando a matriz de transição de probabilidade.

Uma forma geral de descrever essa matriz de transição é:

$$P = \begin{bmatrix} \frac{w_{11}}{W_1} & \frac{w_{12}}{W_1} & \cdots & \frac{w_{1n}}{W_1} \\ \frac{w_{21}}{W_2} & \frac{w_{22}}{W_2} & \cdots & \frac{w_{2n}}{W_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_{m1}}{W_m} & \frac{w_{m2}}{W_m} & \cdots & \frac{w_{mn}}{W_m} \end{bmatrix}$$

Para mostrarmos que este passeio aleatório induz uma cadeia de Markov devemos provar que:

$$\bullet \sum_{j=1}^n p_{ij} = 1$$

Ou seja, a probabilidade de saída tem que somar 1. Tomaremos uma linha qualquer da matriz P :

$$\frac{w_{m1}}{W_m} + \frac{w_{m2}}{W_m} + \cdots + \frac{w_{mn}}{W_m} = \frac{1}{W_m} \cdot \sum_{j=1}^n w_{mj} = 1$$

Assim, vemos que esse grafo com pesos enviesados nas arestas induz uma cadeia de Markov.

Resposta: Determine a distribuição estacionária desta cadeia de Markov (dica: use o método da inspeção).

Uma distribuição estacionária π possui a seguinte propriedade:

$$\pi = \pi P$$

Onde P é a matriz de transição. Pegando elemento a elemento, temos que:

Como $W_i = \sum_j w_{ij}$, isto é, W_i é a soma dos pesos das arestas incidentes ao vértice, $W_i = d(i)$. Assim:

$$\pi_i = \frac{d(i)}{\sum_k d(k)} = \frac{W_i}{\sum_k W_k}$$

$$\pi_j = \sum_i P_{ij} \cdot \pi_i = \sum_i \frac{w_{ij}}{W_i} \cdot \frac{W_i}{\sum_k W_k} = \frac{1}{\sum_k W_k} \sum_i w_{ij} = \frac{W_j}{\sum_k W_k}$$

$$\pi_j = \frac{d(j)}{\sum_k d(k)} = \frac{W_j}{\sum_k W_k}$$

Resposta: Determine se esta cadeia de Markov é reversível no tempo.

Para que uma cadeia de Markov seja reversível, tem que existir um π tal que:

$$\pi_i \cdot P_{ij} = \pi_j \cdot P_{ji}$$

Se existir esse π , esta é a distribuição estacionária!

$$\sum_j \frac{W_j}{\sum_k W_k} \cdot \frac{w_{ij}}{W_j} = \sum_i \frac{W_i}{\sum_k W_k} \cdot \frac{w_{ji}}{W_i}$$

$$\Leftrightarrow \frac{w_{ij}}{\sum_k W_k} = \frac{w_{ji}}{\sum_k W_k}$$

Isto só será verdade se $w_{ij} = w_{ji}$, o que é verdade pois G é um grafo não-direcionado!

20 Convergência de passeios aleatórios

Considere um passeio aleatório preguiçoso (com $p = \frac{1}{2}$) caminhando sobre um grafo com $n = 100$ vértices. Estamos interessados em entender a convergência da distribuição $\pi(t)$ em diferentes grafos. Assuma que o passeio sempre inicia sua caminhada no vértice 1, ou seja, $\pi_1(0) = 1$. Considere os seguintes grafos: grafos em anel, árvore binária cheia, grafo em reticulado com duas dimensões (grid 2D).

1. Para cada grafo, construa a matriz de transição de probabilidade (ou seja, determine P_{ij} para todo vértice i, j do grafo). Atenção com a numeração dos vértices!
2. Determine analiticamente a distribuição estacionária para cada grafo (ou seja, determine π_i para cada vértice i do grafo).
3. Para cada grafo, calcule numericamente a variação total entre $\pi(t)$ e a distribuição estacionária, para $t = 0, 1, \dots$. Trace um gráfico onde cada curva corresponde a um grafo (preferencialmente em escala logarítmica, com $t \in [1, 10^3]$).

Questão feita no notebook: [Google Colab](#)

