NÃO NORMALIDADE E TESTES DE HIPÓTESES

SOBRE VARIÂNCIAS E MÉDIAS

ADILINA MARIA SILVA DA CUNHA

DISSERTAÇÃO APRESENTADA

AO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA

UNIVERSIDADE DE SÃO PAULO

PARA OBTENÇÃO DO GRAU DE MESTRE

EM

ESTATÍSTICA

ORIENTADOR:

Prof. Dr. CLÓVIS DE ARAÚJO PERES

Durante a realização deste trabalho a autora recebeu apoio financeiro do CAPES e FINEP.

- SÃO PAULO, AGOSTO DE 1978 -

Aos meus pais

José Quirino

e

Francisquinha

Ao

Gre

AGRADECIMENTOS

Ao Professor Doutor Clóvis de Araújo Peres, pela sugestão do tema, dedicação e eficiência na orientação desta dissertação.

Ao Professor Doutor Adolpho Walter Pimazoni Canton, pelas sugestões apresentadas.

Aos Professores Gregório Maranguape da Cunha e Roberto Cláudio Frota Bezerra.

Aos colegas do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará.

Aos Professores do Departamento de Estatística do Instituto de Matemática e Estatística da Universidade de São Paulo.

Aos colegas do Curso de Pós-Graduação do Instituto de Matemática e Estatística da Universidade de São Paulo.

Ao Senhor João Baptista Esteves de Oliveira, pela datilografia deste trabalho.

ÍNDICE

I - INTRODUÇÃO 1
II - TESTES DE HIPÓTESES ENVOLVENDO UMA POPULAÇÃO 5
2.1 - Testes de Hipóteses sobre a Variância. 5
2.1.1 - A população tem distribuição normal. 5
2.1.2 - A população não tem distribuição normal. . . . 6
2.2 - Testes de Hipóteses sobre a Média. 12
2.2.1 - A população tem distribuição normal. 12
2.2.2 - A população não tem distribuição normal. . . . 13
2.2.3 - As observações não são independentes 14
III - TESTES DE HIPÓTESES ENVOLVENDO DUAS POPULAÇÕES 19
3.1 - Testes de Hipóteses sobre duas Variâncias 19
3.1.1 - As populações têm distribuição normal. 19
3.1.2 - As populações não têm distribuição normal. . . 21
3.2 - Testes de Hipóteses duas Médias. 29
3.2.1 - As amostras são independentes e as populações normais com variâncias iguais. 29
3.2.2 - As amostras são independentes e as populações normais com variâncias diferent es 31
3.2.3 - As amostras são independentes e as populações não normais com variâncias iguais. 35
3.2.4 - As amostras são independentes e as populações não normais com variâncias diferentes. 40
3.2.5 - As amostras são dependentes e as população são normais. 46
3.2.6 - As amostras são dependentes e as população não normais. 49
IV - TESTES DE HIPÓTESES ENVOLVENDO MAIS DE DUAS POPULAÇÕES . . . 54
4.1 - Testes de Hipóteses sobre Várias Variâncias. 54
4.1.1 - As população são normais 55
4.1.2 - As população não são normais 60
4.2 - Testes de Hipóteses sobre Várias Médias. 64
4.2.1 - As populações são normais e as variâncias homogêneas 64

4.2.2 - As populações são normais e as variâncias heterogêneas 66
4.2.3 - As populações não são normais e as variâncias homogêneas 67
4.2.4 - As populações não são normais e as variâncias heterogêneas 72
4.3 - Comparações Múltiplas. 76
4.3.1 - Método de Scheffé. 76
a) Populações Normais e variâncias iguais. . .
b) Populações Normais e variâncias diferentes.
4.3.2 - Contrastes ortogonais. 79
a) Populações Normais e variâncias iguais. . .
b) Populações Normais e variâncias diferentes.
SUMÁRIO 87
BIBLIOGRAFIA 96

\- o -

CAPÍTULO I

INTRODUÇÃO

A técnica estatística conhecida como Análise de Variância desenvolvida por R.A.Fisher no perído de 1925 a 1944 teve como finalidade a Análise e Interpretação dos dados em pesquisas biológicas e experimentos agrícolas. Sua aplicação estendeu-se rapidamente ao campo de quase todas as ciências.

Em 1947 foram publicados na revista *Biometries* três artigos que vieram de certa maneira alertar os pesquisadores sobre os cuidados que deviam ter no uso da Análise de Variância.

O primeiro escrito por Eisenhart (1947) discute as suposições básicas da Análise de Variância (Normalidade, Independência e Igualdade das Variâncias dos Erros) e fez uma distinção entre modelo fixo e componentes de Variâncias.

O segundo desenvolvido por Cochran (1947), analisa algumas consequências quando as suposições básicas não são mantidas.

O terceiro por Bartlett (1947) propõe transformações na variável dependente que tornam possível o uso da Análise de Variância mesmo nos casos em que as suposições não são satisfeitas.

Embora estes artigos tenham fornecidos alguns caminhos para a solução do problema, as pesquisas no sentido de se examinar consequências das violações das suposições e soluções para estas situações, continuam a ser estudados, e é comum vermos nas revistas especializadas artigos tratando do assunto.

Nessa monografia mostramos que quando as suposições básicas não são mantidas, o nível de significância α quase sempre se altera. Em algumas situações é possível determinar procedimentos que permitem corrigir o nível de significância, em outras situações aplicaremos testes aproximados ou modificados que resistem as violações das suposições.

Para o estudo destas questões tomamos como base o capítulo 10 do livro de Scheffé (1959) e complementamos com vários artigos mais recentes conforme indica a bibliografia. Para melhor entendimento nós dividimos estes resultados considerando o número de populações envolvidas; testes sobre a média e a variância para uma população, para duas populações e várias populações, porque nem todos resultados são facilmente generalizados.

Para os testes sobre a média amostral de uma população vemos que o efeito da não normalidade para n grande, sendo n o tamanho da amostra, é insignificante, no sentido de que o teste é independente da distribuição da variável aleatória X, enquanto que para testes sobre a variância, o nível de significância α se altera acarretando sérios erros nas conclusões do teste. Notemos ainda que o nível de significância é modificado no teste sobre a média, quando existe correlação serial ρ, entre as observações. No capítulo II nós estudamos este efeito e apresentamos uma maneira de corrigí-los. Dois exemplos são usados para ilustrar a teoria.

No caso de duas populações, os testes comumente mais usados são aqueles comparando, as duas variâncias σ_1^2 e σ_2^2, ou as duas médias μ_1 e μ_2. A fim de testar a igualdade das variâncias sob as suposições básicas, nós usamos o Teste F ou o teste de Bartlett, enquanto para igualdade de duas médias, sob as suposições exigidas, utilizamos o teste t de Student. Observamos o efeito causado pela não normalidade, nos testes comparando duas variâncias e ressaltamos que a sensibilidade para não normalidade nestes testes foi primeiro apontado por Pearson (1931), cujos resultados foram confirmados por Geary (1947), Finch (1950) e Gayen (1950a). No caso em que as médias μ_1 e μ_2 são conhecidas, um teste de grande robustez para igualdade das variâncias, foi desenvolvido por Box e Andersen (1955). Este teste consiste em determinar um fator de correção para os graus de liberdade da distribuição F Snedecor. No caso em que μ_1 e μ_2 não são conhecidos, foi desenvolvido um teste aproximado a partir da estatística de Bartlett, para $n_1 = n_2$ e $n_1 + n_2$ grande. Um exemplo com dados reais foi apresentado para ilustrar os referidos técnicas. Com testes sobre duas médias, apresentamos o efeito da desigualdade das variâncias no teste t de Student, da maneira considerada por Welch (1937b). No caso de amostras do mesmo tamanho este efeito é pequeno tornando-se maior para amostras de tamanhos diferentes. Os artigos de Hsu (1938) e Grunw (1951) confirmaram o resultado de Welch (1937b). O efeito da não normalidade em testes da igualdade de duas médias, quando as variâncias são diferentes, foi estudado por Scheffé (1959) que apresentou uma solução para $N = n_1 + n_2$ grande. No caso em que as variâncias são iguais Box e Andersen (1955) desenvolveram um teste robsuto para igualdade das mé dias, baseados na Teoria da Permutação.

No capítulo IV, desenvolvemos testes para igualdade de várias variâncias e várias médias. No caso em que as suposições básicas são mantidas apresentamos o teste de Bartlett e de Cochran para igualdade de variâncias, e o teste F da Análise de Variância para igualdade de médias. Alguns métodos de comparações múltiplas foram também apresentados. O efeito da não normalidade em teste para igualdade das variâncias foi apresentado segundo o artigo de Box (1953). Vimos que a "sensibilidade" para não normalidade aumenta a medida que o número de populações a ser comparados cresce.

Box e Andersen (1955) desenvolveram um teste aproximado para o caso acima. Nos testes para igualdade das médias usando a Análise de Variância, algumas estatísticas foram desenvolvidas por Pitman (1937b) e Welch (1938) para o caso da não normalidade. Box e Andersen (1955) baseados em Pitman (1937b) e Welch (1938 e 1947) desenvolveram um teste aproximado, que consiste em uma modificação na estatística F da Análise de Variância e em uma correção nos graus de liberdade na distribuição da estatística F de Snedecor. Discutimos ainda neste capítulo o estudo de testes para igualdade das médias, Contrastes Ortogonais e Método de Scheffé para o caso em que as variâncias nos grupos diferem baseado em Brown e Forsythe (1974). Alguns exemplos para ilustrar a teoria foram apresentados.

Para que se possa usar as informações contidas neste trabalho, de maneira prática, apresentaremos um sumário contendo as principais recomendações para testes sobre variâncias e médias.

- o -

CAPÍTULO II

TESTES DE HIPÓTESES ENVOLVENDO UMA POPULAÇÃO

Os problemas práticos de testes de hipóteses baseados em uma amostra, se resumem, na maioria das vezes em testes da variância e da média.

Neste capítulo iremos discorrer sobre cada um deles, apresentando soluções nos casos em que suposições de Normalidade e de Independência são satisfeitas, como também naqueles casos em que essas suposições não são verificadas.

2.1 - TESTES DE HIPÓTESES SOBRE A VARIÂNCIA

Aqui dividiremos o nosso desenvolvimento em duas partes: uma primeira, em que a amostra que trabalhamos provem de uma população(1) Normal e uma segunda, na qual a amostra não provem de uma população Normal.

2.1.1 - A População tem Distribuição Normal

Seja $X_1, X_2, \ldots, X_n$, uma amostra aleatória, de uma população, caracterizada pela variável aleatória X que segue a distribuição Normal.

No caso de termos interesse em testar a hipótese H_0,

(1) - A palavra população é entendida como o conjunto de todos valores possíveis de uma variável aleatória.

sobre a variância, dada como

$$H_0: \sigma^2 = \sigma_0^2$$

temos como estatística para o teste

$$\chi^2 = (n-1)\frac{s^2}{\sigma_0^2} \qquad (1.1.1)$$

onde,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

que sob a hipótese H_0 tem distribuição χ^2 (qui-quadrado) central com (n-1) graus de liberdade. Sob a alternativa:

$$H_a: \sigma^2 = \sigma_1^2 \neq \sigma_0^2,$$

a estatística acima tem distribuição proporcional a uma χ^2 com (n-1) graus de liberdade, com coeficiente de proporcionalidade igual a σ_1^2/σ_0^2. Esse resultado nos permite, obviamente, construir curvas de poder.

2.1.2 - A População não tem Distribuição Normal

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma população, caracterizada pela variável aleatória X que não segue a distribuição Normal.

O nosso problema consiste em testar as hipóteses,

$$H_0: \sigma^2 = \sigma_0^2$$
$$H_a: \sigma^2 \neq \sigma_0^2$$

com base na amostra aleatória acima.

De início, estudaremos o comportamento da estatística s^2/σ_0^2, calculando sua média e sua variância, no caso em que X tem distribuição normal, como também na situação em que X não tem distribuição normal.

Na seção anterior, vimos que, se X tem distribuição normal, $(n-1)s^2/\sigma_0^2$ tem sob H_0 distribuição χ^2 central, com (n-1) graus de liberdade. Por conseguinte, a estatística

$$s^2/\sigma_0^2 \tag{1.2.1}$$

tem distribuição $\frac{1}{n-1}\chi^2_{(n-1)}$

Usando o fato que

$$E[\chi^2_{(n-1)}] = n-1 \quad \text{e} \quad Var[\chi^2_{(n-1)}] = 2(n-1)$$

temos

$$E[s^2/\sigma_0^2] = 1$$

e

$$Var[s^2/\sigma_0^2] = \frac{2}{n-1}.$$

Temos pelo Teorema do Limite Central (T.L.C.) que a distribuição limite de

$$\left(\frac{n-1}{2}\right)^{1/2} \cdot \left(\frac{s^2}{\sigma_0^2} - 1\right) \tag{1.2.2}$$

é N(0,1). Em consequência, s^2/σ_0^2 é assintóticamente normal com média 1 (um) e variância 2/n-1, ver Cramér [1946]. ?

