

Estatística Descritiva



cognitiza
Your Data & AI school



Prof. Altino Dantas

Probabilidade

- A teoria de probabilidade é um conjunto de regras, axiomas e teoremas matemáticos que tratam incertezas;
- O mundo real é repleto de incertezas ou imprecisões;
- Problemas do mundo real tendem a ser modelados com imprecisões;
- Modelo simples e incerto vs modelo preciso e complexo;
- Probabilidade ***frequentista*** vs **Bayesiana**.



Probabilidade

Dado um experimento aleatório E e S o espaço amostral, probabilidade de um evento A , que chamamos $P(A)$, é uma função definida em S que associa a cada evento um **número real**, satisfazendo os seguintes axiomas:

- $0 < P(A) < 1$
- $P(S) = 1$
- Se A e B forem eventos mutuamente exclusivos, ($A \cap B = \emptyset$), então:

$$P(A \cup B) = P(A) + P(B)$$

Chamamos de probabilidade de um evento A ($A \subset S$) o número real $P(A)$, tal que:

$$P(A) = \frac{\text{Número de Casos Favoráveis (A)}}{\text{Número Total de Casos}} = \frac{NCF(A)}{NTC}$$

Exemplo

Exemplo 5.2. De um grupo de duas mulheres (M) e três homens (H), uma pessoa será sorteada para presidir uma reunião. Queremos saber as probabilidades de o presidente ser do sexo masculino ou feminino. Observamos que: (i) só existem duas possibilidades: ou a pessoa sorteada é do sexo masculino (H) ou é do sexo feminino (M); (ii) supondo que o sorteio seja honesto e que cada pessoa tenha igual chance de ser sorteada, teremos o modelo probabilístico da Tabela 5.2 para o experimento.

Tabela 5.2: Modelo teórico para o Exemplo 5.2.

Sexo	M	H	Total
Frequência teórica	$2/5$	$3/5$	1



Exemplo

Exemplo 5.1. Queremos estudar as frequências de ocorrências das faces de um dado. Um procedimento a adotar seria lançar o dado certo número de vezes, n , e depois contar o número n_i de vezes em que ocorre a face i , $i = 1, 2, \dots, 6$. As proporções n_i/n determinam a distribuição de frequências do experimento realizado. Lançando o dado um número n' ($n' \neq n$) de vezes, teríamos outra distribuição de frequências, mas com um padrão que esperamos ser muito próximo do anterior.

Tabela 5.1: Modelo para lançamento de um dado.

Face	1	2	3	4	5	6	Total
Frequência teórica	1/6	1/6	1/6	1/6	1/6	1/6	1

Tipos de eventos

- **Evento certo:** $P(A) = 1$ - é aquela que ocorre em qualquer realização do experimento aleatório;
- **Evento impossível:** $P(A) = 0$ é aquela que não ocorre em nenhuma realização de um experimento aleatório. Se $E = \emptyset$, E é chamado evento impossível;
- **Evento complementar:** $P(A) + P(A^c) = 1$ - seja um evento A qualquer, o evento A' (chamado de complementar de A), tal que $A' = S - A$, ou seja, é um outro conjunto formado pelos elementos que pertencem a S e não pertencem a A
- **Eventos equiparáveis:** $P(x) = 1/n$, para x representando qualquer evento do espaço amostral. Quando se associa a cada ponto amostral a mesma probabilidade.



Exemplo

Retira-se uma carta de um baralho completo de 52 cartas. Qual a probabilidade de sair um rei ou uma carta de espadas?

Solução:

Considerando A o evento **rei** e B o evento **espadas**, deveríamos considerar:

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

$$P(A) = 4/52 = 0,076 \quad P(B) = 13/52 = 0,25 \quad P(A \cap B) = 1/52 = 0,019$$

$$P(A \cup B) = 0,076 + 0,25 - 0,019 = \mathbf{0,307}$$



Na Tabela 5.3 temos dados referentes a alunos matriculados em quatro cursos de uma universidade em dado ano

Tabela 5.3: Distribuição de alunos segundo o sexo e escolha de curso.

