Universidade de São Paulo

Escola Superior de Agricultura "Luiz de Queiroz"

Seção Técnica de Informática

ANOVA e Estatísticas não-paramétricas

Marcelo Corrêa Alves

Aprender Fazendo

– Piracicaba / 2016 –

_____ANOVA e Estatísticas não-paramétricas

Sumário

1 Introdução

Tendo sido abordados os conceitos básicos da comparação de médias por meio do teste t para duas amostras e para dados pareados, passa-se ao estudo da análise de variância, uma técnica mais abrangente e largamente utilizada no campo da pesquisa científica, sobretudo no âmbito experimental.

A análise de variância é uma ferramenta versátil, posto que se baseia em um modelo que pode ser modificado de acordo com as especificidades de cada delineamento experimental ou observacional proposto. Apesar de estar mais afeita ao rigor obtido nas condições experimentais, a facilidade de interpretação e o poder das inferências obtidas por meio da técnica fazem com que ela seja aplicada, também, em condições observacionais, apesar de serem necessárias certas concessões do método para esta derivação.

Partindo-se do teste t para duas amostras independentes e, como o nome já especifica, uma condição bastante restrita na qual somente podem haver dois grupos com médias que serão comparadas, temos na análise de variância uma ferramenta bem mais genérica e cujas hipóteses de nulidade são sempre da forma apresentada na equação 1.

$$H_0: \mu_{G1} = \mu_{G2} = \mu_{G3} = ... = \mu_{Gn} \quad (1)$$

Onde G1, G2, ... Gn indicam grupos de números que se sintetizam em médias (μ) que estão sendo comparadas, todas elas entre si.

Rejeitar H_0, como sempre, implica na existência de indícios para aceitação de uma hipótese alternativa (H_a) e que se associa à hipótese científica a qual é enunciada na equação 2.

$$H_a: \mu_{Gi} \neq \mu_{Gj} \ (i \neq j) \quad (2)$$

Rejeitar a hipótese de nulidade com apoio de uma análise de variância corresponde a concluir dentro do nível de significância especificado a priori (α) que há diferença entre, pelo menos duas dentre as médias comparadas, o que fundamenta a aplicação de um teste adicional como os testes de Tukey, de Duncan, Dunnett, Scheffe, Bonferroni, REGWQ, ... para comparar de forma mais pormenorizada quais são as médias que diferem entre si.

Do ponto de vista prático, o que se faz é o ajuste de um modelo matemático e se avalia a aderência desse modelo às distribuições observadas nos dados que compõem as médias havendo, de acordo com os resultados, ajustes de qualidade variáveis, conforme ilustrado na figura 1, onde são representadas as distribuições e o ajuste ao modelo.

Sem ajuste	Ajuste razoável	Excelente ajuste

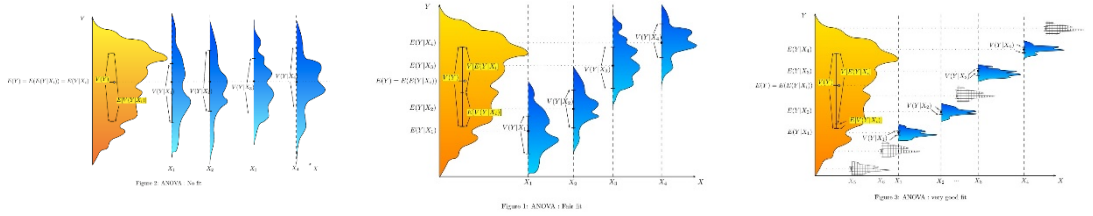

Figura 1. Comparação de ajustes de dados ao modelo de análise de variância. Fonte: By Vanderlindenma - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=31264414

O objetivo desse texto é o de guiar uma aula do tipo “aprender fazendo” no qual o SAS será usado como ambiente de resolução de problemas cujas dificuldades vão crescendo e decisões vão sendo tomadas de maneira conjunta, com professores e alunos participando das decisões.

2 Objetivo

Nesse capítulo objetiva-se que você:

- Reconheça casos de aplicação da técnica de análise de variância;
- Identifique os modelos de análise de variância adequados;
- Ajuste os modelos;
- Avalie as suposições
- Identifique problemas do modelo e dos conjuntos de dados
- Adote soluções apropriadas diante de problemas.

3 Casos para estudo

Em seguida serão descritos casos que poderão ser analisados por meio de modelos de análise de variância para experimentos inteiramente casualizados e para experimento casualizados em blocos, não sendo, em princípio, relevantes a existência de diferenças entre eles, o que já foi abordado nas aulas anteriores.

3.1 Céu de brigadeiro

Um primeiro conjunto de dados se refere a um experimento conduzido por um pesquisador interessado em comparar as médias de durabilidade de três tecidos (Algodão, Seda

e Poliéster) submetidos a 20 ciclos de lavagem, tendo sido medida a perda de massa de cada amostra apresentados na tabela 1.

Tabela 1. Perda de massa de tecidos (mg) submetidos a 20 ciclos de lavagem.

