Distribuições contínuas: Normal

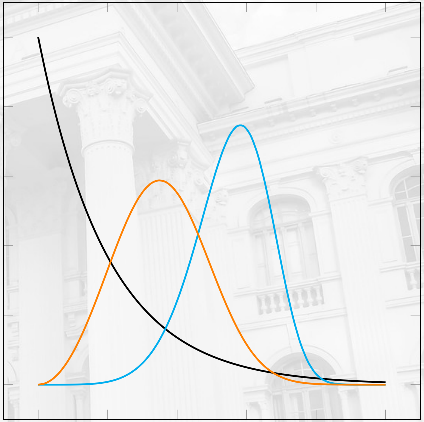
Prof. Walmes M. Zeviani

Departamento de Estatística
Universidade Federal do Paraná

Conteúdo

Neste vídeo

- A distribuição Normal.
    - Sua importância e fundamentação.
    - Definição e propriedades.
    - A Normal Padrão e tabela de probabilidades.
    - Aproximações com a distribuição Normal.
- Exemplos de aplicação.



A Distribuição Normal

Importância da distribuição Normal

A distribuição Normal é a **mais importante** distribuição contínua, dentre outros motivos porque:

1. Modela adequadamente a distribuição de um grande número de variáveis aleatórias contínuas.
2. Serve de aproximação para diversas outras distribuições contínuas e discretas.
3. Tem papel importante na Teoria Estatística, fundamentando a obtenção de inferências em diferentes contextos.

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

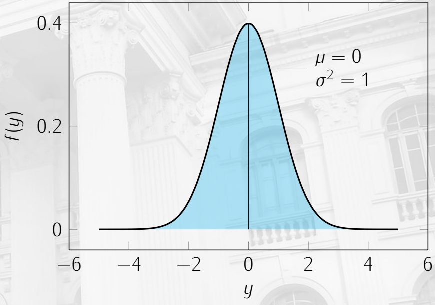


Figura 1. A distribuição Normal.

Características de uma v.a. com distribuição Normal

Características

- Variável aleatória contínua não limitada.
- Comportamento simétrico.
- Resultado do efeito de muitos fatores de pequena contribuição.

Aspectos históricos

- Desenvolvida no século XVIII pelo alemão Carl Friedrich Gauss (1777 - 1855).
- Também chamada de distribuição Gaussiana.
- Também obtida pelo francês Abraham de Moivre (1667 - 1754).

Figura 2. Gauss na nota da moeda Alemã.

Uma justificativa natural para a Normal

- Y: resultado do efeito de **muitos fatores** de contribuição **pequena**.
- Características governadas por muitos **genes**: altura, peso, produtividade, tolerância, etc.
- Características governadas por condições **ambientais**: crescimento de bactérias, decomposição de madeira, etc.
- Processos sujeitos a muitos **erros**: processos de mensuração, processos laboratoriais/industriais, etc.

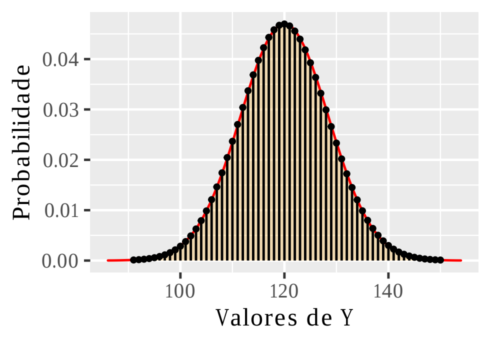


Figura 3. Aproximação da Binomial pela Normal.

Exemplos de v.a. com distribuição Normal

1. Peso ao nascer de um bebê.
2. Produtividade de uma lavoura.
3. Volume de líquido em uma garrafa de vinho.
4. Comprimento de um pão francês.
5. Tempo de deslocamento em um trajeto.
6. Retorno de um investimento.
7. Distância dos tiros ao alvo em provas de pontaria.
8. Altura de uma planta.
9. Produção diária de leite por vaca.
10. Quantidade consumida de ração pelo rebanho.
11. Volume diário de urina de um paciente.

