Especialização em *Data Science* e Estatística Aplicada

Módulo II - Análise estatística de várias populações

Profa. Dra. Tatiane F N Melo

Goiânia, 2024

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN FACULDADE DE ENFERMAGEM
UFG UNIVERSIDADE FEDERAL DE GOIÁS

Aula 4 - Parte 1

1. Análise de variância de um fator
   - Definição
   - Hipóteses do teste de ANOVA
   - Pressupostos
   - Execução da ANOVA

2. Referências Bibliográficas

Análise de Variância de um fator (ANOVA)

Definições

- A **análise de variância (ANOVA)** é um método para se testar a igualdade de três ou mais médias populacionais através da análise das variâncias amostrais.
- A metodologia usada aqui se chama **análise de variância com um fator**, pois usamos uma única propriedade, ou característica, para categorizar as populações.
  - Essa característica é chamada de tratamento ou fator.
- Um **tratamento** (ou **fator**) é uma propriedade, ou característica, que nos permite distinguir populações umas das outras.

Análise de Variância de um fator (ANOVA)

Objetivo

- Verificar se as médias de três ou mais grupos são estatisticamente diferentes entre si.
- Essa técnica é uma extensão do teste t de Student, que é usado para comparar as médias de dois grupos.
- A ANOVA de um fator permite essa comparação quando há mais de dois grupos envolvidos.

Análise de Variância de um fator (ANOVA)

Hipóteses do teste de ANOVA

As hipóteses em um teste de análise de variância de um fator são dadas por:

- H_0: Todas as médias populacionais são iguais. Ou seja, não há diferença significativa entre os grupos em termos da variável de interesse.
- H_1: Pelo menos uma das médias é diferente. Isto é, existe uma diferença significativa entre os grupos, sem especificar quais grupos são diferentes entre si.

Análise de Variância de um fator (ANOVA)

Hipóteses do teste de ANOVA

Matemáticamente:

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$,

  onde k representa o número de grupos ou níveis do fator no teste de ANOVA. Esses grupos são as categorias ou tratamentos cujas médias desejamos comparar no teste.

- $H_1 : \mu_i \neq \mu_j$, para pelo menos um par $i \neq j$.

Análise de Variância de um fator (ANOVA)

Exemplo 1

- Um pesquisador deseja avaliar o impacto de três diferentes tipos de exercícios físicos sobre a pressão arterial sistólica de pacientes hipertensos.
- O objetivo é verificar se há diferença significativa na redução da pressão arterial após um mês de prática desses exercícios.
- **Fator:** Tipo de exercício (único fator, com 3 níveis)
  - Exercício A: Caminhada moderada.
  - Exercício B: Corrida leve.
  - Exercício C: Natação.
- **Interesse:** Redução na pressão arterial sistólica (medida em mmHg após um mês de exercício).

Análise de Variância de um fator (ANOVA)

Continuação do Exemplo 1

Aplicação da ANOVA:

1. **Fator:** O único fator neste exemplo é o tipo de exercício físico (Caminhada, Corrida, Natação). Cada paciente foi aleatoriamente atribuído a um dos três tipos de exercícios.

2. **Objetivo:** Verificar se o tipo de exercício influencia a redução da pressão arterial. Como estamos interessados apenas no efeito do tipo de exercício (fator único), podemos usar uma ANOVA de um fator.

Análise de Variância de um fator (ANOVA)

Exemplo 2

Para descobrir se um novo soro vai interromper a leucemia, nove ratos, todos com um estágio avançado da doença, são selecionados. Cinco ratos recebem o tratamento e quatro, não. O tempo de sobrevida, em anos, a partir do momento em que o experimento foi iniciado, é o seguinte:

Com tratamento	2,1	5,3	1,4	4,6	0,9
Sem tratamento	1,9	0,5	2,8	3,1	

Análise de Variância de um fator (ANOVA)

Continuação do Exemplo 2

- Nesse caso, dizemos que há um *fator*, chamado *tratamento*, e o fator tem dois *níveis*.
- Se diversos tratamentos concorrentes fossem usados no processo amostral, mais amostras de camundongos seriam necessárias.
- Dessa forma, o problema envolveria um fator com mais de dois níveis e, portanto, mais de duas amostras.