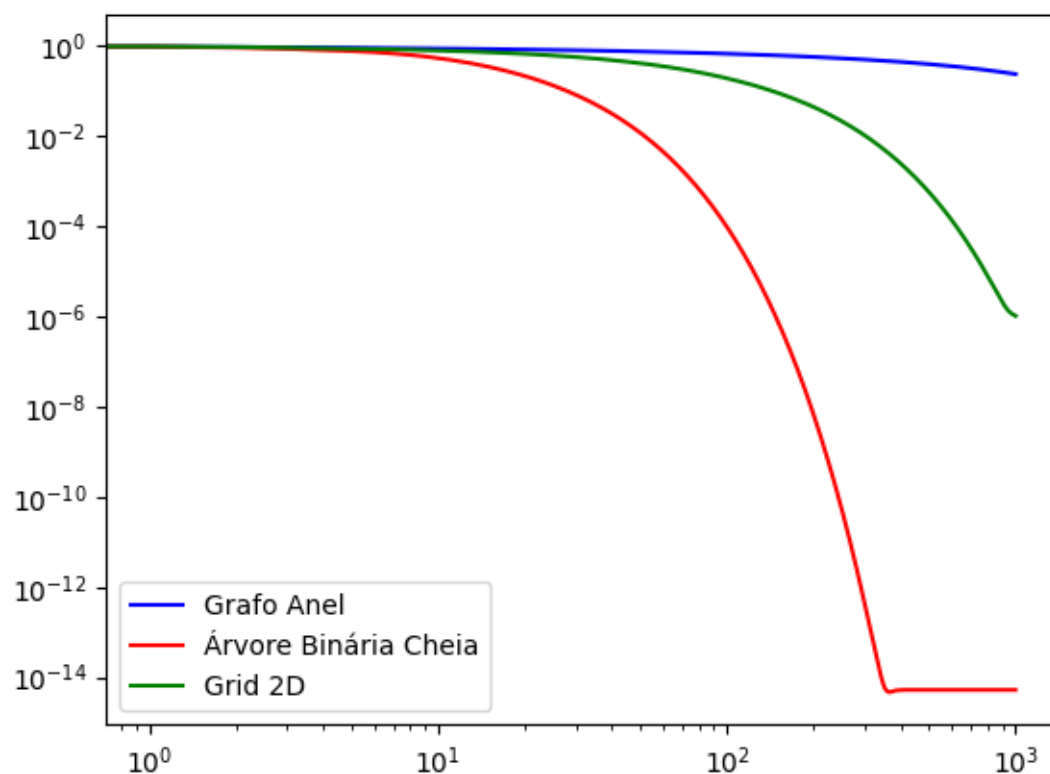


Figura 27: Variacão total

21 Tempo de Mistura

Considere um processo estocástico que inicia no estado 1 e a cada instante de tempo incrementa o valor do estado em uma unidade com probabilidade p ou retorna ao estado inicial com probabilidade $1 - p$. No estado n o processo não cresce mais, e se mantém neste estado com probabilidade p . Assuma que $n = 1$ e que $p \in \{0.25, 0.5, 0.75\}$.

1. Construa a cadeia de Markov deste processo mostrando a matriz de transição de probabilidade em função de p .
2. Determine numericamente o vão espectral da cadeia de Markov para cada valor de p
3. Determine numericamente a distribuição estacionária para cada valor de p , e indique o estado de menor probabilidade.
4. Utilizando os dados acima, determine um limitante inferior e superior para o tempo de mistura quando $\epsilon = 10^{-6}$ para cada valor de p .
5. O que você pode concluir sobre a influência de p no tempo de mistura?

Questão feita no notebook: [Google Colab](#)

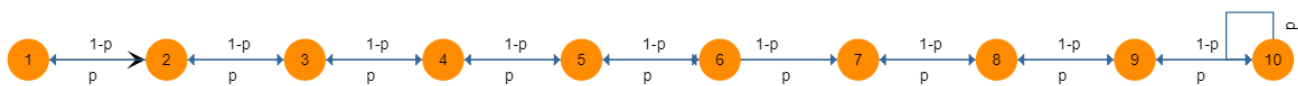


Figura 28: Construção do problema

Resposta: Construa a cadeia de Markov deste processo mostrando a matriz de transição de probabilidade em função de p .

$$T = \begin{bmatrix} (1-p) & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ (1-p) & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ (1-p) & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 & 0 \\ (1-p) & 0 & 0 & 0 & p & 0 & 0 & 0 & 0 & 0 \\ (1-p) & 0 & 0 & 0 & 0 & p & 0 & 0 & 0 & 0 \\ (1-p) & 0 & 0 & 0 & 0 & 0 & p & 0 & 0 & 0 \\ (1-p) & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & 0 \\ (1-p) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 \\ (1-p) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p \\ (1-p) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p \end{bmatrix}$$

Resposta: Determine numericamente o vão espectral da cadeia de Markov para cada valor de p

x	delta
0.25	0.998618
0.5	0.999579
0.75	0.99995

Resposta: Determine numericamente a distribuição estacionária para cada valor de p , e indique o estado de menor probabilidade.

Para calcular a distribuição estacionária, faremos:

$$\pi T = \pi \text{ e } \sum_{i=1}^{10} \pi = 1$$

π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}
0.75	0.1875	0.046875	0.0117188	0.0029297	0.0007324	0.0001831	4.58e-05	1.14e-05	3.8e-06

Figura 29: $p=0.25$

π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}
0.5	0.25	0.125	0.0625	0.03125	0.015625	0.0078125	0.0039062	0.0019531	0.0019531

Figura 30: $p=0.5$

π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}
0.25	0.1875	0.140625	0.105469	0.0791016	0.0593262	0.0444946	0.033371	0.0250282	0.0750847

Figura 31: $p=0.75$

$p=0.25$	$p=0.5$	$p=0.75$
3.8147e-06	0.00195312	0.0250282

Figura 32: O valor estacionário de menor probabilidade

Resposta: Utilizando os dados acima, determine um limitante inferior e superior para o tempo de mistura quando $\epsilon = 10^{-6}$ para cada valor de p .

$$\underbrace{\left(\frac{1}{\delta} - 1\right) \log\left(\frac{1}{2\epsilon}\right)}_{\text{Limitante inferior}} \leq \tau_\epsilon \leq \underbrace{\frac{\log\left(\frac{1}{\pi_0\epsilon}\right)}{\delta}}_{\text{Limitante superior}}$$

p	τ_{inf}	τ_{sup}
0.25	0.0181575	26.3285
0.5	0.00552093	20.0623
0.75	0.000661806	17.5041

Resposta: O que você pode concluir sobre a influência de p no tempo de mistura?

De acordo com o Teorema:

Considere uma CM aperiódica e irredutível com matriz de probabilidade P com distribuição estacionária π . Existem constantes α em $(0, 1)$ e $C > 0$, tal que:

$$\max_{\pi(0)} d_{TV}(\pi(t), \pi) \leq C\alpha^t$$

Distribuição transiente $\pi(t)$ converge exponencialmente rápido em t para distribuição estacionária π , independente de P e $\pi(0)$.

Nosso resultado mostra que independente de p , os limitantes foram bem próximos, apesar de apresentar um comportamento de queda. Na realidade, verifiquei que há um erro numérico no python ao calcular os autovalores e autovetores, que justificam essa distância entre eles.

22 Voltando à origem

Considere uma cadeia de Markov cujo espaço de estados é um látice de duas dimensões sobre os números naturais, ou seja, $S = \{(i, j) | i \geq 1, j \geq 1\}$. Cada estado pode transicionar para um dos seus vizinhos no látice. Entretanto, se afastar da origem (se movimentar para o norte ou para o leste) tem probabilidade $\frac{p}{2}$, e se aproximar da origem tem probabilidade $\frac{(1-p)}{2}$, onde p é um parâmetro do modelo (nas bordas, utilize self-loops). Assuma que $p \in \{0.25, 0.35, 0.45\}$.

