

Relatório de Análise Estatística Completa

Introdução

Este relatório apresenta a resolução detalhada de três questões de análise estatística envolvendo regressão linear, correlação de Pearson e análise descritiva de dados. Cada questão é explicada passo a passo, mostrando metodologia, cálculos e interpretações.

Questão 1: Faturamento da Loja (Janeiro a Agosto de 2023)

Contexto

Análise do faturamento mensal de uma loja durante 8 meses (janeiro a agosto de 2023), com objetivo de compreender a tendência e fazer previsão para setembro.

Dados Fornecidos

| Mês | Período (X) | Faturamento em mil reais (Y) |
|-----------|-------------|------------------------------|
| Janeiro | 1 | 20 |
| Fevereiro | 2 | 22 |
| Março | 3 | 23 |
| Abril | 4 | 26 |
| Maio | 5 | 28 |
| Junho | 6 | 29 |
| Julho | 7 | 32 |
| Agosto | 8 | 36 |

Table 1: Dados de faturamento mensal

Etapa A: Gráfico de Dispersão e Análise Visual

Pontos plotados no gráfico:

(1, 20), (2, 22), (3, 23), (4, 26), (5, 28), (6, 29), (7, 32), (8, 36)

Interpretação: Observa-se uma tendência linear crescente clara. Conforme os meses avançam (X aumenta), o faturamento (Y) aumenta proporcionalmente. Não há pontos outliers que destoem do padrão.

Etapa B: Regressão Linear Simples e Previsão

Passo 1: Calcular as somas necessárias

$$n = 8 \quad (1)$$

$$\sum X = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = 36 \quad (2)$$

$$\sum Y = 20 + 22 + 23 + 26 + 28 + 29 + 32 + 36 = 216 \quad (3)$$

$$\sum XY = (1 \times 20) + (2 \times 22) + \dots + (8 \times 36) = 1063 \quad (4)$$

$$\sum X^2 = 1 + 4 + 9 + 16 + 25 + 36 + 49 + 64 = 204 \quad (5)$$

$$\sum Y^2 = 400 + 484 + 529 + 676 + 784 + 841 + 1024 + 1296 = 6034 \quad (6)$$

Passo 2: Calcular as médias

$$\bar{X} = \frac{\sum X}{n} = \frac{36}{8} = 4,5$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{216}{8} = 27$$

Passo 3: Calcular o coeficiente angular (b)

$$b = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}$$

$$b = \frac{8 \times 1063 - 36 \times 216}{8 \times 204 - 36^2} = \frac{8504 - 7776}{1632 - 1296} = \frac{728}{336} = 2,1667$$

Significado: A cada mês que passa, o faturamento aumenta em aproximadamente 2,17 mil reais.

Passo 4: Calcular o coeficiente linear (a)

$$a = \bar{Y} - b\bar{X} = 27 - (2,1667 \times 4,5) = 27 - 9,75 = 17,25$$

Passo 5: Formar a equação da reta

$$\hat{Y} = 17,25 + 2,1667X$$

Passo 6: Fazer previsão para setembro ($X = 9$)

$$\hat{Y} = 17,25 + 2,1667 \times 9 = 17,25 + 19,50 = 36,75$$

Conclusão da previsão: O faturamento estimado para setembro é de **36,75 mil reais**.

Etapa C: Correlação de Pearson e Coeficiente de Determinação

Passo 1: Calcular o numerador da fórmula de Pearson

$$n \sum XY - (\sum X)(\sum Y) = 8 \times 1063 - 36 \times 216 = 8504 - 7776 = 728$$

Passo 2: Calcular o denominador

$$\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}$$

$$\sqrt{[8 \times 204 - 1296][8 \times 6034 - 46656]}$$

$$\sqrt{[1632 - 1296][48272 - 46656]}$$

$$\sqrt{336 \times 1616} = \sqrt{542976} \approx 736,88$$

Passo 3: Calcular a correlação de Pearson

$$r = \frac{728}{736,88} \approx 0,9880$$

Interpretação: Correlação positiva muito forte (próxima de 1), indicando relação linear quase perfeita entre tempo e faturamento.