Curso \ Sexo	Homens (H)	Mulheres (F)	Total
Matemática Pura (M)	70	40	110
Matemática Aplicada (A)	15	15	30
Estatística (E)	10	20	30
Computação (C)	20	10	30
Total	115	85	200

As letras M, A, E, C, H e F representam eventos de seleção de alunos ao acaso, por exemplo, A representa a seleção de um aluno matriculado em **Matemática Aplicada**.

- Dessa maneira, vemos que $P(E) = 30/200$, ao passo que $P(H) = 115/200$;

Na Tabela 5.3 temos dados referentes a alunos matriculados em quatro cursos de uma universidade em dado ano

Tabela 5.3: Distribuição de alunos segundo o sexo e escolha de curso.

Curso \ Sexo	Homens (H)	Mulheres (F)	Total
Matemática Pura (M)	70	40	110
Matemática Aplicada (A)	15	15	30
Estatística (E)	10	20	30
Computação (C)	20	10	30
Total	115	85	200

Dados os eventos A e H , podemos considerar dois novos eventos:

- $A \cup H$, chamado a *reunião* de A e H , quando pelo menos um dos eventos ocorre;
- $A \cap H$, chamado a *intersecção* de A e H , quando A e H ocorrem simultaneamente.

É fácil ver que $P(A \cap H) = 15/200$, pois o aluno escolhido terá de estar, ao mesmo tempo, matriculado no curso de Matemática Aplicada e ser homem.

Na Tabela 5.3 temos dados referentes a alunos matriculados em quatro cursos de uma universidade em dado ano

Tabela 5.3: Distribuição de alunos segundo o sexo e escolha de curso.

Curso \ Sexo	Homens (H)	Mulheres (F)	Total
Matemática Pura (M)	70	40	110
Matemática Aplicada (A)	15	15	30
Estatística (E)	10	20	30
Computação (C)	20	10	30
Total	115	85	200

Vemos que $P(A) = 30/200$ e $P(H) = 115/200$; suponha que nosso cálculo para $P(A \cup H)$ fosse

$$P(A \cup H) = P(A) + P(H) = \frac{30}{200} + \frac{115}{200} = \frac{145}{200}.$$

Se assim o fizéssemos, estaríamos contando duas vezes os alunos que são homens e estão matriculados no curso de Matemática Aplicada, como destacado na Tabela 5.3. Portanto, a resposta correta é

$$P(A \cup H) = P(A) + P(H) - P(A \cap H) = \frac{30}{200} + \frac{115}{200} - \frac{15}{200} = \frac{130}{200}.$$

Na Tabela 5.3 temos dados referentes a alunos matriculados em quatro cursos de uma universidade em dado ano

Tabela 5.3: Distribuição de alunos segundo o sexo e escolha de curso.

Curso \ Sexo	Homens (H)	Mulheres (F)	Total
Matemática Pura (M)	70	40	110
Matemática Aplicada (A)	15	15	30
Estatística (E)	10	20	30
Computação (C)	20	10	30
Total	115	85	200

Calcule a probabilidade de $P(A \cup C)$ usando $P(A) + P(C) - P(A \cap C)$ e $P(A) + P(C)$, por que são iguais?

Probabilidade

- As operações de reunião, intersecção e complementação entre eventos possuem propriedades análogas àsquelas válidas para operações entre conjuntos.