Figura 4. Foto de Shanice McKenzie no Pexels.

Distribuição Normal

A variável aleatória Y tem distribuição Normal com parâmetros $\mu \in \mathbb{R}$ e $\sigma^2 > 0$ se sua função densidade de probabilidade é dada por

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < y < \infty.$$

Denotamos por $Y \sim \text{N}(\mu, \sigma^2)$.

A média e variância de Y são

$$\text{E}(Y) = \mu \quad \text{e} \quad \text{V}(Y) = \sigma^2,$$

portanto, média e variância são diretamente os parâmetros da distribuição.

Gráfico da distribuição Normal

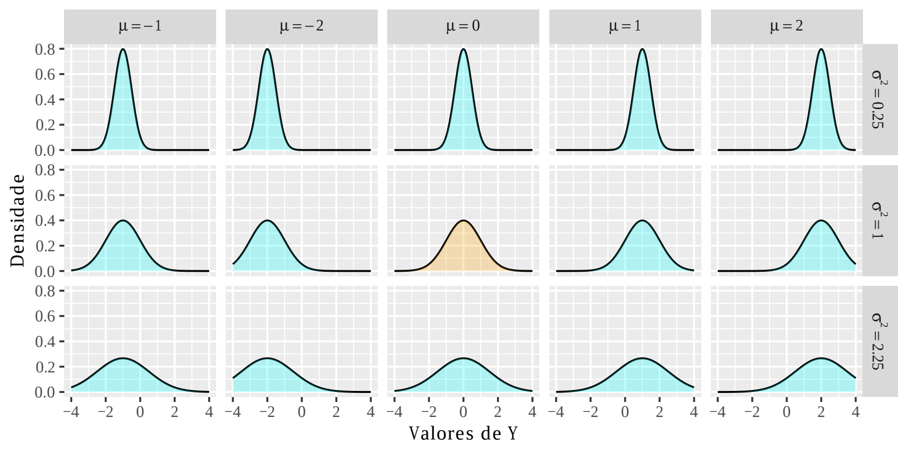


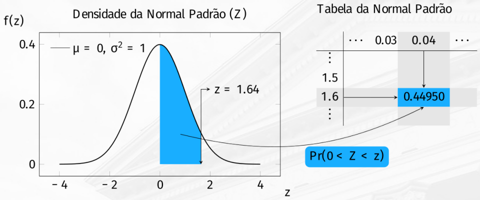
Figura 5. Gráficos para a distribuição Normal.

Cálculo de probabilidades com a distribuição Normal

Distribuição Normal Padrão

- Probabilidades associadas à distribuição Normal não podem ser obtidas analiticamente (mas sim numericamente), pois a integral correspondente **não tem forma fechada**.
- Como recursos, temos os *softwares* estatísticos ou a consulta a **tabelas da distribuição Normal Padrão** ($\mu = 0, \sigma^2 = 1$).
- Ou seja, basta saber determinar probabilidades na Normal Padrão para calcular probabilidades para qualquer v.a. Normal.
- A consulta a tabela da Normal Padrão se justifica pelo fato que se $Y \sim \text{N}(\mu, \sigma^2)$, então
$$Z = \frac{Y - \mu}{\sigma} \sim \text{N}(\mu = 0, \sigma^2 = 1).$$
Então Z representa o caso em que $Y \sim \text{N}(0, 1)$.
- Dessa forma, pelo caminho contrário, tem-se que
$$Y = \mu + \sigma \cdot Z \sim \text{N}(\mu, \sigma^2).$$

A tabela da Normal Padrão



	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891
1	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41309	0.41466	0.41621	0.41774
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408
1.6	0.44520	0.44630	0.44738	0.44845	**0.44950**	0.45053	0.45154	0.45254	0.45352	0.45449
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327

Figura 6. Fragmento de uma tabela de probabilidades da Normal Padrão.