1. Construa um simulador para essa cadeia de Markov.
2. Utilize o simulador para estimar a distribuição estacionária da origem (estado $(1,1)$), ou seja, $\pi_{1,1}$, para cada valor de p . Dica: Utilize os tempos de retorno!
3. Seja $d(t)$ o valor esperado da distância (de Manhattan) entre X_t (o espaço no tempo t) e a origem. Utilize o simulador para estimar $d(t)$ para $t \in \{10, 100, 1000\}$, para cada valor de p . O que você pode concluir?

Resposta: Questão feita no notebook: [Google Colab](#)

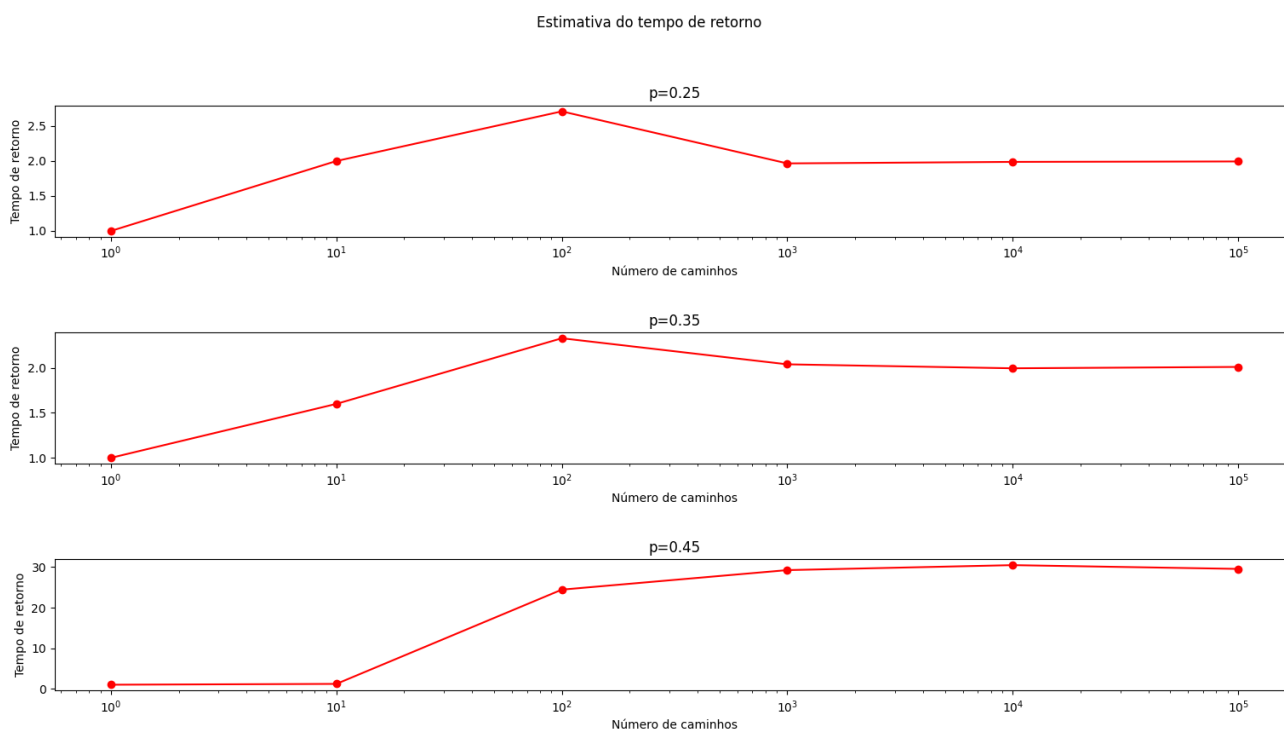


Figura 33:

Resposta: Utilize o simulador para estimar a distribuição estacionária da origem (estado $(1,1)$), ou seja, $\pi_{1,1}$, para cada valor de p . Dica: Utilize os tempos de retorno!

Figura 34: Estimativa para distância de manhattan para $n = 10000$

16 Monte Carlo e Cadeias de Markov

Construiremos e simularemos uma cadeia de Markov cujos os estados correspondem aos **elementos de S**. Além disso, tomaremos como dado uma matriz de transição P que tenha distribuição estacionária π uniforme. Para gerar uma amostra, simulamos um caminho amostral até dar τ_ϵ passos.

Boas práticas:

- Usar como uma cadeia base uma cadeia que possua fácil descrição dos estados vizinhos a partir do estado atual.
- Uma cadeia que possua poucas transições de saída de cada estado (esparsa, $\log|S|$)
- Baixo tempo de mistura, (idealmente *fast mixing*), $\tau_\epsilon \leq \log|S|$

16.1 Metropolis-Hastings

Entrada: Descrição do espaço amostral S , distribuição de probabilidade sobre os estados, π .

Saída: Uma amostra aleatória de S de acordo com π

- Construir uma CM irredutível onde cada estado corresponde a um elemento do espaço. (cadeia base)
- Transformar essa cadeia base em uma outra cadeia que é reversível e possui distribuição estacionária π .
- Simular caminho amostral longo o suficiente e retornar o estado final.

A ideia é modificar P para construir uma nova CM que seja reversível com distribuição estacionária π , que é uma entrada do problema. Na nova cadeia, não será aceito todas as transições da cadeia base, de forma a *induzir* π ficando no mesmo estado ao rejeitar.

Esse aceite é probabilístico e independente para cada transição. Diremos que $a(i, j)$ é a probabilidade de aceitar a transição $i \rightarrow j$.

16.1.1 Caso simétrico

Neste caso, a cadeia base tem **matriz de transição P simétrica**. Em outras palavras, $P_{ij} = P_{ji}$ para todo estado i, j .

