Distribuições

Luan Fiorentin

UFPR - DEST

2019-04-14

Sumário

Introdução

- Os **Experimentos aleatórios** são aqueles que, se repetidos sob uma mesma condição, geram resultados diferentes.

- Os **fenômenos aleatórios** em estudo se comportam de forma diferente em cada amostra.

- A **distribuição de probabilidade** descreve o comportamento de um fenômeno aleatório.

- As distribuições de probabilidade univariadas podem variar em complexidade, mas todas possuem **ingredientes básicos**:
    - Há uma variável aleatória.
    - Há um ou mais parâmetros que indexam o modelo probabilístico.

Distribuições Discretas e Contínuas

- Dada a realização de um experimento aleatório qualquer, com um certo espaço de probabilidade, deseja-se estudar a **estrutura probabilística** de quantidades associadas a esse experimento.

- Antes da realização de um experimento, não sabemos seu resultado, entretanto seu **espaço de probabilidade** pode ser previamente estabelecido.

- Dessa forma, podemos atribuir probabilidades aos eventos desse espaço amostral, dando origem ao conceito de **variável aleatória** (V. A.).

- **Definição**: em probabilidade, variável aleatória é uma função X que associa a cada evento do espaço amostral um número real $X(\omega) \in \mathbb{R}$.

- A variável aleatória pode ser **discreta** ou **contínua**, dependendo do domínio de X.

- Em geral, denota-se a **probabilidade** de uma V. A. X assumir determinado valor x como

$$P[X] \quad \text{ou} \quad P[X = x].$$

- Existem diversos **modelos probabilísticos** que procuram descrever as V. A., denominados de **distribuições de probabilidade** de V. A. (discretas ou contínuas).

- A **distribuição de probabilidade** é a descrição das probabilidades associadas com os possíveis valores de X.

- Os valores que X assume determinam o **suporte** (S) da V. A.:
    - **Variáveis discretas**: suporte em um conjunto de valores enumeráveis (finitos ou infinitos).
    - **Variáveis contínuas**: suporte em um conjunto não enumerável de valores.

Dada uma variável de interesse, denomina-se de **distribuição de probabilidade** de alguma V. A., a regra geral que define a

- **Função de probabilidade** (fp): para variáveis aleatórias discretas;
- **Função densidade de probabilidade** (fdp): para variáveis aleatórias contínuas.

- Devido a importância prática, algumas distribuições são mais estudadas:
  - **Binomial** (discreta);
  - **Poisson** (discreta);
  - **Hipergeométrica** (discreta);
  - **Exponencial** (contínua);
  - **Normal** (contínua).

Distribuição Binomial

Antes de introduzir a distribuição Binomial é preciso apresentar a distribuição de Bernoulli!

A variável aleatória X segue a distribuição Bernoulli se assume apenas os valores 0 ("fracasso") ou 1 ("sucesso"). A função de probabilidade é dada por

$$P[X = x] = p^x (1 - p)^{1-x},$$

onde: $x = 0, 1$; e o parâmetro $0 < p < 1$ é a probabilidade de sucesso.

Notação: $X \sim Ber(p)$

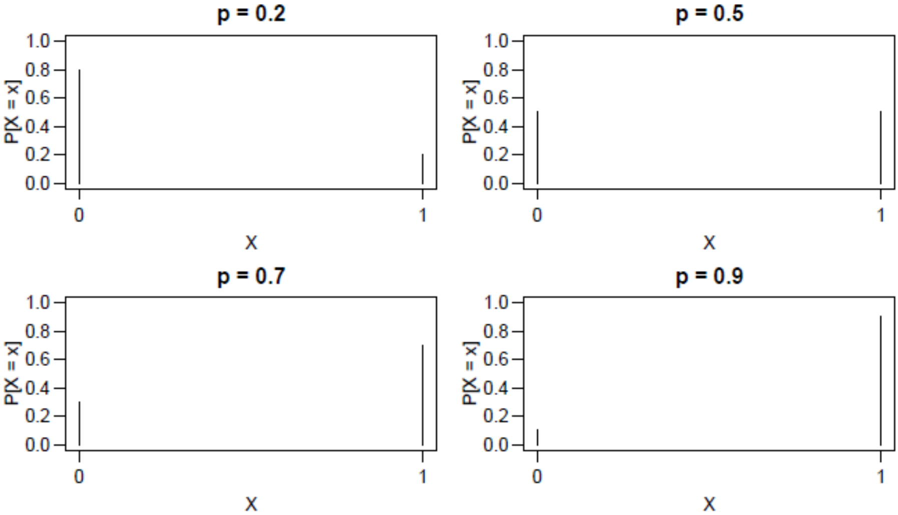

Figura 1: Distribuição Bernoulli.