Suponha agora que a variável aleatória X não tenha distribuição normal. Então aplicando as definições de esperança e variância obtemos:

$$E\left(\frac{s^2}{\sigma_0^2}\right) = 1$$

e

$$\mathrm{Var}\left(\frac{s^2}{\sigma_0^2}\right) = \frac{2}{n-1} + \frac{\gamma_2}{2},$$

onde $\gamma_2 = \beta_2 - 3$ e β_2 é o coeficiente de curtose da população, Bartlett [1959].

Neste caso pelo T.L.C., quando n cresce a variável aleatória (1.2.2) tem distribuição limite $N(0,1+\frac{1}{2}\gamma_2)$ e não $N(0,1)$ como no caso em que $\gamma_2 = 0$, isto é, quando X é Normal. Obviamente, este fato, mesmo para n grande, pode causar um sério erro no nível de significância e função poder, se considerarmos erradamente X como tendo distribuição normal. Mesmo para grandes amostras há necessidade de corrigirmos a efeito de $\gamma_2 \neq 0$, e para isto consideremos um fator de ponderação β, o qual será determinado convenientemente.

Com relação à probabilidade do Erro tipo I consideremos n grande e X não normal. Seja $Z_{\alpha/2}$ um ponto da distribuição Normal padronizada Z, tal que,

$$P_r\{|Z| \geq Z_{\alpha/2}\} = \alpha.$$

Então, a probabilidade do erro tipo I, levando em conta a não normalidade de X é obtido por:

P(|Z| ≥ Z α/2)

$$2\lim_{b\to\infty} \int_{Z_{\alpha/2}}^{b} \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{1+\frac{1}{2}\gamma_2}} \cdot e^{-\frac{1}{2}\left(\frac{1}{\sqrt{1+\frac{1}{2}\gamma_2}}\right)^2} dx = \alpha'$$

Usando a transformação

$$y = \frac{x}{\sqrt{1+\frac{1}{2}\gamma 2}}$$

obtemos

$$2\lim_{b\to\infty} \int_{\frac{Z_{\alpha/2}}{\sqrt{1+\frac{1}{2}\gamma_2}}}^{b} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}y^2} dy =$$

$$= 2\frac{1}{\sqrt{2\pi}} \lim_{b\to\infty} \int_{\frac{Z_{\alpha/2}}{\sqrt{1+\frac{1}{2}\gamma_2}}}^{b} e^{-\frac{1}{2}y^2} dy = \alpha'$$

ou seja

$$\alpha' = 2\cdot\frac{1}{\sqrt{2\pi}} \lim_{b\to\infty} \int_{\frac{Z_{\alpha/2}}{\beta}}^{b} e^{-\frac{1}{2}y^2} dy$$

onde a variável aleatória Y tem distribuição normal padronizada e $\beta = \sqrt{1+\frac{1}{2}\gamma_2}$, $\gamma_2 > -2$.

Observa-se que, quando $\gamma_2 < 0$ (tem distribuição platicúrtica) α' será menor que α. Já no caso de $\gamma_2 > 0$ (X tem distribuição liptocúrtica) α' será maior do que α. E finalmente, quando $\gamma_2 = 0$ temos $\alpha' = \alpha$.

A tabela abaixo ilustra o efeito da não normalidade da variável aleatória X no teste de uma variância. Nesta tabela temos, para alguns valores de γ_2, as correspondentes probabilidades do erro do tipo I, quando testamos

$$\begin{cases} H_0: \sigma^2 = \sigma_0^2 \\ H_a: \sigma^2 \neq \sigma_0^2 \end{cases}$$

para um nível de significância $\alpha = 0,05$.

γ_2	-1	-0,8	-0,5	0	0,5	1	1,5	2	4
Probabilidade	$9\text{x}10^{-5}$	0,008	0,024	0,05	0,08	0,11	0,136	0,17	0,26

O procedimento descrito acima é válido para n grande, pois os resultados são obtidos pelo T.L.C. No caso de desejarmos testar hipóteses sobre a variância com X não normal e valores pequenos de n, sugere-se transformações ou testes não paramétricos.

Sob a hipótese alternativa $H_a: \sigma^2 = \sigma_1^2 \neq \sigma_0^2$ a estatística (1.2.2) é assintoticamente Normal com média e variância dadas respectivamente por:

$$\frac{1-\sigma_1^2/\sigma_0^2}{\sqrt{\frac{2}{n-1}}} \quad \text{e} \quad \left(\frac{\sigma_1^2}{\sigma_0^2}\right)^2 (1 + \frac{1}{2}\gamma_2)$$

Essa informação nos permite construir curvas de poder do teste.

EXEMPLO 2.1.2

Os dados abaixo[1] representam uma amostra aleatória de 79 exemplares de sardinhas nas quais foram contadas os raios da nadadeira dorsal.

---

(1) - Dados fornecidos pelo Instituto Oceanográfico da USP.

Número de raios da nadadeira dorsal	Frequência
16	4
17	39
18	31
19	5

Experiências anteriores baseadas em amostras muito grandes mostraram que a variável acima não segue a distribuição normal, além do fato de apresentar uma variância igual a 0,7. O nosso problema consiste em, a partir dos dados apresentados na tabela acima, testar as hipóteses

$$\begin{cases} H_0: \sigma^2 = 0,7 \\ H_a: \sigma^2 \neq 0,7 \end{cases}$$

Para isto calculamos as estimativas da média, variância e γ_2, encontrando-se respectivamente 17,49; 0,483 e -1,7069.

Levando em conta o efeito da não normalidade a estatística adequada para o teste é dada por,

$$T = \left(\frac{n-1}{2}\right)^{1/2} \left(\frac{s^2}{\sigma_0^2} - 1\right),$$

cujo valor observado é $T_0 = -1,936$.

A região de rejeição é constituída por valores de T, tais que,

$$P(|T|>T_c) = \alpha,$$

onde α é o nível de significância prefixado. Para $\alpha = 0,05$ temos

$\varepsilon_{0,05}$

$$T_c = +1,96 \times \sqrt{1 + \frac{1}{2}\gamma_2} = 1,96 \times 0,3839 = 0,7413$$

Como $|T_0| > T_c$ concluímos então pela rejeição de H_0. Seria interessante observar que sem levar em conta a não normalidade, o valor crítico da estatística seria $T_c = 1,96$ e portanto a hipótese seria aceita,, conclusão contrária a aquela que obtemos.

2.2 - TESTES DE HIPÓTESES SOBRE A MÉDIA

Esta seção tem uma divisão análoga àquela em que tratamos de testes sobre a variância, ou seja, uma parte relativa a testes de hipóteses a partir de uma amostra de uma população normal e uma segunda parte em que temos uma amostra de uma população não normal.

2.2.1 - A População tem Distribuição Normal

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória de uma população caracterizada pela variável aleatória X que segue a distribuição normal.

No caso de termos a variância populacional σ^2 conhecida e desejarmos testar as hipóteses

$$H_0: \mu = \mu_0$$

$$H_a: \mu \neq \mu_0$$

a estatística a ser usada será

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \qquad (2.1.1)$$

que sob a hipótese H_0 terá distribuição normal com média ze ro e variância 1. Sob a hipótese alternativa $H_a: \mu = \mu_1 \neq \mu_0$ a estatística acima tem distribuição

$$N\left\{\frac{\mu_1-\mu_0}{\sigma/\sqrt{n}},\ 1\right\}$$

Essa informação nos permite construir curvas de poder do teste.

Às vezes, a variância populacional σ^2 não é conhecida e temos interesse em testes de hipóteses do tipo visto no início desta seção. O procedimento, nestes casos, consiste em substituírmos σ^2 por s^2 (estimador não viciado de σ^2) e neste caso a estatística para o teste passa a ser,

$$T = \frac{\bar{x}-\mu_0}{\frac{s}{\sqrt{n}}}, \qquad (2.1.2)$$

onde

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i-\bar{x})^2$$

Sob H_0 a estatística acima tem distribuição t de Studen Central com (n-1) graus de liberdade e sob a hipótese alternativa $H_a: \mu = \mu_1 \neq \mu_0$ tem distribuição t de Student não central, com parâmetro de não centralidade igual a

$$\frac{(\mu_1-\mu_0)}{(\sigma/\sqrt{n})}.$$

2.2.2 - A População não tem Distribuição Normal

Seja uma amostra aleatória $(X_1,...,X_n)$ de uma população não normal. Estudaremos aqui o comportamento da estatística (2.1.2) sob a hipótese $H_0: \mu = \mu_0$ como também sob a alternativa H_a: $\mu = \mu_1 \neq \mu_0$.

Sob H_0, usando o T.L.C. e o fato de que s^2 converge em probabilidade para σ^2, temos que a distribuição limite de (2.1.2) é Normal com média zero e variância um, isto é, a mesma do item 2.1.

Sob a hipótese alternativa, pelo mesmo raciocínio a distribuição limite de (2.1.2) é Normal com média

$$\delta = \frac{n^{1/2}(\mu_1 - \mu_0)}{\sigma} \text{ e variância um.}$$

Concluímos assim que para n grande o teste sobre a média *independe* do fato da população ser normal ou não.

2.2.3 - As Observações não são Independentes

Examinaremos um tipo de correlação de particular interesse prático denominado, correlação serial "lag um". Detectaremos esta dependência através do coeficiente de correlação ρ_K, definido abaixo. Veremos de que modo esta dependência influe no teste sobre a média, e apresentaremos uma correção.

DEFINIÇÃO - Coeficiente de correlação serial de "lag k", baseado nas observações $(x_1, x_2, \ldots, x_n)$ é definido por:

$$\rho_K = \frac{\frac{1}{(n-k)} \sum_{i=1}^{n-k} (x_i \cdot x_{i+k}) - \frac{1}{(n-k)^2} \left\{ \left(\sum_{i=1}^{n-k} x_i \right) \left(\sum_{i=1}^{n-k} x_{i+k} \right) \right\}}{\left\{ \frac{1}{n-K} \sum_{i=1}^{n-k} x_i^2 - \frac{1}{(n-k)^2} \left(\sum_{i=1}^{n-k} x_i \right)^2 \right\}^{1/2} \left\{ \frac{1}{n-k} \sum_{i=1}^{n-k} x_{i+k}^2 - \frac{1}{(n-k)^2} \left(\sum_{i=1}^{n-k} x_{i-k} \right)^2 \right\}^{1/2}}$$

Kendall (1951).

Para $k = 1$, obtemos:

$$\rho_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1}(x_i \cdot x_{i+1}) - \frac{1}{(n-1)^2}\left\{\left(\sum_{i=1}^{n-1} x_i\right)\left(\sum_{i=1}^{n-1} x_{i+1}\right)\right\}}{\left\{\frac{1}{n-1}\sum_{i=1}^{n-1} x_i^2 - \frac{1}{(n-1)^2}\left(\sum_{i=1}^{n-1} x_i\right)^2\right\}^{1/2}\left\{\frac{1}{n-1}\sum_{i=1}^{n-1} x_{i+1}^2 - \frac{1}{(n-1)^2}\left(\sum_{i=1}^{n-1} x_{i+1}\right)^2\right\}^{1/2}}$$

Consideremos as n observações extraídas de uma população Normal com $E(x_i) = \mu$, $Var(x_i) = \sigma^2$ e coeficiente de correlação serial de "lag 1" igual a ρ_1. Sem perda de generalidade vamos construir um teste de hipóteses para testar, $H_0: \mu = 0$ contra $H_a: \mu \neq 0$.

Sob H_0,

$$E(\bar{x}) = 0 \tag{2.3.1}$$

$$Var(\bar{x}) = \frac{\sigma^2}{n}\{1+2\rho_1(1-\frac{1}{n})\} \tag{2.3.2}$$

$$E(s^2) = \sigma^2(1-2\frac{\rho_1}{n}) \tag{2.3.3}$$

onde s^2 é definido como antes. Desprezando temos de ordem $\frac{1}{n^2}$ e usando o fato de que s^2 converge em probabilidade para σ^2 temos pelo T.L.C. que a variável,

$$t^* = \frac{\bar{x}}{\frac{s}{\sqrt{n}}\sqrt{1+2\rho_1}} \tag{2.3.4}$$

tem distribuição Normal padronizada. Portanto a estatística usual para o teste, isto é,

$$t = \frac{\sqrt{n}\ \bar{x}}{s} \tag{2.3.5}$$

é assintoticamente Normal com média zero e variância $\{1+2\rho_1\}$ e não Normal padronizada. O fato da variância da distribuição limite de (2.3.5) ser $(1+2\rho_1)$ altera a probabilidade do erro tipo I e observamos que mesmo para n grande, existe necessidade de corrigirmos o efeito da correlação serial, de "lag um" através de um fator de ponderação β, procedendo da seguinte maneira:

"Considere n grande, as observações com coeficiente de correlação serial ρ_1 e seja $Z_{\alpha/2}$ um ponto da distribuição Normal padronizada, tal que:

$$Pr\{|Z| \geq Z_{\alpha/2}\} = \alpha$$

Então, a probabilidade de erro Tipo I, levando em conta a não independência das observações é:

$$\alpha' = 2 \cdot \frac{1}{\sqrt{2\pi}} \lim_{b \to \infty} \int_{\frac{Z_{\alpha/2}}{\sqrt{1+2\rho_1}}}^{b} e^{-\frac{1}{2}y^2} dy$$

onde y segue a distribuição Normal padrão e $\beta = \sqrt{1+2\rho_1}$

Observamos que, quando $-\frac{1}{2} < \rho_1 < 0$(1), α' será menor que α, se $\rho_1 > 0$, α' será maior que α, e se $\rho = 0$ temos $\alpha' = \alpha$.