A matriz de transição modificada fica da seguinte maneira:

$$P'_{ij} = \begin{cases} P_{ij}a(i, j), & \text{se } i \neq j \\ 1 - \sum_{k:k \neq i} P_{ik}a(i, k), & \text{se } i = j \end{cases}$$

Temos que escolhemos $a(i, j)$ de tal forma que P' tenha **distribuição estacionária dada por π** e a CM seja **reversível**, ou seja:

$$\pi_i P'_{ij} a(i, j) = \pi_j P'_{ji} a(j, i)$$

Daí, como $P_{ij} = P_{ji}$ por simetria:

$$\pi a(i, j) = \pi a(j, i) \Leftrightarrow a(i, j) = \frac{\pi_j}{\pi_i} a(j, i)$$

Para o par $a(i, j)$ e $a(j, i)$ temos infinitas soluções. Dessa forma, maximizamos a probabilidade de aceite:

$$a(i, j) = \begin{cases} 1, & \text{se } \pi_i \leq \pi_j \\ \frac{\pi_j}{\pi_i}, & \text{se } \pi_i > \pi_j \end{cases}$$

Logo,

e

$$a(i, j) = \min\{1, \frac{\pi_j}{\pi_i}\}$$

$$a(j, i) = \min\{1, \frac{\pi_i}{\pi_j}\}$$

16.1.2 Caso Geral

A matriz de transição P' da nova cadeia será:

$$P'_{ij} = \begin{cases} P_{ij}a(i, j), & \text{se } i \neq j \\ 1 - \sum_{k:k \neq i} P_{ik}a(i, k), & \text{se } i = j \end{cases}$$

Temos que escolhemos $a(i, j)$ de tal forma que P' tenha **distribuição estacionária dada por π** e a CM seja **reversível**, ou seja:

$$\pi_i P_{ij} a(i, j) = \pi_j P_{ji} a(j, i)$$

Para o par $a(i, j)$ e $a(j, i)$ temos infinitas soluções. Dessa forma, maximizamos a probabilidade de aceite:

$$a(i, j) = \begin{cases} 1, & \text{se } \pi_i P_{ij} \leq \pi_j P_{ji} \\ \frac{\pi_j P_{ji}}{\pi_i P_{ij}}, & \text{se } \pi_i P_{ij} > \pi_j P_{ji} \end{cases}$$

Logo,

$$a(i, j) = \min\{1, \frac{\pi_j P_{ji}}{\pi_i P_{ij}}\}$$

P' definida com essa probabilidade de aceite é chamada de **cadeia de Metropolis-Hastings**, dando origem ao algoritmo

Exemplos:

- Se a cadeia base for um passeio aleatório com probabilidade $\frac{1}{d_i}$ de transição, podemos definir uma outra CM em que $P'_{ij} = \min\{1, \frac{d_i}{d_j}\}$ com distribuição $\pi_v = \frac{1}{Z}$
-

Alguns problemas que são resolvidos com MCMC:

- Gerar amostras de um espaço S grande e complicado com distribuição π : Construir uma CM com espaço de estado S tal que π seja sua distribuição estacionária e gerar amostras através de caminhos amostrais de comprimento τ_ϵ
- Computar $E_\pi[f(X)]$ para alguma função f no espaço amostral S com distribuição π : Estimador de $E_\pi[f(X)]$ utilizando caminho amostral

16.2 Gibbs Sampling

Algoritmo de Markov Chain e Monte Carlo, também chamado de *Glauber Dynamics*. A ideia é construir uma CM com distribuição estacionária π sobre o espaço S (ambos são entrada para o problema). O elemento do espaço amostral é um vetor $V = (V_1, \dots, V_n)$, onde cada variável V_i assume valores do conjunto K .

O algoritmo se baseia na distribuição condicional de uma variável dado o valor de todas as outras, ou seja, é necessário ser conhecido a priori essas condicionais para induzir π .

A CM é construída considerando todos os possíveis valores para uma variável, dado o valor atual de todas as outras. A **probabilidade de transição é proporcional a probabilidade condicional, dividida por n** .

Prova de reversibilidade: slide 10 aula 14.

Algoritmo Gibbs Sampling: Dado um estado atual $X_t = (V_1(t), V_2(t), \dots, V_n(t))$

1. Escolher uma das variáveis uniformemente: $K \sim \text{Unif}(1, n)$.
2. Escolher valor para V_k dado o vetor X_t .
3. Usar a probabilidade condicional para V_k para escolher um valor a_k .
4. Transicionar para o estado X_{t+1} copiando valores de $V_i(t)$ e atualizando apenas V_k com a_k .
5. Repetir por τ_ϵ passos para que a distribuição X_t esteja ϵ próxima de π .

17 Otimização

Considere um espaço discreto S e uma função f que avalia cada elemento $f : S \rightarrow \mathbb{R}$.

Problema: Encontrar um elemento de S que minimiza (ou maximiza) f . Isto é:

$$S^* = \{e | e = \text{argmin}_{s \in S} f(s)\}$$

17.1 Exemplo do Caxeiro Viajante

Considere n pontos no plano onde $v_i = (x_i, y_i)$. Considere a distância euclidiana entre pares de pontos.

Problema: Encontrar o percurso com menor comprimento.

Definimos um grafo onde cada permutação viável é um vértice do grafo. Definiremos as ligações entre essas permutações através de uma regra, como por exemplo:

- Diferir em apenas um par de elementos. $((1,2,3,4) \rightarrow (1,2,4,3))$
- Usar estratégia de n -opt (busca-local) para inverter parte da permutação.

Cada vértice desse grafo artificial define um custo $f(s)$ dado pela soma dos custos de todas as distâncias consideradas na permutação.

Para evitar mínimos locais, a ideia é não ser tão guloso e usar a aleatoriedade para controlar ela. Ou seja, nos permite transicionar para algo pior para que seja possível encontrar pontos críticos globais.

A probabilidade π_s dessa cadeia será inversamente proporcional a $f(s)$, uma vez que queremos soluções com os menores valores de $f(s)$ e, fazendo dessa forma, esses vértices terão mais chances de serem visitados.

O número de transições de saída de um vértice i será o número de n de inversões e isso tem $n(n-1)/2$ possibilidades, não dependendo da permutação. Dessa forma, $P[(i, j)] = \frac{1}{n(n-1)/2}$ para qualquer i, j .