Algodão	Seda	Poliéster
0,71	0,89	0,38
1,02	0,63	0,36
0,70	0,97	0,13

Fonte: Dados fictícios

3.2 Um estranho no ninho

Com objetivo de comparar três 4 tipos de rações em peixes, um pesquisador avaliou o ganho de peso (Kg) de lotes de peixes criados em tanques rede alocados em 4 lagoas de uma propriedade. Objetivando evitar o efeito das lagoas, foram aleatoriamente definidos quatro tanques, suficientemente distantes um dos outros para que não houvesse interferência de uma ração no tanque alocado para outra. Os dados são apresentados na tabela 2.

Tabela 2. Ganho de peso (Kg) de lotes inicialmente similares tratados com 4 rações.

Lago	Ração	Ganho de peso (Kg)
1	R1	55,03
1	R2	28,08
1	R3	22,00
1	R4	19,57
2	R1	30,04
2	R2	27,68
2	R3	22,04
2	R4	25,80
3	R1	24,24
3	R2	23,36
3	R3	16,53
3	R4	20,97
4	R1	23,29
4	R2	24,46
4	R3	17,28
4	R4	17,50

Fonte: Dados fictícios

3.3 E viva a revolução

Um professor resolveu testar a existência de diferença de médias entre três provas, para isso, tomou seis alunos de cada uma de suas cinco classes. Em cada classe dois alunos aleatoriamente selecionados responderam a prova dissertativa, dois alunos a prova em testes e dois alunos a prova virtual.

A hipótese científica é a de que o tipo de prova é determinante de variações nas notas dos alunos e para isso deseja-se fazer uma comparação de médias entre os tipos de prova.

As notas obtidas pelos alunos das três provas são apresentadas na tabela 3.

Tabela 3. Notas obtidas por alunos em três tipos de avaliação.

Tipo de prova	Classe	Aluno	Nota da avaliação
Dissertativa	1	1	11.67
Dissertativa	1	2	24.67
Teste	1	1	18.00
Teste	1	2	34.97
Virtual	1	1	10.48
Virtual	1	2	12.33
Dissertativa	2	1	14.59
Dissertativa	2	2	13.74
Teste	2	1	13.66
Teste	2	2	18.85
Virtual	2	1	22.62
Virtual	2	2	13.48
Dissertativa	3	1	8.37
Dissertativa	3	2	8.53
Teste	3	1	17.28
Teste	3	2	9.64
Virtual	3	1	6.11
Virtual	3	2	16.09
Dissertativa	4	1	12.24
Dissertativa	4	2	14.17
Teste	4	1	11.24
Teste	4	2	8.37
Virtual	4	1	9.42
Virtual	4	2	7.71
Dissertativa	5	1	12.32
Dissertativa	5	2	19.14
Teste	5	1	11.99
Teste	5	2	34.40
Virtual	5	1	7.63
Virtual	5	2	9.78

Fonte: Dados fictícios

4 Estatísticas não-paramétricas

Nos exemplos anteriores são apresentadas situações que podem ser resolvidas por meio da aplicação de análises paramétricas, o primeiro exemplo, sem a necessidade da adoção de nenhuma medida saneadora, o segundo caso, havendo a necessidade de se excluir um dado claramente discrepante e o terceiro caso no qual se pode resolver o problema encontrado por meio de uma transformação de dados.

Existem situações, todavia, em que o modelo resultante do processo de análise de variância não atende os requisitos básicos dentre os quais se destacam a necessidade de que os resíduos sejam aderentes à distribuição gaussiana e a de que os dados sejam homocedásticos.

Nesses casos, há uma desconfiança de que o modelo possa ser aplicado no intuito de gerar estatísticas confiáveis e a adoção da análise paramétrica esbarra na necessidade de se assumir que as estatísticas resultantes do modelo não são exatas.

Quanto menos aderentes são os resíduos, em relação à distribuição gaussiana, menos aproximadas são as estatísticas, chegando-se ao ponto de ser inaceitável a adoção das técnicas paramétricas, ponto em que se recomenda a adoção das técnicas não-paramétricas.

4.1 Nem tudo é normal

Certas áreas e certos tipos de dados são classicamente problemáticos em relação à natureza dos dados gerados e, muitas vezes, até se criam atalhos próprios dada a recorrência dos problemas encontrados.

As análises em seguida serão desenvolvidas com base em um único estudo conduzido com o objetivo de se expor a dados que, de alguma forma, não atendem os requisitos para a realização de uma análise de variância.

Os dados se referem a um estudo com microrganismos no qual se testou 3 métodos de preparo de substratos (tratamentos). Posteriormente ao preparo do substrato, foi feita a inoculação de uma quantidade inicial de um fungo nos substratos preparados.

Três semanas após a inoculação do fungo foram avaliados o pH e a quantidade de unidades formadoras de colônias desse fungo, mesmo momento em que foi feita a inoculação de uma bactéria, cujo crescimento espera-se, seja afetado pela presença do fungo.