Tipos de tabelas da Normal Padrão

1. Tabelas com probabilidades acumuladas até z,

$$\text{P}(Z < z) = \text{P}(-\infty < Z < z) = \int_{-\infty}^{z} f(u)\,\text{d}u.$$

2. Tabelas com probabilidades acumuladas de 0 a $z > 0$,

$$\text{P}(0 < Z < z) = \int_{0}^{z} f(u)\,\text{d}u.$$

3. Tabelas com probabilidades nas caudas da distribuição,

$$\text{P}(Z > |z|) = \text{P}(Z < -z) + \text{P}(Z > z).$$

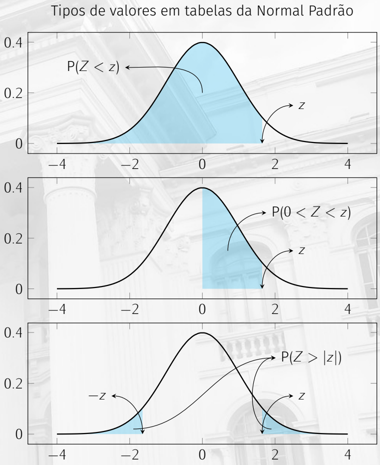


Figura 7. Tipos de valores em tabelas de probabilidades da Normal Padrão.

Uso da tabela da Normal Padrão

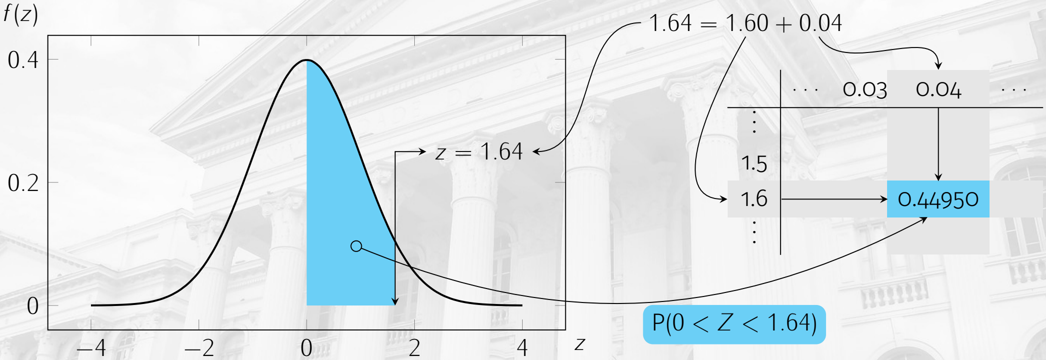


Figura 8. Uso da tabela da Normal Padrão que retorna as probabilidades entre 0 e z.

Exemplo: envase de detergente

Uma máquina envasa detergente líquido de forma que o volume tem distribuição Normal com média $\mu = 5025$ ml e variância 625 ml^2.

1. Qual a proporção de embalagens com menos de 5000 ml (5L, valor do rótulo)?
2. Se a regulamentação diz que tolera apenas 10% das embalagens com volume inferior a 5000 ml, no mínimo em quanto o fabricante deve deslocar a média da distribuição para estar dentro das especificações?

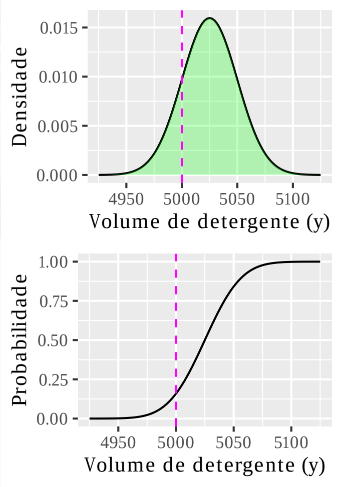


Figura 9. Gráficos da distribuição para o exercício.

