MBA: Data Science e Analytics

Pedro P. Bittencourt

2024–2025

Sumário

Capítulo 1

Fundamentos de estatística

Observação	preço (R$)	O	p	O	p	O	p
1	189,00	26	215,00	51	199,00	76	185,00
2	195,00	27	149,00	52	209,00	77	179,00
3	199,00	28	189,00	53	229,00	78	169,00
4	189,00	29	169,00	54	199,00	79	179,00
5	197,00	30	179,00	55	195,00	80	189,00
6	189,00	31	159,00	56	199,00	81	199,00
7	199,00	32	199,00	57	179,00	82	209,00
8	202,00	33	195,00	58	169,00	83	169,00
9	199,00	34	189,00	59	189,00	84	159,00
10	209,00	35	209,00	60	205,00	85	179,00
11	189,00	36	196,00	61	199,00	86	185,00
12	179,00	37	189,00	62	189,00	87	189,00
13	175,00	38	165,00	63	189,00	88	179,00
14	199,00	39	170,00	64	199,00	89	199,00
15	205,00	40	179,00	65	179,00	90	199,00
16	219,00	41	170,00	66	189,00	91	189,00
17	229,00	42	175,00	67	239,00	92	169,00
18	205,00	43	169,00	68	215,00	93	159,00
19	190,00	44	189,00	69	199,00	94	169,00
20	179,00	45	195,00	70	179,00	95	209,00
21	199,00	46	199,00	71	195,00	96	189,00
22	189,00	47	199,00	72	199,00	97	179,00
23	183,00	48	199,00	73	209,00	98	189,00
24	199,00	49	189,00	74	205,00	99	199,00
25	206,00	50	182,00	75	179,00	100	195,00

Tabela 1.1.1: Preços de um produto coletado em 100 pontos de venda

1.1 Estatística descritiva

Aula 1
Sex 11 out 2023

Estrutura tabular de dados é o mais comum de se utilizar. Dimensão linha e dimensão coluna. Esse cruzamento resulta numa estrutura tabular — igual a de uma planilha de excel. As unidades de observação estão nas linhas e os atributos (variáveis) a serem medidos ou classificados estão nas colunas.

O ponto de partida de uma análise de dados é compreender as características dos dados que estão sendo analisados. Isso implica em entender os tipos de variáveis envolvidas. Essa distinção é fundamental porque vai determinar a escolha da técnica a ser utilizada.

Essa aula tem como foco a **estatística descritiva**, dada por meio de *medidas resumo*. O objetivo principal dessas medidas é representar o comportamento da variável em estudo por meio de seus valores centrais e não centrais e suas dispersões ou formas de distribuição dos seus valores em torno da média. Estudaremos medidas de posição ou localização (medidas de tendência central e medidas separatrizes), medidas de dispersão ou variabilidade e medidas de forma, como assimetria e curtose (FÁVERO; BELFIORE, 2025, p. 112).

Ao longo da Seção 1.1, estudaremos o seguinte caso:

Estudo de caso 1.1. Realizou-se uma coleta de preços de um determinado produto em 100 diferentes locais de venda, ao longo de um certo intervalo de tempo. A Tabela 1.1.1 contém a tabulação desses dados. Na Tabela 1.1.2, temos as mesmas informações, ordenadas de forma crescente. Determine as estatísticas descritivas para a variável preço e construa a tabela resumo.

Posição	preço (R$)	P	p	P	p	P	p
1	149,00	26	179,00	51	189,00	76	199,00
2	159,00	27	179,00	52	189,00	77	199,00
3	159,00	28	179,00	53	190,00	78	199,00
4	159,00	29	179,00	54	195,00	79	199,00
5	165,00	30	182,00	55	195,00	80	199,00
6	169,00	31	183,00	56	195,00	81	199,00
7	169,00	32	185,00	57	195,00	82	199,00
8	169,00	33	185,00	58	195,00	83	202,00
9	169,00	34	189,00	59	195,00	84	205,00
10	169,00	35	189,00	60	196,00	85	205,00
11	169,00	36	189,00	61	197,00	86	205,00
12	169,00	37	189,00	62	199,00	87	205,00
13	170,00	38	189,00	63	199,00	88	206,00
14	170,00	39	189,00	64	199,00	89	209,00
15	175,00	40	189,00	65	199,00	90	209,00
16	175,00	41	189,00	66	199,00	91	209,00
17	179,00	42	189,00	67	199,00	92	209,00
18	179,00	43	189,00	68	199,00	93	209,00
19	179,00	44	189,00	69	199,00	94	209,00
20	179,00	45	189,00	70	199,00	95	215,00
21	179,00	46	189,00	71	199,00	96	215,00
22	179,00	47	189,00	72	199,00	97	219,00
23	179,00	48	189,00	73	199,00	98	229,00
24	179,00	49	189,00	74	199,00	99	229,00
25	179,00	50	189,00	75	199,00	100	239,00

Tabela 1.1.2: Valores de Tabela 1.1.1, ordenados de modo crescente

1.1.1 Medidas de posição

1.1.1.1 Média

A média aritmética simples $\overline{x}$ dos x elementos de um conjunto é dada pela expressão

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{1.1.1}$$

em que n é a quantidade de elementos do conjunto e x_i representa cada um desses valores. No Estudo de caso 1.1, tem-se que a média das observações é dada por

$$\begin{aligned} \overline{x} &= \frac{189 + 195 + 199 + \ldots + 189 + 199 + 195}{100} \\ &= \frac{19077}{100} \\ &= 190.77 \end{aligned}$$

1.1.1.2 Média ponderada

No caso da média aritmética simples, todas as ocorrências têm o mesmo peso, isto é, a mesma frequência (ou importância). Se atribuirmos pesos p_i para cada valor x_i da variável X, podemos calcular a média aritmética ponderada $(\overline{x})$, dada por:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot p_i}{\sum_{i=1}^{n} p_i} \tag{1.1.2}$$

Se o peso estiver expresso de forma relativa ao total (f_i), podemos reescrever a Equação (1.1.2) como:

$$\overline{x} = \sum_{i=1}^{n} x_i \cdot f_i \tag{1.1.3}$$

Todo mundo que passou pela escola já lidou com isso. Imaginemos que, no ensino médio, a nota bimestral era composta por três avaliações, todas valendo entre 0 e 10, mas com diferentes importâncias:

- mensal, de peso 2;
- bimestral, de peso 2;
- contínua, de peso 1.

Um estudante fictício, Pedro, obteve, respectivamente, as seguintes notas: 8, 9 e 7. Utilizando a Equação (1.1.2), temos que sua nota bimestral $\overline{x}$ é dada por:

$$\overline{x} = \frac{2 \cdot 8 + 2 \cdot 9 + 1 \cdot 7}{2 + 2 + 1} = \frac{41}{5} = 8.2$$

Isso é equivalente a dizer que ele teve duas notas bimestrais, duas mensais e uma contínua, totalizando cinco notas. Podemos representar os pesos dessas avaliações como sendo respectivamente iguais a 0.4, 0.4 e 0.2 — 40%, 40% e 20%. Ou seja, a avaliação mensal, por exemplo, compõe 40% da sua nota final. É importante notar que a soma dessas frequências deve sempre ser igual a 1. Deste modo, utilizando a Equação (1.1.3), temos que sua nota bimestral $\overline{x}$ é dada por:

$$\overline{x} = 0.4 \cdot 8 + 0.4 \cdot 9 + 0.2 \cdot 7 = 8.2$$

1.1.1.3 Mediana

Dispondo-se dos n valores do conjunto em rol, tem-se que a mediana Md é dada por:

$$Md = \begin{cases} x_i, \text{ com } i = \dfrac{n+1}{2} & \text{, se } n \text{ for ímpar} \\ \dfrac{x_i + x_{i+2}}{2}, \text{ com } i = \dfrac{n}{2} & \text{, se } n \text{ for par} \end{cases} \tag{1.1.4}$$

De forma menos rigorosa, podemos definir a determinação da mediana da seguinte maneira:

1. Ordena-se todos os elementos do conjunto, de forma crescente — a este ordenamento chamamos de *rol*.
2. Se o conjunto tiver quantidade *ímpar* de elementos, a mediana corresponde ao termo que está no centro da distribuição.
3. Se a quantidade for *par*, dividimos o conjunto ao meio e tomamos os dois elementos que determinam essa divisão — o último do grupo à esquerda e o primeiro do grupo à direita. A média aritmética simples entre esses elementos corresponderá à mediana.

Para o Estudo de caso 1.1, o número de observações é par. A mediana Md é dada pela média entre os 50^{o} e 51^{o} elementos:

$$Md = \frac{189 + 189}{2} = 189$$

1.1.1.4 Moda

A moda de um conjunto de dados corresponde ao elemento que ocorre com a maior frequência — a moda do conjunto $[1, 2, 2, 3, 5]$ é 2, por exemplo.

Analisando os dados do Estudo de caso 1.1, a partir da Tabela 1.1.2, pode-se mostrar que a observação mais frequente e, portanto, a moda do conjunto, é dada por 199.

1.1.1.5 Percentis

A posição P do i-ésimo percentil p_i de uma distribuição é determinada por:

$$P(p_i) = \left[(n-1) \cdot \left(\frac{i}{100}\right)\right] + 1 \tag{1.1.5}$$

Vamos determinar a posição do 25º percentil p_{25} da amostra indicada pela Tabela 1.1.2 — note que o 25º percentil corresponde também ao primeiro quartil da amostra.

$$\begin{aligned} P(p_{25}) &= \left[(n-1) \cdot \left(\frac{i}{100}\right)\right] + 1 \\ &= \left[(100-1) \cdot \left(\frac{25}{100}\right)\right] + 1 \\ &= 99 \cdot 0.25 + 1 \\ &= 25.75 \end{aligned}$$

Isso indica que a observação que determina o primeiro quartil está entre as posições 25 e 26. De acordo com a Tabela 1.1.2, observa-se que em ambas as posições as observações possuem o mesmo valor e, portanto, este corresponde ao primeiro quartil. Ou seja:

$$p_{25} = 179$$

Como interpretamos essa informação: podemos dizer que 25% dos produtos custam R$ 179,00 ou menos, enquanto os outros 75% custam mais do que R$ 179,00.

Nessa amostra, dado que as observações 25 e 26 possuem o mesmo valor, não é necessário nenhum passo adicional para determinar o percentil.

Quando esses valores forem diferentes, precisamos realizar a *interpolação* entre eles. Utilizando a Equação (1.1.5), pode-se mostrar que $P(p_{60}) = 60.4$. Da Tabela 1.1.2, temos que os elementos das posições 60 e 61 são, respectivamente, iguais a 196 e 197. Devemos encontrar o elemento que está, a partir de 196, a 40% da distância entre 196 e 197 — analogamente, ele também está a 60% dessa distância. Matematicamente:

$$\begin{aligned} p_{60} &= 196 + 0.4 \cdot (197 - 196) \\ &= 196 + 197 \cdot 0.4 - 196 \cdot 0.4 \\ &= 196 \cdot 0.6 + 197 \cdot 0.4 \\ &= 196.4 \end{aligned}$$

1.1.2 Medidas de dispersão

1.1.2.1 Amplitude

É a medida mais simples de dispersão. A amplitude A é dada pela diferença entre os valores máximo e mínimo do conjunto de observações:

$$A = x_{\text{máx}} - x_{\text{mín}} \tag{1.1.6}$$

Para o caso do Estudo de caso 1.1 a amplitude A é dada por:

$$A = 239.00 - 149.00 = 90$$

1.1.2.2 Desvio médio

O desvio D_i de um elemento é a diferença entre cada valor do conjunto e a média aritmética deste conjunto. Matematicamente:

$$D_i = x_i - \overline{x}$$

O desvio médio D_m é a média dos desvios **absolutos** de cada elemento, isto é, tratados em módulo:

$$D_m = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \overline{x}| \tag{1.1.7}$$

Em relação ao Estudo de caso 1.1, o desvio médio D_m é dado por:

$$
\begin{aligned}
D_m &= \frac{1}{100} \cdot \Big(\\
&\quad |189.00 - 190.77| + |195.00 - 190.77| + |199.00 - 190.77| + \\
&\quad + \ldots + \\
&\quad |189.00 - 190.77| + |199.00 - 190.77| + |195.00 - 190.77| \\
&\Big) \Rightarrow \\
D_m &= \frac{1.77 + 4.23 + 8.23 + \ldots + 1.77 + 8.23 + 4.23}{100} \Rightarrow \\
D_m &= \frac{1207.62}{100} \therefore \\
D_m &= 12.08
\end{aligned}
$$

1.1.2.3 Variância

Dispersão das observações de uma variável em torno de sua média. A variância σ^2 pode ser determinada pela expressão

$$\sigma^2 = \frac{1}{(n-1)} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{1.1.8}$$

em que n representa o tamanho da amostra, x_i representa o i-ésimo elemento da amostra e $\overline{x}$ indica a média aritmética simples das observações.

Usando dados da Tabela 1.1.1, vemos que a variância das observações no Estudo de caso 1.1 é determinada por

$$
\begin{aligned}
\sigma^2 &= \frac{1}{99} \cdot \Big(\\
&\quad (189.00 - 190.77)^2 + (195.00 - 190.77)^2 + (199.00 - 190.77)^2 + \\
&\quad + \ldots + \\
&\quad (189.00 - 190.77)^2 + (199.00 - 190.77)^2 + (195.00 - 190.77)^2 \\
&\Big) \Rightarrow \\
\sigma^2 &= \frac{3.13 + 17.89 + 67.73 + \ldots + 3.13 + 67.73 + 17.89}{99} \Rightarrow \\
\sigma^2 &= \frac{24157.71}{99} \therefore \\
\sigma^2 &= 244.02
\end{aligned}
$$

1.1.2.4 Desvio-padrão

Interpretar a dispersão a partir da variância é pouco viável, dado que a dimensão está elevada ao quadrado. Utiliza-se, então, o desvio-padrão em relação à amostra, determinado pela expressão

$$\sigma = \sqrt{\sigma^2} \tag{1.1.9}$$

No Estudo de caso 1.1, o desvio-padrão em relação à amostra é dado por

$$\sigma = \sqrt{244.02} = 15.621$$

Isso sugere que a maior parte dos valores da amostra está distribuída dentro de um intervalo de aproximadamente R$ 15.62 reais acima ou abaixo da média. Ou seja, grande parte dos dados está entre $190.77 - 15.62 = 175.15$ e $190.77 + 15.62 = 206.39$.

1.1.2.5 Erro padrão

Corresponde ao desvio-padrão da média e pode ser obtido pela expressão

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \tag{1.1.10}$$

Para o Estudo de caso 1.1, tem-se que o erro padrão é dado por

$$\sigma_{\overline{x}} = \frac{15.621}{\sqrt{100}} = 1.562$$

1.1.2.6 Coeficiente de variação

O coeficiente de variação CV é uma medida de dispersão relativa que fornece a variação dos dados em relação à sua média.

$$CV = \frac{\sigma}{\overline{x}} \tag{1.1.11}$$

Quanto menor o coeficiente de variação, mais homogêneos são os dados, isto é, menor a dispersão em torno da média. Um CV pode ser considerado baixo, indicando um conjunto de dados razoavelmente homogêneo, quando for menor do que 30%. Se esse valor for acima de 30%, o conjunto de dados pode ser considerado heterogêneo (FÁVERO; BELFIORE, 2025, p. 159). Em relação ao Estudo de caso 1.1, tem-se que o coeficiente de variação é dado por

$$CV = \frac{15.621}{190.77} = 0.08188 = 8.19\%$$

Observa-se que os dados são homogêneos, indicando que a média é uma boa medida para representá-los.