Ilustramos o efeito da correlação serial entre as observações, no teste sobre uma média, exibindo a tabela abaixo onde estão calculados para alguns valores de ρ_1, as correspondentes probabilidades do erro tipo I, para o teste de $H_0: \mu = \mu_0$ contra a alternativa $H_a: \mu = \mu_1 \neq \mu_0$ quando o nível de significância fixado é $\alpha = 0,05$.

ρ_1	-0,3	-0,2	-0,1	0	0,1	0,2	0,3
probabilidade	0,002	0,011	0,028	0,05	0,074	0,098	0,12

(1) - Ver Scheffé (1959).

EXEMPLO 2.2.3

Em estudos de pesca é importante o conhecimento da captura máxima sustentável que uma determinada população de de certo organismo aquático pode suportar. Essa captura dividido pelo respectivo esforço de pesca empregado, nos dá o índice de captura média populacional recomendado. No Estado do Ceará, maior produtor de lagostas, estudos anteriores indicam que o índice de captura médio populacional recomendado era de 3,1 lagostas/covo-dia. Os dados anuais de índice de captura são apresentados na tabela a seguir:

ANO	x_i	$y_i=x_i-3,1$	ANO	x_i	$y_i=x_i-3,1$	ANO	x_i	$y_i=x_i-3,1$
1948	4,6	1,5	1958	3,9	0,8	1968	2,0	-1,1
1949	4,5	1,4	1959	3,8	0,7	1969	1,3	-1,8
1950	4,3	1,2	1960	3,7	0,6	1970	1,5	-1,6
1951	4,4	1,3	1961	3,5	0,4	1971	1,0	-2,1
1952	4,2	1,1	1962	3,4	0,3	1972	0,9	-2,2
1953	4,1	1,0	1963	3,2	0,1	1973	0,6	-2,5
1954	4,0	0,9	1964	2,8	-0,3	1974	0,8	-2,3
1955	3,9	0,8	1965	2,6	-0,5	1975	0,8	-2,3
1956	4,2	1,1	1966	2,5	-0,6	1976	0,7	-2,4
1957	4,1	1,0	1967	1,9	-1,2	1977	0,7	-2,4

Os dados foram baseados no trabalho *Algumas tendências da pesca de lagosta no Estado do Ceará.*

A partir dessas informações testaremos a hipótese de que a pesca da lagosta no Ceará está se processando acima da capacidade suportável, e consequentemente o índice de captura média está diminuindo.

Portanto o problema consiste em testar, $H_0: \mu = 3,1$ contra a alternativa $H_a: \mu < 3,1$ onde μ representa o índice de captura média de lagosta covo/dia no Estado do Ceará.

A teoria desenvolvida anteriormente permite a construção do teste para as hipóteses H_0: $\mu = 0$ contra H_a: $\mu < 0$, portanto usaremos a transformação $y_i = x_i - 3{,}1$.

As estimativas do desvio padrão, coeficiente de correlação ρ_1 e da média são respectivamente, 1,43; 0,93 e -0,27.

Considerando a existência da correlação serial entre as observações ρ_1, o valor da estatística para o teste é,

$$t_0 = \frac{\sqrt{(30)} \cdot (-0{,}27)}{1.43} = -1{,}03$$

A região de rejeição é constituída por valores de T, tal que:

$$P\{T < T_c\} = \alpha,$$

onde α é o nível de significância prefixado. Para $\alpha = 0{,}05$, temos

$$T_C = -1{,}64\sqrt{1+2\times 0{,}93} = -2{,}11$$

Como $t_0 > T_c$ aceitamos H_0 e portanto verificamos que o índice de captura média de lagosta covo/dia não diminuiu significantemente.

- o -

CAPÍTULO III

TESTE DE HIPÓTESES ENVOLVENDO DUAS POPULAÇÕES

Um grande número dos problemas práticos de testes de hipóteses, baseados em duas amostras de duas populações se resumem de um modo geral em testes de comparação de duas variâncias e de comparação de duas médias.

Neste capítulo iremos desenvolver os testes, sob a suposição de normalidade das populações envolvidas e também quando esta suposição é verificada.

3.1 - TESTE DE HIPÓTESE SOBRE AS DUAS VARIÂNCIAS

Vamos considerar duas situações:

3.1.1 - As Populações têm Distribuição Normal

Seja x_{ij} a j-ésima observação da i-ésima amostra, ($i=1,2$ e $j=1,2,\ldots,n_i$). As duas amostras são independentes e extraídas de populações normais.

Temos interesse em testar as hipóteses H_0, sobre as duas variâncias, dada como

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2.$$

A estatística para o teste é definida por:

$$F = \frac{s_1^2}{s_2^2} \qquad (1.1.1)$$

onde

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}{n_i - 1} \qquad i=1,2$$

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \qquad i=1,2.$$

Sob a hipótese H_0 a estatística (1.1.1) tem distribuição F Snedecor Central com (n_1-1) e (n_2-1) graus de liberdade. Sob a alternativa H_a $\sigma_1^2 \neq \sigma_2^2$ a estatística tem distribuição proporcional a uma F Snedecor com (n_1-1) e (n_2-1), graus de liberdade e coeficiente de proporcionalidade igual a σ_1^2/σ_2^2.

EXEMPLO 3.1.1

Duas populações humanas são diferentes quanto a algumas características genéticas e ambientais. Gostaríamos a partir de uma amostra de cada uma destas populações, verificar se os fatos acima induzem uma diferença entre as distribuições da estatura das duas populações, com relação a variância das distribuições. Então o nosso problema consiste em testar as hipóteses

$$\begin{cases} H_0: \sigma_1^2 = \sigma_2^2 \\ H_a: \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

A tabela abaixo nos dá as informações obtidas das duas amostras.

	AMOSTRA I	AMOSTRA II
Nº de observações	100	61
s^2.	283	144
$\bar{x}$	157,8	155,2

O valor calculado da estatística para o teste é dado por

$$F_0 = \frac{s_1^2}{s_2^2} = 1,965.$$

O valor de F_c tal que $P(F_{100,60} > F_c) = 0,025$ é 1,58. Portanto a hipótese H_0 é rejeitada ao nível $\alpha = 0,05$. Isto é, as características genéticas e ambientais da população tem influência sobre a variância da variável estatura.

3.1.2 - As populações não tem Distribuição Normal

Sejam duas amostras aleatórias ambas extraídas de populações não normais.

O nosso problema consiste em testar as hipóteses

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_a: \sigma_1^2 \neq \sigma_2^2$$

Para este caso apresentamos um teste aproximado e robusto proposto por Box e Andersen (1955). Consideremos duas situações:

a) Médias Conhecidas

Suponhamos de início que as médias populacionais são conhecidas e sem perda de generalidade, tomemos $E(x_{ij})=0$. Relembremos que, quando as populações são normais, a estatística para o teste é dado por,

$$F = \frac{\sum_{j=1}^{n_1} x_{1j}^2 / n_1}{\sum_{j=1}^{n_2} x_{2j}^2 / n_2} = \frac{n_2}{n_1} \cdot \frac{\sum_{j=1}^{n_1} x_{1j}}{\sum_{j=1}^{n_2} x_{2j}} \qquad (1.2.1)$$

Sob H_0 F segue a distribuição F de Snedecor com n_1 e n_2 graus de liberdade.

As hipóteses acima também podem ser testadas por meio de uma outra estatística W definida por,

$$W = \frac{\sum_{j=1}^{n_1} x_{1j}^2}{\sum_{j=1}^{n_1} x_{1j}^2 + \sum_{j=1}^{n_2} x_{2j}^2} \qquad (1.2.2)$$

a qual sob H_0 e a Normalidade das observações W segue uma distribuição beta com $v_1 = n_1$ e $v_2 = n_2$ graus de liberdade. Com base nesta última estatística vamos obter o teste no caso de não normalidade, para isto apresentaremos a média e a variância da estatística W, levando em consideração dois aspectos; quando as amostras são extraídas de populações normais, onde diremos sob a Teoria Normal e quando as amostras não são extraídas de populações normais e então diremos sob a Teoria das Permutações, Box e Andersen (1955).

Os resultados estão na seguinte tabela:

Sob a Teoria da Permutação	Sob a Teoria Normal
$\underset{P}{E}(w) = \frac{n_1}{N}$	$\underset{N}{E}(w) = \frac{n_1}{N}$
$\underset{P}{Var}(w) = \frac{2n_1 n_2}{N^2(N+2)}\left\{1 + \frac{1}{2}\frac{N}{N-1}(b_2-3)\right\}$	$\underset{M}{Var}(W) = \frac{2n_1 n_2}{N^2(N+2)}$
$N = n_1+n_2$	$N = n_1+n_2$
$b_2 = (N+2)\frac{\left(\sum_{j=1}^{n_1} x_{1j}^4 + \sum_{j=1}^{n_2} x_{2j}^4\right)}{\left(\sum_{j=1}^{n_1} x_{1j}^2 + \sum_{j=1}^{n_2} x_{2j}^2\right)^2}$	

TABELA 1.2.1

Vemos que a média da estatística W, sob a Teoria da Permutação é a mesma que aquela calculada sob a Teoria da Normal pelo fator multiplicativo, $\{1 + \frac{1}{2}\frac{N}{N-1}(b_2-3)\}$. Como sob a normalidade W segue uma distribuição Beta, é portanto, razoável aproximar a distribuição da estatística W sob a Teoria da Permutação por uma distribuição Beta, igualando a média e a variância das duas distribuições e obtendo-se assim os graus de liberdade da distribuição Beta da estatística W sob a Teoria da Permutação.

Assim sob a Teoria Normal temos:

$$\underset{N}{E}(W) = \frac{v_1}{N} = \mu_1$$

$$\underset{N}{Var}(W) = \frac{2v_1 v_2}{N^2(N+2)} = \mu_2$$

segue que,

$$v_2 = \frac{1-\mu_2}{\mu_1}\cdot v_1 \tag{1}$$

e

$$v_1 = \frac{2\mu_1(\mu_1-\mu_1^2-\mu_2)}{\mu_2} \tag{2}$$

Sob a Teoria da Permutação, temos:

$$\underset{P}{E}(W) = \frac{n_1}{N}$$

e

$$\underset{P}{Var}(W) = \frac{2n_1n_2}{N^2(N+2)}\left\{1 + \frac{1}{2}\,\frac{N}{N-1}\,(b_2-3)\right\}$$

Substituindo em (1) e (2) os valores $\underset{P}{E}(W)$ e $\underset{P}{Var}(W)$, obtemos

$$v_2 = \frac{1-\frac{n_1}{N}}{\frac{n_1}{N}}\cdot v_1 = \frac{n_2}{n_1}\cdot v_1$$

e

$$v_1 = \frac{2\frac{n_1^2}{N^2}\left\{1 - \frac{n_1}{N} - \frac{n_2}{N(N+2)}\,\frac{(Nb_2-N-2)}{(N-1)}\right.}{\frac{n_1n_2}{N^2(N+2)}\left\{\frac{Nb_2-N-2}{N-1}\right\}} = n_1\cdot\frac{2(N+2-b_2)}{N(b_2-N-2)}.$$

Portanto,

$$v_1 = n_1d$$

$$v_2 = n_2d$$

onde,

$$d = \left[\frac{N \cdot b_2 - N - 2}{2(N+2-b_2)}\right]^{-1}$$

e b_2 foi definido na Tabela 1.2.1. Logo a distribuição aproximada da estatística w é uma Beta com n_1d e n_2d graus de liberdade.

Para $N = n_1+n_2$ grande, podemos reduzir a expressão d acima em,

$$d = [1+\tfrac{1}{2}(b_2-3)]^{-1}$$

a qual é obtida tomando o limite para N tendendo a infinito a expressão,

$$d = \left[1 + \frac{1}{2}\left\{\frac{N+2}{N-1-(b_2-3)}\right\}(b_2-3)\right]^{-1}$$

Pelo fato da distribuição F Snedecor ser mais usual e seus valores facilmente encontrados, usamos a transformação,

$$U = \frac{w}{1-w} \cdot \frac{n_2}{n_1}, \qquad (1.2.3)$$

a qual coincide com a estatística (1.2.1). Temos que sob H_0, U tem distribuição F Snedecor com n_1d e n_2d graus de liberdade.