Solução

1. A probabilidade corresponde à integral de $f(y, \mu, \sigma^2)$ de $-\infty$ a 5000. Aqui será usada a tabela da Normal Padrão. Para isso, tem-se que

$$z = \frac{y - \mu}{\sigma} = \frac{5000 - 5025}{25} = -1$$
$$\text{P}(Y < 5000) = \text{P}(Z < -1) = 0.1587.$$

2. Pelo uso na tabela no sentido inverso (da célula para as margens) obtém-se os percentis, ou seja, o valor z tal que $F(z)$ é 0.1. Esse valor é $z = -1.2816$. Na escala de Y, o valor é

$$y = z \cdot \sigma + \mu = -1.2816 \cdot 25 + 5025 = 4992.96.$$

Dessa forma, o fabricante deve deslocar a média em no mínimo $m = 5000 - 4992.96 = 7.04$ ml para ter apenas 10% de embalagens com menos de 5000 ml.

Exemplo: classificação de laranjas

Uma cooperativa de produtores rurais de laranja possui uma máquina de beneficiamento das frutas. A máquina lava e escova os frutos, bem como classifica conforme o tamanho. Os produtores destinam a produção conforme o mercado (suco, consumo em natura, etc.) fazendo a classificação pelo diâmetro das frutas (ver figura a seguir). Suponha que o diâmetro dos frutos tenha distribuição Normal com média 9 cm e desvio-padrão 0.5 cm.

1. Se a máquina divide os frutos pelos diâmetros 8.6 e 9.7 cm, qual a proporção esperada de frutos em cada classe?
2. Qual o diâmetro que divide os frutos em grupos de diâmetro menor com 40% deles e 60% em outro grupo?

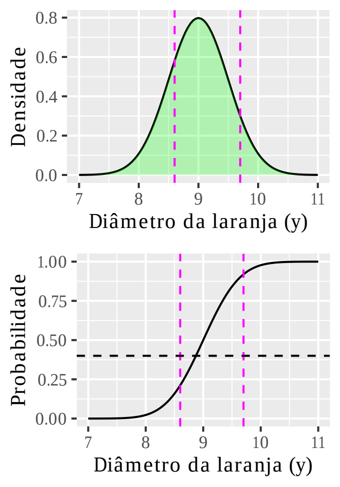


Figura 10. Gráficos da distribuição para o exercício.

Exemplo: classificação de laranjas (ilustração)

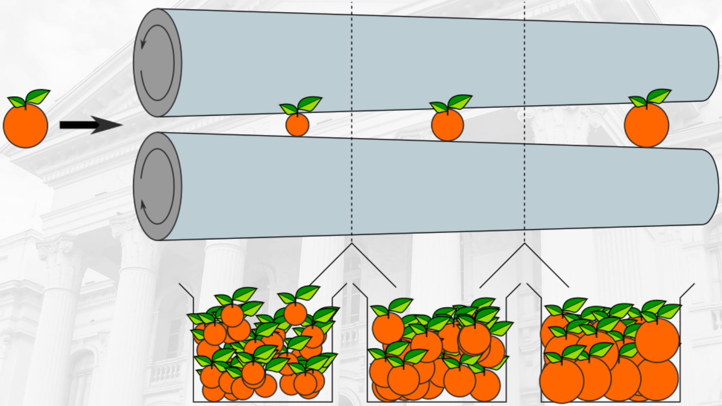


Figura 11. Ilustração da classificação das laranjas pelo diâmetro.

Solução

1. Para determinar as proporções, basta consultar a tabela ou fazer a integração,

$$z_1 = \frac{8.6 - 9}{0.5} = -0.8 \quad \text{e} \quad z_2 = \frac{9.7 - 9}{0.5} = 1.4$$

$$P(Z < -0.8) = 0.2119$$

$$P(-0.8 < Z < 1.4) = P(Z < 1.4) - P(Z < -0.8) = 0.7073$$

$$P(Z > 1.4) = 1 - P(Z \leq 1.4) = 1 - 0.9192 = 0.0808$$

2. Para isso, consulta-se a tabela da forma inversa ou obtém-se o percentil no software,

$$z \text{ tal que } P(Z < z) = 0.4 \text{ é } -0.2533471$$

$$y = z \cdot \sigma + \mu = -0.2533471 \cdot 0.5 + 9 = 8.8733.$$

Propriedades da distribuição Normal

Intervalos simétricos em relação à média

A distribuição Normal é **simétrica**. Seguem algumas características sobre isso.