Análise de Variância de um fator (ANOVA)

Pressupostos

Uma análise de variância só deve ser conduzida se estiverem satisfeitas algumas exigências.

1. As populações devem ter distribuições aproximadamente normais.
    - Para amostras pequenas (geralmente menor que 30 por grupo), a suposição de normalidade se torna mais relevante. Se a amostra for pequena e os dados não forem normais, a ANOVA pode produzir resultados imprecisos.
    - Para um tamanho amostral grande, a ANOVA pode ser usada mesmo quando os dados não seguem uma distribuição normal, pois o Teorema Central do Limite, afirma que, à medida que o tamanho da amostra aumenta, a distribuição da média tende a se aproximar de uma distribuição normal, independentemente da forma da distribuição original dos dados.

Análise de Variância de um fator (ANOVA)

Pressupostos

2. Os grupos devem ser formados por unidades que proveem de populações com variâncias iguais (homogeneidade das variâncias).

    - O estatístico da Universidade de Wisconsin, George E. P. Box, mostrou que, desde que os tamanhos amostrais sejam iguais, as variâncias podem diferir por quantidades que tornem a maior até 9 vezes o valor da menor e os resultados da ANOVA continuarão a ser essenciamente confiáveis.

Análise de Variância de um fator (ANOVA)

Pressupostos

3. As amostras são amostras aleatórias simples (AAS).
    - Ou seja, amostras de mesmo tamanho têm a mesma probabilidade de serem selecionadas.

4. As unidades devem ser independentes, tanto dentro do mesmo grupo como entre os diferentes grupos.
    - As amostras não são pareadas.

Análise de Variância de um fator (ANOVA)

Voltando no Exemplo 1

No exemplo sobre a redução da pressão arterial devido a diferentes tipos de exercícios físicos, as hipóteses nula e alternativa podem ser definidas da seguinte forma:

Análise de Variância de um fator (ANOVA)

Voltando no Exemplo 1

Hipóteses da ANOVA:

H_0: Não há diferença significativa nas médias de redução da pressão arterial entre os três tipos de exercícios (Caminhada, Corrida e Natação).

- Em outras palavras, a redução média da pressão arterial é igual para os três grupos de pacientes, independentemente do tipo de exercício realizado.

H_1: pelo menos uma das médias de redução da pressão arterial é diferente entre os três grupos.

- Isso significa que o tipo de exercício influencia a redução da pressão arterial e que pelo menos um dos exercícios tem um efeito diferente sobre a redução da pressão arterial em relação aos outros.

Análise de Variância de um fator (ANOVA)

Voltando no Exemplo 1

Matematicamente,

$$H_0 : \mu_A = \mu_B = \mu_C,$$

onde

- μ_A é a média de redução da pressão arterial para o grupo que realizou Caminhada;
- μ_B é a média de redução da pressão arterial para o grupo que realizou Corrida;
- μ_C é a média de redução da pressão arterial para o grupo que realizou Natação.

$$H_1 : \text{Pelo menos uma das médias } \mu_i \text{ é diferente.}$$

Ou seja, pelo menos uma das médias μ_A, μ_B ou μ_C é diferente das outras.

Análise de Variância de um fator (ANOVA)

Voltando no Exemplo 1

- Na Análise de Variância (ANOVA), o objetivo é separar a variabilidade observada nos dados em dois tipos principais de variação: a **variação entre os grupos** e a **variação dentro dos grupos**.
- Esses dois tipos de variação ajudam a determinar se as diferenças observadas nas médias dos grupos são significativas ou não.

Análise de Variância de um fator (ANOVA)

Voltando no Exemplo 1 - Variação Entre os Grupos

- Essa variação mede o quanto as médias de cada grupo (Caminhada, Corrida, Natação) diferem umas das outras e da média global dos dados.
- A variação entre os grupos ocorre devido ao efeito do tipo de exercício físico.
- Se as médias de redução da pressão arterial para os diferentes grupos são muito diferentes, isso sugere que o tipo de exercício físico (fator) está tendo um impacto significativo nas respostas dos pacientes.