18 Simulated Annealing

De forma mais elegante, dado um parâmetro T não-negativo, o qual chamaremos de *temperatura*, podemos definir a probabilidade π_s como sendo a *distribuição de boltzman*:

$$\pi_s = \frac{e^{-\frac{f(s)}{T}}}{\sum_{s \in S} e^{-\frac{f(s)}{T}}}$$

Assim, quanto menor $f(s)$, maior π_s . Valores menores de $f(s)$ tem probabilidade exponencialmente maiores. Com relação ao parâmetro T , se T for muito pequeno podemos afirmar que a probabilidade de escolher um elemento ótimo vai a 1.

Assim, a ideia principal é amostrar essa cadeia de Markov utilizando π_s como distribuição estacionária e diminuir T para aumentar as chances de escolher o valor do ótimo.

Formalmente, cada valor de T induz uma CM e uma distribuição π_s . Podemos definir uma *agenda de resfriamento*, a partir do número de passos da CM com temperatura T_i , chamado N_i .

Algoritmo

1. Simular CM com T_1 por N_1 passos a partir de X_0
2. Trocar para T_2 e simular CM por N_2 passos a partir de X_{N_1} (o valor anterior).
3. E assim por diante.

Ideia: Resfriar simultaneamente com a geração de um caminho amostral, guardando sempre o estado de menor valor ao longo de toda simulação.

18.1 Estratégias de Resfriamento

Teorema: Se T_i decresce devagar o suficiente, então $P[X_t \text{ ser ótimo}] \rightarrow 1$ quando $t \rightarrow \text{infinito}$.

Problemas:

- Se resfriar muito rápido, ou seja, T for muito rápido a zero, a CM pode acabar ficando presa em um ótimo local.
- Devagar o suficiente depende do problema.
- Devagar o suficiente pode ser muito lento para ser usado na prática.

Uma ideia é definir a temperatura para cada transição da cadeia. Ou seja, $N_i = 1$ para todo i e T_i . Assim, podemos definir uma função $T(t)$ que representa a temperatura a ser usada no passo t . Algumas funções clássicas usadas:

$$T(t) = T_0 \beta^t$$

Exponencial

$$T(t) = T_0 - \beta t$$

Linear

$$T(t) = \frac{a}{\log(t + b)}$$

Logarítmico. Onde a e b são constantes. Se a for suficientemente grande, prova-se a convergência.

Podemos também gerar um resfriamento dinâmico e adaptativo, usando os valores das amostras para poder resfriar. Uma ideia seria usar a diferença $f(X_t) - f(X_{t-1})$ para definir a redução de T .

18.2 Voltando ao Caixeiro

Para definir a probabilidade da cadeia artificial, usaremos *Metropolis-Hastings*, com $T > 0$. Como a cadeia base é **simétrica**, a probabilidade de transição de s para s' será:

$$P_{s,s'} = \frac{2}{n(n-1)} \min\left\{e^{\frac{f(s)-f(s')}{T}}, 1\right\} \quad \text{se } s \neq s'$$

Se $f(s')$ for menor, então aceita com probabilidade 1. (Escolha uniforme entre os vizinhos s' melhores que s).

Se $f(s')$ for maior, aceita com probabilidade que é inversamente proporcional a diferença, diminuindo a chance de aceitar permutações que aumentam muito o comprimento do caminho.

19 Lista 4 - Cadeias de Markov e Algoritmos de Monte Carlo

23 Sequências binárias restritas

Considere uma sequência de dígitos binários (0s e 1s) de comprimento s . Uma sequência é dita válida se ela não possui 1s adjacentes. Considerando a distribuição uniforme, queremos determinar o valor esperado do número de 1s de uma sequência válida, denotado por μ_s .

1. Considerando $s = 4$, determine todas as sequências válidas e calcule μ_4 .
2. Construa uma cadeia de Markov sobre o conjunto de sequências válidas, deixando claro como funcionam as transições de estado. Argumente que a cadeia é irredutível.
3. Desenhe a cadeia de Markov para o caso de $s = 4$, mostrando todas as transições.
4. Mostre como aplicar Metropolis-Hasting para resolver o problema de estimar μ_s . Deixe claro as probabilidades de aceite, e o funcionamento do estimador.

i

Resposta: Considerando $s = 4$, determine todas as sequências válidas e calcule μ_4 .

As sequências válidas são:

$$S' = \{(0000), (0001), (0010), (0100), (1000), (1001), (1010), (0101)\}$$

Podemos definir uma função $f : \Omega \rightarrow [0, s]$ tal que $f(x)$ representa o número de 1's de uma configuração válida. Assim, o valor esperado pode ser calculado como:

$$E[f(x)] = \mu_s = \frac{1}{|S'|} \sum_{x \in S'} f(x)$$

Ou seja,

$$\mu_4 = \frac{1}{8} \sum_{x \in S'} f(x) = \frac{1}{8} \cdot 10 = \frac{10}{8}$$

i

Resposta: Construa uma cadeia de Markov sobre o conjunto de sequências válidas, deixando claro como funcionam as transições de estado. Argumente que a cadeia é irredutível.

Seja a seguinte cadeia de Markov:

- **Estados:** Cada estado será uma sequência de dígitos binários de comprimento s válida.
- **Transições:** Escolheremos um índice i uniformemente entre $[1, |S'|]$ e mudaremos de 0 para 1 ou de 1 para 0, transicionado entre configurações válidas.

A probabilidade de transicionar entre um estado i e um estado j é dado por $P(i, j) = \frac{1}{d_i}$, onde d_i é o grau de saída de i . Assim, essa cadeia induz um **passeio aleatório simples**.

Uma cadeia é dita *irredutível* se para todo estado s_i, s_j temos que $s_i \leftrightarrow s_j$. Assim, $s_i \leftrightarrow s_j$ se e somente se existe algum caminho de s_i para s_j no grafo direcionado. Ora, como há caminho direcionado entre qualquer par de vértices nessa cadeia, então a CM é irredutível.

Resposta: Desenhe a cadeia de Markov para o caso de $s = 4$, mostrando todas as transições.