Exemplo:

No lançamento de uma moeda, considere cara como o evento de sucesso (1), e coroa como fracasso (0).

Qual a probabilidade de sair cara, sendo que $p = 1/2$?

X	$P[X = x]$	$p = 1/2$
0	$1 - p$	$1/2$
1	p	$1/2$

- Seja um **experimento aleatório** realizado dentro das seguintes condições:
    - São realizados n "ensaios" de **Bernoulli** independentes.
    - Cada ensaio só pode ter dois resultados possíveis: "**sucesso**" (1) ou "**fracasso**" (0).
    - A probabilidade p de sucesso em cada ensaio é **constante**.

- Associa-se a variável aleatória X o número de **sucessos** em n **ensaios independentes de Bernoulli**.

- Portanto, X poderá assumir os valores $0, 1, ..., n$.

Distribuição de probabilidade Binomial para um **caso genérico**:

- Suponha que ocorram **sucessos** (1) apenas nos x primeiros ensaios, e **fracassos** (0) nos $n - x$ ensaios restantes, tal como

$$1, 1, 1, ..., 1, 0, 0, 0, ..., 0.$$

- Dado que os ensaios são independentes, a probabilidade de ocorrência de x sucessos em n tentativas é uma **extensão do modelo de Bernoulli** para n ensaios, ou seja,

$$p.p.p...p.(1-p).(1-p).(1-p)...(1-p) = p^x(1-p)^{n-x}.$$

- No entanto, o evento "x sucessos em n ensaios" pode ocorrer de **diferentes maneiras** (ordens), todas com a mesma probabilidade.

- Como o número de ordens é o número de **combinações** de n elementos tomados x a x, temos que

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

A variável aleatória X, com distribuição Binomial de parâmetros n e p, tem sua função de probabilidade dada por

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

onde: $x = 0, 1, 2, ..., n$.
A média (esperança) e a variância são dadas, respectivamente, pelas expressões

$$\mu = np,$$

$$\sigma^2 = np(1-p).$$

Notação: $X \sim B(n, p)$

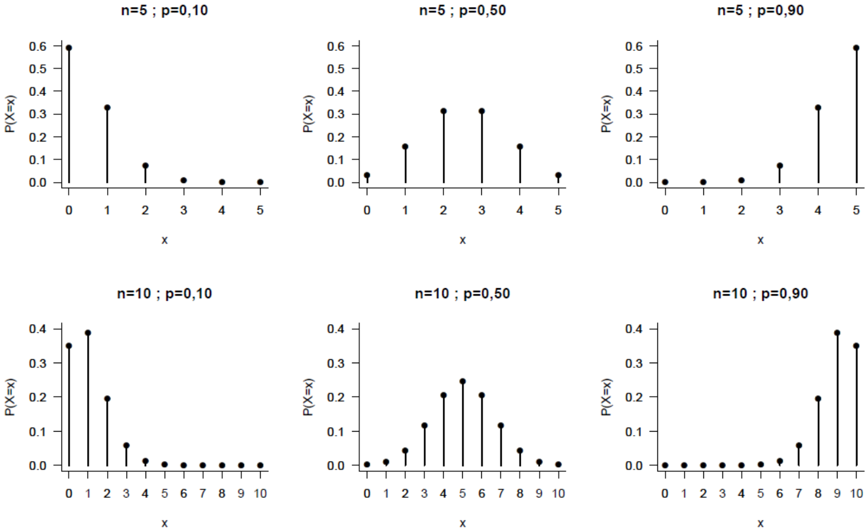

Figura 2: Distribuição Binomial.

Exercício 1

Um processo de produção opera a uma taxa de 5% de peças produzidas não conformes. A cada hora uma amostra de 5 unidades do produto é retirada, e o número de peças não conformes (defeituosas) registrado.

1. Qual a probabilidade de duas peças defeituosas serem verificadas numa realização da inspeção?
2. A empresa será multada se dois ou mais itens defeituosos forem verificados na próxima inspeção. Qual a probabilidade de multa?