1.1.3 Medidas de forma

Estudaremos medidas de assimetria (*skewness*) e curtose (*kurtosis*), que caracterizam como os elementos do conjunto estão distribuídos em torno da média.

1.1.3.1 Medidas de assimetria

Uma distribuição é dita **simétrica** quando sua média, mediana e moda coincidem. A curva de distribuição é bem representada pela Figura 1.1.1; o eixo das abscissas representa o valor da observação, enquanto o eixo das ordenadas indica a frequência desse valor na distribuição.

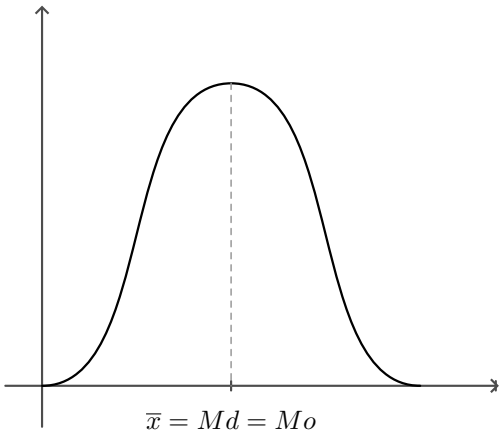
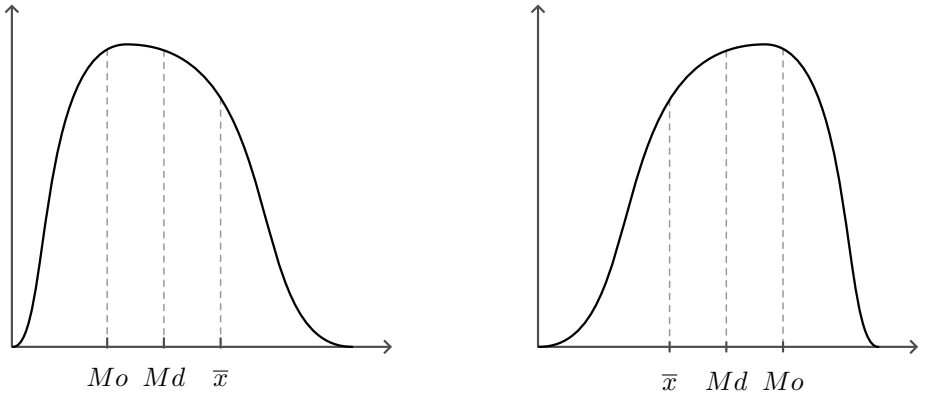


Figura 1.1.1: Representação de uma distribuição simétrica

Se a distribuição se concentrar do lado esquerdo, criando uma "cauda" alongada no lado direito, temos uma **distribuição assimétrica positiva** ou **distribuição assimétrica à direita**—Figura 1.1.2a. Estando a distribuição concentrada do lado direito, portanto com a "cauda" alongada no lado esquerdo, temos uma **distribuição assimétrica negativa** ou **distribuição assimétrica à esquerda**—Figura 1.1.2b.



(a) Distribuição assimétrica à direita (b) Distribuição assimétrica à esquerda

Figura 1.1.2: Representação de distribuições assimétricas

Vejamos maneiras de determinar os coeficientes de assimetria.

Primeiro coeficiente de assimetria de Pearson Medida de assimetria dada pela diferença entre a média e a moda, ponderada pelo desvio-padrão (FÁVERO; BELFIORE, 2025, p. 161). Matematicamente, pode-se expressar o primeiro coeficiente de assimetria de Pearson A_{s_1} como

$$A_{s_1} = \frac{\overline{x} - Mo}{\sigma} \tag{1.1.12}$$

em que:

- se A_{s_1} é igual a 0, a distribuição é simétrica.
- se A_{s_1} é maior do que 0, a distribuição é assimétrica à direita — distribuição positiva.
- se A_{s_1} é menor do que 0, a distribuição é assimétrica à esquerda — distribuição negativa.

Retomando o Estudo de caso 1.1, determinemos A_{s_1}:

$$A_{s_1} = \frac{190.77 - 199}{15.621} = \frac{-8.23}{15.621} = -0.527$$

Observa-se que $A_{s_1} < 0$, indicando uma distribuição assimétrica à esquerda—representada pela Figura 1.1.2b. Na imagem, pode-se notar também que, para esse tipo de distribuição, $\overline{x} < Md < Mo$. Entretanto [1], na situação em estudo, nota-se que $Md < \overline{x} < Mo$

Segundo coeficiente de assimetria de Pearson Medida de assimetria que não é dada em função da moda da distribuição. O segundo coeficiente de assimetria de Pearson A_{s_2} é determinado por:

$$A_{s_2} = \frac{3 \cdot (\overline{x} - Md)}{\sigma} \tag{1.1.13}$$

As relações de assimetria são iguais para segundo e primeiro coeficientes, isto é:

[1] Precisamos entender se isso indica um erro em nossos cálculos, alguma imprecisão da definição fornecida pelo professor ou se a relação apontada nem sempre é verdadeira.

- se $A_{s_2} = 0$, a distribuição é simétrica.
- se $A_{s_2} > 0$, a distribuição é positiva.
- se $A_{s_2} < 0$, a distribuição é negativa.

Determinando o segundo coeficiente de assimetria de Pearson para o Estudo de caso 1.1, temos:

$$A_{s_2} = \frac{3 \cdot (190.77 - 189)}{15.621} = \frac{5.31}{15.621} = 0.340$$

Podemos notar que $A_{s_2} > 0$, denotando uma assimetria positiva, isto é, à direita—representada pela Figura 1.1.2a. Temos um conflito com o cálculo do primeiro coeficiente, que indicada assimetria em sentido contrário. Acredito que o segundo seja mais confiável — na apresentação de slides, o professor sugere a relação apenas entre média e mediana, desconsiderando a moda. Isto é:

- quando $\overline{x} > Md$, a distribuição é assimétrica à direita (positiva);
- quando $\overline{x} < Md$, a distribuição é assimétrica à esquerda (negativa).

Coeficiente de assimetria de Bowley Também conhecido como **coeficiente quartílico de assimetria**, o coeficiente de assimetria de Bowley A_{s_B} é dado em função das medidas separatrizes—primeiro, segundo e terceiro quartis:

$$A_{s_B} = \frac{Q_3 + Q_1 - 2 \cdot Q_2}{Q_3 - Q_1} \tag{1.1.14}$$

De forma análoga aos coeficientes anteriores, tem-se que:

- se $A_{s_B} = 0$, a distribuição é simétrica.
- se $A_{s_B} > 0$, a distribuição é assimétrica positiva (à direita).
- se $A_{s_B} < 0$, a distribuição é assimétrica negativa (à esquerda).

Usando a Equação (1.1.14) para determinar A_{s_B} no Estudo de caso 1.1, temos:

$$\begin{cases} Q_1 = 179 \\ Q_2 = 189 \\ Q_3 = 199 \end{cases} \Rightarrow A_{s_B} = \frac{199 + 179 - 2 \cdot 189}{199 - 179} \Rightarrow$$

$$A_{s_B} = \frac{378 - 378}{20} \therefore$$

$$A_{s_B} = 0$$

Nota-se que, para o terceiro coeficiente de assimetria determinado, chega-se à terceira [2] conclusão diferente.

Coeficiente de assimetria de Fisher É determinado a partir do terceiro momento em torno da média (M_3) (FÁVERO; BELFIORE, 2025, p. 165):

$$g_1 = \frac{n^2 \cdot M_3}{(n-1) \cdot (n-2) \cdot \sigma^3} \tag{1.1.15}$$

em que

$$M_3 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^3 \tag{1.1.16}$$

[2] Não sabemos interpretar o que está acontecendo.

Façamos isso "no braço" para o Estudo de caso 1.1, iniciando com a determinação de M_3 a partir da Equação (1.1.16):

$$
\begin{aligned}
M_3 = \frac{1}{100} \cdot \Big(& \\
& (189.00 - 190.77)^3 + (195.00 - 190.77)^3 + (199.00 - 190.77)^3 + \\
& + \ldots + \\
& (189.00 - 190.77)^3 + (199.00 - 190.77)^3 + (195.00 - 190.77)^3 \\
& \Big) \Rightarrow \\
M_3 = & \frac{-5.55 + 75.69 + 557.44 + \ldots + -5.55 + 557.44 + 75.69}{100} \Rightarrow \\
M_3 = & \frac{33249.04}{100} \therefore \\
M_3 = & \ 332.49
\end{aligned}
$$

Utilizando a Equação (1.1.15), temos que g_1 é dado por:

$$
\begin{aligned}
g_1 &= \frac{23^2 \cdot 332.49}{22 \cdot 21 \cdot 15.621^3} \\
&= \frac{529 \cdot 332.49}{462 \cdot 3811.768} \\
&= \frac{175887.21}{1761036.97} \\
&= 0.09988 \approx 0.1
\end{aligned}
$$

Em resumo, o que pudemos determinar até então:

- Primeiro coeficiente de assimetria de Pearson: $A_{s_1} = -0.527 < 0$ — indica assimetria à esquerda.
- Segundo coeficiente de assimetria de Pearson: $A_{s_2} = 0.340 > 0$ — indica assimetria à direita.
- Coeficiente de assimetria de Bowley: $A_{s_B} = 0$ – indica distribuição simétrica.
- Coeficiente de assimetria de Fisher: $g_1 = 0.1 > 0$ — indica assimetria à direita.

Não há convergência total, mas os resultados de A_{s_2} e g_1 mostram uma certa consistência, indicando que há, de fato, uma distribuição assimétrica positiva. Me parece [3] que, para o Estudo de caso 1.1, tanto o segundo coeficiente de Pearson quanto o coeficiente de assimetria de Fisher são os mais adequados.

1.1.3.2 Medidas de curtose

A curtose pode ser definida como o grau de achatamento de uma distribuição de frequências em relação a uma distribuição teórica que geralmente corresponde à distribuição normal (FÁVERO; BELFIORE, 2025, p. 166). Podemos ter:

- **Mesocúrtica**: assemelha-se à distribuição normal — não é muito achatada nem alongada.
- **Platicúrtica**: apresenta uma curva de frequências mais *achatada* do que a curva normal.
- **Leptocúrtica**: apresenta uma curva de frequências mais *alongada* do que a curva normal.

[3] Confirmar se essa premissa é verdadeira.

Coeficiente de curtose O coeficiente de curtose k, também chamado de *coeficiente percentílico de curtose* é um dos mais utilizados para medir o grau de achatamento da curva de distribuição. Ele é dado por

$$k = \frac{Q_3 - Q_1}{2 \cdot (p_{90} - p_{10})} \tag{1.1.17}$$

em que Q_1 e Q_3 representam o primeiro e o terceiro quartis, respectivamente, enquanto p_{90} e p_{10} representam o 90^{o} e o 10^{o} percentis, respectivamente. Interpretamos o resultado da seguinte maneira:

- $k < 0.263$ indica curva *leptocúrtica*;
- $k = 0.263$ indica curva *mesocúrtica*;
- $k > 0.263$ indica curva *platicúrtica*.

Vamos determinar k para o Estudo de caso 1.1. Iniciamos com a determinação dos percentis p_{90} e p_{10}, usando a Equação (1.1.5):

$$\begin{aligned} P(p_{90}) &= \left[(100-1) \cdot \left(\frac{90}{100}\right)\right] + 1 \\ &= 99 \cdot 0.9 + 1 \\ &= 90.1 \therefore \\ p_{90} &= 209 \end{aligned}$$

$$\begin{aligned} P(p_{10}) &= \left[(100-1) \cdot \left(\frac{10}{100}\right)\right] + 1 \\ &= 99 \cdot 0.1 + 1 \\ &= 10.9 \therefore \\ p_{10} &= 169 \end{aligned}$$

Dado que $Q_1 = 179$ e $Q_3 = 199$, tem-se que:

$$\begin{aligned} k &= \frac{Q_3 - Q_1}{2 \cdot (p_{90} - p_{10})} \\ &= \frac{199 - 179}{2 \cdot (90.1 - 10.9)} \\ &= \frac{20}{158.4} \therefore \\ k &= 0.126 \end{aligned}$$

De acordo com o critério anteriormente mencionado, a curva de distribuição é *leptocúrtica*.

Coeficiente de curtose de Fisher O coeficiente de curtose de Fisher g_2 é também bastante utilizado para medir o grau de achatamento de uma curva de distribuição—no excel, a função `CURT` utiliza tal coeficiente. Ele é calculado a partir do quarto momento em torno da média (M_4)— *apud Maroco, 2014* (FÁVERO; BELFIORE, 2025, p. 168):

Tenho que mexer nessa parte de *apud*

$$g_2 = \frac{n^2 \cdot (n+1) \cdot M_4}{(n-1) \cdot (n-2) \cdot (n-3) \cdot \sigma^4} - 3 \cdot \frac{(n-1)^2}{(n-2) \cdot (n-3)} \tag{1.1.18}$$

em que

$$M_4 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^4 \tag{1.1.19}$$

Para a Equação (1.1.18), tem-se que:

- $g_2 < 0$ indica curva *platicúrtica*.

- $g_2 = 0$ indica curva *mesocúrtica*;
- $g_2 > 0$ indica curva *leptocúrtica*;

Determinemos M_4 para o Estudo de caso 1.1:

$$\begin{aligned}
M_4 &= \frac{1}{100} \cdot \Big(\\
&\quad (189.00 - 190.77)^4 + (195.00 - 190.77)^4 + (199.00 - 190.77)^4 + \\
&\quad + \ldots + \\
&\quad (189.00 - 190.77)^4 + (199.00 - 190.77)^4 + (195.00 - 190.77)^4 \\
&\quad \Big) \Rightarrow \\
M_4 &= \frac{9.82 + 320.16 + 4587.75 + \ldots + 9.82 + 4587.75 + 320.16}{100} \Rightarrow \\
M_4 &= \frac{2087524.18}{100} \therefore \\
M_4 &= 208754.24
\end{aligned}$$

Com isso, pode-se utilizar a Equação (1.1.18) para determinar g_2:

$$\begin{aligned}
g_2 &= \frac{100^2 \cdot (100 + 1) \cdot 208754.24}{(100 - 1) \cdot (100 - 2) \cdot (100 - 3) \cdot 15.621^4} - 3 \cdot \frac{(100 - 1)^2}{(100 - 2) \cdot (100 - 3)} \Rightarrow \\
g_2 &= \frac{210,841,782,400.00}{56,036,155,804.04} - 3 \cdot \frac{9,801}{9,506} \Rightarrow \\
g_2 &= 3.76 - 3.09 \therefore \\
g_2 &= 0.67
\end{aligned}$$

De acordo com o critério acima mencionado, a curva de distribuição é *leptocúrtica*. Isso está de acordo com a determinação feita a partir do coeficiente percentílico de curtose.