Análise de Variância de um fator (ANOVA)

Voltando no Exemplo 1 - Variação Dentro dos Grupos

- Essa variação mede o quanto as reduções da pressão arterial variam dentro de cada grupo de pacientes submetidos ao mesmo tipo de exercício (Caminhada, Corrida ou Natação).
- A variação dentro dos grupos ocorre devido a diferenças individuais entre os pacientes e a variabilidade aleatória.
- Mesmo que todos os pacientes de um grupo realizem o mesmo tipo de exercício, haverá variações nas respostas devido a fatores individuais (idade, condições de saúde, etc.).

Análise de Variância de um fator (ANOVA)

Estatística de teste

- A estatística de teste usada na ANOVA é a F, que é calculada como a razão entre a variabilidade entre os grupos e a variabilidade dentro dos grupos.
- Se o valor de F for suficientemente grande, a hipótese nula pode ser rejeitada, indicando que há uma diferença significativa entre os grupos.

Análise de Variância de um fator (ANOVA)

Notação:

- y_{ij}: j-ésima observação do i-ésimo tratamento;
- $Y_{i.}$: total de observações na amostra do i-ésimo tratamento;
- $\bar{y}_{i.}$: média de todas as observações na amostra do i-ésimo tratamento;
- $Y_{..}$: total de todas as nk observações;
- $\bar{y}_{..}$: média de todas as nk observações.

Análise de Variância de um fator (ANOVA)

Tabela 1: Amostras aleatórias.

Tratamento	1	2	$\dots$	i	$\dots$	k	
	y_{11}	y_{21}	$\dots$	y_{i1}	$\dots$	y_{k1}	
	y_{12}	y_{22}	$\dots$	y_{i2}	$\dots$	y_{k2}	
	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	
	y_{1n}	y_{2n}	$\dots$	y_{in}	$\dots$	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	$\dots$	$Y_{i.}$	$\dots$	$Y_{k.}$	$Y_{..}$
Média	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\dots$	$\bar{y}_{i.}$	$\dots$	$\bar{y}_{k.}$	$\bar{y}_{..}$

Análise de Variância de um fator (ANOVA)

- O teste que realizaremos aqui será baseado na comparação de duas estimativas independentes da variância populacional σ^2.

- Essas estimações serão obtidas dividindo-se a variabilidade total de nossos dados, atribuída pela soma dupla

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{..})^2$$

em dois componentes.

Análise de Variância de um fator (ANOVA)

Somas de quadrados

É possível mostrar que

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2 = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$$

Ou seja,

$$SQT = SQ_{Trat} + SQE, \tag{1}$$

com

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2, \quad SQ_{Trat} = n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2, \quad SQE = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2.$$

Análise de Variância de um fator (ANOVA)

Somas de quadrados

- $SQT-$ soma de quadrados total: é uma medida da variação total (em torno de $\bar{y}_{..}$) em todos os dados amostrais combinados;
- $SQ_{Trat}-$ soma de quadrados do tratamento ou SQ(fator) ou SQ(entre grupos) ou SQ(entre amostras): é uma medida da variação entre as médias amostrais. Representa a variabilidade devido aos diferentes níveis do fator A;
- $SQE-$ soma de quadrados do erro ou SQ(dentro dos grupos) ou SQ(dentro das amostras): é uma soma de quadrados que representa a variação que se supõe comum a todas as populações em consideração. Representa a variabilidade dentro de cada nível do fator.

Análise de Variância de um fator (ANOVA)

Graus de liberdade

Temos que graus de liberdade (g.l.) está diretamente ligado a soma de quadrados. Considere, por exemplo, a amostra $y_1, \ldots, y_k$. Sabemos que

$$\bar{y} = \frac{\sum_{i=1}^{k} y_i}{k} \quad \text{e} \quad \sum_{i=1}^{k} (y_i - \bar{y}) = 0,$$

para encontrarmos todos os desvios $y_i - \bar{y}$, basta conhecermos apenas $(k-1)$ deles, pois o k-ésimo desvio pode ser obtido a partir dos $(k-1)$ anteriores. Logo, dizemos que a soma quadrática $\sum_{i=1}^{k} (y_i - \bar{y})^2$ tem $(k-1)$ graus de liberdade.