- Média, mediana e moda **coincidem**.
- Se $Y \sim \mathrm{N}(\mu, \sigma^2)$, então
  - $\mathrm{P}(\mu - \sigma < Y < \mu + \sigma) \approx 0.6826$.
  - $\mathrm{P}(\mu - 2\sigma < Y < \mu + 2\sigma) \approx 0.9546$.
  - $\mathrm{P}(\mu - 3\sigma < Y < \mu + 3\sigma) \approx 0.9973$.
- Pela simetria em relação a μ
  - $\mathrm{P}(Y < \mu - \delta) = \mathrm{P}(Y > \mu + \delta), \quad$ para qualquer δ.
  - $\mathrm{P}(Y > |\mu + \delta|) = \mathrm{P}(Y < \mu - \delta) + \mathrm{P}(Y > \mu + \delta) = 2 \cdot \mathrm{P}(Y > \mu + \delta)$.

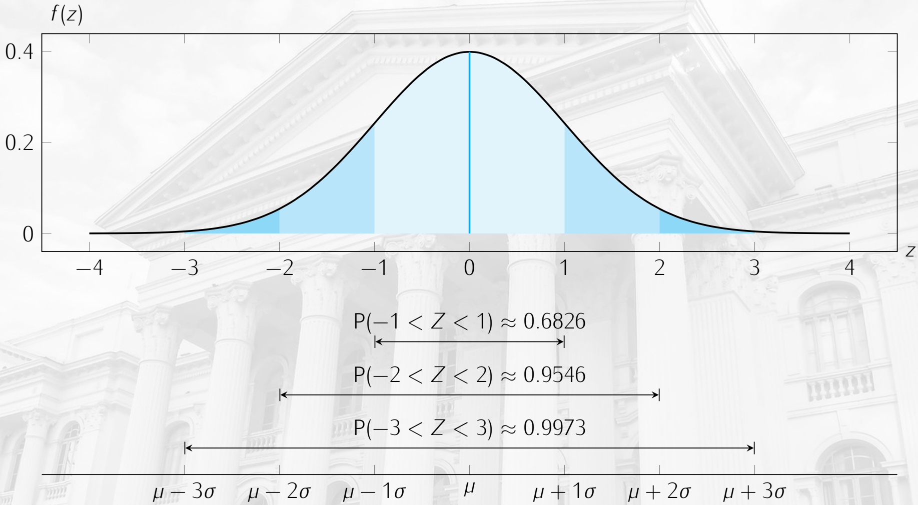


Figura 12. Probabilidades em intervalos simétricos em relação à média.

Combinações lineares de v.a. Normais

- **Combinações lineares** de variáveis aleatórias com distribuição Normal também têm distribuição Normal.
- Considere então k v.a. de distribuição Normal,

$$Y_1 \sim \text{N}(\mu_1, \sigma_1^2), Y_2 \sim \text{N}(\mu_2, \sigma_2^2), \cdots, Y_k \sim \text{N}(\mu_k, \sigma_k^2),$$

  todas **independentes entre si**, e $a_1, a_2, \cdots, a_k$ um conjunto de constantes. Então, a v.a.

$$Y_{\text{comb}} = a_1 Y_1 + a_2 Y_2 + \cdots + a_k Y_k$$

  tem distribuição Normal com média e variância

$$\mu_{\text{comb}} = a_1 \mu_1 + a_2 \mu_2 + \cdots + a_k \mu_k,$$
$$\sigma^2_{\text{comb}} = a_1^2 \sigma_1^2 + a_2^2 \sigma_2^2 + \cdots + a_k^2 \sigma_k^2.$$

Exemplo: duração de atendimentos

A duração do atendimento de cada cliente no caixa de um supermercado tem distribuição Normal com tempo médio de atendimento de 8 minutos e desvio-padrão de 2 minutos.

1. Se um cliente entra na fila tendo 5 clientes antes dele, considerando independência entre os atendimentos, em média, quanto tempo ele irá esperar para ser atendido?
2. Qual a probabilidade de ser atendido, após os 5 clientes, antes de 30 minutos?