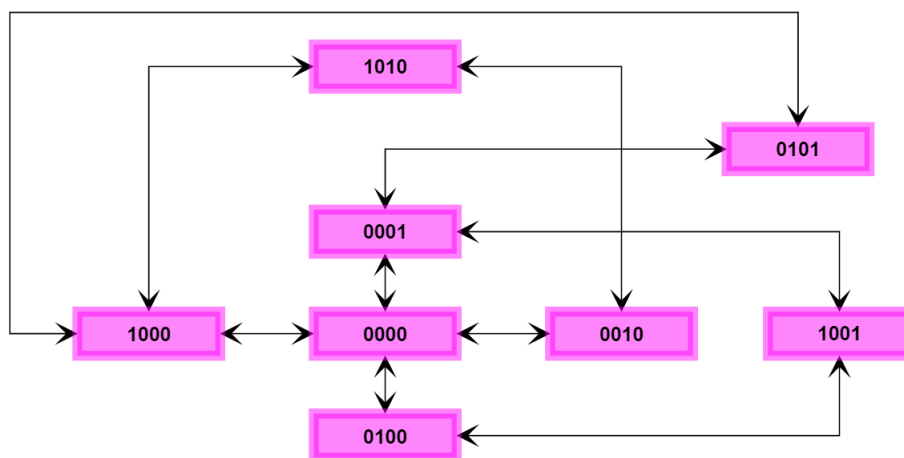


Figura 35: Cadeia com $s = 4$

Resposta: Mostre como aplicar Metropolis-Hasting para resolver o problema de estimar μ_s . Deixe claro as probabilidades de aceite, e o funcionamento do estimador.

Como o passeio aleatório não é uniforme, gostaríamos de criar uma cadeia de *Metropolis-Hastings* tal que $\pi = \frac{1}{|V|}$, onde $|V|$ é o número de vértices, induzindo uniformidade a cadeia.

- **Cadeia base:** Os estados são elementos de S' , onde fazemos um passeio aleatório simples com $P(i, j) = \frac{1}{d_i}$ e $P(j, i) = \frac{1}{d_j}$, onde d_i é o grau da configuração i .
- **Probabilidade de aceite:** $a(i, j) = \min\{1, \frac{\pi_j P(j, i)}{\pi_i P(i, j)}\} = \min\{1, \frac{d_i}{d_j}\}$

Logo, para $i \neq j$:

$$P'(i, j) = \frac{1}{d_i} \cdot \min\{1, \frac{d_i}{d_j}\}$$

E para o self-loop:

$$P'(i, i) = 1 - \sum_{k \neq i} P(i, k) \cdot \min\{1, \frac{d_i}{d_k}\}$$

Após isso, para estimar o valor esperado do número de 1's, geraremos um **caminho amostral** pela CM onde X_k é o estado da CM no passo k . Assim, um estimador do número médio de 1's após k amostras é:

$$\bar{f}(k) = \frac{1}{k} \sum_{i=1}^k f(X_k)$$

Pelo teorema ergódico, esse somatório converge para o valor esperando quando k tende ao infinito.

24 Amostras de Modelos de Mistura (Mixture Models)

Considere a seguinte função de probabilidade:

$$p(x) = \alpha p_B(x; n, p_1) + (1 - \alpha) p_B(x; n, p_2)$$

onde $p_B(x, n, p)$ é a probabilidade associada ao valor x da binomial com parâmetros n e p , e $\alpha \in [0, 1]$ é um peso. Repare que $p(x)$ é um modelo de mistura de duas binomiais com diferentes valores de p com pesos dado por α e $1 - \alpha$. Considere duas variáveis aleatórias X e K , representando o valor de $X \in [0, n]$ e a binomial utilizada $K \in [1, 2]$. Queremos gerar amostras de acordo com $p(x)$.

1. Determine as distribuições de probabilidade condicionais de $P(X|K)$ e $P(K|X)$. Dica: utilize regra de Bayes no segundo caso.
2. Determine a distribuição de probabilidade conjunta $P(X, K)$.
3. Utilize a técnica de Gibbs Sampling para gerar amostras de X . Mostre como construir a cadeia de Markov e determine a transição entre os estados.
4. Para $n = 2$, $p_1 = 0.2$, $p_2 = 0.8$ e $\alpha = 0.3$, desenhe a cadeia de Markov com todas as transições.
5. Descreva como usar a cadeia de Markov para gerar amostras.

1

Resposta: Determine as distribuições de probabilidade condicionais de $P(X | K)$ e $P(K | X)$. Dica: utilize regra de Bayes no segundo caso.

Podemos particionar nosso conjunto em dois, uma partição para cada K . Pela lei da probabilidade total, temos:

$$P(X) = \sum_{k=1}^2 P(X|K=k)P(K=k)$$

Substituindo,

$$P(X) = \alpha p_B(x; n, p_1) + (1 - \alpha) p_B(x; n, p_2) = P(K=1)P(X | K=1) + P(K=2)P(X | K=2)$$

Dessa equação, podemos tirar que:

- $P(K=1) = \alpha$
- $P(K=2) = 1 - \alpha$
- $P(X|K=1) = p_B(x; n, p_1)$
- $P(X|K=2) = p_B(x; n, p_2)$

Agora, usaremos a regra de bayes para determinar $P(K | X)$. Por definição:

$$P(K | X) = \frac{P(X=x | K=k)P(K=k)}{P(X=x)}$$

Substituindo para $k=1$:

$$P(K=1 | X) = \frac{p_B(x; n, p_1)\alpha}{\alpha p_B(x; n, p_1) + (1 - \alpha)p_B(x; n, p_2)}$$

Substituindo para $k=2$:

$$P(K=2 | X) = \frac{p_B(x; n, p_2)(1 - \alpha)}{\alpha p_B(x; n, p_1) + (1 - \alpha)p_B(x; n, p_2)}$$

Resposta: Determine a distribuição de probabilidade conjunta $P(X, K)$.