A Tabela 1.1.3 indica todas as medidas descritivas abordadas na Seção 1.1, para Estudo de caso 1.1.

	Preço (R$)
Média	190.77
Mediana	189.00
Moda	199.00
Primeiro quartil	179.00
Terceiro quartil	199.00
Amplitude	90.00
Desvio médio	12.08
Variância	244.02
Desvio-padrão	15.62
Erro padrão	1.56
Coeficiente de variação	0.0819
Coeficiente de assimetria (g_1)	0.1
Coeficiente de curtose (g_2)	0.67

Tabela 1.1.3: Estatísticas descritivas para o estudo de caso da Seção 1.1

1.2 Relação entre variáveis

Aula 2
Qua 16 out 2023

Na seção anterior, trabalhamos com análises de apenas uma variável, o que denominamos de estatística descritiva univariada. Supondo que tenhamos um banco de dados com n variáveis métricas, podemos determinar as medidas-resumo para cada uma dessas n variáveis.

Nessa seção, nosso foco será a análise da *relação* entre **duas** variáveis — a isto denominamos **análise bivariada**. Inicia-se conhecendo o *tipo de variável*:

- Se as variáveis são **qualitativas**, utiliza-se análise da **associação** pelo teste qui-quadrado (χ^2);
- Se as variáveis são **quantitativas**, utiliza-se a análise da **correlação** por meio da covariância e do coeficiente de correlação de Pearson.

Para maiores detalhes, conferir o capitulo 3 de (FÁVERO; BELFIORE, 2025), a partir da página 221.

1.2.1 Variáveis qualitativas

1.2.1.1 Teste qui-quadrado

De um modo geral, iniciamos com uma tabela de *distribuição conjunta* de frequências, também chamada de **tabela de contingência** ou **tabela de classificação cruzada**, representada pela Tabela 1.2.1. Nela, apresentam-se as *frequências absolutas observadas* para cada par de categorias das variáveis.

		Variável Y					
		Cat 1	Cat 2	Cat 3	...	Cat J	**Total**
Variável X	Cat 1	n_{11}	n_{21}	n_{31}	...	n_{1j}	$\sum_{k=1}^{j} n_{1k}$
	Cat 2	n_{21}	n_{22}	n_{32}	...	n_{2j}	$\sum_{k=1}^{j} n_{2k}$
	Cat 3	n_{31}	n_{32}	n_{33}	...	n_{3j}	$\sum_{k=1}^{j} n_{3k}$
	...	...	...	...	...	...	...
	Cat I	n_{i1}	n_{i2}	n_{i3}	...	n_{ij}	$\sum_{k=1}^{j} n_{ik}$
	Total	$\sum_{k=1}^{i} n_{k1}$	$\sum_{k=1}^{i} n_{k2}$	$\sum_{k=1}^{i} n_{k3}$	...	$\sum_{k=1}^{i} n_{kj}$	N

Tabela 1.2.1: Representação genérica de uma tabela de classificação cruzada

Vejamos um exemplo para ilustrar melhor essa ideia (FÁVERO; BELFIORE, 2025, p. 224).

Exemplo 1.1 Um estudo com 200 indivíduos de três operadoras de planos de saúde pretendia analisar o comportamento conjunto das variáveis X — operadora de plano de saúde — e Y — nível de satisfação. A tabela de classificação cruzada é dada pela Tabela 1.2.2. Determine o valor de χ^2 total e avalie se há relação estatisticamente significativa entre as variáveis.

Solução

Uma análise rápida indica, por exemplo, que 40 indivíduos têm satisfação baixa com a operadora Total Health, enquanto 16 indivíduos têm satisfação alta com a operadora

	Nível de satisfação			
Operadora	Baixo	Médio	Alto	**Total**
Total Health	40	16	12	**68**
Viva Vida	32	24	16	**72**
Mena Saúde	24	32	4	**60**
Total	**96**	**72**	**32**	**200**

Tabela 1.2.2: Nível de satisfação de 200 indivíduos com suas operadoras de plano de saúde

Vida Vida, ao passo que 32 indivíduos têm satisfação média com a operadora Mena Saúde. Pode-se observar também que, dos 200 indivíduos, 68 possuem contrato com a operadora Total Health, independentemente de seu nivel de satisfação. A Tabela 1.2.2 também mostra que, dos 200 indivíduos, 32 têm alto nível de satisfação, independentemente da operadora de saúde.

Verificaremos se existe associação entre essas variáveis. Utilizaremos, para tanto, a análise da estatística qui-quadrado (χ^2). Ela mede a discrepância entre uma tabela de contingência *observada* e uma tabela de contingência *esperada*, supondo que não há associação entre as variáveis estudadas. Se a distribuição de frequências observadas é igual à distribuição de frequências esperadas, o resultado do teste qui-quadrado é zero. Conclui-se que um valor baixo de χ^2 indica **independência entre as variáveis** (FÁVERO; BELFIORE, 2025, p. 237).

Pode-se determinar χ^2 a partir da expressão

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{1.2.1}$$

em que:

- O_{ij} representa a quantidade de observações na i-ésima categoria da variável X e da j-ésima categoria da variável Y;
- E_{ij} representa a frequência esperada de observações na i-ésima categoria da variável X e da j-ésima categoria da variável Y;
- I indica a quantidade de categorias (linhas) da variável X;
- J indica a quantidade de categorias (colunas) da variável Y.

Uma maneira de interpretar a frequência esperada de observações em uma célula da tabela é a seguinte: pudemos perceber, da Tabela 1.2.2, que 96 de 200 indivíduos têm nível de satisfação baixo em relação à sua operadora de saúde, seja ela qual for. Isso representa 48% do total. Contudo, quando analisamos a correspondência entre o baixo nível de satisfação e cada operadora de saúde, verificamos diferentes percentuais: 58,8% para Total Health, 44,4% para Viva Vida e 24,0% para Mena Saúde — verificar Tabela 1.2.3. **Supondo** que *não houvesse associação* entre as variáveis, *seria de se esperar* a **mesma proporção** de 48% em relação ao total de clientes (total da linha), para cada operadora (cada linha). A esse suposto valor podemos dar o nome de frequência esperada de observação.

Isso sugere que, para determinar a frequência esperada de uma dada observação, pode-se prosseguir da seguinte maneira:

1. dividimos o total da coluna — total da categoria I na variável Y — pela quantidade de observações (N), obtendo a frequência relativa daquela categoria em relação ao total;
2. multiplicamos o valor encontrado pelo total da linha — total da categoria J na variável X.

	Nível de satisfação			
Operadora	Baixo	Médio	Alto	Total
Total Health	40 (58,8%)	16 (23,5%)	12 (17,6%)	**68 (100%)**
Viva Vida	32 (44,4%)	24 (33,3%)	16 (22,2%)	**72 (100%)**
Mena Saúde	24 (40,0%)	32 (53,3%)	4 (6,7%)	**60 (100%)**
Total	**96 (48%)**	**72 (36%)**	**32 (16%)**	**200 (100%)**

Tabela 1.2.3: Valores observados na Tabela 1.2.2 com as respectivas proporções em relação ao total geral da linha

Matematicamente, isto equivale a:

$$E_{ij} = \frac{1}{N} \cdot \left(\sum_{k=1}^{I} n_{kj} \cdot \sum_{k=1}^{J} n_{ik} \right) \tag{1.2.2}$$

Utilizando a Equação (1.2.2), pode-se determinar as frequências esperadas para o Exemplo 1.1 e obter os resultados indicados pela Tabela 1.2.4. De forma concreta, com alguns exemplos:

- Total Health, baixo: $\dfrac{96 \cdot 68}{200} = 32.64$
- Viva Vida, alto: $\dfrac{32 \cdot 72}{200} = 11.52$
- Mena Saúde, médio: $\dfrac{72 \cdot 60}{200} = 21.60$

	Nível de satisfação		
Operadora	Baixo	Médio	Alto
Total Health	32,64	24,48	10,88
Viva Vida	34,56	25,92	11,52
Mena Saúde	28,80	21,60	9,60

Tabela 1.2.4: Valores esperados para a Tabela 1.2.2

De posse dos resultados da Tabela 1.2.4, podemos usar a Equação (1.2.1) e determinar χ^2 para cada observação. A Tabela 1.2.5 também indica o valor total, isto é, a soma de todos os χ^2 encontrados.

	Nível de satisfação		
Operadora	Baixo	Médio	Alto
Total Health	1,66	2,94	0,12
Viva Vida	0,19	0,14	1,74
Mena Saúde	0,80	5,01	3,27
Total	$\chi^2 = 15,861$		

Tabela 1.2.5: Determinação de χ^2 para as observações da Tabela 1.2.2

Precisamos, agora, avaliar se o resultado encontrado para χ^2 *rejeita* ou *não rejeita* a chamada **hipótese nula**. Essa parte está bem confusa e demanda maior entendimento que podemos conferir posteriormente em outros capítulos do livro e complementar nosso caderno. Mas, em linhas gerais, isso está no contexto do chamado *teste de hipóteses*, no qual objetiva-se determinar H_0, chamado de **hipótese nula**, e H_1, chamado de **hipótese alternativa**. O nível de significância α de um teste indica a probabilidade

Quando possível, editar essa parte.

de rejeitar determinada hipótese quando ela for verdadeira. O *p-value* representa a probabilidade associada ao valor observado da amostra, indicando o menor nível de significância que levaria à rejeição da hipótese suposta—quanto mais baixo seu valor, menos se pode acreditar da hipótese suposta (FÁVERO; BELFIORE, 2025, p. 239).

Esse teste de hipóteses demanda também o entendimento dos *graus de liberdade* da distribuição. Eles se referem à quantidade de observações da amostra que podem variar de forma independente e aleatória, obtendo, ainda assim, o valor em análise. Cada teste estatístico tem um cálculo específico dos graus de liberdade. Para o teste qui-quadrado, tem-se que distribuição possui $(i-1)\cdot(j-1)$ graus de liberdade. No caso do Exemplo 1.1, com i e j iguais a 3, temos 4 graus de liberdade.

Na Figura 1.2.1 temos uma representação de uma distribuição qui-quadrado genérica. O valor crítico x_c indica o valor da observação que separa a distribuição em duas áreas: à sua esquerda, uma região de não rejeição de H_0, enquanto, à sua direita, uma região de rejeição de H_0. Essa última, também chamada de região crítica, pode ser denotada a partir do nível de significância α. Observe que, na figura, tal região está hachurada. A área dessa região corresponde ao percentual α da área total. Ou seja, para um nível de significância de 5%, tem-se uma área de 0,05.

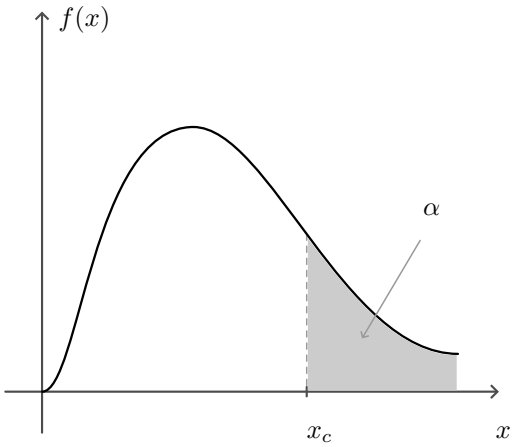


Figura 1.2.1: Representação de uma distribuição qui-quadrado

Isso nos permite duas maneiras de avaliar o resultado de nosso teste:

- Estabelecer a relação entre χ^2 e x_c, ou seja, se χ^2 está à direita ou à esquerda de x_c;
- Estabelecer a relação entre *p-value* e α, isto é, se a área determinada por χ^2 é maior ou menor do que o nível de significância.

Referenciar onde isso pode ser feito

É possível mostrar que, para o Exemplo 1.1, o valor crítico é $x_c = 9.48$ e *p-value* é igual a 0,003. Dado que $\chi^2 > x_c$ e p-value $< \alpha$, conclui-se, portanto, que pode-se rejeitar H_0, indicando que há relação estatisticamente significativa entre as variáveis observadas; isto é, a associação não se dá de forma aleatória. □

1.2.2 Variáveis quantitativas

Para o estudo de correlação entre variáveis quantitativas, precisamos entender os conceitos de **covariância** e **coeficiente de correlação de Pearson**.

1.2.2.1 Covariância

A covariância mede a variação conjunta entre duas variáveis quantitativas. Para duas variáveis X e Y, tem-se que a covariância *cov* é dada por:

$$cov(X,Y) = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) \tag{1.2.3}$$

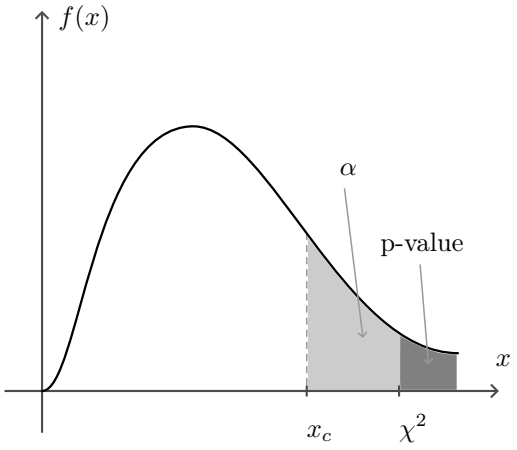


Figura 1.2.2: Interpretação do *p-value* em uma distribuição qui-quadrado

O objetivo dessa medida é indicar se existe alguma variação em uma medida conforme outra varia. Isto é:

- se x aumenta, y aumenta também?
- se x aumenta, y diminui?
- se x aumenta, y pode tanto aumentar quanto diminuir?

Esse tipo de pergunta é mais facilmente respondido a partir da análise do coeficiente de correlação de Pearson.

1.2.2.2 Coeficiente de correlação de Pearson

A partir da Equação (1.2.3), define-se o coeficiente de correlação de Pearson ρ, que pode ser determinado por:

$$\rho = \frac{cov(X,Y)}{\sigma_x \cdot \sigma_y} = \frac{1}{(n-1)\cdot \sigma_x \cdot \sigma_y} \cdot \sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y}) \tag{1.2.4}$$

Vimos, com a Equação (1.1.8) e com a Equação (1.1.9) — Sub-seção 1.1.2 — que o desvio padrão pode ser expresso como:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Substituindo-se em 1.2.4, pode-se mostrar que ρ é também dado por:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{1.2.5}$$

Como indicado anteriormente, tanto a covariância quanto o coeficiente de correlação de Pearson pretendem investigar de que modo X e Y variam. Por estar padronizado com os desvios médios, o coef. de correlação é mais simples de interpretar, dado que é sempre um valor entre -1 e $+1$. Deste modo, tem-se que:

- Se ρ for positivo, dizemos que há relação diretamente proporcional entre X e Y. Para $\rho = +1$, temos uma correlação linear positiva perfeita.
- Se ρ for negativo, dizemos que há relação inversamente proporcional entre X e Y. Para $\rho = -1$, temos uma correlação linear negativa perfeita.
- Se ρ for nulo, não existe correlação entre as variáveis.