Exercício 1

1. $$P(X = 2) = \binom{5}{2} 0,05^2 (1 - 0,05)^{5-2} = 0,021.$$

2. $$1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1)]$$

$$1 - P(X < 2) = 1 - [0,774 + 0,204] = 0,023.$$

```r
# P(X = 2)
dbinom(x = 2, size = 5, prob = 0.05)
```

```
## [1] 0.02143438
```

```r
# P(X >= 2) = 1-[P(X < 2)] = 1-[P(X = 0) + P(X = 1)]
1 - sum(dbinom(x = 0:1, size = 5, prob = 0.05))
```

```
## [1] 0.0225925
```

Distribuição Poisson

- Seja um **experimento** realizado nas seguintes condições:
  - As ocorrências são **independentes**.
  - As ocorrências são **aleatórias**.
  - A variável aleatória X é o número de ocorrências de um evento **ao longo de algum intervalo** (de tempo, espaço, volume, área, ...).

- Denominamos esse experimento de **Processo de Poisson**.

- Vamos associar a V. A. X o **número de ocorrências em um intervalo**. Portanto X poderá assumir os valores $0, 1, 2, ...,$ (sem limite superior).

- Aplicável no caso de eventos que ocorrem sob **taxa constante**.

Exemplos:

- Usuários de computador ligados à internet.
- Clientes chegando ao caixa de um supermercado.
- Acidentes com automóveis em uma determinada estrada.
- Número de carros que chegam a um posto de gasolina.
- Número de aviões sequestrados em um dia.
- Número de falhas em componentes por unidade de tempo.
- Número de árvores atacadas por pragas em um dia.
- Número de peças defeituosas substituídas num veículo durante o primeiro ano de vida.

A variável aleatória X, com distribuição Poisson de parâmetro λ, tem sua função de probabilidades dada por

$$P[X = x] = \frac{\lambda^x e^{-\lambda}}{x!},$$

onde: $x = 0, 1, 2, ..., n$ e $\lambda > 0$.
A média (esperança) e a variância são dadas, respectivamente, pelas expressões

$$\mu = \lambda,$$

$$\sigma^2 = \lambda.$$

Notação: $X \sim P(\lambda)$

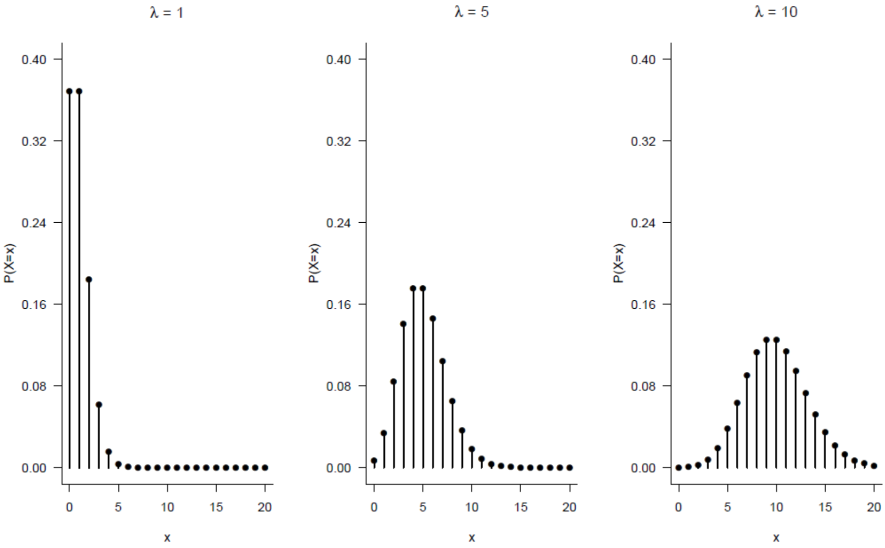

Figura 3: Distribuição Poisson.

Exercício 2

A emissão de partículas radioativas têm sido modelada através de uma distribuição de Poisson, com o valor do parâmetro dependendo da fonte utilizada.

Suponha que o número de partículas alfa, emitidas por minuto, seja uma variável aleatória seguindo o modelo Poisson com parâmetro 5, isto é, a taxa média de ocorrência é de 5 emissões a cada minuto.

1. Qual a probabilidade de haver duas emissões em um minuto?
2. Calcule a probabilidade de haver mais de 2 emissões em um minuto?