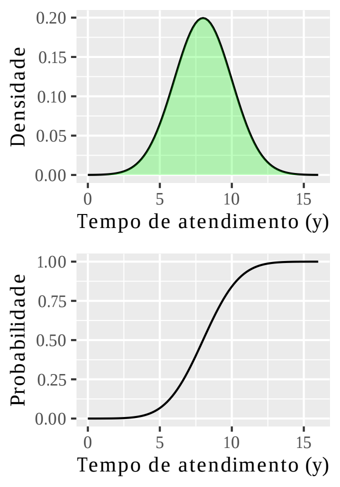


Figura 13. Gráficos da distribuição para o exercício.

Solução

1. Usa-se a propriedade da média. O tempo de atendimento de cada é Y_i ($i \in \{1, \ldots, 5\}$), assim

$$\mathrm{E}(Y_1 + \cdots + Y_5) = \mathrm{E}(Y_1) + \cdots + \mathrm{E}(Y_5) = 5 \cdot \mathrm{E}(Y) = 40.$$

2. A soma de k v.a. $Y \sim \mathrm{N}(\mu, \sigma^2)$ independentes é uma v.a. Normal, como visto, ou seja $Y_{\text{comb}} = Y_1 + \cdots + Y_k \sim \mathrm{N}(k\mu, k\sigma^2)$. Dessa forma,

$$\begin{aligned} Y_{\text{comb}} &\sim \mathrm{N}(5 \cdot 8, 5 \cdot 4) = \mathrm{N}(40, 20) \\ z &= \frac{30 - 5 \cdot 8}{\sqrt{5} \cdot 2} = \frac{30 - 40}{4.472136} = -2.236068 \\ \mathrm{P}(Z &< -2.236068) = 0.0127. \end{aligned}$$

Aproximações usando a distribuição Normal

Aproximação da Binomial e Poisson pela Normal

Binomial

- Problemas de ordem computacional.
- Cálculo de $\binom{n}{y}$ envolve **fatorial**.
- Cálculo de p^y e $(1-p)^{n-y}$ envolve **potenciação**.
- Cálculo de $\text{P}(Y \leq y) = \sum_0^y \binom{n}{y} \cdot p^y \cdot (1-p)^{n-y}$ pode envolver **muitos termos**.
- Solução: aproximação.
- Com os computadores atuais, deixa de ser justificável.

Poisson

- Problemas de ordem computacional.
- Cálculo de $\exp\{-\lambda\}$ e λ^y envolvem **potenciação**.
- Cálculo de $y!$ envolve **fatorial**.
- Cálculo de $\text{P}(Y \leq y) = \sum_0^y \frac{\exp\{-\lambda\} \cdot \lambda^y}{y!}$ pode envolver **muitos termos**.
- Solução: aproximação.

OBS: Com o atual poder computacional, o uso das aproximações é menos justificável.

Aproximação da Binomial pela Normal

Seja $Y_\text{B} \sim \text{Bin}(n, p)$ para a qual $\mu_B = np$ e $\sigma_B^2 = np(1-p)$. Pode-se aproximar probabilidades da v.a. Y_B usando a variável $Y \sim \text{N}(np, np(1-p))$. Equivalentemente, usando a Normal Padrão,

$$Z = \frac{Y_\text{B} - np}{\sqrt{np(1-p)}} \overset{\text{aprox}}{\sim} \text{N}(0, 1).$$

Como a variável aproximada é discreta, aplica-se a **correção de continuidade** para uso da aproximação,

$$\text{P}(Y_\text{B} \leq y) = \text{P}(Y_\text{B} \leq y + 0.5) = \text{P}\left(Z \leq \frac{y + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$\text{P}(y \leq Y_\text{B}) = \text{P}(y - 0.5 \leq Y_\text{B}) = \text{P}\left(\frac{y - 0.5 - np}{\sqrt{np(1-p)}} \leq Z\right)$$

A aproximação é boa para $n \cdot \min(p, 1-p) > 5$.