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

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

$$(c) A \cap \emptyset = \emptyset, A \cap \Omega = A$$

$$(d) \emptyset^c = \Omega, \Omega^c = \emptyset$$

$$(e) A \cap A^c = \emptyset$$

$$(f) A \cup A^c = \Omega$$

$$(g) A \cup \emptyset = A, A \cup \Omega = \Omega$$

$$(h) A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$



Probabilidade condicional

- Nesse caso, a chance de ocorrências de um evento é alterada pela ocorrência de outro evento;
- Para dois eventos quaisquer A e B, sendo $P(B) > 0$, definimos a probabilidade condicional de A dado B, $P(A|B)$, como sendo:

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

- Para o exemplo mencionado, se B e A indicam, respectivamente, os eventos “aluno matriculado em Estatística” e “aluno é mulher”, então

$$P(A|B) = \frac{20/200}{30/200} = \frac{2}{3}$$

Probabilidade condicional

Tabela 5.3: Distribuição de alunos segundo o sexo e escolha de curso.

Curso \ Sexo	Sexo		Total
	Homens (H)	Mulheres (F)	
Matemática Pura (M)	70	40	110
Matemática Aplicada (A)	15	15	30
Estatística (E)	10	20	30
Computação (C)	20	10	30
Total	115	85	200

- Para o exemplo mencionado, se B e A indicam, respectivamente, os eventos “aluno matriculado em Estatística” e “aluno é mulher”, então

$$P(A|B) = \frac{20/200}{30/200} = \frac{2}{3}$$

Regra de Bayes

- Sejam $A_1, A_2, A_3, \dots, A_n$, n eventos mutuamente exclusivos tais que $A_1 \cup A_2 \cup \dots \cup A_n = S$;
- Sejam $P(A_i)$ as probabilidades conhecidas dos vários eventos e B um evento qualquer de S , tal que são conhecidas todas as probabilidades condicionais $P(B/A_i)$

$$P(A_i / B) = \frac{P(A_i).P(B / A_i)}{P(A_1).P(B / A_1) + P(A_2).P(B / A_2) + \dots + P(A_n).P(B / A_n)}$$

Exemplo

Três máquinas, A, B e C produzem respectivamente 40%, 50% e 10% do total de peças de uma fábrica. As porcentagens de peças defeituosas nas respectivas máquinas são 3%, 5% e 2%. Uma peça é sorteada ao acaso e verifica-se que é defeituosa. Qual a probabilidade de que a peça tenha vindo da máquina B? E da máquina A?

Exemplo

Três máquinas, A, B e C produzem respectivamente 40%, 50% e 10% do total de peças de uma fábrica. As porcentagens de peças defeituosas nas respectivas máquinas são 3%, 5% e 2%. Uma peça é sorteada ao acaso e verifica-se que é defeituosa. Qual a probabilidade de que a peça tenha vindo da máquina B? E da máquina A?

Solução:

$$P(A) = 0,4$$

$$P(B) = 0,5$$

$$P(C) = 0,10$$

$$P(D|A) = 0,03$$

$$P(D|B) = 0,05$$

$$P(D|C) = 0,02$$

$$P(B/D) = ?$$

$$P(B/D) = \frac{P(B) * P(D/B)}{P(A) * P(D/A) + P(B) * P(D/B) + P(C) * P(D/C)}$$

$$P(B/D) = \frac{0,5 * 0,05}{0,4 * 0,03 + 0,5 * 0,05 + 0,1 * 0,02}$$

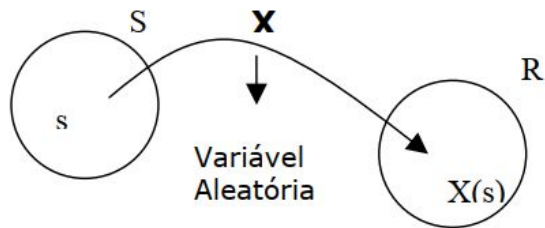
$$P(B/D) = \frac{0,025}{0,039} = 0,641 = 64,1\%$$

Variáveis aleatórias

- Muitos experimentos aleatórios produzem resultados não-numéricos;
- Variável aleatória é uma regra de associação de um valor numérico a cada ponto do espaço amostral;
- Assim, são variáveis numéricas associadas a modelos probabilísticos;
- Uma distribuição de probabilidades associa uma probabilidade a cada resultado numérico do experimento;
- As variáveis aleatórias podem ser discretas ou contínuas.

