Trabalho final da disciplina

Prof. Dr. Marcelo de Oliveira Rosa

27 de maio de 2021

1 Introdução

Este trabalho consiste de uma série de exercícios usando ferramenta `R` para melhor compreensão dos princípios de análise estatística vistos durante o curso. Alguns dados usados neste trabalho encontram-se na página eletrônica do curso.

Mais do que usar a ferramenta, é importante entender os princípios estatísticos subjacentes, para evitar erros de interpretação durante a análise de quaisquer dados.

2 Apresentação dos resultados

Os resultados gráficos devem apresentar um título e nomes claros para os eixos X e Y, pensando em um leitor externo, que possa interpretar corretamente as variáveis envolvidas nas informações gráficas. Esses elementos são gerados diretamente por funções e parâmetros apropriados do `R`.

Os resultados textuais provenientes da ferramenta `R` podem ser copiados diretamente desta ferramenta para facilitar sua apresentação.

3 Preliminares

1. Identifique o tipo de escala (nominal, ordinal, ou contínua) das seguintes variáveis:

   - Níveis de dificuldade de um jogo
   - Partidos políticos de uma eleição
   - Produção mensal de aço
   - Idade das tartarugas
   - Preço da picanha
   - Número do RG
   - *Ranking* do mundial de F1
   - Coeficiente de inteligência (Q.I)

2. Carregue os dados `entrega_pizza` no `R` e obtenha seu resumo estatístico;

3. Acrescente uma nova variável (coluna) chamada `temperatura_far` como sendo a temperatura em Fahrenheit da pizza (Celsius para Fahrenheit) ao *data frame* `entrega_pizza`. Qual o resumo estatístico do novo *data frame*?

4. Obtenha o nome das variáveis de `entrega_pizza`.

4 Probabilidade

1. A Copa do Mundo do Qatar (Fifa 2022) terá 32 seleções. Quantas combinações existem para os três primeiros lugares, considerando:

   - a ordem ser relevante;
   - a ordem ser irrelevante.

2. Uma moeda é jogada 2 vezes. Qual a probabilidade de obter cara nas duas jogadas assumindo que:

   - obteve-se cara na primeira jogada;
   - obteve-se cara em uma das jogadas.

3. Quantas placas distintas de carro com 7 caracteres podem ser formadas se os 2 primeiros caracteres forem letras e os demais forem números?

4. Uma loja aceita cartões de crédito Visa e Mastercard. 24% dos clientes possui Mastercard, 61% possui Visa, e 11% possui cartões de ambas as bandeiras. Qual o percentual de clientes possui cartão?

5 Estatística descritiva

1. Calcule as medidas de tendência central (média, mediana e moda) da variável `temperatura_far`. APENAS pelos valores obtidos, diga se você considera que estes dados apresentam uma distribuição simétrica, assimétrica à direita (mais concentrada à direita) ou assimétrica à esquerda;

2. Calcule as medidas de dispersão (desvio absoluto e desvio padrão) da variável `temperatura_far`, com atenção para o uso de n ao invés de $n-1$ (sim, será uma medida com viés);

3. Calcule as medidas de assimetrica e curtose da variável `temperatura_far`, indicando agora qual o tipo de distribuição tempo para a variável `temperatura_far`;

4. Monte um função em `R` que retorne indique o tipo de distribuição, calculando a assimetria e curtose usando laços - sem chamar uma função pronta de `R`. A função deve receber um vetor de dados e seu retorno deve ser um texto indicativo, ou seja:

   ```
   texto <- tipo_distribuicao(dados)
   ```

5. Plote o histograma usando o agrupamento *default* da função interna de `R`;

6. Plote o histograma com dois agrupamentos distintos: um mais curto e outro mais comprido. Sua percepção da distribuição dos dados mudou?

7. Usando o comando `ping` do seu computador (disponível para Windows, Linux e MacOS), gere um conjunto de 100 medições da latência do seu computador para cada um dos seguintes endereços:

   (a) `www.utfpr.edu.br` ou `www.ufpr.br` (ou outra instituição paranaense)

   (b) `www.ucla.edu`

   (c) `www.epfl.ch`

(d) `www.tsinghua.edu.cn`

(e) `www.monash.edu`

Cada sistema operacional tem um parâmetro para definir a quantidade de "medições" de latência e para salvar o conteúdo em arquivo. Após salvá-los, isole os valores numéricos da latência (em qualquer linguagem, ou até manualmente em programas como Excel);

Carregue os dados de cada endereço em `R` e determine seus resumos estatísticos, variância (com $n-1$), histograma e diagrama de caixa (*boxplot()*);

8. Comente os resultados: latência maior e menor, nível de variabiliade e diferenças na distribuição de dados;

9. Usando gráficos Q-Q, verifique individualmente se os dados de latência dos endereços seguem uma distribuição normal;

10. Usando gráficos Q-Q, verifique dois-a-dois se as distribuições dos dados de latência dos endereços são as mesmas;

11. Usando os dados de `entrega_pizza`, verifique se as entregas do motorista Luigi apresentam o mesmo padrão de distribuição das entregas do motorista Domenico;

12. Usando os dados de `entrega_pizza`, faça o mesmo para os motoristas Mario e Salvatore;

6 Tabelas e correlações

1. Construam uma tabela de contingência envolvendo o tempo de entrega de pizzas (*data frame* `entrega_pizza`) e as filiais da pizzaria. Use o teste χ^2 para avaliar se há associação entre elas ou não (para facilitar a avaliação, você pode usar o coeficiente V de Craemer);