Observação	Matemática	Física	Literatura
1	5,50	7,50	9,00
2	9,00	8,50	5,50
3	4,50	5,00	6,50
4	6,50	8,00	6,50
5	7,50	6,00	5,00
6	3,00	6,00	10,00
7	10,00	8,00	5,50
8	9,00	8,00	6,50
9	4,50	5,50	8,00
10	5,00	5,00	5,50
11	3,50	5,00	7,50
12	7,50	9,00	4,50
13	6,50	7,50	8,50
14	8,00	9,00	5,00
15	4,00	5,00	6,50
16	7,00	6,00	8,50
17	7,50	7,50	6,00
18	6,00	9,00	3,00
19	10,00	7,50	5,00
20	9,00	10,00	5,50
21	8,00	9,00	9,00
22	5,00	5,00	5,00
23	4,00	3,00	7,50
24	9,50	8,00	8,50
25	6,50	7,00	4,50
26	7,00	7,50	8,00
27	5,00	4,50	9,00
28	6,50	8,00	5,00
29	8,50	6,00	6,00
30	9,75	5,00	6,50

Tabela 1.2.6: Notas de 30 estudantes em matemática, física e literatura

Exemplo 1.2 O coordenador de um curso deseja analisar se existe correlação entre as notas dos alunos nas disciplinas de matemática, física e literatura. Para isso, montou uma base com as notas de 30 estudantes nos referidos componentes curriculares — ver Tabela 1.2.6. Determine os coeficientes de correlação de Pearson para os pares de correlação matemática-física, matemática-literatura e física-literatura.

Solução

Podemos usar as equações 1.1.1 e 1.1.9 para determinar as médias e os desvios padrões das notas em cada disciplina. A Tabela 1.2.7 traz esses resultados.

Iniciemos com a correlação entre as notas de matemática e física, que chamaremos de ρ_{mf}. Usando a Equação (1.2.4), temos:

Descritivas	Nota Matemática	Nota Física	Nota Literatura
Média	6,78	6,87	6,57
Desvio Padrão	2,05	1,72	1,72

Tabela 1.2.7: Medidas descritivas para as notas dos estudantes

$$
\begin{aligned}
\rho_{mf} &= \frac{cov(M,F)}{\sigma_m \cdot \sigma_f} \\
&= \frac{1}{(n-1)\cdot\sigma_m\cdot\sigma_f} \cdot \sum_{i=1}^{n} (m_i - \overline{m}) \cdot (f_i - \overline{f}) \\
&= \frac{1}{29 \cdot 2.05 \cdot 1.72} \cdot \Big(\\
&\quad (5.50 - 6.78) \cdot (7.50 - 6.87) + \\
&\quad (9.00 - 6.78) \cdot (8.50 - 6.87) + \\
&\quad (4.50 - 6.78) \cdot (5.00 - 6.87) + \\
&\quad \ldots + \\
&\quad (6.50 - 6.78) \cdot (8.00 - 6.87) + \\
&\quad (8.50 - 6.78) \cdot (6.00 - 6.87) + \\
&\quad (9.75 - 6.78) \cdot (5.00 - 6.87) \\
&\Big) \\
&= \frac{-0.807 + 3.634 + 4.247 + \ldots + -0.312 + -1.495 + -5.553}{102.25} \therefore \\
\rho_{mf} &= 0.602
\end{aligned}
$$

O valor $\rho_{mf} = 0.602$ sugere uma correlação linear positiva, indicando uma proporção direta e relativamente grande entre as notas de matemática e física. Ou seja, bons estudantes numa disciplina tendem para um bom desempenho na outra, enquanto estudantes com dificuldade numa tendem a manifestar dificuldade também na outra.

Determinemos agora a correlação ρ_{ml} entre as notas de matemática e de literatura:

$$
\begin{aligned}
\rho_{ml} &= \frac{cov(M,L)}{\sigma_m \cdot \sigma_l} \\
&= \frac{1}{(n-1)\cdot\sigma_m\cdot\sigma_l} \cdot \sum_{i=1}^{n} (m_i - \overline{m}) \cdot (l_i - \overline{l}) \\
&= \frac{1}{29 \cdot 2.05 \cdot 1.72} \cdot \Big(\\
&\quad (5.50 - 6.78) \cdot (9.00 - 6.57) + \\
&\quad (9.00 - 6.78) \cdot (5.50 - 6.57) + \\
&\quad (4.50 - 6.78) \cdot (6.50 - 6.57) + \\
&\quad \ldots + \\
&\quad (6.50 - 6.78) \cdot (5.00 - 6.57) + \\
&\quad (8.50 - 6.78) \cdot (6.00 - 6.57) + \\
&\quad (9.75 - 6.78) \cdot (6.50 - 6.57) \\
&\Big) \\
&= \frac{-3.103 - 2.373 + 0.152 + \ldots + 0.431 - 0.978 - 0.198}{102.25} \therefore \\
\rho_{ml} &= -0.309
\end{aligned}
$$

Temos uma correlação linear negativa, sugerindo que bons estudantes numa disciplina não tendem a um bom desempenho também na outra. Em módulo, o valor de ρ_{ml} não é tão alto, sugerindo que tal correlação não é muito forte.

Determinemos, por fim, a correlação ρ_{fl} entre as notas de física e de literatura:

$$
\begin{aligned}
\rho_{fl} &= \frac{cov(F,L)}{\sigma_f \cdot \sigma_l} \\
&= \frac{1}{(n-1)\cdot\sigma_f\cdot\sigma_l}\cdot\sum_{i=1}^{n}(f_i-\overline{f})\cdot(l_i-\bar{l}) \\
&= \frac{1}{29\cdot 1.72\cdot 1.72}\cdot\Big(\\
&\quad (7.50-6.87)\cdot(9.00-6.57)+ \\
&\quad (8.50-6.87)\cdot(5.50-6.57)+ \\
&\quad (5.00-6.87)\cdot(6.50-6.57)+ \\
&\quad \ldots + \\
&\quad (8.00-6.87)\cdot(5.00-6.57)+ \\
&\quad (6.00-6.87)\cdot(6.00-6.57)+ \\
&\quad (5.00-6.87)\cdot(6.50-6.57) \\
&\Big) \\
&= \frac{1.541-1.742+0.124+\ldots-1.776+0.491+0.124}{85.79} \therefore \\
\rho_{ml} &= -0.288
\end{aligned}
$$

Assim como na comparação matemática-literatura, a correlação entre as notas de física e de literatura é linear negativa, com módulo um pouco menor do que a anterior — isso sugere que a correlação não é tão forte. □

1.3 Distribuições de probabilidades

Nas seções anteriores, descrevemos variáveis qualitativas e quantitativas a partir de suas estatísticas descritivas. Vimos também como relacionar variáveis a partir de medidas de correlação. Essa seção tem como objetivo verificar como a frequência de uma dada observação está distribuída em relação ao todo. Dizendo de outro modo, poderemos observar quais valores são mais frequentes em nosso conjunto de dados e se essas frequências estão *distribuídas* de alguma maneira em particular. Conferir Capítulo 5 de (FÁVERO; BELFIORE, 2025), a partir da página 305.

Estudaremos o comportamento de variáveis aleatórias discretas e contínuas. Para o primeiro grupo, podemos utilizar as seguintes distribuições:

- uniforme discreta
- Bernoulli
- binomial
- geométrica
- binomial negativa
- hipergeométrica
- Poisson

Para o segundo grupo, teremos:

- uniforme
- normal
- exponencial
- Gama
- qui-quadrado (χ^2)
- t de Student
- f de Snedecor

1.3.1 Definições gerais

Espaço amostral É o conjunto de todos os resultados possíveis de um experimento aleatório. Convenciona-se descrever esse experimento a partir da associação de valores numéricos aos elementos do espaço amostral. A variável aleatória pode ser caracterizada como aquela que apresenta um valor único para cada elemento, sendo esse valor determinado aleatoriamente. Uma variável aleatória x representa um valor numérico associado a cada resultado de um experimento probabilístico (ou aleatório).

Definição 1.1 (Espaço amostral). Consideremos ε um experimento aleatório e S o espaço amostral associado ao experimento. Podemos entender a **variável aleatória** X como a função f que associa cada elemento $s \in S$ a um número real x.

Variáveis aleatórias Uma variável aleatória é dita **discreta** quando tem um número finito ou *contável* de resultados possíveis que podem ser enumerados. Uma variável aleatória é **contínua** quando tem um número *incontável* de resultados possíveis, representados por um intervalo na reta numerada (LARSON; FARBER, 2016).

x vendas	$p(x)$
0	0.2
1	0.4
2	0.3
3	0.1

Tabela 1.3.1: Distribuição de probabilidades da venda mensal de imóveis

Esperança de uma variável aleatória discreta Seja X uma variável aleatória discreta que pode assumir os valores x, em que

$$x \in [x_1, x_2, x_3, \ldots, x_n]$$

As respectivas probabilidades de ocorrência de cada valor são dadas por $p(x)$, em que

$$p(x) \in [p(x_1), p(x_2), p(x_3), \ldots, p(x_n)]$$

Define-se *esperança* (valor esperado ou médio) de X como:

$$E(X) = \sum_{i=1}^{n} x_i \cdot p(x_i) \tag{1.3.1}$$

Pode-se notar que isso se parece muito com uma média ponderada pelas probabilidades. Vejamos com um exemplo (FÁVERO; BELFIORE, 2025, p .307):

Exemplo 1.3 A venda mensal de imóveis de um determinado corretor segue a distribuição de probabilidades indicada pela Tabela 1.3.1. Determine o valor esperado de sua venda mensal.

Solução

Usando a Equação (1.3.1), temos que $E(X)$ é dado por:

$$E(X) = 0 \cdot 0.2 + 1 \cdot 0.4 + 2 \cdot 0.3 + 3 \cdot 0.1 = 1.3$$

Isso significa que o *valor médio* dessa distribuição de probabilidades é igual a 1.3. Podemos entender isso como uma média das probabilidades. Apesar da similaridade com a Equação (1.1.3), façamos uma distinção aqui: no caso da média aritmética ponderada, cada valor é multiplicado pela sua frequência **observada**, isto é, aquilo que se verifica no experimento ocorrido. Já para o cálculo de $E(X)$, o valor da variável aleatória é multiplicado pela sua *probabilidade de ocorrência*, no contexto de um experimento probabilístico. □

Esperança de uma variável aleatória contínua Uma variável aleatória contínua X está associada a uma *função densidade de probabilidade* $f(x)$ que satisfaz à seguinte condição:

$$\int_{-\infty}^{+\infty} f(x)\, dx = 1 \tag{1.3.2}$$

Isso pode ser interpretado da seguinte maneira: vimos, na seção anterior, que a soma de todas as probabilidades é sempre igual a 1. Dada sua natureza contínua, tal soma é feita a partir da integração da função $f(x)$ em todo o intervalo para o qual existe x. Nesse cenário, é importante também definir que

$$f(x) \geq 0$$

para todo $x \in \mathbb{R}$.

Define-se a probabilidade da variável assumir valores aleatórios entre um intervalo a e b como

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx \tag{1.3.3}$$

com

$$-\infty < a < b < +\infty$$

Isso nos permite definir a esperança de uma variável aleatória contínua como:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx \tag{1.3.4}$$

Função de distribuição acumulada A função de distribuição acumulada (FDA) ou função de distribuição cumulativa (CDF) — *cumulative distribuction function* — é uma função que descreve a probabilidade de uma variável aleatória ser menor ou igual a um certo valor. Em outras palavras, para uma variável aleatória X, a função de distribuição acumulada $F(x)$ representa a probabilidade de X assumir um valor menor ou igual a x, isto é:

$$F(x) = P(X \leq x)$$

A FDA é útil porque oferece uma visão completa de como as probabilidades se acumulam ao longo dos valores possíveis da variável aleatória.

1.3.2 Distribuições para variáveis aletórias discretas

1.3.2.1 Distribuição uniforme discreta

Nesse tipo de distribuição, todos os possíveis valores da variável aleatória têm a mesma probabilidade de ocorrência. A função de probabilidade é dada por:

$$P(X = x_i) = p(x_i) = \frac{1}{n} \tag{1.3.5}$$

com

$$i = 1, 2, 3, \ldots, n$$

A Figura 1.3.1a representa graficamente tal distribuição. Observe que cada valor x_i da variável X tem a mesma probabilidade $1/n$ de ocorrência.

A esperança $E(X)$ pode ser determinada por:

$$E(X) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{1.3.6}$$

A variância $Var(X)$ é dada por:

$$Var(X) = \frac{1}{n} \cdot \left[\sum_{i=1}^{n} {x_i}^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \right] \tag{1.3.7}$$

A função de distribuição acumulada é dada por:

$$F(X) = P(x_i \leq x) = \sum_{x_i < x}^{n} \frac{1}{n} = \frac{n(x)}{n} \tag{1.3.8}$$

em que $n(x)$ é o número correspondente a $x_i \leq x$ — ver representação na Figura 1.3.1b.

Exemplo 1.4 No lançamento de um dado não viciado, a variável aleatória X representa o valor da face voltada para cima. Determine a distribuição de X, a esperança $E(X)$ e a variância $Var(X)$.

x	$p(x)$
1	$1/6$
2	$1/6$
3	$1/6$
4	$1/6$
5	$1/6$
6	$1/6$
Soma	**1**

Tabela 1.3.2: Distribuição de probabilidades para o Exemplo 1.4

Solução

Sabendo que existem seis faces, todas com a mesma chance de ocorrência — uma vez que o dado não é viciado — tem-se que a probabilidade é igual a $1/6$ para todas as variáveis, como indicado na Tabela 1.3.2.

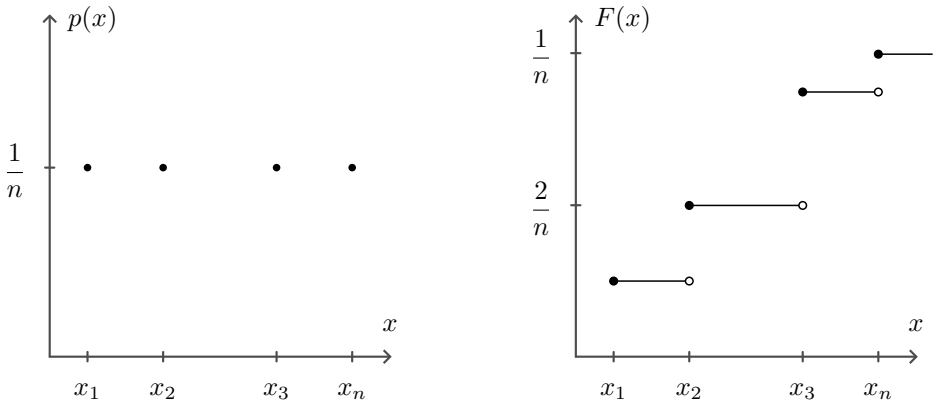
A esperança $E(X)$ é dada por:

$$E(X) = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$$

A variância $Var(X)$ é dada por:

$$\begin{aligned} Var(X) &= \frac{1}{6} \cdot \left[(1^2 + 2^2 + \ldots + 6^2) - \frac{(1+2+\ldots+6)^2}{6} \right] \\ &= \frac{1}{6} \cdot \left(91 - \frac{441}{6} \right) \\ &= 2.917 \end{aligned}$$

□



(a) Distribuição uniforme discreta

(b) Função de distribuição acumulada

Figura 1.3.1: Distribuição de probabilidades para variáveis aleatórias discretas

1.3.2.2 Distribuição de Bernoulli

Esse é um tipo de distribuição binária, ou seja, que admite apenas dois possíveis resultados — cara ou coroa, par ou ímpar, aparelho ligado ou desligado, etc — convencionalmente denominados *sucesso* ou *fracasso*.