Concluímos que quando as populações não são Normais, um teste sobre duas variâncias para populações com médias conhecidas, pode ser feito, usando a estatística usual F definida em (1.2.1) a qual sob H_0 segue a distribuição F Snedecor com n_1d e n_2d graus de liberdade, onde:

i) para n_1 e n_2 pequenos

$$d = \left[1 + \frac{1}{2}\left\{\frac{N+2}{N-1-(b_2-3)}\right\}(b_2-3)\right]^{-1}$$

ii) para n_1 e n_2 grandes

$$d = [1 + \frac{1}{2}(b_2-3)]^{-1}$$

e

$$b_2 = \frac{(N+2)\left(\sum_{j=1}^{n_1} x_{1j}^4 + \sum_{j=1}^{n_2} x_{2j}^4\right)}{\left(\sum_{j=1}^{n_1} x_{1j}^2 + \sum_{j=1}^{n_2} x_{2j}^2\right)^2}$$

b) Médias Desconhecidas

Comumente encontramos situações que queremos testar as hipóteses

$$H_0:\ \sigma_1^2 = \sigma_2^2$$

$$H_a:\ \sigma_1^2 \neq \sigma_2^2$$

porém temos que as médias populacionais μ_1 e μ_2 respectivamente são desconhecidas. Sob a teoria Normal e sob H_0 vimos que a estatística para o teste, é dado em (1.1.1).

Podemos ainda testar as mesmas hipóteses dadas através da estatística de Bartlett,

$$\chi^{*2} = V|ns_p^2 - \sum_{i=1}^{2} v_i|ns_i^2 \qquad (1.2.4)$$

onde,

$$V = \sum_{i=1}^{2} v_i \quad \text{e} \quad v_i = (n_i - 1) \qquad i=1,2$$

$$s_p^2 = \frac{\sum_{i=1}^{2} (n_i - 1) s_i^2}{\sum_{i=1}^{2} (n_i - 1)}, \qquad i=1,2$$

e

$$s_i^2 = \frac{\sum_{j=1}^{2} (x_{ij} - \bar{x}_{i.})}{(n_i - 1)} \qquad i=1,2$$

Suponhamos agora que as populações não são Normais e que μ_1 e μ_2 são desconhecidas. Neste caso, Box e Andersen (1955) provaram que sob H_0, a estatística

$$\chi^{*2'} = \frac{\chi^{*2}}{1 + \frac{1}{2} c_2} \qquad (1.2.5)$$

para N grande, tem distribuição χ^2 (qui-quadrado) com um (1) grau de liberdade, onde para $n_1 = n_2$, c_2 é dado por:

$$c_2 = \frac{2 \sum_{i=1}^{2} K_{4i}}{\left(\sum_{i=1}^{2} K_{2i} \right)^2}$$

$$K_{4i} = \frac{n_i (n_i + 1) s_{4i} - 3 (n_i - 1) s_{2i}^2}{(n_i - 1)(n_i - 2)(n_i - 3)}, \qquad i=1,2$$

$$K_{2i} = \frac{s_{2i}}{(n_i-1)}, \qquad i=1,2$$

$$s_{ri} = \sum_{j=1}^{n_i} (x_{ij}-\bar{x}_{i.})^r, \qquad i=1,2.$$

No caso em que $n_1 \neq n_2$, a expressão c_2 torna-se complicada, Box e Andersen (1955).

EXEMPLO 3.1.2

Os dados abaixo representam o número de raios da nadadeira dorsal, em duas amostras de sardinhas, colhidas em duas áreas B e D na primavera.

Nº de raios da nadadeira dorsal	Frequência da Área B	Frequência da Área D
16	4	-
17	39	5
18	31	54
19	5	20
	79	79

* Dados fornecidos pelo Inst.Oceanográfico da USP.

Baseados em amostras foi verificada a não normalidade das distribuições. O problema consiste em a partir dos dados observados verificar se a variabilidade, do número de ráios da nadadeira dorsal nas duas áreas são iguais. Temos portanto as hipóteses:

$$\begin{cases} H_0: \ \sigma_1^2 = \sigma_2^2 \\ H_a: \ \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

que serão testadas, ao nível de significância $\alpha = 0,05$. Para isto calculamos as estimativas das médias e das variâncias, encontrando-se respectivamente $\bar{x}_B = 17,4684$, $\bar{x}_D = 18,1898$, $S_B^2 = 0,4829$ e $S_D^2 = 0,2833$.

Considerando o efeito da não normalidade a estatística adequada para o teste é dada por,

$$\chi^{*2'} = \frac{\chi^{*2}}{1 + \frac{1}{2}c_2}$$

O valor calculado é

$$\chi_o^{*2'} = 5,5797.$$

A estatística $\chi^{*2'}$, de acordo com a teoria descrita segue a distribuição χ_1^2 (qui-quadrado com um grau de liberdade). O valor crítico ao nível de 0,05 é 3,84. Então concluímos, pela rejeição de H_0 e consequentemente dizemos que existe diferença entre a variabilidade do número de raios da nadadeira dorsal, nas duas áreas.

3.2 - TESTES DE HIPÓTESES SOBRE DUAS MÉDIAS

Iremos desenvolver nesta seção, testes para igualdade de duas médias, apenas quando as variâncias σ_1^2 e σ_2^2 respectivas de cada população são desconhecidas.

Consideremos as situações de amostras independentes e amostras dependentes levando em conta, a normalidade e não normalidade, variâncias iguais ($\sigma_1^2 = \sigma_2^2$) e variâncias desiguais ($\sigma_1^2 \neq \sigma_2^2$).

3.2.1 - As Amostras são Independentes, as populações normais com variâncias iguais

Consideremos x_{ij}, a j-ésima observação da i-ésima amostra, ($i=1,2$ e $j=1,2,\ldots,n_i$). E suponhamos $\sigma_1^2=\sigma_2^2$ desconhecidas.

Se estamos interessados em testar as hipóteses:

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

a estatística a ser usada será,

$$T = \frac{(\bar{x}_1-\bar{x}_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}, \qquad (2.1.1)$$

onde

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-1}$$

$$s_i^2 = \frac{\sum\limits_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2}{n_i-1} \quad \text{e} \quad \bar{x}_{i'} = \frac{\sum\limits_{j=1}^{n_i}x_{ij}}{n_i}, \qquad i=1,2,$$

Sob a hipótese H_0: $\mu_1=\mu_2$, (2.1.1) segue a distribuição t de Student Central com (n_1+n_2-2) graus de liberdade. Sob uma alternativa H_a: $\mu_1 \neq \mu_2$, a estatística acima tem distribuição t de Student não Central.

<u>EXEMPLO 3.2.1</u>

Duas populações humanas são diferentes quanto a algumas características genéticas e ambientais. Queremos verificar a partir de uma amostra de cada uma destas populações, se os fatos acima implicam uma diferença entre as distribui ções da estatura das duas populações, com relação a média da distribuição.

O problema consiste em verificar as seguintes hipóteses,

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_a: \mu_1 \neq \mu_2 \end{cases}$$

E para isto consideremos as seguintes informações:

	Amostra I	Amostra II
Nº de Observações	101	61
s_i	138	144
$\bar{x}_i$	157,8	155,2

Tomemos para o teste de hipóteses acima, $\alpha = 0,05$. E com base nas informações temos a quantidade

$$T_0 = \frac{157,8 - 155,2}{11,8427\sqrt{0,1622}} = 1,3539 \quad .$$

Sob a H_0, a estatística T segue a distribuição t de Student com $n_1+n_2-2 = 160$ graus de liberdade, cujo valor de T_c, tal que $P(|t|>T_c) = 0,05$ é $T_c = 1,9$.

Como $|T_0| < 1,9$, aceitamos H_0 e podemos dizer que as duas populações não diferem significantemente em relação as estaturas médias.

3.2.2 - As Amostras são Independentes, as Populações Normais com Variâncias Desiguais

Consideremos as duas amostras e as mesmas hipóteses como definidas em (2.1). Através de testes sobre duas variâncias baseadas em outras amostras, ou por experimentos anteriores, comprovamos que $\sigma_1^2 \neq \sigma_2^2$.

Neste caso um teste de hipótese sobre as duas médias pode ser feito, através da estatística t'v dada por:

$$t'v = \frac{\bar{x}_1 - \bar{x}_2}{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^{1/2}} , \qquad (2.2.1)$$

a qual sob H_0: $\mu_1 = \mu_2$, segue aproximadamente uma distribuição t de Student, Welch (1937) e (1947) com v graus de liberdade, onde,

$$v = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Quando o valor de v não é um número inteiro, arredondamos para o número inteiro mais próximo.

Observamos que

1) Se $n_1 = n_2$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} \quad \frac{1}{n_2}}} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = t'v$$

contudo para a estatística t temos associados 2(n-1) graus de liberdade enquanto que para estatística t'v os graus de liberdade associados são sempre menor ou igual a 2(n-1). A igualdade só se verifica se $s_1^2 = s_2^2$. Brownlee (1960).

2) O maior valor que v assume é (n_1+n_2-2) e isto ocorre quando

$$\frac{s_1^2}{n_1(n_1-1)} = \frac{s_2^2}{n_2(n_2-1)},$$

e o menor valor de v é o mínimo $(n_1-1,\ n_2-1)$ isto acontece quando,

$$\frac{s_2^2/n_2}{s_1^2/n_1} \text{ ou } \frac{s_1^2/n_1}{s_2^2/n_2}$$

aproxima-se de zero. Brownlee (1960).

3) Se $n_1 = n_2$, e colocamos $s = s_1^2/s_2^2$, temos a informação que quando,

i) $0,5 < s < 2$, a redução nos graus de liberdade v em relação a $2(n-1)$ não excede a 10%.

ii) $0,333<s<3$, a redução nos graus de liberdade v em relação a $2(n-1)$ não excede a 20%,

assim, se $n_1 = n_2$ e $s_1^2 \simeq s_2^2$, existe pouca diferença entre o procedimento descrito em 2.1 e o em 2.2. Brownlee (1960).

<u>EXEMPLO</u> <u>3.2.2</u>

Duas populações de sardinhas são diferentes quanto as características ambientais.

Queremos verificar, se as médias de comprimento das sardinhas, são iguais nas duas áreas. Sabemos por experiências anteriores que $\sigma_B^2 \neq \sigma_C^2$. E temos as seguintes informações

	Amostra I	Amostra II
Nº de Observações	23	79
s_i^2	$(13,71)^2$	$(21,48)^2$
$\bar{x}_i$	185,4	174,3

Os dados foram obtidos pelo Instituto Oceanográfico da USP.

Nosso problema consiste em, a partir dos dados da tabela acima testar as hipóteses:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_a: \mu_1 \neq \mu_2, \end{cases}$$

ao nível de significância $\alpha = 0,05$. E baseado nas informações temos o valor da estatística,

$$t'_0 v = \frac{174,3 - 185,4}{\sqrt{18,7561 + 2,3793}} =$$

$$= -1,979.$$

Os graus de liberdade $\underline{v}$ é calculado abaixo

$$v = \frac{446,705}{16,071} \simeq 28.$$

Observamos que sem levar em conta a diferença das variâncias o número de graus de liberdade seria $n_1+n_2-2=100$. Esta diferença se deve ao fato de que as amostras são de tamanho bem diferentes e as variâncias amostrais são também muito diferentes.

Temos que t'v segue sob H_0 uma distribuição t de Student com (28) graus de liberdade, cujos valores são dados por $|T| > 2{,}048$ como $-2{,}048 < t_0 < 2{,}048$ concluímos pela aceitação a hipótese H_0 e dizemos que não existe em média diferença entre as estaturas das duas populações.

3.2.3 - <u>As Amostras são Independentes, as Populações não são Normais com Variâncias Iguais</u>

Sejam $(x_{11},\dots,x_{1n_1})$ e $(x_{21},\dots,x_{2n_2})$ duas amos-amostras aleatórias, independentes, extraídas de populações com média μ_1 e μ_2 respectivamente e $\sigma_1^2 = \sigma_2^2$.

No caso de populações normais as hipóteses

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

podem ser testadas por meio do teste t de Student.

Uma solução alternativa para o mesmo problema é dada pela distribuição F da Análise de Variância. Outra, solução menos usada, mais útil para o desenvolvimento desta seção, é usando a estatística W definida abaixo

$$W = \frac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (\bar{x}_{i.} - \bar{x}_{..})^2}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2}$$

Usando o mesmo procedimento da seção 1.2 vamos calcular a média e a variância da estatística W sob a teoria Normal e sob a teoria da Permutação. Os resultados estão na Tabela 2.3.1.