Da definição de probabilidade condicional,

$$P(X = x | K = k) = \frac{P(X = x \cap K = k)}{P(K = k)}$$

Podemos isolar $P(X = x \cap K = k)$, que é a distribuição conjunta. Assim:

$$P(X = x \cap K = k) = P(X | K = k) \cdot P(K = k)$$

Substituindo para $k = 1$:

$$P(X = x \cap K = 1) = P(X | K = 1) \cdot P(K = 1) = p_B(x; n, p_1) \cdot \alpha$$

Substituindo para $k = 2$:

$$P(X = x \cap K = 2) = P(X | K = 2) \cdot P(K = 2) = p_B(x; n, p_2) \cdot (1 - \alpha)$$

Resposta: Utilize a técnica de Gibbs Sampling para gerar amostras de X . Mostre como construir a cadeia de Markov e determine a transição entre os estados.

Um amostrador de Gibbs é um algoritmo de MCMC para obter uma sequência de observações que são aproximadas de uma distribuição de probabilidade multivariada especificada, quando a amostragem direta é difícil. Esta sequência pode ser usada para aproximar a distribuição conjunta, que é o nosso caso aqui. Assim, definiremos uma CM onde os estados serão os possíveis valores da distribuição conjunta. Para construir iterativamente as transições, fazer:

1. Escolher arbitrariamente um ponto inicial qualquer $S = (X_t, K_t)$, por exemplo $S_0 = (0, 1)$
2. Escolher uniformemente X em $[1, N]$ ou K em $[1, 2]$, utilizando a distribuição condicional dado valor atual da variável, o qual chamaremos de a_t .
3. Faça $S_{t+1} = (X_t, a_t)$ ou $S_{t+1} = (a_t, K_t)$.

Resposta: Para $n = 2$, $p_1 = 0.2$, $p_2 = 0.8$ e $\alpha = 0.3$, desenhe a cadeia de Markov com todas as transições.

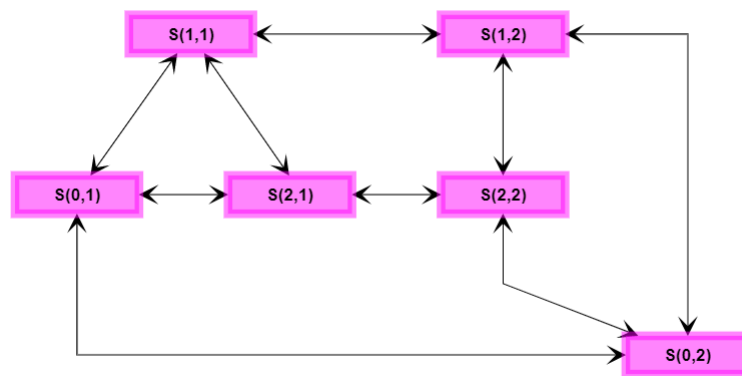
Como a distribuição conjunta é:

$$P(X, 1) = \alpha p_B(x; n, p_1) \text{ e } P(X, 2) = (1 - \alpha) p_B(x; n, p_2)$$

Podemos substituir os valores, e calcular a binomial, criando a seguinte tabela:

$P(X, K)$	$X = 0$	$X = 1$	$X = 2$	$P(K = k)$
$K = 1$	0.243	0.054	0.003	0.3
$K = 2$	0.028	0.224	0.448	0.7
$P(X = x)$	0.271	0.278	0.451	1

O desenho da cadeia de Markov para estas configurações é a seguinte:



Adicione a cada estado dois self-loops indicando a escolha quando a_t se mantém no mesmo valor, para cada variável.

Resposta: Descreva como usar a cadeia de Markov para gerar amostras.

Repetir τ_ϵ (tempo de mistura) vezes o algoritmo descrito no subitem 3, retornando S_t como sendo a amostra.

25 Amostrando triângulos

Considere um grafo conexo qualquer. Desejamos gerar amostras de triângulos deste grafo (cliques de tamanho 3), tal que todo triângulo tem igual probabilidade de ser amostrado. Ou seja, uma distribuição uniforme sobre o conjunto de triângulos do grafo.

1. Mostre como gerar amostras de forma direta, utilizando a distribuição uniforme (dica: pense em amostragem por rejeição). Determine a eficiência desse método.
2. Mostre como gerar amostras utilizando *Metropolis-Hastings*. Determine os estados da CM, as transições da cadeia base (que deve ser irredutível), e a probabilidade de aceite na cadeia modificada pelo método *Metropolis-Hastings*.
3. Intuitivamente, discuta quando a abordagem via *Metropolis-Hastings* é mais eficiente (do ponto de vista computacional) do que a abordagem via amostragem por rejeição.

i

Resposta: Mostre como gerar amostras de forma direta, utilizando a distribuição uniforme (dica: pense em amostragem por rejeição). Determine a eficiência desse método.

Considere um espaço amostral com todas as possíveis combinações de três vértices distintos de G . Assim, $\Omega = \{i, j, k \in V \mid i \neq j \neq k\}$. Definiremos uma função indicadora $f : \Omega \rightarrow \{0, 1\}$ onde $f(p) = 1$ se os vértices $p = (i, j, k)$ formam uma clique de tamanho 3. Caso contrário, $f(p) = 0$.

Para gerar amostras, podemos fazer uso da técnica de rejeição da seguinte maneira:

1. Gerar um elemento p de Ω uniformemente.
2. Se $f(p) = 0$, volte para 1). Se não, retorna p .

O tempo médio para gerar amostra será de $\frac{1}{q}$, onde q é a fração de amostras que atendem a restrição. Note que será, em geral, pouco eficiente, principalmente quando o grafo for mais esparso, uma vez que teria que testar diversas combinações de três vértices com reposição e poucos seriam as amostras que atendem a restrição. Se G for um grafo completo, intuitivamente a chance de gerar um triângulo é bem maior.

i

Resposta: Mostre como gerar amostras utilizando *Metropolis-Hastings*. Determine os estados da CM, as transições da cadeia base (que deve ser irredutível), e a probabilidade de aceite na cadeia modificada pelo método *Metropolis-Hastings*.

Criaremos uma cadeia base da seguinte maneira:

- **Estados:** Cada elemento de Ω define um vértice.
- **Transições:** Aresta entre s e s' se e somente se diferem de exatamente um vértice.