Variáveis aleatórias

Sejam E um experimento e S o espaço associado ao experimento. Uma função X , que associe a cada elemento $s \in S$ um número real $X(s)$, é denominada variável aleatória.



Exemplo:

E : lançamento de duas moedas;

X : nº de caras obtidas nas duas moedas;

$S = \{(C,C), (C,R), (R,C), (R,R)\}$

$X = 0 \rightarrow$ corresponde ao evento (r,r) com probabilidade $\frac{1}{4}$

$X = 1 \rightarrow$ corresponde ao evento $(r,c), (c,r)$ com probabilidade $\frac{2}{4}$

$X = 2 \rightarrow$ corresponde ao evento (c,c) com probabilidade $\frac{1}{4}$.

Empregamos a termo **variável aleatória** para descrever o valor que corresponde ao resultado de determinado experimento.

Distribuição de probabilidade

- Uma vez definida a variável aleatória, existe interesse no cálculo dos valores das probabilidades correspondentes;
- DP é o conjunto das variáveis e das probabilidades correspondentes;

$$\{(x_i, p(x_i)), i=1, 2, \dots, n\}$$

- Essa atribuição de valores é dada por uma função de densidade de probabilidade (FDP);

$$P(X=x_i) = P(A_i), i=1, 2, \dots, n$$

Distribuição de probabilidade - Discreta

- Histogramas - frequência de valores da variável aleatória no conjunto de dados;
- Qual procedimento para definir a probabilidade de um aluno da turma, tomado ao acaso, ter mais de 22 anos?
- Pelo definição básica de probabilidade, o número de casos favoráveis pelo número de casos totais;
- Distribuição de probabilidades.