TABELA 2.3.1

SOB A TEORIA DA PERMUTAÇÃO (Populações não normais)	SOB A TEORIA NORMAL
$E_P(W) = \frac{1}{N-1}$	$E_N(W) = \frac{1}{N-1}$
$\underset{P}{\mathrm{Var}}(W) = \frac{2(N-2)}{(N-1)^2(N+1)}\left[1 - \frac{K_4}{NK_2^2}\right] - \frac{K_4}{(N-1)^2K_2^2}\left[\frac{4}{N} - \sum_{i=1}^{2} \frac{1}{n_i}\right]$	$\underset{N}{\mathrm{Var}}(W) = \frac{2(N-2)}{(N-1)^2(N+1)}$
$N = n_1 + n_2$	$N = n_1 + n_2$

Sob a normalidade das observações, W segue uma distribuição Beta com 1 (um) e N-2 graus de liberdade.

Sob a não normalidade, vamos aproximar a distribuição da estatística W por uma distribuição β e usaremos raciocínio análogo a (1.2) para determinar os graus de liberdade da distribuição Beta da estatística W. Consideremos,

a) $n_1 = n_2$ - Neste caso, a variância de W sob a Teoria da Permutação se reduz a

$$\underset{P}{\mathrm{Var}}(W) = \frac{2(N-2)}{(N-1)\cdot(N+1)}\left(1 - \frac{K_4}{NK_2^2}\right),$$

já que

$$\frac{4}{N} = \sum_{i=1}^{2} \frac{1}{n_i},$$

e então, W tem distribuição Beta, com os graus de liberdade,

$$v_1 = 1\cdot d$$

e

$$v_2 = (N-2)d$$

onde

$$d = 1 + \frac{(N+1)}{(N-1)} \cdot \frac{c_2}{(N-c_2)}$$

$$c_2 = \frac{K_4}{K_2^2}$$

$$K_4 = \frac{\{N(N+1)S_4 - 3(N-1)S_2^2\}}{(N-1)(N-2)(N-3)}$$

$$K_2 = \frac{S_2}{N-1}$$

$$S_r = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^r$$

$$K = 2$$

e

$$n_1 = n_2 = n, \text{ portanto } N = 2n.$$

Como os valores tabela da distribuição F Snedecor são facilmente encontrados, usaremos a transformação

$$U = \frac{w}{1-w} (N-2), \qquad (2.3.2)$$

a qual com a F da Análise de Variância sob H_0, U tem distribuição F Snedecor com d e (N-2)d graus de liberdade.

b) $n_1 \neq n_2$ - Neste caso a variância W sob a Teoria da Permutação é dada por

$$\underset{P}{\mathrm{Var}}\ (W) = \frac{2(N-2)}{(N-1)^2(N+1)}\left[1 - \frac{c_2}{N}\right] - \frac{c_2}{(N-1)^2}\left[\frac{4}{N} - \sum_{i=1}^{K}\frac{1}{n_i}\right],$$

e temos que a estatística (2.3.3) usada para testar, as hipóteses definidas no início de (2.3), segue uma distribuição F Snedecor com d e (N-2)d graus de liberdade, e neste caso d é dado por

$$d = 1 + \frac{N+1}{N-1} \cdot \frac{c_2}{(N^{-1}+A)^{-1} - c_2}$$

onde

$$A = \frac{N+1}{2(N-1)}\left(\frac{K^2}{N} - \sum_{i=1}^{2}\frac{1}{n_i}\right)$$

e

$$c_2 = \frac{K_4}{(K_2)^2},$$

como vimos em (a).

EXEMPLO 3.2.3

Os dados abaixo representam duas amostras aleatórias, de sardinhas, nas quais foram contados os raios da nadadeira dorsal.

Nº de raios da nadadeira dorsal	Frequência da Área D	Frequência da Área E
16	1	-
17	2	3
18	16	2
19	3	1
	22	6

Os dados foram fornecidos pelo Inst.Oceanográfico da USP.

Experiências anteriores mostraram que a variável acima não segue a distribuição Normal, além de nos informar que $\sigma_E^2 = \sigma_D^2$. Nosso problema consiste em verificar se o número médio de raios da nadadeira dorsal são iguais nas duas áreas. Temos então as hipóteses,

$$\begin{cases} H_0: \mu_D = \mu_E \\ H_a: \mu_D \neq \mu_E \end{cases}$$

Para isto calculamos as estimativas das médias e variâncias, encontrando respectivamente

$$\bar{x}_D = 17{,}96,\ \bar{x}_E = 17{,}67,\ s_D^2 = 0{,}407 \text{ e } s_E^2 = 0{,}266.$$

Considerando o efeito da não Normalidade a estatística para o teste é dada por:

$$U = (N-2)\frac{w}{1-w} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot})^2}{\frac{1}{(N-2)}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\bar{x}_{i\cdot}-\bar{x}_{\cdot\cdot})^2}$$

Temos que o valor observado de U, é

$$U_0 = 1{,}2454.$$

A estatística U, sob H_0, tem distribuição F Snedecor com $v_1 = d$ e $v_2 = (N-2)d$ graus de liberdade.

O valor de F_c tal que $P\{F_{1,26} > F_c\} = 0{,}25$ é 5,72. Logo concluímos pela aceitação de H_0 ou seja, não existe diferença entre o número de raios médios da nadadeira dorsal, nas duas Áreas.

3.2.4 - As Amostras são Independentes, as Populações não Normais com Variâncias Diferentes

Para testarmos as hipóteses, $H_0: \mu_1 = \mu_2$ contra $H_a: \mu_1 \neq \mu_2$ neste caso em que as populações não são normais e as variações diferentes, usaremos uma solução limite dada que por Scheffe (1959).

Para isto, seja $\{x_{i1} \ldots x_{in_i}\}$ uma amostra aleatória de uma população com média μ_i, variância σ_i^2 e curtose γ_{2i} finito $i=1,2$, e as duas amostras independentes. Seja também $\bar{x}_{i.}$ e s_i^2, $i=1,2$ a média e a variância amostral como definida em (2.1). Sob estas condições e sob H_0, vamos determinar a distribuição limite da usual estatística

$$T = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \qquad (2.4.1)$$

onde

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

Dividiremos nosso estudo em duas partes, a primeira relativa ao numerador de T e a outra relativa ao denominador.

Com relação ao numerador, temos que a distribuição assintótica de $(\bar{x}_1 - \bar{x}_2)$ é Normal com média zero e variância,

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma_2^2 \frac{(1+R)(\theta+R)}{RN} \qquad (2.4.2)$$

onde,

$$\theta = \frac{\sigma_1^2}{\sigma_2^2}, \quad R = \frac{n_1}{n_2} \text{ e } N = n_1+n_2.$$

Com relação ao denominador verificamos que a estatística s_p^2 converge em probabilidade para $(1+R)^{-1}(R\theta+1)\sigma_2^2$, Scheffe (1959).

Portanto para N grande e negligenciando os termos de ordem N^{-1}, temos,

$$\frac{\sigma_2^2 \dfrac{(1+R)(\theta+R)}{RN}}{s_p\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)} = \frac{\sigma_2^2 \dfrac{(1+R)(\theta+R)}{RN}}{\sigma_2^2 \dfrac{(R\theta+1)(1+R)^{-1}(1+R)^2}{RN}} = \frac{\theta+R}{R\theta+1}$$

De acordo com o desenvolvimento acima, podemos dizer que,

$$\frac{(\bar{x}_1 - \bar{x}_2)}{s_p\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)^{1/2}},$$

tem distribuição assintótica Normal com média zero e variância $\frac{\theta+R}{R\theta+1}$ e não normal padronizada.

Um estudo mais detalhado sobre a variância $\left(\frac{\theta+R}{R\theta+1}\right)$, pode ser feito e concluímos que, se o tamanho das amostras são iguais $n_1 = n_2 = n$, temos que $R = 1$ e portanto a estatística 2.4.1, para n grande, segue a distribuição Normal Padronizada, como no caso em que *supomos* $\sigma_1^2 = \sigma_2^2$. Isto implica que para $n_1 = n_2 = n$ grande, a distribuição limite da estatística (2.4.1) segue a distribuição $N(0,1)$, *ainda que as Populações* não sejam Normais e *tenham variâncias diferentes.*

Agora se $n_1 \neq n_2$, como $\sigma_1^2 \neq \sigma_2^2$ temos que $\theta \neq 1$ e portanto $\left(\frac{\theta + R}{\theta R + 1}\right) \neq 1$ mesmo para $N = n_1 + n_2$ grande, o efeito de $\theta \neq 1$ é significante na probabilidade do Erro Tipo I, já que a estatística (2.4.1) usada para o teste de hipóteses sobre as duas médias segue a distribuição $N\left(0, \frac{\theta + R}{R\theta + 1}\right)$ em vez de $N(0,1)$.

Podemos corrigir o efeito da variância da distribuição Normal limite, determinando α' como efeito no capítulo II, itens (1.2), (2.3), isto é,

$$\alpha' = 2 \lim_{b \to \infty} \int_{\frac{Z_{\alpha/2}}{B}}^{b} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2} y^2} dy$$

onde $Z_{\alpha/2}$ é um ponto da distribuição Normal padronizada Z, tal que, $P_r\{|Z| \geq Z_{\alpha/2}\} = \alpha$

$$B = \left(\frac{\theta + R}{R\theta + 1}\right)^{1/2}$$

e

y segue a distribuição normal com média zero e variância um.

A tabela 2.4.1 mostra o efeito de $\sigma_1^2 \neq \sigma_2^2$ e $n_1 \neq n_2$, na verdadeira probabilidade do Erro Tipo I, para testar as hipóteses H_0: $\mu_1 = \mu_2$ contra a alternativa H_a: $\mu_1 \neq \mu_2$, quando o nível de significância fixado é $\alpha = 0,05$. Consideremos N grande.

O processo acima só é válido para $N = n_1 + n_2$ grande, portanto para N pequeno, recomendamos o uso da transformação ou testes não paramétricos.

Contudo para n_1 e n_2 pequenos, ilustraremos com a tabela 2.4.2, o efeito de $\sigma_1^2 \neq \sigma_2^2$ e $n_1 \neq n_2$, na verdadeira probabilidade do Erro Tipo I, para testar a H_0: $\mu_1 = \mu_2$ versus H_a: $\mu_1 \neq \mu_2$, quando o nível de significância fixado é $\alpha = 0{,}05$.

TABELA 2.4.1

$R = \frac{n_1}{n_2}$ \ $\theta = \frac{\sigma_1^2}{\sigma_2^2}$	0*	$\frac{1}{5}$	$\frac{1}{2}$	1	2	5	∞*
1	0,050	0,050	0,050	0,050	0,050	0,050	0,050
2	0,17	0,12	0,080	0,050	0,029	0,014	0,006
5	0,38	0,22	0,12	0,050	0,014	0,002	1×10^{-1}
∞*	1,00	0,38	0,17	0,050	0,006	1×10^{-1}	0

TABELA 2.4.2

(n_1, n_2) \ $\theta = \frac{\sigma_1^2}{\sigma_2^2}$	0	0,1	0,2	0,5	1	2	5	10	∞
(15,5)	0,32	0,23	0,18	0,098	0,05	0,025	0,008	0,005	0,002
(5,3)	0,22	0,14	0,10	0,072	0,05	0,038	0,031	0,030	0,031
(7,7)	0,072	0,07	0,063	0,058	0,05	0,051	0,058	0,063	0,072

* - Calculado por Hsv.P.L. (1938).

Sob H_a: $\mu_1 \neq \mu_2 \Longrightarrow \mu_1 - \mu_2 = \Delta \neq 0$. Procedendo como antes temos que para n_1 e n_2 grandes, a distribuição limite da razão (2.4.1) é normal com média

$$\frac{\Delta}{\sqrt{(\sigma^2)_W}} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1/2}$$

e variância $\frac{\theta + R}{R\theta+1}$ onde $(\sigma^2_{\cdot})w = \frac{n_1\sigma_1^2+n_2\sigma_2^2}{n_1+n_2}$ é denominada de variância ponderada, sendo obtida da expressão $\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ como segue: para n_i grande s_i^2 converge em probabilidade para σ_i^2 e pode-se aceitar o fato que $n_1+n_2-2 \simeq n_1+n_2$ portanto a expressão acima fica:

$$\frac{n_1\sigma_1^2+n_2\sigma_2^2}{n_1+n_2} - \frac{\sigma_1^2+\sigma_2^2}{n_1+n_2},$$

como $\sigma_1^2 + \sigma_2^2$ é finito temos que para n_1 e n_2 grande $\frac{\sigma_1^2+\sigma_2^2}{n_1+n_2}$ aproxima-se de zero e então

$$\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \simeq \frac{n_1\sigma_1^2+n_2\sigma_2^2}{n_1+n_2} = (\sigma^2_{\cdot})w$$

Se a suposição, $\sigma_1^2 = \sigma_2^2 = \sigma^2$ é verdadeira, a média de (2.4.1) é $\frac{\Delta}{\sigma}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)^{-1/2}$ e variância um.

Contudo se a suposição da igualdade de variâncias não é mantida, $\theta \neq 1$, o resultado do cálculo do poder será correto se só se $n_1 = n_2$ e $R = 1$.