Considere a variável aleatória X que assume o valor 1 no caso de sucesso e 0 no caso de fracasso. A probabilidade de sucesso é representada por p e a probabilidade de

fracasso é representada por $q = 1-p$. A distribuição de Bernoulli fornece a probabilidade de sucesso ou fracasso de X na realização de um único experimento — ver Figura 1.3.2a. Matematicamente:

$$P(X = x) = p(x) = \begin{cases} p & , \texttt{se } x = 1 \\ q = 1 - p & , \texttt{se } x = 0 \end{cases} \tag{1.3.9}$$

O valor esperado $E(X)$ é dado por:

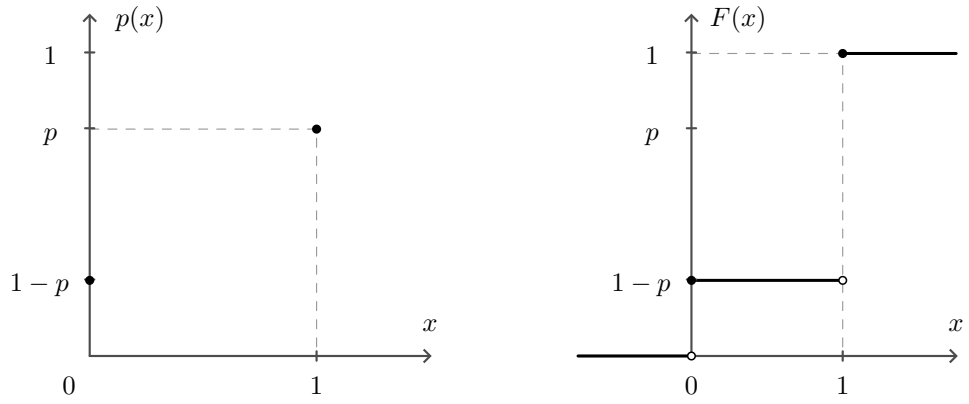
$$E(X) = p \tag{1.3.10}$$

A variância $Var(X)$ é dada por:

$$Var(X) = p \cdot (1 - p) \tag{1.3.11}$$

A função de distribuição acumulada de Bernoulli $F(x)$ é dada por:

$$F(x) = P(X \leq x) = \begin{cases} 0 & , \texttt{se } x < 0 \\ 1 - p & , \texttt{se } 0 \leq x < 1 \\ 1 & , \texttt{se } x \geq 1 \end{cases} \tag{1.3.12}$$



(a) Distribuição de Bernoulli

(b) Função de distribuição acumulada

Figura 1.3.2: Distribuição de Bernoulli para variáveis discretas

Vamos analisar a Equação (1.3.12), a partir de sua representação gráfica dada pela Figura 1.3.2b:

- Para $x < 0$, a probabilidade é sempre 0, pois a variável x só assume valores a partir de 0.
- Para x entre 0 e 1, com x podendo ser 0 mas não chegando a se igualar a 1, a probabilidade é igual à de $x = 0$, ou seja, $1 - p$.
- Para x maior do que ou igual a 1, a probabilidade é a soma das probabilidades $p(x = 0)$ e $p(x = 1)$ — ou seja, $(1 - p) + p = 1$, uma vez que o experimento não assume valores acima de 1.

Exemplo 1.5 Duas equipes de futebol, A e B, se enfrentam na final da Copa Libertadores da América. A probabilidade de A se sagrar campeã é igual a 0.60. A variável X representa o time vencedor dessa partida. Determine a distribuição de X, seu valor esperado $E(X)$ e a variância $Var(X)$.

Solução
Analisando a partir da perspectiva de A, temos que os valores possíveis a serem assumidos por X são dados por:

$$X = \begin{cases} 1 & \texttt{, se A vencer} \\ 0 & \texttt{, se B vencer} \end{cases}$$

A função de probabilidade $p(x)$, portanto, é dada por:

$$p(x) = \begin{cases} 0.60 & \texttt{, se A vencer} \\ 0.40 & \texttt{, se B vencer} \end{cases}$$

O valor esperado, quando analisamos sob a perspectiva da equipe A, é justamente sua vitória (sucesso). Ou seja:

$$E(X) = 0.60$$

Logo, a variância é dada por:

$$Var(X) = 0.60 \cdot 0.40 = 0.24$$

□

1.3.2.3 Distribuição binomial

Um experimento binomial consiste em n repetições independentes de um experimento binário (experimento de Bernoulli) com probabilidade constante p de sucesso em todas as repetições. Como exemplo prático temos o lançamento sucessivo de uma mesma moeda, n vezes, todas as quais possuem a mesma probabilidade de ocorrência de um dado evento — sair cara, por exemplo. A variável aleatória discreta X corresponde ao número k de sucessos nas n repetições do experimento. A função $f(k)$ de distribuição de probabilidade é dada por:

$$f(k) = p(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \tag{1.3.13}$$

onde

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Exemplo 1.6 No lançamento de uma moeda, a probabilidade de sair "cara" é igual a $p = 0.50$. Determine a probabilidade de, em três lançamentos, obtermos "cara" duas vezes.

Solução
O total de combinações possíveis é dado por $2^3 = 8$. Podemos representar "cara" e "coroa" por C e O (c**O**roa). Já que estamos considerando como sucesso a obtenção de "cara", a respectiva representação também pode ser dada por 1 e 0. A Tabela 1.3.3 sistematiza os 8 resultados possíveis desse experimento aleatório. Nota-se que em três deles podemos obter "cara" duas vezes, o que implica numa probabilidade $p = 3/8$.

Vamos obter essa mesma probabilidade a partir da Equação (1.3.13):

$$p(X = 2) = \frac{3!}{2! \cdot 1!} \cdot (0.5)^2 \cdot (0.5)^1 = 3 \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{3}{8}$$

□

Por conta da natureza binomial do experimento, podemos demonstrar que, no lançamento de uma moeda, a distribuição é simétrica em torno da média. Basta reparar em dois pontos importantes.

O termo

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Lançamento	Resultado			Resultado			Sucessos
1	C	C	C	1	1	1	3
2	C	C	O	1	1	0	2
3	C	O	C	1	0	1	2
4	O	C	C	0	1	1	2
5	C	O	O	1	0	0	1
6	O	C	O	0	1	0	1
7	O	O	C	0	0	1	1
8	O	O	O	0	0	0	0

Tabela 1.3.3: Resultados em três lançamentos de uma mesma moeda

corresponde a uma **combinação simples** de k elementos dentre todos os n elementos de um determinado subconjunto, também denotada $C_{n,k}$. É simples mostrar que a seguinte igualdade é verdadeira:

$$\binom{n}{k} = \binom{n}{n-k}$$

Basta observar que:

$$\binom{n}{n-k} = \frac{n!}{(n-k)![n-(n-k)]!} = \frac{n!}{(n-k)!k!}$$

Nota-se que os denominadores resultam no mesmo produto, uma vez que seus fatores apenas se apresentam em ordens distintas. Desse modo, supondo 10 lançamentos consecutivos de uma mesma moeda, a quantidade distinta de combinações que resulta em 3 sucessos é idêntica à quantidade de combinações distintas que resulta em 7 sucessos.

Aliado a esse ponto, tem-se também o fato de, em todos os eventos do experimento, a probabilidade de sucesso ser igual à probabilidade de fracasso. Observe a Tabela 1.3.4. Como sinalizado no parágrafo anterior, existem 120 combinações que resultam em 3 sucessos, valor igual ao de combinações para resultar em 7 sucessos. No caso de uma moeda convencional, as **probabilidades** desses eventos ocorrerem também é a mesma, como denotado na terceira coluna. Observa-se que a maior quantidade de combinações é aquela que resulta em 5 sucessos; portanto, esse é também o evento com a maior probabilidade. Por essa razão, como indicado na Figura 1.3.3, a distribuição de frequências é simétrica em relação à média e corresponde a uma curva normal.

		$p = 0.5$	$p = 0.3$	$p = 0.7$
k	$C_{10,k}$	$f(k)$	$f(k)$	$f(k)$
0	1	0.0010	0.0282	0.0000
1	10	0.0098	0.1211	0.0001
2	45	0.0439	0.2335	0.0014
3	120	0.1172	0.2668	0.0090
4	210	0.2051	0.2001	0.0368
5	252	0.2461	0.1029	0.1029
6	210	0.2051	0.0368	0.2001
7	120	0.1172	0.0090	0.2668
8	45	0.0439	0.0014	0.2335
9	10	0.0098	0.0001	0.1211
10	1	0.0010	0.0000	0.0282

Tabela 1.3.4: Distribuições binomiais para $n = 10$ em três cenários de probabilidades p

Contudo, suponhamos que a moeda seja viciada, pendendo mais para um resultado do que para outro. Nesse caso, as probalidades de sucesso seriam diferentes de 50%. Na Tabela 1.3.4 e na Figura 1.3.3 simulamos cenários considerando probabilidades $p = 0.30$ e $p = 0.70$. Novamente, nota-se que, apesar de haver os mesmos números de

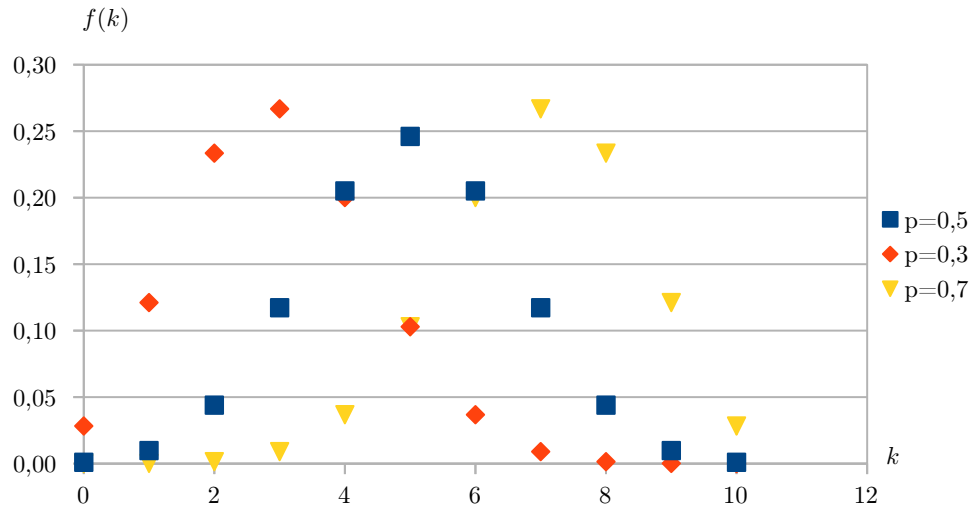


Figura 1.3.3: Distribuições binomiais para $n = 10$ em três cenários de probabilidades p

combinações que correspondam a um determinado resultado, as funções probabilidades $f(k)$ são diferentes. Por exemplo, há 120 combinações que resultam em 3 sucessos. Com uma probabilidade $p = 0.30$ a função de distribuição de probalidade é igual a 0.2668. Considerando uma probabilidade $p = 0.70$, essa função passa a ter valor 0.090.

Nota-se, na Figura 1.3.3, que a curva de distribuição para $p = 0.30$ é assimétrica à direita, enquanto a curva para $p = 0.70$ é assimétrica com uma cauda alongada à esquerda. Podemos interpretar isso da seguinte maneira: com uma probabilidade de sucesso independente igual a 30%, é maior a chance de obter mais fracassos — menos sucessos.

Exemplo 1.7 Em uma indústria, a probabilidade p de encontrar peças defeituosas em cada lote produzido é 6.50%. São produzidos 12 lotes a cada mês. Determine a probabilidade de encontrar peças defeituosas em

a) 2 lotes no mês.

b) 4 lotes no mês.

c) no máximo 2 lotes no mês.

Solução

Para determinar a probabilidade de encontrar peças defeituosas em 2 lotes no mês, considerando $n = 12$ e $p = 0.065$, podemos utilizar a Equação (1.3.13), com $k = 2$:

$$p(X = k) = f(2) = \frac{12!}{2! \cdot 10!} \cdot (0.065)^2 \cdot (0.935)^{10} = 0.142 = 14.2\%$$

A probabilidade de encontrar peças defeituosas em 4 lotes é dada por:

$$p(X = k) = f(4) = \frac{12!}{4! \cdot 8!} \cdot (0.065)^4 \cdot (0.935)^8 = 0.00516 = 0.52\%$$

Por fim, a probabilidade de encontrar peças defeituosas em *no máximo* 2 lotes é dada pelas somas das probabilidades de encontrar peças defeituosas em 0, 1 e 2 lotes. Ou

seja:

$$\begin{aligned} p(x \leq 2) &= f(0) + f(1) + f(2) = \\ &= \frac{12!}{0! \cdot 12!} \cdot (0.065)^0 \cdot (0.935)^{12} + \\ &+ \frac{12!}{1! \cdot 11!} \cdot (0.065)^1 \cdot (0.935)^{11} + \\ &+ \frac{12!}{2! \cdot 10!} \cdot (0.065)^2 \cdot (0.935)^{10} = \\ &= 0.44641 + 0.37240 + 0.14222 = \\ &= 0.96103 = 96.10\% \end{aligned}$$

□

1.3.2.4 Distribuição geométrica

Corresponde a uma distribuição de Bernoulli, que considera experimentos aleatórios sucessivos, todos com probabilidade p de ocorrência. Neste caso, porém, o experimento é realizado até que **o primeiro sucesso** seja atingido.

Tal distribuição apresenta duas parametrizações distintas:

- Consirar sucessivos ensaios até obter o primeiro sucesso.
  - Nesse cenário, não incluímos o zero como um possível resultado.
  - O domínio é dado pelo conjunto $\mathbb{N}* = [1, 2, 3, \ldots]$.
- Contar o número de fracassos antes do primeiro sucesso.
  - Nesse cenário, considera-se o zero como um possível resultado.
  - O domínio é dado pelo conjunto $\mathbb{N} = [0, 1, 2, 3, \ldots]$.