Duas semanas após a inoculação da bactéria foi avaliada a existência de bactérias sobreviventes e a porcentagem de crescimento dessas bactérias.

Observe na tabela 4 que alguns dados podem ser considerados estranhos quando avaliados visualmente. O pH até não é tão estranho, entretanto, as medidas de UFC são valores muito grandes e com uma variabilidade igualmente grande.

As medidas de sobrevivência são de natureza binária já que houve (1) ou não ouve crescimento bacteriano no substrato preparado com o fungo. Por fim, as medidas de porcentagem guardam uma relação com os dados de sobrevivência de forma que, onde não houve sobrevivência também não houve crescimento, entretanto, onde foi observado crescimento há valores numa escala de grandeza maior.

Nesses casos, por vezes, adota-se por questões que não se relacionam aos pressupostos, a análise não paramétrica e em outros casos é feita a avaliação das suposições e diante da inobservância das suposições que embasam as técnicas paramétricas, adotam-se técnicas não-paramétricas.

Tabela 4. Avaliação do desenvolvimento de fungos e bactérias em substratos preparados por meio de três métodos de controle.

Tratamento	Inoculação do fungo		Bactéria	
	pH	UFC	Sobrevivência	Porcentagem
Controle	4,480	225200000	0	0
Controle	4,551	2162000000	0	0
Controle	4,527	284600000	0	0
Controle	3,503	240050000	1	12
Controle	3,592	258950000	0	0
Controle	3,672	213950000	0	0
Controle	4,700	231500000	0	0
Controle	4,780	227450000	0	0
Controle	4,787	235550000	1	22
Controle	3,584	226100000	0	0
Controle	3,689	231950000	0	0
Controle	3,777	222050000	0	0
Térmico	4,483	27200000	0	0
Térmico	4,483	276950000	0	0
Térmico	4,464	240950000	1	35
Térmico	3,762	240050000	1	36
Térmico	3,806	213050000	1	32
Térmico	3,709	238250000	0	0
Térmico	3,644	217100000	0	0
Térmico	3,623	211250000	1	41
Térmico	3,446	230150000	1	28
Térmico	4,734	255350000	1	45
Térmico	4,571	2270000	0	0
Térmico	4,543	224750000	0	0
Químico	3,690	21278000000	1	55
Químico	4,614	21116000000	1	58
Químico	4,639	22385000000	1	62
Químico	4,614	21800000000	1	61
Químico	4,721	21530000000	0	0
Químico	4,628	21575000000	1	58
Químico	3,405	23600000000	1	50
Químico	3,440	27560000000	1	46
Químico	3,481	224750000	0	0
Químico	4,611	222950000	0	0
Químico	4,860	220700000	1	49
Químico	4,861	28100000	1	45

Fonte: Dados fictícios

No SAS o procedimento npar1way é usado para a aplicação de testes não-paramétricos tradicionais quando se trata de experimentos com um único fator (One-way ANOVA), o que é verdadeiro para o corrente estudo.

Sendo assim, para executar a análise a partir do SAS data set, partiremos da premissa de que uma técnica não-paramétrica deve ser aplicada, entretanto, sugerimos a realização dos estudos por meio da análise de variância convencional e o estudo de suas suposições, suposições assim como os resultados observados nas duas análises.

O formato de uso do proc npar1way é muito parecido com o proc ttest quando usado para efetuar análises de duas amostras independentes;

```
proc npar1way data=_________ wilcoxon|median|anova|…;
     class _________;
     var _________;
run;
```

4.2 Os postos são a chave

Um passo importante na aplicação das técnicas não paramétricas é o cálculo dos postos, fundamento que embasa muitas das técnicas não paramétricas aplicadas e uma solução por vezes usada quando se deseja a análise de dados de experimentos com mais de um fator no SAS.

O cálculo dos postos é feito por meio do *proc rank* e uma solução híbrida, que mescla as técnicas paramétricas e não-paramétricas é a análise de variância baseada em postos sobre os quais se aplica a análise de variância convencional (Teste F e eventualmente teste t para comparações múltiplas de médias).

O proc rank é muito similar ao proc sort e podemos representar seus comandos mais frequentemente utilizados da seguinte forma:

```
proc rank data=_________ out=_________ normal=blom|tukey|VW;
     var _________;
run;
```

O cálculo dos postos é feito por meio do proc rank e uma solução híbrida, que mescla as técnicas paramétricas e não-paramétricas é a análise de variância baseada em postos sobre os quais se aplica a análise de variância convencional (Teste F e eventualmente teste t para comparações múltiplas de médias).

No caso de experimentos casualizados em blocos, os postos são calculados dentro dos blocos e, em seguida é aplicada a análise de variância nos postos, mas como se o experimento fosse inteiramente casualizado já que o efeito dos blocos é eliminado. Nesse caso, usa-se o comando by no procedimento rank.

```
proc sort data=_________;
     by bloco; * Bloco é a variável que identifica o bloco;
run;
proc rank data=_________ out=_________ normal=blom|tukey|VW;
     var _________;
     by bloco; * Bloco é a variável que identifica o bloco;
run;
```