Exercício 2

1. $$P(X = 2) = \frac{5^2 e^{-5}}{2!} = 0,084.$$

2. $$1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1)]$$

$$1 - P(X < 2) = 1 - [0,007 + 0,034] = 0,959.$$

```r
# P(X = 2)
dpois(x = 2, lambda = 5)
```

```
## [1] 0.08422434
```

```r
# P(X >= 2) = 1-[P(X < 2)] = 1-[P(X = 0) + P(X = 1)]
1 - sum(dpois(x = 0:1, lambda = 5))
```

```
## [1] 0.9595723
```

Distribuição Hipergeométrica

- A **distribuição Hipergeométrica** tem sido utilizada com frequência em:
  - Aplicações no pôquer.
  - Controle estatístico de qualidade.
  - Amostragem de aceitação.
  - Tamanho de populações (animais, plantas, . . .).

- Considere um conjunto de N objetos dos quais D são do tipo I e $N - D$ são do tipo II.

- Para um sorteio de n objetos $n < N$, feito ao acaso e sem reposição, defina X como o número de objetos de tipo I selecionados.

A variável aleatória X, com distribuição Hipergeométrica de parâmetros D, N e n, tem sua função de probabilidades dada por

$$P[X = x] = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}},$$

onde: $x = 0, 1, 2, ..., min(n, D)$ e $x = max\{0, n - (n - D)\} \leq x \leq min\{n, D\}$.
A média (esperança) e a variância são dadas, respectivamente, pelas expressões

$$\mu = \frac{nD}{N},$$

$$\sigma^2 = \frac{nD}{N}\left(\frac{N - D}{N}\right)\left(\frac{N - n}{N - 1}\right).$$

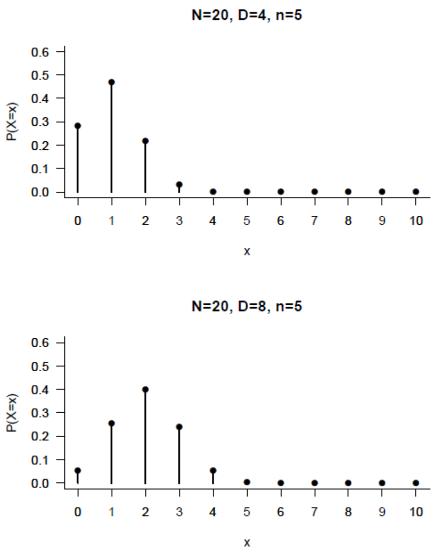

Figura 4: Distribuição Hipergeométrica.

Exercício 3

Considere que, num lote de 20 peças, existam o total de 4 peças defeituosas.
Selecionando-se 5 dessas peças, sem reposição, qual seria a probabilidade de 2 defeituosas terem sido escolhidas?

Exercício 3

$$P[X=2] = \frac{\binom{4}{2}\binom{16}{3}}{\binom{20}{5}} = 0,217.$$

```r
# x = x = 2
# m = D = 4
# n = N-D = 16
# k = n = 5

# P(X = 2)
dhyper(x = 2, m = 4, n = 16, k = 5)
```

```
## [1] 0.2167183
```

Distribuição Exponencial

- A **distribuição exponencial** é frequentemente empregada na modelagem do **tempo de vida**.

- Em geral, as **aplicações** ocorem na análise do:
  - Tempo de vida de de equipamentos.
  - Tempo de vida de de componentes.
  - Tempo de vida de de sistemas.
  - Tempo de vida de populações.

Uma variável aleatória contínua X, assumindo valores não negativos, segue o modelo exponencial com parâmetro $\alpha > 0$ se sua densidade é dada por

$$f(x) = \alpha e^{-\alpha x},$$

onde $x \geq 0$, e assume valor 0 caso contrário.
A média (esperança) e a variância são dadas, respectivamente, pelas expressões