Seja X a variável aleatória que representa o número de tentativas até o primeiro sucesso, com distribuição geométrica de parâmetro p (probabilidade independente de sucesso em cada tentativa). Sua função de probabilidade $f(x)$ é dada por:

$$f(x) = P(X = x) = p \cdot (1 - p)^{x-1} \tag{1.3.14}$$

com $x \in \mathbb{N}*$.

Para o segundo caso, consideremos Y a variável aleatória que representa o número de fracassos antes do primeiro sucesso, com distribuição geométrica de parâmetro p (probabilidade independente de sucesso em cada tentativa). Sua função de probabilidade $f(y)$ é dada por:

$$f(y) = P(Y = y) = p \cdot (1 - p)^{y} \tag{1.3.15}$$

com $y \in \mathbb{N}$.

O valor esperando (esperança) é dado por:

$$E(X) = \frac{1}{p} \tag{1.3.16}$$

$$E(Y) = \frac{1 - p}{p} \tag{1.3.17}$$

A variância é dada por:

$$Var(X) = \frac{1 - p}{p^2} \tag{1.3.18}$$

$$Var(Y) = \frac{1 - p}{p^2} \tag{1.3.19}$$

Exemplo 1.8 Suponha um jogador de basquete *extremamente* regular: para todo e qualquer arremesso livre, ele tem **sempre** a mesma probabilidade de acerto: 40%. Determine a probabilidade do primeiro acerto desse jogador ocorrer no quinto arremesso livre.

Solução

Para o Exemplo 1.8, temos que X denota o experimento aleatório "arremesso livre", com distribuição geométrica de parâmetro $p = 0.40$. Utilizando a Equação (1.3.14), vamos determinar a probabilidade do primeiro acerto ocorrer no quinto ($x = 5$) arremesso livre:

$$\begin{aligned} P(X = 5) &= 0.40 \cdot (1 - 0.40)^{5-1} \\ &= 0.40 \cdot (0.60)^4 \\ &= 0.0518 \\ &= 5.18\% \end{aligned}$$

Isso é equivalente a determinar a probabilidade de ocorrerem $y = 4$ fracassos, utilizando a Equação (1.3.15):

$$P(Y = 4) = (0.40) \cdot (1 - 0.40)^4 = 5.18\%$$

A Figura 1.3.4 e a Tabela 1.3.5 ilustram a distribuição de probabilidades para o Exemplo 1.8, considerando até $x = 10$ arremessos. É fácil notar que a sequência de probabilidades denota uma *progressão geométrica*. □

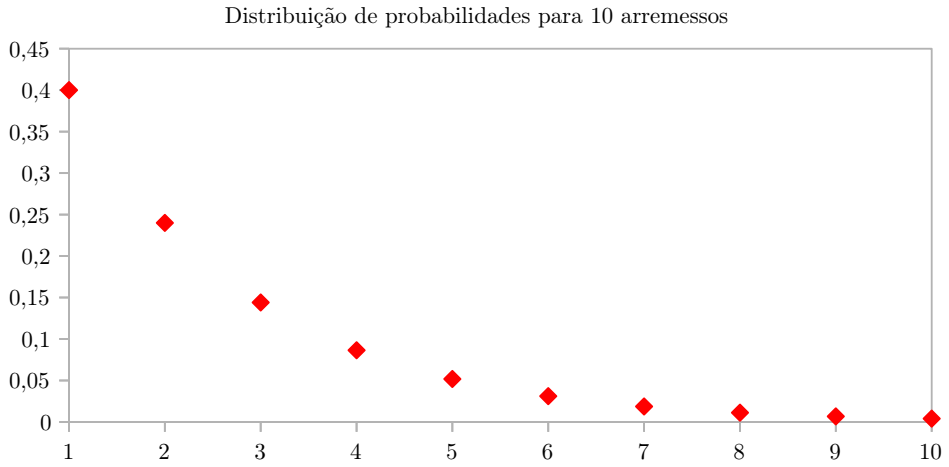


Figura 1.3.4: Distribuição geométrica com $p = 0.40$, até $x = 10$

x	f(x)
1	40,00%
2	24,00%
3	14,40%
4	8,64%
5	5,18%
6	3,11%
7	1,87%
8	1,12%
9	0,67%
10	0,40%

Tabela 1.3.5: Distribuição geométrica com $p = 0.40$, até $x = 10$

1.3.2.5 Distribuição binomial negativa

Também conhecida como **distribuição de Pascal**, ela se assemelha bastante, em conceito, à distribuição binomial. A diferença, nesse caso, é que se realizam x ensaios

de Bernoulli independentes — com probabilidade p de sucesso constante em todas as tentativas — até atingir um número k de sucessos.

A função $f(x)$ de probabilidade em uma distribuição binomial negativa é dada por

$$\begin{aligned} f(x) = P(X = x) = \binom{x-1}{k-1} \cdot p^k \cdot (1-p)^{x-k} \Rightarrow \\ f(x) = \frac{(x-1)!}{(k-1)! \cdot (x-k)!} \cdot p^k \cdot (1-p)^{x-k} \end{aligned} \tag{1.3.20}$$

com $\{x \in N | x \geq k\}$.

O valor esperado $E(X)$ é dado por

$$E(X) = \frac{k}{p} \tag{1.3.21}$$

A variância $Var(X)$ é dada por

$$Var(X) = \frac{k \cdot (1-p)}{p^2} \tag{1.3.22}$$

Como mencionado no início deste tópico, a distribuição binomial negativa relaciona-se com a distribuição binomial. Nesta última, fixa-se o tamanho n da amostra — ou seja, a quantidade de ensaios realizados — e observa-se o número k de sucessos (variável aleatória). Na binomial negativa, fixa-se o número k de sucessos e observa-se a quantidade x de ensaios realizados (variável aleatória) para obter tal quantidade de sucessos.

Podemos também notar que uma distribuição binomial negativa com $k = 1$ é equivalente a uma distribuição geométrica.

Exemplo 1.9 Suponha que um jogador de futebol converte 3 a cada 5 pênaltis. Sendo X o número de tentativas até décimo segundo gol, determine a probabilidade de que esse jogador precise cobrar 20 pênaltis para atingir essa marca.

Solução

Dado que o jogador converte 3 a cada 5 pênaltis, tem-se que a probabilidade dele marcar um gol é $p = 0.60$. Vamos usar a Equação (1.3.20) e determinar a função probabilidade desse jogador precisar de $x = 20$ pênaltis para obter $k = 12$ sucessos (gols):

$$\begin{aligned} f(20) &= \frac{(20-1)!}{(12-1)! \cdot (20-12)!} \cdot 0.60^{12} \cdot (1-0.60)^{20-12} \\ &= \frac{19!}{11! \cdot 8!} \cdot 0.60^{12} \cdot 0.40^8 \\ &= 0.1078 \\ &= 10.78\% \end{aligned}$$

A curva de distribuição de probabilidades para esse caso está representada na Figura 1.3.5. Destaca-se que a maior probabilidade ocorre para converter o 12º pênalti na tentativa $x = 19$. □

Exemplo 1.10 Em um parque de diversões, existe uma máquina em que o jogador deve capturar algum item utilizando os comandos de um braço mecânico. Considere que a probabilidade p de que o jogador consiga capturar algum item em cada jogada é 11%. Identifique as seguintes probabilidades:

a) De que o jogador necessite de 10 jogadas para capturar 3 itens.

b) De que o jogador necessite de 20 jogadas para capturar 3 itens.

c) De que o jogador necessite de 5 jogadas para capturar 1 item.

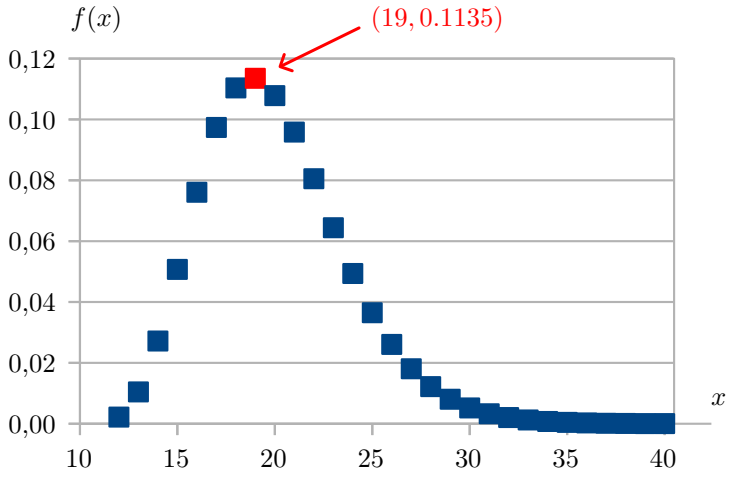


Figura 1.3.5: Distribuição binomial negativa para $k = 12$ e $p = 0.60$, considerando de $x = 12$ a $x = 40$ tentativas

Solução

Vamos usar a Equação (1.3.20) e determinar a probabilidade do jogador necessitar de $x = 10$ jogadas para capturar 3 itens ($k = 3$ sucessos), dada uma probabilidade constante $p = 0.11$ a cada tentativa:

$$\begin{aligned} f(10) &= \frac{(10-1)!}{(3-1)! \cdot (10-3)!} \cdot 0.11^3 \cdot (1-0.11)^{10-3} \\ &= \frac{9!}{2! \cdot 7!} \cdot 0.11^3 \cdot 0.89^7 \\ &= 0.0212 \\ &= 2.12\% \end{aligned}$$

Agora, determinemos a probabilidade do jogador necessitar de $x = 20$ jogadas para obter os mesmos $k = 3$ sucessos:

$$\begin{aligned} f(20) &= \frac{(20-1)!}{(3-1)! \cdot (20-3)!} \cdot 0.11^3 \cdot (1-0.11)^{20-3} \\ &= \frac{19!}{2! \cdot 17!} \cdot 0.11^3 \cdot 0.89^{17} \\ &= 0.0314 \\ &= 3.14\% \end{aligned}$$

Por fim, determinemos a probabilidade do jogador necessitar de $x = 5$ jogadas para obter $k = 1$ sucesso:

$$\begin{aligned} f(5) &= \frac{(5-1)!}{(1-1)! \cdot (5-1)!} \cdot 0.11^1 \cdot (1-0.11)^{5-1} \\ &= \frac{4!}{0! \cdot 4!} \cdot 0.11^1 \cdot 0.89^4 \\ &= 0.0690 \\ &= 6.90\% \end{aligned}$$

□

1.3.2.6 Distribuição hipergeométrica

Essa distribuição também se relaciona com o experimento de Bernoulli. Diferentemente da distribuição binomial, com probabilidade constante de sucesso, na distribuição hipergeométrica a amostragem é *sem reposição*. Desse modo, conforme os elementos são

retirados da população para formar a amostra, o tamanho da população diminiu, fazendo variar a probabilidade de sucesso. Podemos citar, como exemplos, o sorteio de uma loteria — a cada bolinha sorteada a quantidade de bolinhas disponíveis na caixa diminui — e a formação de uma "mão" no pôquer — diminui-se a quantidade de cartas disponíveis na pilha conforme retiram-se cartas, uma a uma, para destiná-las ao jogador.

Seja uma população de N elementos, da qual se retiram, sem reposição, n elementos para a formação da amostra. A distribuição hipergeométrica descreve a probabilidade de obter k sucessos na amostra, sabendo que a população possui um total K de sucessos.

Sendo X a variável que representa o número de sucessos obtidos a partir dos n elementos retirados da amostra, a função de probabilidade de uma distribuição hipergeométrica é dada por

$$f(k) = P(X = k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}} \tag{1.3.23}$$

com $0 \leq k \leq \texttt{min}(K, n)$.

O valor esperado $E(X)$ é dado por:

$$E(X) = \frac{n \cdot K}{N} \tag{1.3.24}$$

A variância $Var(X)$ é dada por:

$$Var(X) = \frac{n \cdot K}{N} \cdot \frac{(N-K) \cdot (N-n)}{N \cdot (N-1)} \tag{1.3.25}$$

Exemplo 1.11 Imagine que você tem um baralho de 52 cartas, das quais 13 são ases (sucessos) e 39 são outras cartas (fracassos). Você quer saber a probabilidade de tirar exatamente 2 ases em uma mão de 5 cartas, sem reposição.

Solução

Temos, para o Exemplo 1.11:

- Tamanho da população: $N = 52$;
- Quantidade de sucessos na população: $K = 13$;
- Tamanho da amostra: $n = 5$;
- Quantidade de sucessos na amostra: $k = 2$.

Utilizando a Equação (1.3.23), tem-se que $f(k = 2)$ é dada por:

$$\begin{aligned}
f(k) &= \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}} \Rightarrow \\
f(2) &= \frac{\binom{13}{2} \cdot \binom{52-13}{5-2}}{\binom{52}{5}} = \frac{\binom{13}{2} \cdot \binom{39}{3}}{\binom{52}{5}} = \\
&= \frac{\frac{13!}{2! \cdot 11!} \cdot \frac{39!}{3! \cdot 36!}}{\frac{52!}{5! \cdot 47!}} = \\
&= \frac{78 \cdot 9139}{2598960} = \\
&= 0.2743 \\
&= 27.43\%
\end{aligned}$$

□

Exemplo 1.12 Uma urna contém 15 bolas, das quais 5 são vermelhas. São escolhidas 7 bolas ao acaso, sem reposição. Determine:

a) A probabilidade de que exatamente duas bolas vermelhas sejam sorteadas.

b) A probabilidade de que pelo menos duas bolas vermelhas sejam sorteadas.

c) O número esperado de bolas vermelhas sorteadas.

d) A variância do número de bolas vermelhas sorteadas.

Solução

Temos, para o Exemplo 1.12:

- Tamanho da população: $N = 15$;
- Quantidade de sucessos na população: $K = 5$;
- Tamanho da amostra: $n = 7$.

No caso do item (a), tem-se que a quantidade de sucessos na amostra é $k = 2$. Sendo assim, utilizemos a Equação (1.3.23) para determinar a probabilidade de ocorrência deste evento:

$$\begin{aligned} f(k) &= \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}} \Rightarrow \\ f(2) &= \frac{\binom{5}{2} \cdot \binom{15-5}{7-2}}{\binom{15}{7}} = \frac{\binom{5}{2} \cdot \binom{10}{5}}{\binom{15}{7}} = \\ &= \frac{\frac{5!}{2! \cdot 3!} \cdot \frac{10!}{5! \cdot 5!}}{\frac{15!}{7! \cdot 8!}} = \\ &= \frac{10 \cdot 252}{6435} = \\ &= 0.3916 = \\ &= 39.16\% \end{aligned}$$

No caso do item (b), a probabilidade de que *pelo menos* duas bolas vermelhas sejam sorteadas implica em aceitar a probabilidade de serem sorteadas 0, 1 ou 2 bolas. Ou seja, devemos determinar

$$f(0) + f(1) + f(2)$$

Pode-se mostrar que $f(0) = 0.0186$ e $f(1) = 0.1632$. Deste modo, tem-se que

$$f(0) + f(1) + f(2) = 0.0186 + 0.1632 + 0.3916 = 0.5734 = 57.34\%$$

O número esperado de bolas vermelhas pode ser determinado pela Equação (1.3.24):

$$E(X) = \frac{n \cdot K}{N} = \frac{7 \cdot 5}{15} = 2.33$$

A variância é dada pela Equação (1.3.25):

$$\begin{aligned} Var(X) &= \frac{n \cdot K}{N} \cdot \frac{(N-K) \cdot (N-n)}{N \cdot (N-1)} = \\ &= \frac{7 \cdot 5}{15} \cdot \frac{(15-5) \cdot (15-7)}{15 \cdot (15-1)} = \\ &= 0.889 \end{aligned}$$

□

1.3.2.7 Distribuição Poisson

A distribuição de Poisson é uma distribuição de probabilidade que descreve o número de ocorrências de um evento em um intervalo fixo de tempo ou espaço, desde que esses eventos ocorram com uma *taxa média constante* e sejam *independentes* entre si. Ela é útil para modelar a ocorrência de eventos que são relativamente raros ($p \to 0$) ou esparsos em relação ao tempo ou espaço.