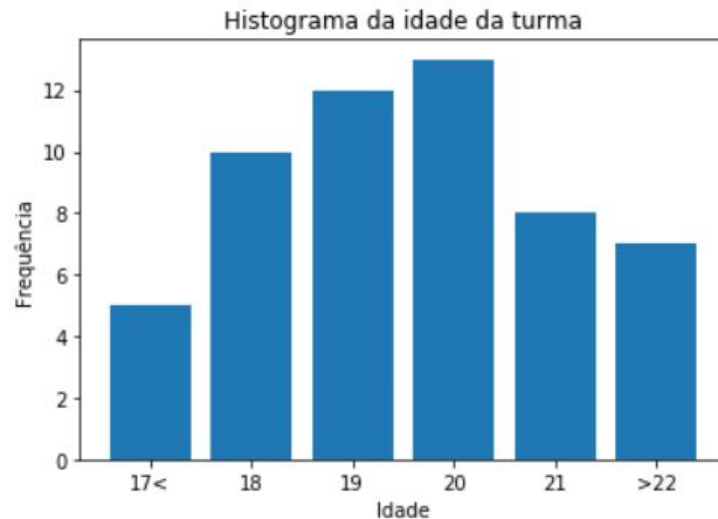


Figura 1: exemplo de um histograma de idades de uma turma

Função de Distribuição de Probabilidade - FDP

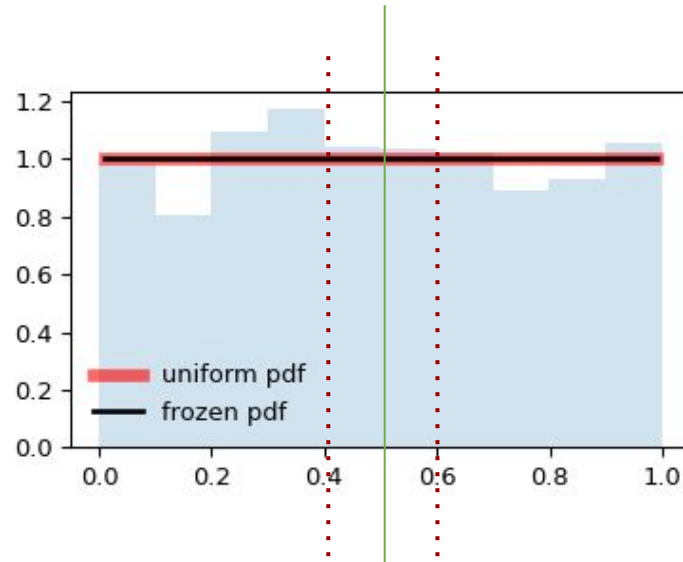
- O domínio de P deve ser o conjunto de todos possíveis estados de x .
- $\forall x \in \mathbf{x}, 0 \leq P(x) \leq 1$, além disso, um evento impossível deve ter probabilidade 0 e um evento garantido 1.
- $\sum_{x \in \mathbf{x}} P(x) = 1$. Essa é a propriedade de normalização. Nós fizemos isso no exemplo do histograma ao dividir as frequências pelo somatório total.

$$P(x = x_i) = \frac{1}{k} \quad , \text{para } k \text{ estados.}$$

distribuição uniforme.

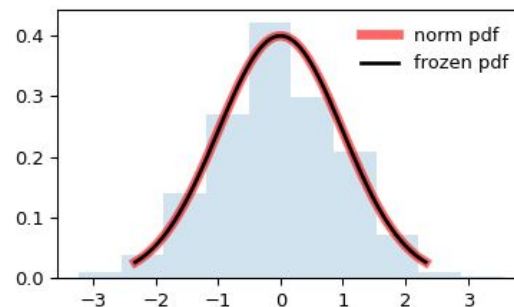
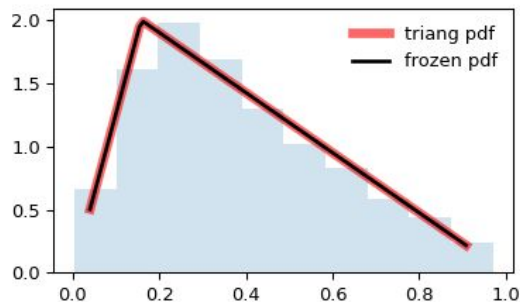
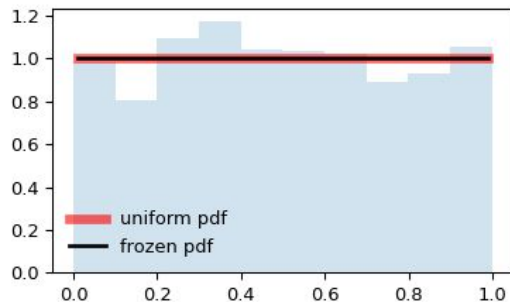
Distribuição de probabilidade - Contínua

- Princípio básico da probabilidade não se aplica ao espaço contínuo;
- Probabilidade de ocorrência em um ponto é zero. $P(x=0,5) = 0$



- Probabilidade $P(x < 0,5) = \frac{1}{2}$
-

Exemplos de distribuições conhecidas



<https://docs.scipy.org/doc/scipy/reference/stats.html>




cognitiza
Your Data & AI school

Inferência da distribuição

- Observação de histograma: analisar as distribuição de frequência para tentar identificar uma distribuição conhecida;
- **Inferência paramétrica:** escolher* uma das distribuições conhecidas a partir do aspecto do histograma e gerar a Função de Densidade de Probabilidade com os parâmetros da amostra.
- Inferência não paramétrica: por exemplo, estimar a densidade de *Kernel*.

*Além de comparação visual, pode-se usar um teste estatístico para checar a proximidade da amostra real e de uma gerada pela distribuição inferida. Não deveria haver diferença significativa entre elas em caso da distribuição ser adequada. _____



Your Data & AI school

Função de Distribuição Acumulada (FDA)

- Calcula a probabilidade acumulada para um determinado valor de x ;
- Utilizada para calcular probabilidade de intervalos em vez de pontos fixos;
- Matematicamente, esta distribuição é um valor infinitesimal e é dada pela integral do intervalo desejado;
- O método **`cdf()`** da biblioteca **`scipy.stats`** é implementação de FDA de cada distribuição suportada na biblioteca.