$$\mu = \frac{1}{\alpha},$$

$$\sigma^2 = \frac{1}{\alpha^2}.$$

Notação: $X \sim E(\alpha)$

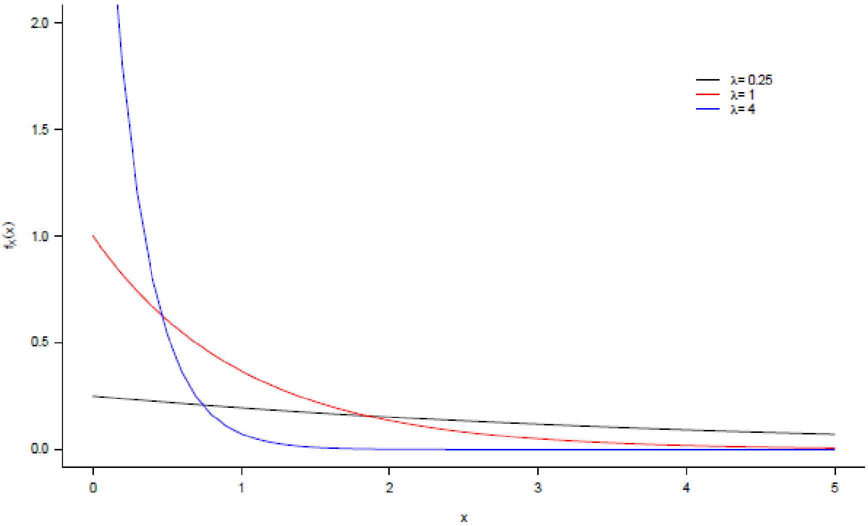

Figura 5: Distribuição Exponencial

- No entanto, para calcular probabilidades com a **distribuição Exponecial** é necessário resolver a **integral** correspondente, pois não tem-se figuras geométricas simples na maior parte dos exemplos.

- Assim,

$$P(a < X < b) = \int_a^b \alpha e^{-\alpha x} dx = -e^{-\alpha x}\big|_a^b = e^{-\alpha a} - e^{-\alpha b}.$$

Exercício 4

Uma indústria fabrica lâmpadas especiais que ficam em operação continuamente.
A empresa oferece a seus clientes a garantia de reposição, caso a lâmpada dure menos de 50 horas. A vida útil dessas lâmpadas é modelada através da distribuição Exponencial com parâmetro 1/8000.
Determine a proporção de troca por defeito de fabricação.

Exercício 4

$$P(0 < X < 50) = \int_0^{50} \frac{1}{8000} e^{-\frac{x}{8000}} dx$$

$$P(0 < X < 50) = -e^{-\frac{x}{8000}} \Big|_0^{50} = e^{-\frac{0}{8000}} - e^{-\frac{50}{8000}}$$

$$P(0 < X < 50) = 0,00623.$$

```r
# P(0 < X < 50)
pexp(q = 50, rate = 1/8000)
```

```
## [1] 0.006230509
```

Distribuição Normal

- A distribuição Normal é a mais **importante** distribuição de probabilidade para descrever variáveis aleatórias contínuas.
- Justifica-se pelo grande número de **aplicações** que a utilizam:
  - Altura.
  - Pressão arterial.
  - Medidas de testes psicológicos.
  - Tempo de vida útil de um dispositivo eletrônico.
  - Índices climáticos.
- Possui capacidade de **aproximar** outras distribuições e, também, pela grande aplicação na **inferência estatística**.

Uma variável aleatória contínua X tem distribuição normal com parâmetros μ e σ^2, se sua função de densidade é dada por

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

onde $-\infty < x < +\infty$, $\mu \in \Re$ é a média da população e $\sigma \in \Re^+$ é o desvio padrão populacional.
A média (esperança) e a variância são dadas, respectivamente, pelas expressões

$$E[X] = \mu,$$

$$Var[X] = \sigma^2.$$

Notação: $X \sim N(\mu, \sigma^2)$

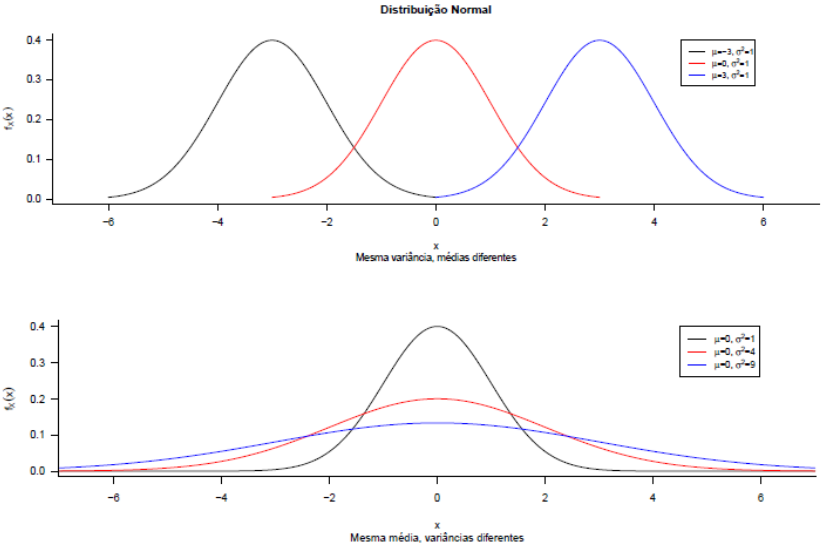

Figura 6: Distribuição Normal

Características da **curva normal**:

- É **simétrica** em relação à μ.
- O **ponto máximo** (moda) de $f(x)$ é o ponto $x = \mu$.
- Os **pontos de inflexão** da função são $\mu - \sigma$ e $\mu + \sigma$.
- A **área total** sob a curva é 1.
- A curva é **assintótica** em relação ao eixo x.

- Para qualquer **variável aleatória normal** X, valem as seguintes relações:

$$P[X > \mu] = P[X < \mu]$$

$$P[\mu - \sigma < X < \mu + \sigma] \cong 0,6827$$

$$P[\mu - 2\sigma < X < \mu + 2\sigma] \cong 0,9545$$

$$P[\mu - 3\sigma < X < \mu + 3\sigma] \cong 0,9973$$

Portanto, 6σ é frequentemente referida como a largura de uma distribuição normal.

- Métodos mais avançados de **integração** podem ser utilizados para mostrar que a área sob a função densidade de probabilidade normal de $-\infty < x < +\infty$ é igual a 1.

- Para obter uma **probabilidade do modelo normal**, devemos calcular a área entre os pontos a e b, ou seja,

$$P[a < X < b] = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] dx.$$

- Porém, essa função não possui forma fechada, e o cálculo de probabilidades pode ser feito apenas por aproximações numéricas.

- Para contornar esse problema, os valores de probabilidade são obtidos para uma distribuição **normal padrão** (Z), com $\mu = 0$ e $\sigma^2 = 1$,

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1).$$

- Pode-se fazer uma única tabela com as integrais aproximadas de Z, ao invés de uma tabela para cada par (μ, σ^2).

- Se $Z \sim N(0,1)$, então a fdp é

$$f(z) = \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}z^2\right].$$

- Para se obter a probabilidade de Z estar entre a e b,

$$P[a < Z < b] = \int_a^b \frac{1}{\sqrt{2\pi}} exp\left[-\frac{1}{2}z^2\right]dz.$$

- As **integrais** (áreas) para valores de Z entre 0,00 e 3,99 estão em tabela. Portanto, para qualquer valor de X entre a e b, podemos calcular a probabilidade correspondente através da transformação

$$P[a < X < b] = P\left[\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right].$$

Exercício 5A

As vendas diárias de um confeitaria no centro de uma cidade tem distribuição normal, com média igual R\$ $450,00$ por dia e desvio padrão igual a R\$ $95,00$.
Qual é a probabilidade das vendas excederem R\$ $700,00$ em determinado dia?

Exercício 5A

$$Z = \frac{X - \mu}{\sigma} = \frac{700 - 450}{95} = 2,63.$$

$$P(Z > 2,63) = 0,0043.$$

```r
# P(Z > 2,63)
1 - pnorm(q = 2.63, mean = 0, sd = 1, lower.tail = T)
```

```
## [1] 0.004269243
```

Exercício 5B

Suponha que entre pacientes de um determinado hospital o nível de glicose tenha uma distribuição aproximadamente Normal de média 105 mg por 100 ml e um desvio padrão 9 mg por 100 ml. Qual a proporção de diabéticos que tem níveis entre 90 e 125 mg por 100 ml?

Exercício 5B

$$Z = \frac{X - \mu}{\sigma} = \frac{90 - 105}{9} = -1,67.$$

$$Z = \frac{X - \mu}{\sigma} = \frac{125 - 105}{9} = 2,22.$$

$$P(-1,67 < Z < 2,22) = P(Z < 2,22) - P(Z < -1,67)$$

$$P(-1,67 < Z < 2,22) = 0,9868 - 0,0475 = 0,9393.$$

```r
# P(-1,67 < Z < 2,22) = P(Z < 2,22) - P(Z < -1,67)
pnorm(q = 2.22) - pnorm(q = -1.67)
```

```
## [1] 0.9393309
```