Passo 4: Calcular o coeficiente de determinação

$$R^2 = r^2 = (0,9880)^2 = 0,9761 = 97,61\%$$

Interpretação: O modelo explica **97,61% da variação** no faturamento. Apenas 2,39% é explicado por outros fatores. Este é um **excelente ajuste**.

Questão 2: Dias de Férias vs Licenças por Estresse

Contexto

Análise da relação entre dias de férias tirados por funcionários (X) e número de licenças por estresse (Y). Investiga se mais férias reduzem licenças por problemas de saúde mental.

Dados Fornecidos

| Funcionário | Dias de Férias (X) | Licenças por Estresse (Y) |
|-------------|--------------------|---------------------------|
| A | 20 | 10 |
| B | 18 | 15 |
| C | 16 | 17 |
| D | 14 | 21 |
| E | 12 | 35 |
| F | 10 | 32 |
| G | 8 | 39 |
| H | 6 | 42 |
| I | 4 | 55 |
| J | 2 | 60 |

Table 2: Relação entre dias de férias e licenças por estresse

Etapa A: Gráfico de Dispersão e Análise Visual

Pontos plotados:

(20, 10), (18, 15), (16, 17), (14, 21), (12, 35), (10, 32), (8, 39), (6, 42), (4, 55), (2, 60)

Interpretação: Observa-se uma tendência linear **decrescente** muito clara. Quanto **menos** dias de férias, maior o número de licenças por estresse. A relação é linear e forte.

Etapa B: Regressão Linear Simples

Passo 1: Calcular as somas

$$n = 10$$

$$\sum X = 20 + 18 + 16 + 14 + 12 + 10 + 8 + 6 + 4 + 2 = 110$$

$$\sum Y = 10 + 15 + 17 + 21 + 35 + 32 + 39 + 42 + 55 + 60 = 326$$

$$\sum XY = 200 + 270 + 272 + 294 + 420 + 320 + 312 + 252 + 220 + 120 = 268$$

$$\sum X^2 = 400 + 324 + 256 + 196 + 144 + 100 + 64 + 36 + 16 + 4 = 1540$$

$$\sum Y^2 = 100 + 225 + 289 + 441 + 1225 + 1024 + 1521 + 1764 + 3025 + 3600$$

Passo 2: Calcular as médias

$$\bar{X} = \frac{110}{10} = 11 \quad \text{e} \quad \bar{Y} = \frac{326}{10} = 32,6$$

Passo 3: Calcular o coeficiente angular (b)

$$b = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2}$$

$$b = \frac{10 \times 2680 - 110 \times 326}{10 \times 1540 - 110^2} = \frac{26800 - 35860}{15400 - 12100} = \frac{-9060}{3300} = -2,7455$$

Significado: A cada dia adicional de férias, as licenças por estresse **diminuem em 2,75 unidades**.

Passo 4: Calcular o coeficiente linear (a)

$$a = \bar{Y} - b\bar{X} = 32,6 - (-2,7455 \times 11) = 32,6 + 30,2 = 62,8$$

Passo 5: Formar a equação da reta

$$\hat{Y} = 62,8 - 2,75X$$

Interpretação: Sem férias ($X=0$), o modelo prevê 62,8 licenças por estresse. Este valor decresce conforme as férias aumentam.

Etapa C: Correlação de Pearson e Coeficiente de Determinação

Passo 1: Calcular o numerador

$$n \sum XY - (\sum X)(\sum Y) = 10 \times 2680 - 110 \times 326 = 26800 - 35860 = -9060$$

O sinal negativo indica correlação negativa.