Análise de Variância de um fator (ANOVA)

Quadrados médios

Temos que

$$\widehat{\sigma}^2_{Trat} = QM_{Trat} = \frac{SQ_{Trat}}{k-1} \quad \text{e} \quad \widehat{\sigma}^2_{E} = QME = \frac{SQE}{k(n-1)}$$

são estimativas de σ^2.

- QM_{Trat}: é chamado de **Quadrado Médio do Tratamento**.
- QME: é chamado de **Quadrado Médio do Erro**.

Análise de Variância de um fator (ANOVA)

- A decomposição da SQT em duas somas de quadrados nos fornece duas estimativas para a variância.
- A primeira baseada na variabilidade dentro dos níveis e a segunda baseada na variabilidade entre os níveis.

Análise de Variância de um fator (ANOVA)

Razão F para testar a igualdade de médias

Considere as hipóteses

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

contra

$$H_1 : \text{Pelo menos uma das médias } \mu_i \text{ é diferente,}$$

com $i = 1, 2, \ldots, k$.

Análise de Variância de um fator (ANOVA)

Razão F para testar a igualdade de médias

- Quando H_0 é verdadeira, o valor da estatística de teste é dado pela razão

$$f_{obs} = \frac{QM_{Trat}}{QME}.$$

Este é o valor da variável aleatória F que tem a distribuição F com $k-1$ e $k(n-1)$ graus de liberdade.

- Assim, o valor-p é calculado por:

$$\widehat{\alpha} = P\left(F > f_{obs}\right),$$

com $F \sim F_{k-1,k(n-1)}$.

Análise de Variância de um fator (ANOVA)

Tabela da ANOVA

Tabela 2: ANOVA

Fonte de variação	Soma dos quadrados	Graus de liberdade	Quadrado Médio	f_{obs}	Valor-p
Tratamentos	SQ_{Trat}	$k-1$	$QM_{Trat} = \frac{SQ_{Trat}}{k-1}$	$\frac{QM_{Trat}}{QME}$	$P(F > f_{obs})$
Erro	SQE	$k(n-1)$	$QME = \frac{SQE}{k(n-1)}$		
Total	SQT	$kn-1$	$-$		

Análise de Variância de um fator (ANOVA)

Voltando no Exemplo 1

Matematicamente,

$$H_0 : \mu_A = \mu_B = \mu_C,$$

onde

- μ_A é a média de redução da pressão arterial para o grupo que realizou Caminhada;
- μ_B é a média de redução da pressão arterial para o grupo que realizou Corrida;
- μ_C é a média de redução da pressão arterial para o grupo que realizou Natação.

$$H_1 : \text{Pelo menos uma das médias } \mu_i \text{ é diferente.}$$

Ou seja, pelo menos uma das médias μ_A, μ_B ou μ_C é diferente das outras.

Análise de Variância de um fator (ANOVA)

Voltando no Exemplo 1

Considere que os três grupos têm 10 pacientes (em cada grupo). Neste caso, temos

- **Fator:** Tipo de exercício (Caminhada, Corrida, Natação).
- **Variável de interesse:** Redução na pressão arterial (em mmHg).
- **Grupos:** Três grupos de pacientes, com 10 pacientes em cada grupo.
- Na segunda parte da nossa aula, faremos este exemplo no software `R`.

Referência bibliográfica

1. Daniel, W. W., & Cross, C. L. *Biostatistics: A Foundation for Analysis in the Health Sciences* (11th ed.). Hoboken, NJ: Wiley, 2018.

Especialização em *Data Science* e Estatística Aplicada

Módulo II - Análise estatística de várias populações

Profa. Dra. Tatiane F N Melo

tmelo@ufg.br

IME Instituto de Matemática e Estatística

FEN Faculdade de Enfermagem

UFG Universidade Federal de Goiás