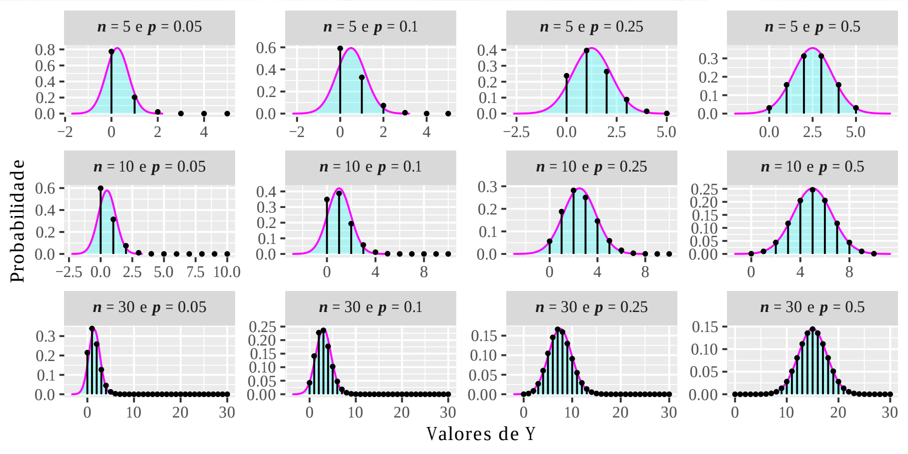


Figura 14. Aproximação da Binomial pela Normal.

Aproximação da Poisson pela Normal

Seja $Y_{\text{P}} \sim \text{Pois}(\lambda)$ para a qual $\mu_P = \lambda$ e $\sigma_P^2 = \lambda$. Pode-se aproximar probabilidades da v.a. Y_{P} usando a variável $Y \sim \text{N}(\lambda, \lambda)$. Equivalentemente, usando a Normal Padrão,

$$Z = \frac{Y_{\text{P}} - \lambda}{\sqrt{\lambda}} \overset{\text{aprox}}{\sim} \text{N}(0, 1).$$

Como a variável aproximada é discreta, aplica-se a **correção de continuidade** para uso da aproximação,

$$\text{P}(Y_{\text{P}} \leq y) = \text{P}(Y_{\text{P}} \leq y + 0.5) = \text{P}\left(Z \leq \frac{y + 0.5 - \lambda}{\sqrt{\lambda}}\right)$$

$$\text{P}(y \leq Y_{\text{P}}) = \text{P}(y - 0.5 \leq Y_{\text{P}}) = \text{P}\left(\frac{y - 0.5 - \lambda}{\sqrt{\lambda}} \leq Z\right)$$

A aproximação é boa para $\lambda \geq 5$.

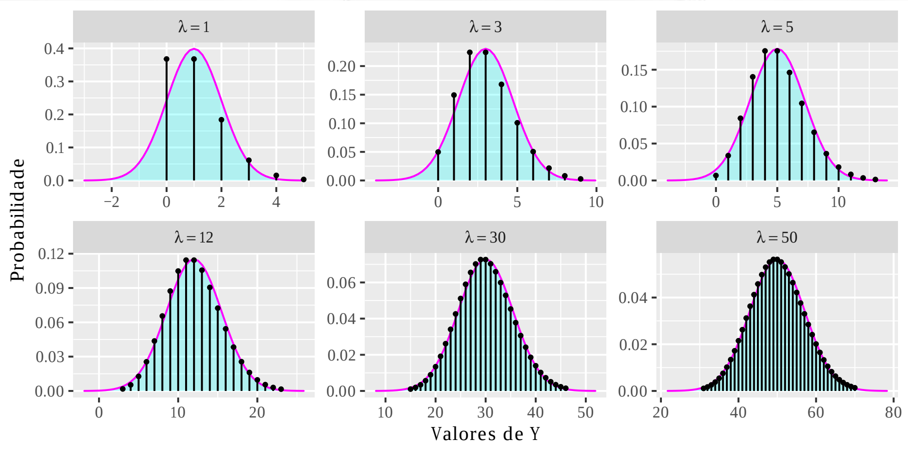


Figura 15. Aproximação da Poisson pela Normal.

