ANÁLISE DE VARIÂNCIA COM UM FATOR: UM ESTUDO DIRIGIDO

Lisbeth Kaiserlian Cordani / Julio da Motta Singer
Departamento de Estatística
Universidade de São Paulo

Amostras de diferentes populações são muitas vezes selecionadas com a finalidade de se avaliarem possíveis diferenças entre as distribuições populacionais de alguma característica de interesse. Exemplos de tais situações são:

1. Seleção de lotes de vacinas com datas de fabricação diferentes para avaliar se a distribuição (populacional) do número médio de anticorpos produzidos varia com o tempo de armazenamento.
2. Determinação da pressão diastólica em grupos de indivíduos com diferentes níveis de colesterol sérico para avaliar se nas subpopulações das quais os indivíduos são oriundos a distribuição da pressão diastólica depende do nível de colesterol sérico.
3. Análise de amostras de água de diversos pontos da cidade para verificar se existe variação na distribuição de alguma característica (concentração de um agente mutagênico, por exemplo) associada à qualidade da água entre bairros.

Em muitos casos, uma análise descritiva dos dados indica que modelos *Gaussianos (normais)* são compatíveis com suas distribuições. Em outras palavras, sob o ponto de vista estatístico, podemos considerar as k amostras disponíveis como provenientes de populações normais com médias $\mu_1,\mu_2,...,\mu_k$. Se não existirem razões contrárias, podemos também supor que as amostras são independentes. Adicionalmente, a análise descritiva, muitas vezes sugere que as k populações têm a mesma variância σ^2 (desconhecida). O problema proposto pode então ser encarado como um teste da hipótese H: $\mu_1=\mu_2=...=\mu_k$.

A técnica para resolver problemas desse tipo (e muitos outros) é chamada de **Análise de Variância**.

A idéia subjacente pode ser ilustrada por meio de um caso particular com 2 amostras provenientes de distribuições normais independentes, com variâncias desconhecidas (mas iguais), em que queremos testar a hipótese

$$H: \mu_1 = \mu_2.$$

Evidentemente, neste caso, k = _ _ _.

Como você já sabe resolver problemas desse tipo por meio de outra metodologia, considere o seguinte exercício:

Exercício 1: Os valores da resistência à tração (Mpa) de corpos de prova de plástico preparados por dois processos diferentes são:

PROCESSO 1	PROCESSO 2
6,1	9,1
7,1	8,2
7,8	8,6
6,9	6,9
7,6	7,5
8,2	7,9

Use α=5% para testar se os dados apresentam evidência suficiente para concluir que as elasticidades médias (populacionais) para os dois processos são iguais.

Quais as suposições necessárias para resolver este problema com os métodos da teoria normal?

_____
_____
_____
_____

A que conclusão você chegou?

_____
_____
_____

Agora vamos resolver o mesmo problema por intermédio de Análise de Variância.

Seja y a variável correspondente à resistência à tração no exercício anterior. Usando os mesmos valores observados, complete o seguinte quadro:

PROCESSO 1	PROCESSO 2
y_{11}= 6,1	y =
= 7,1	=
y_{13}=	=
= 6,9	y_{24}=
y_{15}= 7,6	= 7,5
=	=

Explique o significado da notação y_{23}?

_____
_____

Idem para y_{ij}.

No nosso problema, que valores i pode assumir?

Idem para j.

Se tivéssemos n_1 observações sob o Processo 1 e n_2 observações sob o Processo 2, que valores i poderia assumir? E j?

Se tivéssemos k amostras, com n_i observações cada, que valores i poderia assumir? E j?

Voltando ao nosso problema particular em que k=2 (isto é, dispomos de duas amostras), vamos considerar o caso em que $n_1=n_2$. Esta última igualdade representa simplesmente o quê?

Vamos pensar em todas as observações em conjunto (no exemplo dado seriam 12); poderíamos calcular uma média geral e chamá-la $\bar{y}$.

Assim, em termos de y_{ij}, complete a expressão abaixo

$$\bar{y} = \frac{y_{11} + y_{12} + \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots}{n_1 +}$$

Tente condensar a expressão acima usando notação de somatório.

Se pensássemos em termos de uma variação total das observações y_{ij} em relação à média geral $\bar{y}$, como você poderia quantificá-la? (Dê inicialmente sua resposta em palavras e em seguida tente resumi-la na forma de uma expressão).

Chamemos esta expressão de SQT (soma de quadrados total). Então

SQT= (1)

Vamos agora pensar em cada amostra individualmente; como podemos quantificar a variabilidade dentro de cada amostra? Tente apresentar uma expressão para isso, tanto no caso do Processo 1 quanto no caso do Processo 2 (em termos de y_{ij}).

Então você poderia englobar as duas expressões encontradas (uma para cada amostra) numa expressão só, que significaria a variação total **dentro** das amostras.