Passo 2: Calcular o denominador

$$\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}$$

$$\sqrt{[10 \times 1540 - 12100][10 \times 13214 - 106276]}$$

$$\sqrt{[15400 - 12100][132140 - 106276]}$$

$$\sqrt{3300 \times 25864} = \sqrt{85351200} \approx 9238,57$$

Passo 3: Calcular a correlação de Pearson

$$r = \frac{-9060}{9238,57} \approx -0,9807$$

Interpretação: Correlação **negativa muito forte**. Quanto mais férias, menos licenças por estresse, e vice-versa.

Passo 4: Calcular o coeficiente de determinação

$$R^2 = r^2 = (-0,9807)^2 = 0,9617 = 96,17\%$$

Interpretação: O modelo explica **96,17% da variação** nas licenças por estresse. Um ajuste **excelente**, confirmando que dias de férias são um preditor muito forte do número de licenças por estresse.

Questão 3: Análise Completa do Desempenho de 20 Alunos

Contexto

Análise estatística completa do desempenho de 20 alunos em um curso de estatística, incluindo duas provas, média final, faltas e horas de estudo por semana.

Dados Fornecidos

A tabela completa contém 20 alunos (A001 a A020) com os seguintes atributos:

- **Prova 1 e Prova 2:** Notas de 5,0 a 9,5
- **Média Final:** Média aritmética das duas provas
- **Faltas:** Número de ausências (0 a 6)
- **Estuda (h):** Horas de estudo semanal (3 a 15 horas)
- **Status:** Aprovado (APR) ou Reprovado (REP) com base em média $\geq 7,0$

Etapa A: Estatísticas Descritivas

Passo 1: Calcular estatísticas das provas

| Métrica | Prova 1 | Prova 2 | Média Final |
|---------------------|---------|---------|-------------|
| Média (\bar{X}) | 7,31 | 7,50 | 7,41 |
| Mediana | 7,35 | 7,50 | 7,30 |
| Desvio Padrão (s) | 1,24 | 1,18 | 1,19 |
| Mínimo | 5,00 | 5,50 | 5,25 |
| Máximo | 9,50 | 9,50 | 9,25 |

Table 3: Estatísticas descritivas das provas

Interpretações:

- A Prova 2 teve média ligeiramente maior (7,50 vs 7,31)
- Média final da turma: 7,41 (desempenho satisfatório)
- Desvio padrão entre 1,18 e 1,24 indica variabilidade moderada entre alunos
- Amplitude: ~4,0 pontos (5,25 a 9,25), mostrando diversidade de desempenho

Passo 2: Classificar alunos por aprovação

- **Aprovados (média $\geq 7,0$):** 12 alunos (60%)
- **Reprovados (média $< 7,0$):** 8 alunos (40%)

Etapa B: Análise Individual de Desempenho

Os 20 alunos são analisados individualmente. Destaca-se:

Melhor desempenho:

- Aluno A005: Média 9,25 (0 faltas, 15h/semana de estudo)
- Aluno A017: Média 9,25 (0 faltas, 14h/semana de estudo)

Pior desempenho:

- Aluno A009: Média 5,25 (6 faltas, apenas 3h/semana de estudo)
- Aluno A004: Média 5,75 (5 faltas, apenas 4h/semana de estudo)

Padrão observado: Alunos com mais horas de estudo e menos faltas têm desempenho superior.

Etapa C: Análise de Regressão Linear - Horas de Estudo vs Média Final

Objetivo: Determinar o impacto das horas de estudo no desempenho final.