Diferentemente do modelo binomial, que fornece a probabilidade do número de sucessos em um intervalo *discreto* (n repetições de um experimento), o modelo Poisson fornece a probabilidade do número de sucessos em determinado intervalor **contínuo** — tempo, área, dentre outras possibilidades de exposição (FÁVERO; BELFIORE, 2025, p. 324).

A distribuição de Poisson é adequada para situações em que queremos contar eventos que acontecem de forma esporádica em um intervalo de tempo, área ou volume, como:

- Chamadas de um Call Center: Modelar o número de chamadas recebidas em um call center por hora.
- Chegada de Clientes em um Banco: Contar quantos clientes chegam a uma agência bancária em intervalos fixos de tempo, como a cada 10 minutos.
- Defeitos em Produtos: Modelar o número de defeitos encontrados por metro quadrado de tecido em uma linha de produção.
- Acidentes de Trânsito: Estimar o número de acidentes em um cruzamento específico durante um dia.
- Vazamento de Radiação: Contar partículas detectadas por um sensor em um determinado tempo em um laboratório de física experimental.

Em todos esses casos, os eventos são raros ou ocorrem com baixa frequência dentro do intervalo analisado, e a média de ocorrência é conhecida e constante, tornando a distribuição de Poisson uma ferramenta útil para modelar esses cenários.

Seja X uma variável aleatória discreta que representa a quantidade k de sucessos em determinada unidade contínua. A função de probabilidade dessa variável, quando apresentada uma distribuição Poisson, com parâmetro $\lambda \geq 0$, é dada por

$$f(k) = P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \tag{1.3.26}$$

em que:

- $e \approx 2.718$ representa a base do logaritmo neperiano (ou natural);
- λ representa a taxa média estimada de ocorrência do evento de interesse para dada exposição;

A distribuição Poisson apresenta as seguintes hipóteses (FÁVERO; BELFIORE, 2025, p. 324):

I) Eventos definidos em intervalos não sobrepostos são independentes.

II) Em intervalos de mesmo comprimento, as probabilidades de ocorrência de um mesmo número de sucesso são iguais — p. ex.: se dizemos que há probabilidade de atender 8 pacientes por hora em uma clínica, o número de sucessos é igual na hora 1, na hora 5, em qualquer hora de funcionamento.

III) Em intervalos muito pequenos, a probabilidade de ocorrência de mais de um sucesso é desprezível.

IV) Em intervalos muito pequenos, a probabilidade de um sucesso é proporcional ao comprimento do intervalo.

A média e a variância, na distribuição de Poisson, são iguais e dadas por:

$$E(X) = Var(X) = \lambda \tag{1.3.27}$$

Exemplo 1.13 Suponha que o número de clientes que chegam a um banco siga uma distribuição Poisson. Verifica-se que, em média, chegam 12 clientes por minuto. Determine:

a) a probabilidade de chegada de 10 clientes no próximo minuto;

b) a probabilidade de chegada de 40 clientes nos próximos 5 minutos;

c) a média e a variância de X.

Solução

Assumindo que o caso implica em uma distribuição Poisson, com uma taxa média $\lambda = 12$ de clientes por minuto, vamos usar a Equação (1.3.26) para determinar a probabilidade de chegarem $k = 10$ clientes no próximo minuto:

$$f(10) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} = \frac{e^{-12} \cdot 12^{10}}{10!} = 0.1048 = 10.48\%$$

A probabilidade de chegada de 40 clientes nos próximos 5 minutos é equivalente à probabilidade de chegarem $k = 8$ clientes a cada minuto — vide o segundo item das hipóteses previamente apresentadas. Logo, temos que:

$$f(8) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} = \frac{e^{-12} \cdot 12^{8}}{8!} = 0.0655 = 6.55\%$$

A média (valor esperado da distribuição) e a variância são ambos dados por $\lambda = 12$. □

Exemplo 1.14 Um médico notou que a taxa média de ocorrência de pacientes com certa doença rara em seu consultório é de 2 por ano. Aceitando que esta variável tenha distribuição Poisson, estime:

a) A probabilidade de que o médico receba 1 paciente com a doença em um ano.

b) A probabilidade de que o médico receba 3 pacientes com a doença em um ano.

c) A probabilidade de que o médico não receba pacientes com a doença em um ano.

d) A probabilidade de que o médico receba 6 pacientes com a doença em um ano.

Solução

Aceitando que a variável "pacientes com doença rara no consultório" tenha distribuição Poisson, com taxa média $\lambda = 2$ pacientes por ano, utilizemos a Equação (1.3.26) para determinar a probabilidade de obter $k = 1$ sucesso nessa distribuição — ou seja, a probabilidade desse médico receber 1 paciente com tal doença em um ano:

$$f(1) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} = \frac{e^{-2} \cdot 2^{1}}{1!} = 0.2707 = 27.07\%$$

A probabilidade desse médico receber 3 pacientes com tal doença em um ano ($k = 3$ sucessos) é dada por:

$$f(3) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} = \frac{e^{-2} \cdot 2^{3}}{3!} = 0.1804 = 18.04\%$$

Para determinar a probabilidade do médico não receber pacientes com a doença rara em um ano, basta utilizar a Equação (1.3.26) adotando $k = 0$ sucessos:

$$f(0) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} = \frac{e^{-2} \cdot 2^{0}}{0!} = 0.1353 = 13.53\%$$

Por fim, a probabilidade de ocorrência de $k = 6$ sucessos — receber 6 pacientes com a doença em um ano — é dada por:

$$f(6) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} = \frac{e^{-2} \cdot 2^{6}}{6!} = 0.0120 = 1.20\%$$

É possível notar como as maiores probabilidades concentram-se nos valores k próximos a λ, como indica a Tabela 1.3.6, na qua consideramos até $k = 10$ sucessos.

□

k	$P(X = k)$
0	13,534%
1	27,067%
2	27,067%
3	18,045%
4	9,022%
5	3,609%
6	1,203%
7	0,344%
8	0,086%
9	0,019%
10	0,004%

Tabela 1.3.6: Distribuição de Poisson para $\lambda = 12$

1.3.3 Distribuições para variáveis aletórias contínuas

Aula 3
Sex 18 out 2023

1.3.3.1 Distribuição uniforme

Anotar depois, conferir (FÁVERO; BELFIORE, 2025, p. 327).

1.3.3.2 Distribuição normal

Também conhecida como **distribuição Gaussiana**, é a mais utilizada e importante dentre todas as distribuições, pois permite modelar diversos fenômenos naturais, econômicos e sociais.

A distribuição normal é frequentemente usada para modelar variáveis que se distribuem de maneira próxima ao padrão médio e simétrico. Alguns exemplos incluem:

1. Altura e peso de uma população: as alturas e pesos de indivíduos de uma população tendem a seguir uma distribuição normal.

2. Erros de medição: os erros de medição em experimentos, como os pequenos desvios entre medições reais e valores teóricos, costumam se distribuir normalmente.

3. Desempenho de estudantes em testes: em um grande grupo, as notas de estudantes em testes padronizados geralmente seguem uma distribuição normal, com a maioria concentrada em torno da média.

4. Flutuações no mercado financeiro: em algumas condições, pequenas variações diárias nos preços de ações e índices financeiros podem ser modeladas como uma distribuição normal.

5. Processos biológicos e físicos: características biológicas, como a pressão arterial, e medições físicas, como a intensidade do som em um ambiente, frequentemente seguem uma distribuição normal devido a sua variabilidade natural.

Seja X uma variável aleatória que segue uma distribuição normal, com média $\mu \in \mathbb{R}$ e desvio-padrão $\sigma > 0$. Sua função de densidade de probabilidade $f(x)$ é dada por

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.3.28}$$

com $-\infty < x < +\infty$.

A figura Figura 1.3.6 ilustra graficamente uma distribuição normal, a partir da qual pode-se notar sua curva característica em formato de sino, com simetria em torno da média. Na figura Figura 1.3.7, observa-se outra propriedade particular e importante dessa distribuição: aproximadamente 68% dos valores estão situados 1 desvio-padrão acima ou abaixo da média, enquanto 95% situam-se em até 2 desvios-padrões da média e, por fim, quase 99,7% dos valores concentram-se na região de 3 desvios-padrões em

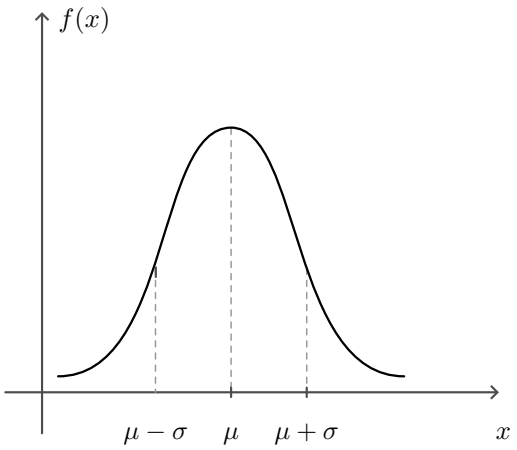


Figura 1.3.6: Distribuição normal

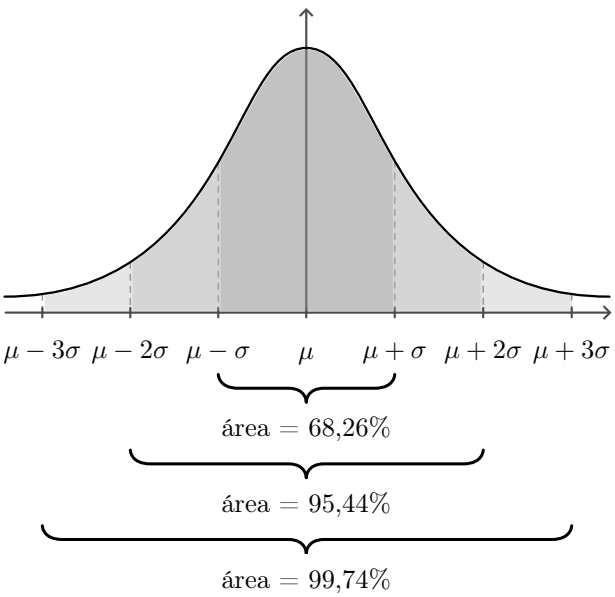


Figura 1.3.7: Frequência de probabilidades em uma distribuição normal

torno da média. Também pode-se observar que, quanto menor o desvio-padrão σ, mais concentrada é a curva em torno da média μ.

Também da Figura 1.3.7, nota-se que o valor esperado é $E(X) = \mu$, de tal forma que também pode-se mostrar que a variância é dada por $Var(X) = \sigma^2$.

Contudo, é mais conveniente e comum trabalhar com a **distribuição normal padrão** — ou distrbuição normal reduzida. Ta distribuição é obtida a partir da transformação da variável X em uma nova variável aleatória Z, conhecida como *z-score* (escore padrão), com média $\mu = 0$ e variância $\sigma^2 = 1$, determinada por:

$$Z = \frac{X - \mu}{\sigma} \tag{1.3.29}$$

Interpretamos a equação Equação (1.3.29) da seguinte maneira:

- Um *z-score* igual a 0 indica que o valor x da variável correspondente é exatamente igual à média — afinal, tem-se neste caso que $x = \mu$, de posse que $x - \mu = 0$.
- Um *z-score* positivo indica que o valor x da variável correspondente está acima da média. Mais especificamente, para $Z = 1$ tem-se que x está 1 desvio-padrão acima da média; para $Z = 2$, x está 2 desvios-padrões acima da média, e assim sucessivamente.
- Um *z-score* negativo indica que o valor x da variável correspondente está abaixo da média. Mais especificamente, para $Z = -1$ tem-se que x está 1 desvio-padrão abaixo da média; para $Z = -2$, x está 2 desvios-padrões abaixo da média, e assim sucessivamente.

De posse do que foi discutido anteriormente — e conforme apresentado na Figura 1.3.7 — podemos dizer também que aproximadamente 68% dos valores em uma distribuição normal padrão possuem *z-score* entre -1 e 1, 95% possuem *z-score* entre -2 e 2, enquanto 99% possuem *z-score* entre -3 e 3.

Esse tipo de transformação permite comparar distribuições de variáveis diferentes, com distintas métricas ou ordens de grandeza, porque não altera as formas das distribuições originais e gera novas variáveis com iguais valores de média e variância. É fácil perceber como pode-se transformar a representação de uma distribuição normal, dada pela Figura 1.3.6, com a aplicação de escores padrões — conferir a Figura 1.3.8.

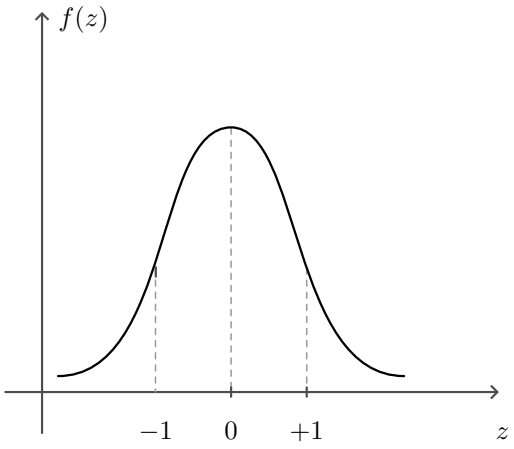


Figura 1.3.8: Distribuição normal padrão

A função de distribuição de probabilidade $f(z)$ é dada por:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-z^2/2} \tag{1.3.30}$$

Podemos obter a frequência acumulada em uma distribuição normal a partir da expressão seguinte:

$$F(x_c) = P(X \leq x_c) = \int_{-\infty}^{x_c} f(x)\, dx \tag{1.3.31}$$

Matematicamente, a Equação (1.3.31) corresponde ao cálculo da área sob a curva de $f(x)$, no intervalo $-\infty$ a x_c. Interpreta-se isso como a determinação de todas as probabilidades acumuladas de ocorrências de valores para a variável aleatória x, até um valor de referência x_c.