Tente explicitar essa expressão geral

Vamos chamar esta expressão de SQD (soma de quadrados dentro). Então

SQD = (2)

Você poderia dizer se SQT é maior ou menor do que SQD? Justifique.

Pense novamente em todos os grupos (ou amostras), que no nosso caso são dois. Se você desejar comparar cada grupo com a média geral $\overline{y}$ por meio de um único valor (um para cada grupo) que valor você usaria?

Mas especificamente, para o 1º grupo quem seria esse valor? E para o 2º?

_____
_____

Como você usaria esses valores para medir a variabilidade dos grupos em relação à média geral?

_____
_____

Nas respostas anteriores, você utilizou um único valor como representante do grupo, isto é como representante das n_i observações do i-ésimo grupo. Como você poderia ponderar a medida acima para levar em conta este fato? (Você acha que grupos com maior número de observações deveriam ter um peso maior?)

_____
_____

Na realidade, o que estamos tentando com isso é definir uma medida de variabilidade **entre** os grupos (ou amostras). Experimente explicitar uma expressão para essa variabilidade.

_____
_____
_____

Chamemos esta expressão de SQE (soma de quadrados entre grupos)

SQE =	(3)

Observe atentamente SQT, SQD e SQE e tente descrever (sem utilizar símbolos matemáticos) o significado de cada uma delas:

SQT_____
_____
SQD_____
_____
SQE_____
_____

Dê um palpite para relacionar SQT com SQD e com SQE. (obs. – a teoria prova essa relação, mas foge aos nossos objetivos demonstrar esse fato aqui).

_____
_____

Conclusão

SQ__= SQ__+ SQ__

Usando a terminologia estatística, isto corresponde a **particionar** a SQ__ em duas parcelas, SQ__ e SQ__.

Vamos examinar novamente SQD, cuja expressão é

_____

Desenvolva esta expressão com relação a i (no nosso caso, i=1,2).

_____
_____
_____

Tal expressão lhe parece familiar?

_____
_____
_____

Faça o quociente entre ela e n_1+n_2-2, isto é

$$\frac{SQD}{n_1+n_2-2} = \qquad (4)$$

Onde você usou algo com essa forma?

_____
_____
_____

Vamos chamar a expressão em (4) de QMD (**quadrado médio dentro** de grupos). Então

$$QMD = \qquad (5)$$

QMD é estimativa de _____
_____

Voltando a SQE, mostre que, para o caso especial em que k=2 e $n_1=n_2$ a expressão SQE pode ser escrita como

$$SQE = \frac{n_1}{2}(\bar{y}_1 - \bar{y}_2)^2 \qquad (6)$$

_____
_____
_____
_____
_____

Voltemos ao início do problema: O que estamos querendo testar? Quem é a hipótese H para o nosso caso particular?

_____
_____

Será que você poderia ligar a expressão em (6) com um possível critério para rejeição de H? Justifique.

_____
_____
_____
_____

A teoria demonstra (mais uma vez!) que se fizermos o quociente entre SQE e um número apropriado, chamado de graus de liberdade (no nosso caso de 2 amostras, este número é 1; no caso de k amostras este número é k-1), então este quociente SQE/1 será a estimativa de $\sigma^2+n_1(\mu_1-\overline{\mu})^2+n_2(\mu_2-\overline{\mu})^2$, ou seja, de

$$\sigma^2 + \sum_{i=1}^{2} n_i(\mu_i - \overline{\mu})^2 \qquad (7)$$

Então podemos definir

$$QME = \frac{SQE}{1} \qquad (8)$$

O que acontece com (7) quando H: $\mu_1=\mu_2$ é verdadeira? Lembre-se que $\overline{\mu}$ é $(\mu_1 + \mu_2)/2$

_____
_____
_____

Sob estas circunstâncias, isto é, se H é verdadeira, QME seria estimativa de quem? Justifique

_____
_____
_____

Nós já vimos outra estimativa para σ^2, independentemente de H ser ou não verdadeira. Qual é essa estimativa?

_____
_____
_____
_____

O que você poderia dizer que “espera” do quociente QME/QMD se H for verdadeira?

_____
_____
_____

E se H for falsa o que você “espera” de tal quociente? [Sugestão: lembre-se de (7)],

_____
_____
_____

É evidente que QME/QMD varia de experimento para experimento, pois depende das observações. Então podemos associar a esse quociente, uma distribuição, que é a distribuição F, cujos parâmetros (graus de liberdade) são (no nosso caso) 1 e (n_1+n_2 –2). Mais especificamente

$$\frac{\frac{SQE}{1}}{\frac{SQD}{n_1+n_2-2}} = \frac{QME}{QMD} = F_{observado} \qquad (9)$$

Para um certo nível de significância α e com os parâmetros (1) e (n_1+n_2-2) podemos determinar na tabela adequada o valor crítico de F. [Observe, na expressão (9), de onde vêm os parâmetros de F]

_____

Como se faz usualmente, rejeitamos H se

$$F_{observado} \geq F_{crítico}$$

Seria razoável esperar $F_{observado} < 1$? Por quê?