EXEMPLO 3.2.4

Os dados no quadro abaixo representam duas amostras

	Amostra 1	Amostra 2
Nº de Observações	161	82
$\bar{x}_i$	115,3	119,3
s_i^2	284,597	104,04

Os dados foram fornecidos pelo Instituto Oceanográfico da USP.

aleatórias de sardinhas, selecionadas em duas estações do ano em uma mesma área de estudo, nas quais foram contados o número de Rastro do Ramo superior.

Experiências anteriores mostraram que a variável acima não segue a distribuição Normal, além do fato de

$$\theta = \sigma_1^2/\sigma_2^2 = 2{,}735.$$

Nosso objetivo é testar a hipótese de igualdade entre as estações outono e verão, com relação ao número médio de rastros do ramo superior. Então as hipóteses são:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_a: \mu_1 \neq \mu_2 \end{cases}$$

Adotamos $\alpha = 0{,}05$ e usamos para o teste a estatística,

$$T = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

cujo valor observado é T_0 1,9718. Considerando o efeito da não Normalidade das Populações e de $\theta = 2{,}735$, a estatística T tem distribuição $N(0, \frac{\theta + R}{R\theta+1})$. Para os dados acima

$$\frac{\theta + R}{R\theta+1} = 0{,}738.$$

A região de rejeição é constituída por valores de T, tais que:

$$P(|T|>T_c) = 0{,}05,$$

onde

Como $|T_0| > T_c$, concluímos pela rejeição de H_0 e verificamos assim que não existe diferença entre o número médio de Rastro Superior da Dorsal, nas duas estações do ano.

Observe que, sem considerar a Não Normalidade das Populações, o valor crítico da estatística seria $T_c = 1,96$

3.2.5 - As Amostras são Dependentes e as Populações são Normais

Abordaremos aqui a situação em que desejamos comparar duas médias populacionais, com base em uma amostra de pares.

Suponhamos uma amostra aleatória de n pares $(x_{11}, x_{12}), (x_{21}, x_{22}), \ldots, (x_{1n}, x_{2n})$, onde a primeira coordenada é extraída de uma população Normal com média μ_1 e variância σ_1^2 e a segunda coordenada é extraída de uma população Normal com média μ_2 e variância σ_2^2.

Denotemos por

$$y_j = x_{1j} - x_{2j}, \qquad j=1,2,\ldots,n,$$
$$\bar{y} = \bar{x}_1 - \bar{x}_2$$

ou seja

$$\bar{y} = \frac{\sum_{j=1}^{n} y_j}{n}$$

e

$$s_y^2 = \frac{\sum_{j=1}^{n} (y_j - \bar{y})^2}{n-1}$$

Temos agora $Y_1, Y_2, \ldots, Y_n$, uma amostra aleatória de uma população normal.

Com base nesta amostra queremos testar as hipóteses,

$$H_0: \mu_y = 0$$

$$H_a: \mu_y \neq 0$$

onde $\mu_y = \mu_1 - \mu_2$.

A estatística para o teste é,

$$T = \frac{\bar{y}}{\frac{s_y}{\sqrt{n}}} \qquad (2.5.1)$$

que sob a hipótese H_0, tem distribuição t de Student central com (n-1) graus de liberdade. Sob a alternativa

$$H_a: \mu_y = \mu'_y \neq 0,$$

a estatística (2.5.1) tem distribuição t de Student não central.

EXEMPLO 3.2.5

Os dados abaixo[1] representam medidas do nível médio de insulina de 14 pacientes em duas situações. Na primeira (antes do tratamento) os pacientes apresentavam um distúrbio na glândula Tireoide na segunda (depois do tratamento) os pacientes foram considerados normais com relação a anomalia acima.

Experiências anteriores mostraram que a variável acima segue a distribuição Normal. Nosso objetivo é verificar se o tratamento diminui o nível médio de insulina dos pacientes, e portanto a partir dos dados apresentados, testa-

---

(1) - Os dados foram fornecidos pelo Setor de Estatística Aplicada (SEA) do Departamento de Estatística do IME-USP.

taremos,

$$H_0:\ \mu_y = 0$$
$$H_a:\ \mu_y > 0,$$

OBSERVAÇÕES MÉDIAS DE TAXAS DE INSULINA NO TEMPO BASE		$y_j = x_{1j} - x_{2j}$
ANTES TRATAMENTO	APÓS TRATAMENTO	
23,00	9,00	14,00
24,20	7,20	17,00
23,60	8,80	14,80
30,40	6,40	24,00
22,80	8,80	14,00
7,20	5,80	1,40
18,60	6,40	12,20
20,80	7,00	13,80
15,40	16,40	-1,00
5,40	7,00	-1,60
21,80	14,20	7,60
23,00	17,40	5,60
24,40	14,00	10,40
8,00	10,00	-2,00

ao nível de significância $\alpha = 0,05$.

O valor

$$T_0 = \frac{9,3}{\sqrt{\frac{62,32}{14}}} = 4,408$$

e o valor tabelado é $T_c = 1,771$. Como $T_0 > 1,771$, rejeitamos H_0 e concluímos que houve decréscimo na taxa média de Insulina após o tratamento.

3.2.6 - As Amostras são Dependentes e as Populações não Normais

Suponhamos como em (2.5) uma situação experimental com n pares de observações $(x_{11},x_{12}),(x_{21},x_{22}),\ldots,(x_{1n},x_{2n})$ e $y_j = x_{1j}-x_{2j}$, $j = 1,2,\ldots,n$.

Neste caso em que as populações são não normais, iremos testar as hipóteses,

$$H_0:\ \mu_y = 0$$
$$H_a:\ \mu_y \neq 0$$

por meio da estatística

$$t^2 = \left(\sqrt{n}\ \frac{\bar{y}}{s_y}\right)^2 \qquad (2.6.1)$$

Para estudarmos a distribuição de t^2 sob H_0 e não normalidade das populações, seguiremos Welch (1938) considerando a estatística,

$$w = \left[1 + \frac{t^2}{n-1}\right]^{-1} \qquad (2.6.2)$$

a qual sob a Teoria Normal tem distribuição Beta com $v_1 = 1$ e $v_2 = (n-1)$ graus de liberdade. Iremos aproximar a distribuição de w sob a não normalidade por uma distribuição Beta e para determinar os correspondentes graus de liberdade considere a Tabela 2.6.1.

Da Tabela 2.6.1 obtemos as duas expressões

$$v_1 = \frac{(1-\mu_1)}{\mu_1}\ v_2 \qquad (1)$$

e

$$v_2 = \frac{2\mu_1(\mu_1-\mu_1^2-\mu_2)}{\mu_2}\ . \qquad (2)$$

TABELA 2.6.1

SOB A TEORIA DA PERMUTAÇÃO	SOB A TEORIA DA NORMAL
$\underset{P}{E}(w) = \frac{n-1}{n} = \mu_1$ $\underset{P}{Var}(w) = \frac{2(n-1)}{n^2(n+2)}\left(1 - \frac{b_2-3}{n-1}\right) = \mu_2$ e $b_2 = (n+2)\frac{\sum_{j=1}^{n} y_j^2}{\left(\sum_{j=1}^{n} y_j^2\right)^2}$	$\underset{N}{E}(w) = \frac{n-1}{n} = \mu_1$ $\underset{N}{Var}(w) = \frac{2(n-1)}{n^2(n+2)} = \mu_2$

Substituindo em (1) e (2) os valores $\underset{P}{E}(w)$ e $\underset{P}{Var}(w)$ temos,

$$v_1 = \frac{\left(1 - \frac{n-1}{n}\right)}{\frac{n-1}{n}} \cdot v_2 \implies v_1 = \frac{1}{n-1} \cdot v_2$$

e

$$v_2 = \frac{2(n-1)}{n}\left[\frac{n-1}{n} - \frac{(n-1)^2}{n} - \frac{2(n-1)}{n^2(n+2)}\left(1 - \frac{b_2-3}{n-1}\right)\right] = (n-1)d$$

Portanto,

$$v_1 = d \quad e \quad v_2 = (n-1)d$$

onde

$$d = \left[1 + \frac{b_2 - 3}{n(1 - \frac{b_2}{n+2})}\right].$$

Logo a distribuição da estatística w, sob a teoria da Permutação, é aproximado por uma distribuição Beta com $v_1 = d$ e $v_2 = (n-1)d$ graus de liberdade. Como nas seções anteriores podemos expressar nosso teste em termos de estatística (2.6.1) o qual sob H_0 tem distribuição F Snedecor Central com $v_1 = d$ e $v_2 = (n-1)d$ graus de liberdade.

Observemos que para n grande, o valor de d aproximar-se de 1 (um). Concluímos que para n grande,o efeito das populações não serem normais, no teste de igualdade de médias é insignificante, no sentido de podermos testar as hipóteses dadas pela estatística (2.6.1) independentes da distribuição da população ser normal ou não.

EXEMPLO 3.2.6

Os dados abaixo representam uma amostra de 5 pacientes, que apresentavam uma deficiência no número de núcleos cromativo positivo. Usando o corante de Feulgen, foi contado o número de núcleos cromativos positivos em cada 100 núcleos, antes e depois do tratamento, cujo objetivo é controlar esta deficiência. Queremos verificar com base nas observações acima se o tratamento aumenta em média o número de nú

NÚMERO CROMALINA ORAL		$y_j = x_{2j} - x_{1j}$
SEM TRATAMENTO	APÓS TRATAMENTO	
16	8	-8
6	14	8
15	12	-3
11	19	8
27	29	2

Dados fornecidos pelo Deptº de Clínica Médica, Hospital das Clínicas, Faculdade de Medicina da USP.

cleos cromatinos positivos.. Para este problema temos as hipóteses

$$H_0: \mu_y = 0$$
$$H_a: \mu_y > 0$$

as quais serão testados ao nível de significância $\alpha = 0{,}05$.

Baseada em outras amostras verificamos que a variável acima não segue a distribuição normal. Considerando o efeito da não normalidade da população, temos que a estatística para o teste é dada por

$$t^2 = \frac{n\,\bar{y}^2}{s_y^2}$$

cujo valor observado é igual a $T_0 = 0{,}2021$.

A estatística t^2 tem sob H_0 distribuição F Snedecor com $v_1 = d$ e $v_2 = (n-1)d$ graus de liberdade onde

$$d = 1 + \frac{b_2 - 3}{n\left\{1 - \frac{b_2}{n+2}\right\}} = 1 + \frac{7\times(-0{,}9371)}{24{,}6855} = 0{,}734.$$

$$b_2 = \frac{\sum_{i=1}^{5} y_1^2}{\left(\sum_{i=1}^{4} y_1^2\right)^2} = \frac{7\times 12385}{(205)^2} = 2{,}0623$$

logo,

$$v_1 = d = 0{,}7342 \simeq 1$$
$$v_2 = (n-1)d = 4\times 0{,}734 \simeq 3.$$

Temos portanto o valor $F_{1,3,0,95} = 10{,}13$.

Como $t_0^2 < 10,13$ decidimos pela aceitação de H_0 e verificamos então que não existe um aumento relativo a média do número de núcleos comatinos positivos em cada 100 núcleos após tratamento.

Observação

O valor de d aproxima-se de 1 pelo fato de b_2 ser aproximadamente igual a 3.

- o -

CAPÍTULO IV

TESTES DE HIPÓTESES ENVOLVENDO MAIS DE DUAS POPULAÇÕES

Como nos capítulos anteriores resumiremos os problemas em testes da igualdade de várias variâncias e da igualdade de várias médias.

Examinaremos o teste da igualdade de várias variâncias de várias populações, no caso em que as populações são Normais e no caso em que as populações não são Normais. Em seguida desenvolveremos testes sobre a igualdade das médias no caso em que as suposições, básicas da Análise de Variâncias (Normalidade, Independência e Homogeneidade das Variâncias) são satisfeitas e no caso em que as suposições de Normalidade nas Populações e Homogeneidade das variâncias não se verificam. Posteriormente discutiremos alguns métodos de comparações múltiplas (Método de Scheffé e Contrastes Ortogonais), quando as variâncias são homogêneas e quando são heterogêneas.

4.1 - TESTES DE HIPÓTESES SOBRE VÁRIAS VARIÂNCIAS

Consideremos as seguintes situações:

4.1.1 - As Populações são Normais

Seja x_{ij}, a j-ésima observação da i-ésima amostra correspondente ao i-ésimo grupo, $i=1,2,\ldots,k$ e $j=1,2,\ldots,n_i$. Os dados observados podem ser colocados na Tabela abaixo.

TABELA 1.1.1

	GRUPOS				
	1	...2	i	k	
	x_{11}	x_{21}	x_{i1}	x_{k1}	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	x_{1n_1}	x_{2n_2}	x_{in_i}	x_{kn_k}	
Totais	$T_{1\cdot}$	$T_{2\cdot}$	$T_{i\cdot}$	$T_{k\cdot}$	$T_{\cdot\cdot}$
Médias	$\bar{x}_{1\cdot}$	$\bar{x}_{2\cdot}$	$\bar{x}_{i\cdot}$	$\bar{x}_{k\cdot}$	$\bar{x}_{\cdot\cdot}$

As k amostras são independentes.