Assim, para um dado estado $s = (i, j, k)$, onde $i \neq |V|$, $j \neq |V|$ e $k \neq |V|$ seus vizinhos podem ser:

- $s' = (i + 1, j, k)$
- $s' = (i, j + 1, k)$
- $s' = (i, j, k + 1)$

E para $i \neq 1$, $j \neq 1$ e $k \neq 1$:

- $s' = (i - 1, j, k)$
- $s' = (i, j - 1, k)$

- $s' = (i, j, k - 1)$

Para as bordas (canto), definiremos *self-loops*. Essa CM induz um passeio aleatório simples com $P(i, j) = \frac{1}{d_i}$, onde d_i é o grau do vértice i na CM. Como há caminho direcionado entre qualquer par de vértices nessa cadeia então a CM é irredutível.

Note, no entanto, que a CM não será uniforme. Assim, criaremos uma CM via *Metropolis-Hastings* com distribuição estacionária sendo $\frac{1}{|V|}$, onde $|V|$ é o número de vértices do grafo G .

A probabilidade de aceite é calculada como: $a(i, j) = \min\{1, \frac{\pi_j P(j, i)}{\pi_i P(i, j)}\} = \min\{1, \frac{d_i}{d_j}\}$
Logo, para $i \neq j$:

$$P'(i, j) = \frac{1}{d_i} \cdot \min\{1, \frac{d_i}{d_j}\}$$

E para o self-loop:

$$P'(i, i) = 1 - \sum_{k \neq i} P(i, k) \cdot \min\{1, \frac{d_i}{d_k}\}$$

Após isso, para gerar uma amostra nessa cadeia, simularemos iterativamente seguindo os passos:

- Descobrir os vizinhos do vértice atual
- Descobrir o grau de cada vizinho do vértice atual
- Determinar as probabilidade de transição para cada vizinho
- Fazer escolha aleatória, atualizar o vértice atual e repetir.

Com isso, geraremos um **caminho amostral** pela CM como amostra e esse caminho terá tamanho igual ao primeiro triângulo encontrado na cadeia.

i

Resposta: Intuitivamente, discuta quando a abordagem via *Metropolis-Hastings* é mais eficiente (do ponto de vista computacional) do que a abordagem via amostragem por rejeição.

Utilizar *Metropolis-Hastings* será mais eficiente computacionalmente do que a técnica de rejeição quando o grafo for esparso, uma vez que as transições entre vizinhos na cadeia independem da quantidade de arestas do grafo original. Além disso, sempre conseguimos garantir transicionar entre vértices adjacentes diferentes, ajudando, intuitivamente, na convergência dos triângulos. Para transicionar, a complexidade é $O(1)$ e basta tomarmos um caminho amostral de, pelo menos, o tempo de mistura τ_ϵ para conseguirmos efetivamente garantir estar ϵ -próximo da estacionariedade.

26 Quebrando o código

Você encontrou uma mensagem que foi cifrada com o código da substituição (neste código, cada letra é mapeada em outra letra, de forma bijetora). Você deseja encontrar a chave do código para ler a mensagem. Repare que a chave é um mapeamento σ entre as letras, por exemplo $\sigma(a) = x, \sigma(b) = h, \sigma(c) = e, \dots$. Considere uma função $f : \Omega \rightarrow [0, 1]$ que avalia a capacidade de uma pessoa entender a mensagem cifrada dada um mapeamento $\sigma \in \Omega$. Repare que $f(\sigma) = 1$ significa que é possível entender por completo a mensagem decifrada com mapeamento σ , e $f(\sigma) = 0$ se o mapeamento não revela nenhuma informação sobre a mensagem. Utilize o método *Simulated Annealing* para resolver este problema! Mostre todos os passos necessários.

Resposta:

Criaremos uma cadeia de Markov base para este problema.

- **Estados:** Cada conjunto de letras de comprimento s é um estado.
- **Transições:** Transacionaremos entre mensagens não cifradas. Escolhemos índice i, j em $[1, |s|]$ com $i < j$ e invertemos a permutação atual entre os índices i, j .

Exemplo, se tomarmos $i = 3$ e $j = 5$:

savored \rightarrow saroved

O número de transições da cadeia base será exatamente o número de escolhas para i, j , com $i < j$. Já que $1 \leq i \leq |s|$ e $1 \leq j \leq |s|$, teremos $\frac{|s|(|s|-1)}{2}$ possibilidades de transições de saída. Assim:

$$P[(i, j)] = \frac{2}{|s|(|s|-1)} \quad \forall i, j$$

Note que é uma **cadeia simétrica**, uma vez que todo estado tem o mesmo grau de saída e mesmas probabilidades de transição. Agora, gostaríamos de definir uma cadeia de Markov com distribuição π_s de acordo com a *distribuição de Boltzmann* para maximizar $f(\sigma(s))$, que é a função de entendimento da mensagem. Cada estado dessa cadeia será uma mensagem com comprimento dado por s .

Definiremos essa nova cadeia via *Metropolis-Hastings*, com $T > 0$. A probabilidade de aceite é dada por:

$$\max(e^{\frac{f(\sigma(s)) - f(\sigma(s'))}{T}}, 1)$$

- Para um T fixo, maior $f(\sigma(s))$, maior π_s .
- Valores maiores de $f(\sigma(s))$ tem probabilidade exponencialmente maiores.

Assim, a escolha de (i, j) define a transição (ou seja, quem será s'):

$$P_{s, s'} = \frac{2}{|s|(|s|-1)} \max(e^{\frac{f(\sigma(s)) - f(\sigma(s'))}{T}}, 1) \quad s \neq s'$$

Em outras palavras:

- Se $f(\sigma(s'))$ for maior, então aceita com probabilidade 1 (escolha uniforme entre os vizinhos). Nesse caso, como estamos lidando com o entendimento da mensagem (decifrar todas as letras), gostaríamos de aceitar vizinhos com a maior quantidade de letras decifradas.
- Se $f(\sigma(s'))$ for menor ou igual, aceita com probabilidade diretamente proporcional a diferença entre as funções, dependendo da distribuição de boltzmann e da temperatura T .

Definindo uma agenda de resfriamento, conseguimos possivelmente decifrar a mensagem, ou seja, convergir para um estado (uma mensagem) em que $f(\sigma(s)) = 1$.