_____
_____

Como você justifica o nome Análise de Variância se o nosso teste é sobre médias?

_____
_____
_____
_____

Resolva novamente o problema proposto inicialmente, agora por meio da Análise de Variância. Use α=5% e compare esta análise com o primeiro método de resolução.

GENERALIZAÇÃO

Se quisermos agora fazer o teste para comparar k populações (normais e com mesma variância) em termos de suas médias qual seria a hipótese H?

_____
_____

Quais os possíveis valores de k?

_____
_____

As suposições feitas inicialmente, ainda valem para este caso geral, e são:

1:_____
_____

2:_____
_____

3:_____
_____

No caso particular de k=2 (isto é, duas amostras) usamos $n_1=n_2$ (ou seja, número de observações da primeira amostra = número de observações na segunda amostra). Isto não é necessário, e, para este caso mais geral de k amostras, vamos considerar cada uma tenha n_i observações. Neste caso, quais os valores possíveis para i?

_____
_____
_____

Evidentemente o número total de observações seria:

n=

Do mesmo modo descrito anteriormente, temos que obter expressões para calcular F_{obs} = QME/QMD.

Como esta técnica vale para qualquer $k \geq 2$, será interessante introduzir uma nova terminologia na tentativa de simplificar as fórmulas para cálculo. Então:

1. y_{ij} denota (como sempre)_____
_____

2. T_i= soma de todas as observações na i-ésima amostra. Quanto seria T_2 no nosso exemplo especial de k=2?
_____
_____

3. $\overline{y_i}$ = média das observações na i-ésima amostra. Quanto seria $\overline{y}_1$ no exemplo especial de k=2?
_____
_____

Aqui, também pode-se mostrar que

$$SQT = SQD + SQE \qquad (10)$$

em que, para este caso de k amostras, cada uma com n_i observações,

SQT =

SQE =

SQD =

Para efeito de cálculo podemos usar as seguintes expressões:

$$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - n\bar{y}^2 \qquad (11)$$

$$SQE = \sum_{i=1}^{k} \frac{T_i^2}{n_i} - n\bar{y}^2 \qquad (12)$$

SQD = (13)

$$QME = \frac{SQE}{k - 1} \qquad (14)$$

$$QMD = \frac{SQD}{n_1 + n_2 + n_k - k} \qquad (15)$$

$$F_{observado} = \frac{\frac{SQE}{k - 1}}{\frac{SQD}{\sum_{i=1}^{k} n_i - k}} = \frac{QME}{QMD} \qquad (16)$$

Então, a hipótese H que estamos testando e que é H: _____ será rejeitada se

$$F_{obs} \_ \_ \_ F_{crítico}$$

em que $F_{crítico}$ encontra-se tabelado para valores usuais de α. Como sabemos, os parâmetros da distribuição F são os números de graus de liberdade associados, respectivamente, ao numerador e denominador da expressão (16). Então, quais são, neste caso, tais parâmetros?

_____
_____
_____

Resolva os exercícios abaixo usando o método exposto, indicando as suposições necessárias e as hipóteses.

Exercício 1 – Um psicólogo clínico deseja comparar três métodos para reduzir o nível de hostilidade em estudantes universitários. Um certo teste (HLT) foi usado para medir o grau de hostilidade (quanto maior a nota no teste, maior o grau de hostilidade). Onze alunos que obtiveram notas altas e aproximadamente iguais foram usados no estudo. Cinco deles escolhidos ao acaso foram tratados pelo método A. Três escolhidos ao acaso entre os restantes foram tratados pelo método B e os outros, pelo método C. O tratamento teve a duração de um semestre. Ao fim do semestre, o grupo de alunos foi submetido novamente ao teste HLT e os resultados foram

MÉTODO A	73	83	76	68	80
MÉTODO B	54	74	71		
MÉTODO C	79	95	87		

Você acha que há evidência suficiente para indicar uma diferença (populacional) entre os resultados médios para os três métodos, após o tratamento?

Exercício 2 – Vinte e uma crianças de 1 ano foram divididas em 3 grupos e cada grupo foi submetido durante um período de 2 anos a uma dieta rica em vitaminas A, B e C, respectivamente. Mediu-se o peso ganho em quilogramas e os dados são apresentados abaixo:

A	B	C
5,1	4,2	4,7
4,4	5,4	5,2
3,7	4,3	4,0
4,1	4,6	3,6
5,0	4,7	4,9
3,3	4,1	3,8
3,7	3,8	4,6

Há evidência estatística de que as dietas são homogêneas quanto ao ganho de peso médio?