Nosso interesse é testar a hipótese sobre igualdade das variâncias, dadas por,

$$H_0: \sigma_1^2 = \sigma_2^2 = \quad = \sigma_k^2 = \sigma^2$$

$$H_a:^{(1)}$$

a) Teste de Bartlett

O teste de Bartlett (1937) é uma modificação do teste de Neyman e Pearson (1931), baseado na Razão de Verossimilhança, Bronlee (1960). Tomando por base a estatística,

(1) - Nesta seção assim como nas subsequentes, H_a significa que: nem todas as k variâncias (médias) são iguais.

$$-2 \ln \lambda = - \sum_{i=1}^{k} n_i \ln \frac{\hat{\sigma}_i^2}{\hat{\sigma}^2} \qquad (1.1.1)$$

Bartlett (1937) substituiu as estimativas viciadas de Máxima Verossimilhança

$$\hat{\sigma}_i^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$$

e

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2}{N} \quad \text{e} \quad N = \sum_{i=1}^{k} n_i,$$

pelos convencionais estimadores não viciados

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}{n_i - 1}$$

$$s_p^2 = \frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{\sum_{i=1}^{k}{}' (n_i - 1)}$$

e n_i por (n_i-1). Obtendo-se assim a estatística,

$$\chi^{*2} = \sum_{i=1}^{k} (n_i - 1) \ln \frac{s_i^2}{s_p^2} = v \ln s_p^2 - \sum_{i=1}^{k} v_i \ln s_i^2 \qquad (1.1.2)$$

onde,

$v_i' = n_i - 1$, os graus de liberdade para cada amostra.

$v = \sum_{i=1}^{k} v_i$, soma total dos graus de liberdade.

Bartlett (1937) mostrou que sob H_0 e a Teoria Normal a estatística,

$$\frac{\chi^{*2}}{1+A} \qquad (1.1.3)$$

tem distribuição assintótica a uma qui-quadrado com (k-1) graus de liberdade. E A é definido como,

$$A = \frac{1}{3(k-1)} \left\{ \sum_{i=1}^{k} v_i^{-1} - v^{-1} \right\} \qquad k > 1.$$

Observamos que A tende a zero para grandes valores de n_i, $i=1,\ldots,k$. Portanto, concluímos que,

i) para grandes amostras, a estatística para o teste é χ^{*2}, o qual sob H_0 tem distribuição χ^2 (qui-quadrado central) com (k-1) graus de liberdade.

ii) para pequenas amostras, a estatística para o teste é dada por, $\frac{\chi^{*2}}{(1+A)}$, e tem distribuição χ^2 com (k-1) graus de liberdade.

b) Teste de Cochran

Este teste é especialmente usado se $n_2 = n_1 = n_k = n$ e para situações em que uma das variâncias amostrais é muito maior que as restantes.

Temos então que para testar as mesmas hipóteses dadas em (1.1.a), a estatística é dada por,

$$R_{n,k} = \frac{\text{maior variância amostral}}{\sum_{i=1}^{k} s_i^2} \qquad (1.1.4)$$

onde

$$s_i = \frac{\sum_{i=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}{(n_i - 1)}$$

e

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \qquad i=1,2,3,\ldots,k$$

Sob H_0, a estatística R_{n_k}, segue uma distribuição $R_{n,(1-\alpha)}$ cujos valores tabelados para $\alpha = 0{,}01$ ou $\alpha = 0{,}05$ são encontrados em Dixon e Massey (1969) ou em Guenther (1961).

A hipótese H_0 é rejeitada ao nível de significância α se $R_0 n,k > Rn,k,(1-\alpha)$.

EXEMPLO 4.1.1 a)

Quatro áreas marítimas A, B, C e D são diferentes quanto a algumas características climáticas. Queremos à partir de uma amostra de sardinhas de cada uma destas áreas, verificar se os fatos acima induzem uma diferença entre a distribuição do comprimento da cabeça de sardinhas das quatro populações, com relação a variância das distribuições.

Por informações da literatura sabemos que a variável, comprimento da cabeça, segue a distribuição normal. Nosso problema consiste em testar a hipótese,

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$$

ao nível de significância $\alpha = 0{,}05$.

A partir dos dados obtivemos as seguintes informações, na tabela que se encontra na página seguinte.

A estatística para o teste é dada por,

$$\chi^{*2} = \frac{1}{1+A}\left(v \ln s_p^2 - \sum_{i=1}^{k} v_i \ln s_i^2\right)$$

onde o valor observado é $\chi_0^{*2} = 81,341$.

	AREA			
	A	B	C	D
Nº de observações	79	22	98	99
s_i^2	12,517	12,264	1,943	4,494

Os dados foram fornecidos por pesquisadores do Instituto Oceanográfico da Universidade de São Paulo.

Sob H_0 χ^{*2} tem distribuição aproximada χ_3^2 cujo valor tabelado $\chi^2_{3;0,95} = 7,815$. Como $\chi_0^{*2} > 7,815$ rejeitamos H_0 ao nível $\alpha = 0,01$. Portanto as características climáticas das populações têm influência sobre a variância da variável comprimento da cabeça de sardinhas.

EXEMPLO 4.1.1 b)

Os dados abaixo representam medidas de resistência

VÃO→	1080	1690	2300	2910	3520
	279	244	138	118	75
	292	205	141	92	78
	301	197	129	91	86
	278	215	133	103	80
	286	180	140	98	79
	263	164	145	103	83
	270	225	112	114	83
	312	220	133	98	82
	219	187	141	94	88
	257	209	121	101	88
	281	191	121	102	86
	284	196	112	85	83
	264	192	144	100	82
	232	195	117	85	81
	208	178	138	90	78

de telhas, quando sujeitas a pressão de baixo para cima. Foram utilizadas no experimento 5 diferentes vãos de fixação, no sentido do comprimento da telha. Os vãos têm as seguintes medidas (1080 cm, 1690 cm, 2300 cm, 2910 cm e 3.50 cm). Queremos testar as hipõteses,

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2,$$

ao nível de significância $\alpha = 0,01$.

Por experiências anteriores, sabemos que a medida de resitência tem distribuição Normal.

A fim de obtermos o valor observado de Rn,k calculamos as variâncias amostrais; 858,49; 416,16; 134,86;88,36 e 14,44, respectivamente.

O valor observado da estatística é $R_0nk = 0,5$.

A região de rejeição é constituído por valores de Rnk, tais que

$$P(Rn,k>R_c) = 0,01,$$

onde α é o nível de significância prefixado. Para $\alpha = 0,01$ temos $R_c = 0,3645$. Portanto a hipõtese H_0 deverá ser rejeitada e concluímos que a distribuição das resistências das telhas nos diferentes vãos possuem variâncias diferentes.

4.1.2 - As Populações não são Normais

Sejam as amostras como descritas na Tabela 1.1.1 extraídas de populações não normais.

Para testarmos a hipõtese,

$$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$$

desenvolvemos um teste aproximado proposto por Box e Andersen

(1955) e dividiremos em dois casos;

a) as médias populacionais são conhecidas e sem perda de generalidade suponhamos $E(x_{ij}) = 0$.

Estudaremos neste caso, a distribuição da estatística (1.1.2) χ^{*2} sob H_0 e sob a Teoria da Permutação. Com esta finalidade calcularemos a esperança e a variância da estatística χ^{*2} até os termos de ordem N^{-1} e obtemos:

$$E_P(\chi^{*2}) = (k-1)[\tfrac{1}{2}(b_2'-1)+N^{-1}\{\tfrac{1}{2}(b_2'-1) + (NA-3)(3b_2'-b_4'-2)+\tfrac{9}{4}(3NA-2)(b_2'-1)^2\}].$$

onde,

$$A = \frac{1}{3(k-1)}\left\{\sum_{i=1}^{k}\frac{1}{n_i}-\frac{1}{N}\right\},$$

$$\begin{aligned}\mathrm{Var}_P(\chi^{*2}) = {} & \frac{1}{2}(k-1)(b_2'-1)^2+\frac{1}{36N}\Big\{N\sum_{i=1}^{k}\frac{1}{n_i}(9b_6+72b_4 - \\ & -108b_4b_2+168b_2^3-207b_2^2+72b_2-6)+k^4(9b_6+36b_2^3+ \\ & +36b_4b_2+9b_2^2)+k(-18b_6-144b_4+216b_4b_2+ \\ & +288b_2^3+306b_2^2-72b_2)+18b_6+72b_4-144b_4b_2+ \\ & +156b_2^3-108b_2^2+6\Big\},\end{aligned}$$

e

$$b_r' = \frac{N^{\frac{1}{2}r}\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^{r+2}}{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{ij}^2\right)^{\frac{r+2}{2}}}$$

Para N grande,

$$E_p(\chi^{*2}) = (k-1)\{1+\frac{1}{2}(b_2'-3)\}$$

e

$$Var_p(\chi^{*2}) = 2(k-1)\{1+\frac{1}{2}(b_2'-3)\}^2.$$

Portanto a estatística para o teste é dada por,

$$\frac{\chi^{*2}}{\{1+\frac{1}{2}(b_2'-3)\}}, \qquad (1.2.1)$$

a qualsob H_0 tem distribuição assintótica $\chi^2_{(k-1)}$ (qui-quadrado) com k-1 graus de liberdade.

b) as médias populacionais são desconhecidas

Neste caso, Box e Andersen (1955) provaram que estatística para o teste é,

$$\chi^{*2'} = \frac{'\chi^{*2}}{1+\frac{1}{2}c_2} \qquad (1.2.2)$$

o qual sob H_0 e para N grande, tem distribuição χ^2 com (k-1) graus de liberdade. Para $n_1 = n_2 = \ldots = n_k = n$, c_2 é definido como,

$$c_2 = \frac{k\sum_{i=1}^{k} k_{4i}}{\left(\sum_{i=1}^{k} k_{2i}\right)^2}$$

onde

$$k_{4i} = \frac{n_i(n_i-1)s_{4i} - 3(n_i-1)s_{2i}}{(n_i-1)(n_i-2)(n_i-3)}, \; i=1,2,\ldots,k,$$

$$k_{2i} = \frac{s_{2i}}{(n_i - 1)} \quad \text{e} \quad s_{ri} = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2$$

EXEMPLO 4.1.2 b)

Foram colhidas na primavera, nas áreas marítimas A, B, C, D e E, amostras de sardinhas nas quais foram contados os números de raios da dordal. Desconhecemos as médias populacionais e com base em grandes amostras verificamos que a variável número de raios da dorsal não segue a distribuição Normal. Usando os dados abaixo queremos testar a hipótese

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 = \sigma^2$$

ao nível de significância $\alpha = 0,05$.

VALORES OBSERVADOS x_i	FREQUÊNCIA da área A	FREQUÊNCIA da área B	FREQUÊNCIA da área C	FREQUÊNCIA da área D	FREQUÊNCIA da área E
16	0	1	1	0	0
17	1	11	2	1	1
18	12	8	16	15	13
19	9	2	3	6	8
n_i	22	22	22	22	22

Levando em conta a não normalidade das populações e o fato de $n_i = 22$, $(i=1,2,3,4,5)$, temos que a estatística é dada por

$$\chi^{*2\prime} = \frac{\chi^{*2}}{1 + \frac{1}{2}c_2},$$

cujo valor observado é $\chi_0^{*2} = 4,1565$ onde, $(1 + \frac{1}{2}c_2) = 41,3868$.

A estatística $\chi^{*2'}$, segue a distribuição $\chi^2_{(k-1)}$, onde o valor tabelado é $\chi^2_{4;0,95} = 9,49$. Como $\chi^{*2}_{\sigma} < 9,49$, não rejeitamos H_0 e concluímos que a distribuição não difere com relação a variância.

4.2 - TESTES DE HIPÓTESES SOBRE VÁRIAS MÉDIAS

Nesta seção estudaremos as situações, quando as populações são normais e quando as populações não são normais, considerando a homogeneidade e a heterogeneidade das variâncias.

4.2.1 - As Populações são Normais e as Variâncias Homogêneas

Seja x_{ij} a j-ésima observações da i-ésima amostra correspondente ao i-ésimo grupo ($i=1,2,\ldots,k$ e $j=1,2,\ldots,n_i$), como dispostas na Tabela 1.1.1.