Exemplo: registros incorretos

Um grande problema nas empresas é na qualidade de registro dos dados: informações incorretas, mudanças de endereço, duplicidade e ausência de informação, etc. Uma grande empresa tem um número grandíssimo de clientes cadastrados. Acredita-se que a proporção de cadastros com problemas de endereço incorreto é 10%.

1. Numa coleção de 1000 registros, qual a probabilidade de ter 120 registros com endereço errado?
2. Numa coleção de 1000 registros, qual a probabilidade de ter 120 ou menos registros com endereço errado?

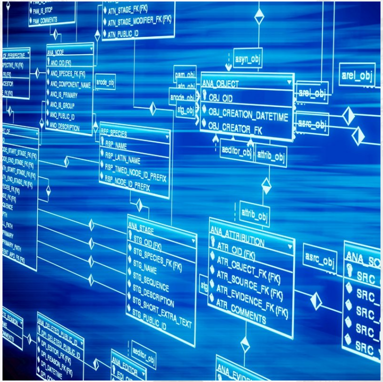


Figura 16. Extraído de www.silwoodtechnology.com.

Solução

1. Usando a função de probabilidade da Binomial, tem-se

$$P(Y_B = y) = \binom{n}{y} \cdot p^y \cdot (1-p)^{n-y}$$

$$\binom{n}{y} = 8.7552905 \times 10^{157}, \quad p^y = 10^{-120}, \quad (1-p)^{n-y} = 3.229246 \times 10^{-6}$$

$$P(Y_B = 120) = 0.0047.$$

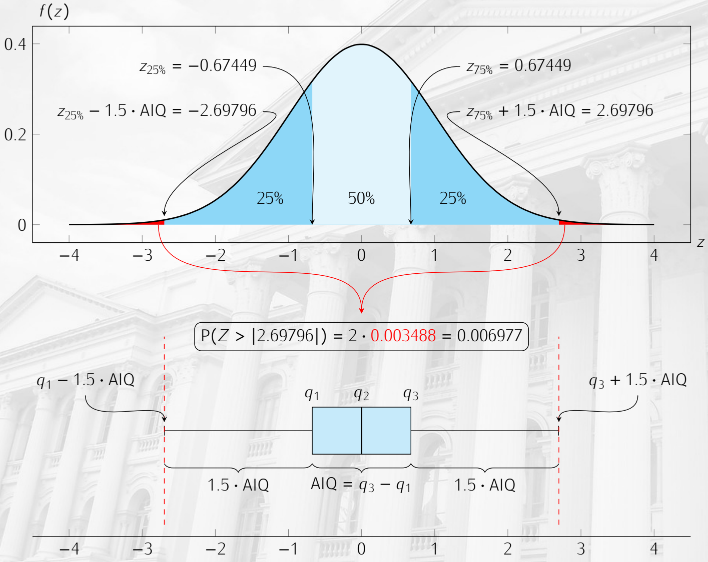
2. Usando a aproximação da Binomial pela Normal, tem-se

$$P(Y_B \leq 120) = P(Y_B \leq 120 + 0.5) = P\left(Z \leq \frac{120 + 0.5 - 1000 \cdot 0.1}{\sqrt{1000 \cdot 0.1 \cdot (1 - 0.1)}}\right)$$

$$P(Z \leq 2.1608897) = 0.9846.$$

Usando a Binomial para fazer a conta no computador, o resultado exato é $P(Y_B \leq y) = 0.9827$.

Probabilidades no gráfico de caixas para v.a. Normal



Em termos teóricos, para uma v.a. Normal Padrão (válido para qualquer Normal) a probabilidade de valores além das extremidades dos bigodes, indicados em vermelho, é menor que 1%.

Figura 17. A distribuição Normal e o gráfico de caixas e bigodes.

Considerações finais

Considerações finais

Revisão

- A distribuição Normal.
  - Sua importância e fundamentação.
  - Definição e propriedades.
  - A Normal Padrão e tabela de probabilidades.
  - Aproximações com a distribuição Normal.
- Exemplos de aplicação.