No caso de uma distribuição normal padrão, a frequência acumulada $F(z_c)$ pode ser obtida por:

$$F(z_c) = P(Z \leq z_c) = \int_{-\infty}^{z_c} f(z)\, dz = \frac{1}{2\pi} \cdot \int_{-\infty}^{z_c} e^{-z^2/2}\, dz \tag{1.3.32}$$

Dado que a Equação (1.3.32) não tem solução analítica simples — pessoalmente não tenho ideia de como *iniciar* a resolução — geralmente se recorre a tabelas de *z-score*, como apresentado em (FÁVERO; BELFIORE, 2025, p. 331) e (FÁVERO; BELFIORE, 2025, p. 2008–2009).

1.4 Exercícios complementares

Exercicio 1.1 Na análise de concessão de empréstimos, uma variável potencialmente importante é a renda da pessoa. O gerente de um banco coleta uma base de dados de seus correntistas e extrai a variável "renda mensal (R$)" para 50 pessoas. Embora se trate de uma variável quantitativa, deseja realizar uma análise por meio de tabela de frequências. Neste sentido, pede-se:

a) Classifique os correntistas em faixas de renda, sendo: 0-2.000; 2.001-4.000; 4.001-6.000; 6.001-8.000; 8.001-10.000 e 10.001-12.000.

b) Em seguida, elabore a tabela de frequências para as faixas de renda acima. O banco de dados está na planilha Lista de Exercício Complementares: aba Exercício 1.

Resolução

A Tabela 1.4.1 traz, para cada observação, a renda mensal da pessoa e a respectiva faixa de renda, de acordo com os critérios solicitados. A tabela de frequências 1.4.2 refere-se a essas faixas de renda.

Obs	Renda (R$)	Faixa	O	R	F
1	2.894,00	2.001–4.000	26	7.665,00	6.001–8000
2	3.448,00	2.001–4.000	27	3.890,00	2.001–4.000
3	1.461,00	0–2.000	28	6.590,00	6.001–8000
4	2.224,00	2.001–4.000	29	1.241,00	0–2.000
5	2.501,00	2.001–4.000	30	1.720,00	0–2.000
6	1.100,00	0–2.000	31	2.556,00	2.001–4.000
7	3.560,00	2.001–4.000	32	4.730,00	4.001–6000
8	5.511,00	4.001–6000	33	4.745,00	4.001–6000
9	2.901,00	2.001–4.000	34	8.550,00	10.001–12.000
10	10.128,00	10.001–12.000	35	3.860,00	2.001–4.000
11	1.855,00	0–2.000	36	11.320,00	10.001–12.000
12	3.161,00	2.001–4.000	37	6.125,00	6.001–8000
13	8.630,00	10.001–12.000	38	5.606,00	4.001–6000
14	6.201,00	6.001–8000	39	3.250,00	2.001–4.000
15	4.130,00	4.001–6000	40	1.500,00	0–2.000
16	2.736,00	2.001–4.000	41	9.216,00	10.001–12.000
17	4.448,00	4.001–6000	42	4.999,00	4.001–6000
18	2.150,00	2.001–4.000	43	3.900,00	2.001–4.000
19	4.595,00	4.001–6000	44	7.000,00	6.001–8000
20	5.561,00	4.001–6000	45	3.508,00	2.001–4.000
21	2.800,00	2.001–4.000	46	1.130,00	0–2.000
22	9.538,00	10.001–12.000	47	4.121,00	4.001–6000
23	2.000,00	0–2.000	48	2.601,00	2.001–4.000
24	3.226,00	2.001–4.000	49	2.901,00	2.001–4.000
25	1.900,00	0–2.000	50	4.871,00	4.001–6000

Tabela 1.4.1: Renda e faixa de renda dos correntistas

Faixa de renda	Frequência absoluta	Frequência relativa
0-2.000	9	18%
2.001-4.000	19	38%
4.001-6.000	11	22%
6.001-8.000	5	10%
8.001-10.000	4	8%
10.001-12.000	2	4%
Total	**50**	**100%**

Tabela 1.4.2: Tabela de frequências para as faixas de renda

□

Exercicio 1.2 Um analista do mercado acionário coletou os retornos mensais de duas ações que pretende indicar aos seus clientes. Calcule as estatísticas descritivas para as duas variáveis, incluindo o coeficiente de correlação entre os retornos. O banco de dados com os retornos percentuais mensais está na planilha Lista de Exercício Complementares: aba Exercício 2 — Tabela 1.4.3.

Meses	Ação 1	Ação 2
1	-0,0212	0,2645
2	0,2438	0,2086
3	0,2296	0,1248
4	-0,2018	0,0209
5	0,1296	0,2055
6	0,0615	0,6260
7	-0,1591	-0,1490
8	-0,1001	0,2580
9	-0,0265	0,1722
10	0,0776	0,0199
11	0,0370	0,4331
12	0,1116	0,5482
13	-0,0667	0,0452
14	-0,0082	-0,1410
15	0,0119	-0,1059
16	0,1205	0,4074
17	0,0477	-0,0056
18	0,2814	0,1482
19	-0,0674	0,0753
20	0,0762	0,0899
21	-0,1111	0,0160
22	-0,0557	0,1805
23	0,1991	0,0334

Tabela 1.4.3: Retornos mensais de duas ações

Resolução

No contexto dessa atividade, temos as seguintes estatísticas descritivas a serem determinadas:

Média Chamando de X o conjunto dos x retornos mensais da "Ação 1" e Y o conjunto dos y retornos mensais da "Ação 2", pode-se usar a Equação (1.1.1) para mostrar que os

valores médios $\overline{x}$ e $\overline{y}$ são iguais a:

$$\begin{aligned}\overline{x} &= \frac{-0.0212 + 0.2438 + 0.2296 + \ldots - 0.1111 - 0.0557 + 0.1991}{23} \\ &= \frac{0.8097}{23} \\ &= 0.0352\end{aligned}$$

$$\begin{aligned}\overline{y} &= \frac{0.2645 + 0.2086 + 0.1248 \ldots + 0.0160 + 0.1805 + 0.0334}{23} \\ &= \frac{3.4761}{23} \\ &= 0.1551\end{aligned}$$

Mediana Ordenando os elementos de X, temos que:

$$\begin{aligned}X = [&-0.2018, -0.1591, -0.1111, -0.1001, -0.0674, -0.0667, \\ &-0.0557, -0.0265, -0.0212, -0.0082, 0.0119, \\ &0.0370, \\ &0.0477, 0.0615, 0.0762, 0.0776, 0.1116, \\ &0.1205, 0.1296, 0.1991, 0.2296, 0.2438, 0.2814]\end{aligned}$$

Ordenando os elementos de Y, temos que:

$$\begin{aligned}Y = [&-0.1490, -0.1410, -0.1059, -0.0056, 0.0160, 0.0199, \\ &0.0209, 0.0334, 0.0452, 0.0753, 0.0899, \\ &0.1248, \\ &0.1482, 0.1722, 0.1805, 0.2055, 0.2086, \\ &0.2580, 0.2645, 0.4074, 0.4331, 0.5482, 0.6260]\end{aligned}$$

É possível observar que o termo central e, portanto, a mediana de X, é igual a 0.0370. De forma análoga, o termo central de Y, sua mediana, é dado por 0.1248.

Moda É possível observar que todos os elementos dos conjuntos X e Y possuem a mesma frequência—logo, não há moda.

Quartis Usando a Equação (1.1.5), determinaremos, para as variáveis x e y — referente aos valores de "Ação 1" e "Ação 2", respectivamente — as posições do segundo e do terceiro quartis — respectivamente iguais a p_{25} e p_{75} — os conjuntos possuem o mesmo número de observações; logo, os percentis encontram-se nas mesmas posições.

$$\begin{aligned}P(p_{25}) &= \left[(23-1) \cdot \left(\frac{25}{100}\right)\right] + 1 \\ &= 22 \cdot 0.25 + 1 \\ &= 6.5\end{aligned}$$

$$\begin{aligned}P(p_{75}) &= \left[(23-1) \cdot \left(\frac{75}{100}\right)\right] + 1 \\ &= 22 \cdot 0.75 + 1 \\ &= 17.5\end{aligned}$$

Nota-se que o primeiro quartil está localizado entre as observações 6 e 7 do rol. No cálculo da mediana, ordenamos as observações, de tal modo que pode-se determinar esses valores como sendo, respectivamente, iguais a -0.0667 e -0.0557, para a variável

x, e 0.0199 e 0.0209, para a variável y. Realizando a interpolação desses dados — como indicado na Sub-sub-seção 1.1.1.5 — tem-se que:

$$\begin{aligned} p_{x25} &= (-0.0667 \cdot 0.5) + (-0.0557 \cdot 0.5) \\ &= -0.0334 - 0.0279 \\ &= -0.0612 \end{aligned}$$

$$\begin{aligned} p_{y25} &= (0.0199 \cdot 0.5) + (0.0209 \cdot 0.5) \\ &= 0.0010 + 0.0105 \\ &= 0.0204 \end{aligned}$$

O terceiro quartil está localizado entre as observações 17 e 18. Interpolando-se esses dados para ambas as variáveis, temos que:

$$\begin{aligned} p_{x75} &= (0.1116 \cdot 0.5) + (0.1205 \cdot 0.5) \\ &= 0.0558 + 0.0603 \\ &= 0.1161 \end{aligned}$$

$$\begin{aligned} p_{y75} &= (0.2086 \cdot 0.5) + (0.2580 \cdot 0.5) \\ &= 0.1043 + 0.1290 \\ &= 0.2333 \end{aligned}$$

Amplitude As amplitudes A_x e A_y são dadas por:

$$A_x = x_{\text{máx}} - x_{\text{mín}} = 0.2814 - (-0.2018) = 0.4832$$

$$A_y = y_{\text{máx}} - y_{\text{mín}} = 0.6260 - (-0.1490) = 0.7750$$

Desvio médio Usando a Equação (1.1.7), podemos determinar os desvios médios D_{mx} e D_{my}:

$$\begin{aligned} D_{mx} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \overline{x}| = \frac{1}{23} \cdot \Big(\\ &|-0.0212 - 0.0352| + |0.2438 - 0.0352| + |0.2296 - 0.0352| \\ &+ \ldots + \\ &|-0.1111 - 0.0352| + |-0.0557 - 0.0352| + |0.1991 - 0.0352| \\ \Big) \Rightarrow \\ D_{mx} = \frac{2.3863}{23} \Rightarrow \\ D_{mx} = 0.1038 \end{aligned}$$

$$\begin{aligned} D_{my} = \frac{1}{n} \cdot \sum_{i=1}^{n} |y_i - \overline{y}| = \frac{1}{23} \cdot \Big(\\ &|0.2645 - 0.1551| + |0.2086 - 0.1551| + |0.1248 - 0.1551| \\ &+ \ldots + \\ &|0.0160 - 0.1551| + |-0.1805 - 0.1551| + |0.0334 - 0.1551| \\ \Big) \Rightarrow \\ D_{my} = \frac{3.5853}{23} \Rightarrow \\ D_{my} = 0.1559 \end{aligned}$$

Variância Utilizando a Equação (1.1.8), determinaremos as variâncias ${\sigma_x}^2$ e ${\sigma_y}^2$ para as variáveis x e y, respectivamente.

$$
\begin{aligned}
{\sigma_x}^2 &= \frac{1}{(n-1)} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{22} \cdot \Big(\\
&\quad (-0.0212 - 0.0352)^2 + (0.2438 - 0.0352)^2 + (0.2296 - 0.0352)^2 \\
&\quad + \ldots + \\
&\quad (-0.1111 - 0.0352)^2 + (-0.0557 - 0.0352)^2 + (0.1991 - 0.0352)^2 \\
&\quad \Big) \Rightarrow \\
{\sigma_x}^2 &= \frac{0.3674}{22} \Rightarrow \\
{\sigma_x}^2 &= 0.0167
\end{aligned}
$$

$$
\begin{aligned}
{\sigma_y}^2 &= \frac{1}{(n-1)} \cdot \sum_{i=1}^{n} (y_i - \overline{y})^2 = \frac{1}{22} \cdot \Big(\\
&\quad (0.2645 - 0.1551)^2 + (0.2086 - 0.1551)^2 + (0.1248 - 0.1551)^2 \\
&\quad + \ldots + \\
&\quad (0.0160 - 0.1551)^2 + (-0.1805 - 0.1551)^2 + (0.0334 - 0.1551)^2 \\
&\quad \Big) \Rightarrow \\
{\sigma_y}^2 &= \frac{0.9140}{22} \Rightarrow \\
{\sigma_y}^2 &= 0.0415
\end{aligned}
$$

Desvio-padrão Os desvios-padrões σ_x e σ_y são determinados a partir da Equação (1.1.9):

$$\sigma_x = \sqrt{{\sigma_x}^2} = \sqrt{0.0167} = 0.1292$$

$$\sigma_y = \sqrt{{\sigma_y}^2} = \sqrt{0.0415} = 0.2037$$

Erro padrão Usando a Equação (1.1.10), tem-se que os erros padrões $\sigma_{\overline{x}}$ e $\sigma_{\overline{y}}$ são dados por:

$$\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{0.1292}{\sqrt{23}} = 0.0269$$

$$\sigma_{\overline{y}} = \frac{\sigma_y}{\sqrt{n}} = \frac{0.2038}{\sqrt{23}} = 0.0425$$

Coeficiente de variação Os coeficientes de variação CV_x e CV_y podem ser determinados pela Equação (1.1.11):

$$CV_x = \frac{\sigma_x}{\overline{x}} = \frac{0.0167}{0.0352} = 0.4743$$

$$CV_y = \frac{\sigma_y}{\overline{y}} = \frac{0.0415}{0.1511} = 0.2749$$

Resumo A tabela 1.4.4 resume as estatísticas descritivas para as variáveis "Ação 1" e "Ação 2". □

Exercicio 1.3 Em certo jogo, probabilidade de vitória (sucesso) a cada nova jogada é 1/6. Se forem feitas 10 jogadas, quais são as seguintes probabilidades:

a) Ter vitória em 4 jogadas.

	Ação 1 (x)	Ação 2 (y)
Média	0.0352	0.1511
Mediana	0,0370	0,1248
Moda	não há	não há
Primeiro quartil	-0,0612	0,0204
Terceiro quartil	0,1161	0,2333
Amplitude	0,4832	0,775
Desvio médio	0,1038	0,1559
Variância	0,0167	0,0415
Desvio-padrão	0,1292	0,2038
Erro padrão	0,0269	0,0425
Coeficiente de variação	0,4743	0,2749

Tabela 1.4.4: Estatísticas descritivas para as variáveis “Ação 1” e “Ação 2”

b) Ter vitória em pelo menos 7 jogadas.

Resolução

No aguardo ... □

Capítulo 2

Introdução à programação com Python

Conferir documento em: `modulo-02/notes.md`

Referências

FÁVERO, Luiz Paulo; BELFIORE, Patricia. **Manual de Análise de Dados: Estatística e Machine Learning com Excel, SPSS, Stata, R e Pyhton**. 2ª ed. [S.l.]: LTC, 2025. ISBN 978-85-9515-993-8.

LARSON, Ron; FARBER, Betsy. **Estatística Aplicada**. 6ª ed. [S.l.]: Pearson, 2016. ISBN 978-85-4301-811-9.

Notas