Inferência de distribuição em Python

```
# Obtenção dos dados da amostra - estes são aleatórios por comodidade
```

```
sample = normal(loc=50, scale=5, size=1000)
```

```
# calcula-se os parâmetros da amostra (média e desvio padrão)
```

```
sample_mean = mean(sample)
```

```
sample_std = std(sample)
```

```
# define a distribuição que será testada
```

```
dist = norm(sample_mean, sample_std)
```

```
from matplotlib import pyplot
```

```
from numpy.random import normal
```

```
from numpy import mean
```

```
from numpy import std
```

```
from scipy.stats import norm
```

```
# captura a densidade de probabilidades para os limites da amostra
```

```
values = [value for value in range(30, 70)]
```

```
probabilities = [dist.pdf(value) for value in values]
```

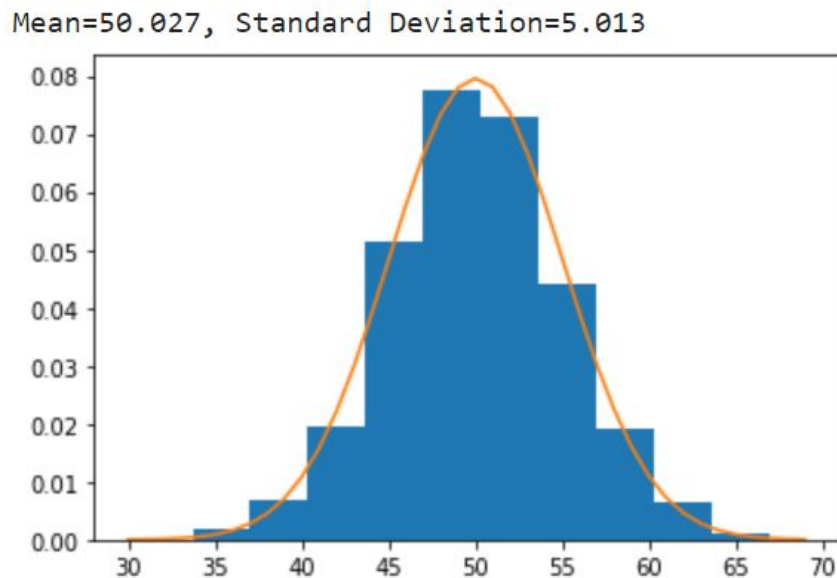
```
# gráfico do histograma e função de densidade de probabilidade - pdf
```

```
pyplot.hist(sample, bins=10, density=True)
```

```
pyplot.plot(values, probabilities)
```

```
pyplot.show()
```

Inferência de distribuição em Python



Saída do código anterior

Inferência de distribuição em Python

- Uma vez inferida a distribuição, o método pode ser usado para calcular a probabilidade acumulada;
- Chance de um valor da população/amostra ser menor do que 60:

```
norm.cdf(60, loc=sample_mean, scale=sample_std)
```

- Chance de um valor da população/amostrar estar entre 35 e 45:

```
a = norm.cdf(45, loc=sample_mean, scale=sample_std)
```

```
b = norm.cdf(35, loc=sample_mean, scale=sample_std)
```

```
a - b
```

Exercício

1. Qual a probabilidade de uma entrega ocorrer em 15 minutos?
2. Qual a chance de uma entrega atrasar se o pedido ocorrer na sexta-feira?
3. Qual a probabilidade do produto D ser entregue em 30 minutos?
4. Qual a chance da distância da entrega ser de até 3km?
5. Qual a chance da distância da entrega ser entre 5 e 7km?

Obrigado.