2. Avalie se há associação entre o dia da semana e o gasto efetuado pelos clientes da pizzaria;

3. Considerando o dia de maior número de pizzas entregues, avalie se há associação entre o tempo de entrega e o atendente/operador da pizzaria;

4. Carregue os dados `decatlo` em um *data frame* e obtenha seu resumo estatístico;

5. Plote um gráfico de dispersão entre os resultados das provas de 100 metros rasos e salto em distância (variáveis `X.100m` e `X.Salto.distancia` do *data frame*, respectivamente);

6. Calcule o coeficiente de correlação entre estas variáveis de `decatlo`: há indícios de algum tipo de associação entre elas? Qual seria?

7. Selecione as 5 (cinco) primeiras observações de `decatlo` e verifique, usando o coeficiente de correlação de Spearman se há alguma associação entre suas variáveis `X.100m` e `X.Salto.distancia`.

7 Distribuições

1. Carregue os dados `entrega_pizza` no `R` e obtenha seu resumo estatístico;

2. Considerando as distribuições uniforme, normal e exponencial e usando gráficos Q-Q, verifique qual delas retrata melhor a distribuição do gasto com cultura e do gasto com teatro na Suíça;

3. Considere uma moeda que é lançada 10 vezes. Qual a probabilidade de obtermos 4 caras? Qual a probabilidade de obtermos ATÉ 4 caras (inclusive)?

4. Considere que o tempo de entrega de pizzas em Curitiba obedeça a uma distribuição normal, de média (μ) igual a 30 minutos, com variância (σ^2) igual a 30^2 minutos^2. Qual a probabilidade de se receber uma pizza em casa em um tempo igual ou menor a 40 minutos? E menor ou igual a 15 minutos? E entre 15 e 45 minutos?

8 Inferência e testes de hipótese

1. Com os dados de pizza carregados, compare os intervalos de confiança das temperaturas em graus Celsius e em Fahrenheit? O que você observa nesses resultados?

2. Com os dados de teatro carregados, quais as estimativas pontuais da média e variância das variáveis idade, renda anual e gastos com cultura e teatro dos moradores da Suíça?

3. Qual seus respectivos intervalos de confiança considerando que suas variâncias populacionais são desconhecidas?

4. Com os dados de decatlo carregados, qual o tamanho de amostra necessário para calcular um intervalo de confiança de 95% para o tempo médio dos 100 metros rasos com acurácia de $\pm$0,1 segundos ($\Delta = 0{,}2$), ASSUMINDO que a variância populacional é conhecida e igual à variância da amostra?

5. Calcule o intervalo de confiança de 95% das performances nas 10 provas do decatlo considerando que suas variâncias populacionais são desconhecidas;

6. Há diferença estatística nos tempos de entrega de pizzas a partir das diferentes filiais (tomados 2 a 2), considerando um nível de significância de 5%? (Pode fazer uma tabela simples com os resultados de cada par, incluindo `p.values`)

7. Há diferença estatística nos tempos de entrega de pizzas para os diferentes motoristas (tomados 2 a 2), considerando um nível de significância de 2%? (Pode fazer uma tabela simples com os resultados de cada par, incluindo `p.values`)

8. Homens gastam estatisticamente em eventos culturais do menos do que mulheres ($\alpha = 5\%$)?

9. No *data frame* `teatro`, há duas variáveis de gastos com teatro do indivíduos: do ano atual e do ano anterior. Tais variáveis permitem avaliar mudanças de comportamento pareado. Logo, houve aumento, manutenção ou queda de gastos do ano anterior em relação ao atual, estatisticamente falando ($\alpha = 5\%$)? Plote o diagrama de caixas dessas duas variáveis para comparar com o resultado do teste estatístico.

10. Considerando os dados de latência, avalie se há diferença significativa ($\alpha = 5\%$) entre os diferentes endereços fornecidos. Use o teste t e o teste de Mann-Whitney (veja que teste é este) para avaliação dos resultados. Se aumentarmos ou diminuirmos o nível de significância dos testes, o resultado melhora ou piora? E o que significa esses aumento e diminuição no nível de falsos positivos e negativos?

9 Regressão Linear

1. Plote a reta de regressão linear que relaciona a performance dos atletas em arremesso de peso e arremesso de dardo, avaliando se este modelo é significativo para representar tal associação.

2. Plote a reta de regressão linear que relaciona a performance dos atletas em arremesso de peso e arremesso de disco, avaliando se este modelo é significativo para representar tal associação.

3. Qual a relação entre o tempo de entrega de pizza e as filiais? Essa relação é significativa ou parcialmente significativa ($\alpha = 5\%$)?

4. Interprete a relação entre o tempo de entrega e as variáveis que representam o gasto para a compra das pizzas e o operador envolvido, se um modelo linear fosse adotado.

5. Avalie por regressão linear que variáveis associam-se significativamente com o gasto para a compra das pizzas. Remova-as sistematicamente até obter um modelo "mínimo" de maior significância.

6. Cheque a heteroscedasticidade do modelo para a associação entre tempo de entrega de pizza e o gasto do consumidor considerando apenas a filial "Centro". O que é mesmo heteroscedasticidade?