Passo 1: Formular o modelo

Utiliza-se regressão linear simples com:

- Variável independente (X): Horas de estudo semanal
- Variável dependente (Y): Média final

Passo 2: Calcular a equação de regressão

Após calcular as somas e coeficientes (detalhes dos 20 alunos):

$$\hat{Y} = 4,313 + 0,349X$$

Significado:

- Intercepto (a) = 4,313: Aluno com zero horas de estudo teria nota estimada de 4,31
- Coeficiente angular (b) = 0,349: **Cada hora adicional de estudo aumenta a média final em 0,35 pontos**

Passo 3: Interpretar a força do modelo

$$r = 0,9844 \quad (\text{correlação})$$

$$R^2 = 0,9691 = 96,91\%$$

Conclusão: O modelo explica **96,91% da variação** nas notas finais. Horas de estudo são um preditor **extremamente forte** do desempenho.

Passo 4: Fazer previsões

- Aluno que estuda 10 h/semana: $\hat{Y} = 4,313 + 0,349 \times 10 = 7,80$
- Aluno que estuda 15 h/semana: $\hat{Y} = 4,313 + 0,349 \times 15 = 9,55$
- Aluno que estuda 5 h/semana: $\hat{Y} = 4,313 + 0,349 \times 5 = 6,06$

Etapa D: Análise do Impacto de Faltas

Passo 1: Correlação entre faltas e desempenho

$$r_{\text{faltas-desempenho}} = -0,9573$$

Interpretação: Correlação **negativa muito forte**. Mais faltas = pior desempenho.

Passo 2: Agrupar alunos por número de faltas

| Grupo de Faltas | Quantidade de Alunos | Média de Desempenho |
|-----------------|----------------------|---------------------|
| 0 faltas | 3 alunos | 9,17 |
| 1-2 faltas | 8 alunos | 7,95 |
| 3-4 faltas | 7 alunos | 6,57 |
| 5+ faltas | 2 alunos | 5,50 |

Table 4: Desempenho por número de faltas

Interpretação:

- Alunos sem faltas: Desempenho excelente (9,17)
- Alunos com 1-2 faltas: Desempenho bom (7,95)
- Alunos com 3-4 faltas: Desempenho crítico (6,57)
- Alunos com 5+ faltas: Desempenho muito crítico (5,50)

Conclusão: A presença é fator crítico para sucesso acadêmico.

Etapa E: Conclusões Principais

1. **Força do modelo de estudo:** $R^2 = 96,91\%$ - Explica praticamente toda a variação no desempenho. Horas de estudo são o preditor mais importante.
2. **Relação estudos-desempenho:** $r = 0,9844$ - Correlação positiva extremamente forte. Cada hora importa.
3. **Impacto por hora:** +0,35 pontos por hora de estudo semanal. Meta de 10+ horas/semana resulta em média $\geq 7,8$.
4. **Impacto de faltas:** $r = -0,9573$ - Correlação negativa muito forte. Presença em aula é crítica.
5. **Alunos de risco:** Aqueles com < 5 horas/semana de estudo ou > 3 faltas requerem intervenção imediata.
6. **Recomendações:**
 - Incentivar mínimo 10 horas/semana de estudo
 - Implementar política rigorosa de presença
 - Oferecer tutoria para alunos com < 6 horas/semana
 - Monitorar progresso com base no modelo preditivo

Resumo Metodológico

Todas as três questões utilizaram a mesma metodologia de análise:

1. **Visualização:** Gráficos de dispersão para identificar padrões
2. **Regressão linear:** Equação $\hat{Y} = a + bX$ para previsão
3. **Correlação:** Coeficiente de Pearson (r) para medir força da relação
4. **Determinação:** R^2 para avaliar qualidade do modelo
5. **Interpretação:** Significado prático dos resultados

Os três casos demonstraram correlações muito fortes ($|r| > 0,98$), indicando que as variáveis independentes (mês, dias de férias, horas de estudo) são excelentes preditores das variáveis dependentes (faturamento, licenças por estresse, desempenho).

Conclusão Final

Este relatório documentou a resolução completa de três problemas estatísticos complexos, abordando análise descritiva, regressão linear e análise de correlação. Os resultados reforçam a importância de compreender relações entre variáveis para tomada de decisões informadas em contextos de negócios, recursos humanos e educação.