Temos então um conjunto de k amostras e desejamos testar a hipótese

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

A estatística para o teste é dada por,

$$F = \frac{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_{i.} - \bar{x}_{..})^2}{(k-1)}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})^2}{(N-k)}} \qquad (2.1.1)$$

onde $N = \sum_{i=1}^{n} n_i$

$$\bar{x}_{i.} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i} \quad \text{e} \quad \bar{x}_{..} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{N}.$$

Os valores observados da estatística são convenientemente obtidos a partir do quadro da Análise de Variância,

A N O V A

CAUSA DA VARIAÇÃO	GRAUS DE LIBERDADE	SOMA DE QUADRADOS	QUADRADOS MÉDIAS (QM)	F
Entre Amostras	k-1	SQ_E	$QM_E = \frac{SQ_E}{k-1}$	
Dentre Amostras	N-k	SQ_D	$QM_D = \frac{SQ_D}{N-1}$	$\frac{QM_E}{QM_D}$
	N-1			

onde

$$QM_E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(\bar{x}_{i.}-\bar{x}_{..})^2}{k-1}$$

e

$$QM_D = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{(x_{ij}-\bar{x}_{i.})^2}{N-k}$$

Sob H_0, a estatística F tem distribuição F de Snedecor Central com (k-1) e (N-k) graus de liberdade. Sob a hipótese alternativa, F tem distribuição F de Snedecor não central. Tabelas para o cálculo do Poder do teste para alguns valores de (k-1) e (N-k), encontra-se em Scheffé (1955) ao nível de significância 0,01 ou 0,05.

4.2.2 - As Populações são Normais e as Variâncias Heterogêneas

Neste caso, temos uma solução aproximada, proposta por Brown e Forsythe (1974).

Consideremos as amostras como na Tabela 1.1.1, onde cada observação tem média μ_i e variância σ_i^2.

Para testar a hipótese,

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k = \mu$$

a estatística adequada é dada por,

nº de comparações

$$F^* = \frac{\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2}{\sum_{i=1}^{k} (1 - \frac{n_i}{N}) s_i^2}, \qquad (2.1.1)$$

a qual sob H_0 segue aproximadamente uma distribuição F Snedecor central (k-1) e f graus de liberdade, onde f é determinador pelo método de Satterthwaite (1941),

$$\frac{1}{f} = \sum_{i=1}^{k} \frac{d_i^2}{(n_i - 1)}$$

onde

$$d_i = \frac{(1 - \frac{n_i}{N}) s_i^2}{\sum_{i=1}^{k} (1 - \frac{n_i}{N}) s_i^2}$$

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}{(n_i - 1)} \quad \text{e} \quad \bar{x}_{i.} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$$

EXEMPLO 4.2.2

O quadro abaixo apresenta informações relativas a amostra de sardinhas colhidas em 5 diferentes áreas, nas quais foram medidas o comprimento standart.

AMOSTRAS	ÁREA A	ÁREA B	ÁREA C	ÁREA D	ÁREA E
n_i	23	6	82	79	90
$\bar{x}_i$	37,09	37,50	39,74	38,18	39,28
s_i^2	12,264	35,772	2,983	12,517	16,403

Dados fornecidos pelo Inst.Oceanográfico da USP.

Baseadas em grandes amostras sabemos que as variâncias diferem. E a partir das informações acima queremos verificar a hipótese,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5,$$

ao nível de significância $\alpha = 0{,}01$.

A estatística para o teste é F^*, cujo valor calculado $F_0^* = 3{,}903$. Sob H_0 F^* segue a distribuição F de Snedecor com (k-1) e f graus de liberdade, para este caso $(k-1) = 4$ e $f = 19$. O valor tabelado de $F_{4,19,0,01} = 4{,}50$ e como $F_0^* < 4{,}50$, aceitamos H_0 e concluímos que não existe diferença significante entre o comprimento médio padrão das sardinhas nas 5 áreas.

4.2.3 - As Populações não são Normais e as Variâncias ~~Heterogêneas~~ HOMOGÊNEAS

O trabalho apresentado por Box e Andersen (1955), propõe para este caso uma solução aproximada, a qual se baseia em uma estatística w desenvolvida por Welch (1938).

Considere a Tabela 1.1.1 e as populações não normais. Para testar a hipótese,

$$H_0: \mu_1 = \mu_2 = ,,, = \mu_k = \mu,$$

estudaremos o comportamento da estatística (2.1.1). Para isto vamos primeiramente considerar a estatística,

$$w = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{x}_{i.} - \bar{x}_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2}, \qquad (2.3.1)$$

a qual é equivalente a estatística (2.1.1).

Sob a H_0 e sob a Teoria Normal w tem distribuição Beta com $v_1 = (k-1)$ e $v_2 = (N-k)$ graus de liberdade e

$$N = \sum_{i=1}^{k} n_i .$$

Aproximaremos a distribuição de w sob a Teoria da Permutação pela distribuição de w sob a Teoria Normal, porém com os graus de liberdade determinados pela relação, graus de liberdade, média e variância. Para isto apresentaremos a seguinte tabela e procederemos como segue:

Com base nessa Tabela (2.3.1), temos que determinar os graus de liberdade v_1 e v_2, considerando as seguintes relações:

$$v_1 = \frac{2\mu_1(\mu_1 - \mu_1^2 - \mu_2)}{\mu_2}$$

TABELA 2.3.1

SOB A TEORIA DA PERMUTAÇÃO	SOB A TEORIA NORMAL
$E_P(w) = \frac{k-1}{N-1} = \mu_1$	$E_N(w) = \frac{k-1}{N-1} = \mu_1$
$\underset{P}{\text{Var}}(w) = \frac{2(k-1)(N-k)}{(N-1)^2(N+1)}\left[1 - \frac{k_4}{Nk_2^2}\right] - \frac{k_4}{(N-1)^2 k_2^2}\left[\frac{k^2}{N} - \sum_{i=1}^{k} \frac{1}{n_i}\right] = \mu_2$	$\underset{N}{\text{Var}}(w) = \frac{2(k-1)(N-k)}{(N-1)^2(N+1)} = \mu_2$
onde	$N = \sum_{i=1}^{k} n_i$
$k_4 = \frac{[N(N+1)S_4 - 3(N-1)s_2^2]}{(N-1)(N-2)(N-3)}$	
$k_2 = \frac{s_2}{(N-1)}$	
$s_r = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2$	
k = nº de amostras	

e

$$v_2 = \frac{1-\mu_2}{\mu_1} v_1$$

a) para amostras do mesmo tamanho $(n_1 = n_2 = \ldots = n_k = n)$, a variância de W sob a Teoria da Permutação é,

$$\underset{P}{\text{Var}}(W) = \frac{2(k-1)(N-k)}{(N-1)^2(N+1)}\left[1 - \frac{c_2}{N}\right] \quad \text{e} \quad c_2 = \frac{k_4}{k_2^2}$$

e temos que

$$v_1 = (k-1)d \quad \text{e} \quad v_2 = (N-k)d$$

onde

$$d = 1 + \frac{(N+1)c_2}{(N-1)(N-c_2)}$$

Concluímos nesta caso que a estatística para o teste é dada por (2.1.1) a qual sob H_0 segue um distribuição F Snedecor com $(k-1)d$ e $(N-k)d$ graus de liberdade.

b) para amostras de tamanhos diferentes, temos que a variância de W sob a teoria de Permutação é dada por,

$$\operatorname{Var}_P(W) = \frac{2(k-1)(N-k)}{(N-1)^2(N+1)}\left\{1 - \frac{c_2}{N}\right\} - \frac{c_2}{(N-1)^2}\left\{\frac{k^2}{N} - \sum_{i=1}^{k}\frac{1}{n_i}\right\}.$$

Usando o raciocínio anterior, a estatística (2.1.1) tem distribuição F Snedecor com $v_1 = (k-1)d$ e $v_2 = (N-k)d$ graus de liberdade e neste caso d é dado por

$$d = 1 + \frac{N+1}{(N-1)}\,\frac{c_2}{(N^{-1}+A)^{-1}-c_2},$$

$$A = \frac{(N+1)}{2(k-1)(N-k)}\left(\frac{k^2}{N} - \sum_{i=1}^{k}\frac{1}{n_i}\right) \quad \text{e} \quad c_2 = \frac{k_4}{k_2^2},$$

k_4 e k_2^2 como definidos no item a).

Observamos que quando $N = \Sigma\, n_i$ cresce, A tende a zero e d tende a 1 (um).

EXEMPLO 4.2.3

Os dados abaixo representam o número de raios da nadadeira dorsal, em cinco amostras de sardinhas, colhidas em cinco áreas A, B, C, D e E na primavera.

Nº DE RAIOS DA NADADEIRA DORSAL	FREQUÊNCIAS				
	ÁREA A	ÁREA B	ÁREA C	ÁREA D	ÁREA E
16	-	4	1	-	-
17	5	39	2	2	2
18	49	31	16	58	55
19	36	5	3	38	42
	90	79	22	98	99

Os dados foram fornecidos pelo Inst.Oceanográfico da USP

Baseada em uma outra amostra e através de um teste de hipótese, verificamos que a variável aleatória acima não segue a distribuição normal.

A partir dos dados observados queremos verificar as seguintes hipóteses,

$$H_0: \mu_A = \mu_B = \mu_C = \mu_D = \mu_E$$

ao nível de significância $\alpha = 0{,}05$.

Considerando o efeito da não normalidade temos que a estatística

$$F = \frac{(N-k)}{(k-1)} \cdot \frac{w}{1-w}, \text{ cujo valor } F_0 = 37{,}7038.$$

F, segue a distribuição F Snedecor com 4d e 383d graus de liberdade, onde

$$d = 1 + \frac{N+1}{N-1} \cdot \frac{c_2}{(N^{-1}+A)^{-1} - c_2} \simeq 1$$

e

$$A \simeq 0{,}003185$$

neste caso

$$4d = 4 \quad \text{e} \quad 383d = 383.$$

O valor tabelado de $F_{4,383,0,95} = 2{,}37$. Como $F_0 > 2{,}37$ concluímos pela rejeição de H_0 e verificamos que existe significante diferença entre o número médio de raios da dorsal, nas cinco áreas especificadas.

O motivo dos graus de liberdade não serem modificados é devido a tamanho do N de acordo com a observação da parte teórica.

4.2.4 - As Populações são não Normais e as Variâncias Heterogêneas

Com base em Scheffé (1955) apresentaremos uma solução limite desenvolvida abaixo.

Seja as amostras como definidas na Tabela 1.1.1 extraídas de populações não normais. Para testar a hipótese sobre a igualdade das médias estudaremos o comportamento da estatística

$$F = \frac{\dfrac{\sum\limits_{i=1}^{k} n_i(\bar{v}_{i.} - v)^2}{k-1}}{\dfrac{\sum\limits_{i=1}^{k}(n_i-1)s_i^2}{\sum\limits_{i=1}^{k}(n_i-1)}} \qquad (2.4.1)$$

onde

$$\bar{v}_i = \bar{x}_{i\cdot} - \mu_i \quad \text{e} \quad \bar{v} = \frac{\sum_{i=1}^{k} n_i \bar{v}_{i\cdot}}{N}$$

Consideremos n_i grande e consequentemente:

i) $\sum_{i=1}^{k} (n_i - 1) = \sum_{i=1}^{k} n_i - k \simeq \sum_{i=1}^{k} n_i$

ii) s_i^2 converge em probabilidade para σ_i^2 e portanto

$$\sum_{i=1}^{k} s_i^2 \text{ para } \sum_{i=1}^{k} \sigma_i^2$$

iii) e pode-se aceitar o fato que,

$$\frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{k} (n_i - 1)} \simeq \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{k} n_i}$$

é desprezível por (i), (ii) e (iii) e para n_i grande,

$$\frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{\sum_{i=1}^{k} (n_i - 1)} \text{ converge para } \frac{\sum_{i=1}^{k} n_i \sigma_i^2}{N} = (\sigma_{\bullet}^2) w$$

Substituindo o denominador de (2.4.1) por $(\sigma_{\bullet}^2)w$, temos,

$$F' = \frac{\dfrac{\sum_{i=1}^{k} n_i (v_i - \bar{v}_{\cdot})^2}{k - 1}}{(\sigma_{\bullet}^2) w} \qquad (2.4.2)$$

A Tabela 2.4.1 representa para n_i grande a média e a variância de F e F'.

SOB A NÃO NORMALIDADE E VARIÂNCIAS HETEROGÊNEAS	SOB A TEORIA NORMAL E VARIÂNCIAS HOMOGÊNEAS
$E(F')= \frac{1}{k-1} \cdot \left\{ \frac{k(\sigma^2_\bullet)u}{(\sigma^2_\bullet)w} - 1 \right\}$	$E(F) = 1$
$Var(F')= \frac{2}{k-1} \left\{ k(1+Vu) \left[\frac{(\sigma^2_\bullet)u}{(\sigma^2_\bullet)w} \right] - 2Vw = 1 \right\}$	$Var(F) = \frac{2}{k-1}$
onde	
$(\sigma^2_\bullet)u$ é a média não ponderada de σ^2_i	
$(\sigma^2_\bullet)w$ é a média ponderada de σ^2_i	
$Vu = [(\sigma^2_\bullet)u]^{-2} \frac{\sum_{i=1}^{k} [\sigma^2_i - (\sigma^2_\bullet)u]^2}{k}$	
$Vw = [(\sigma^2_\bullet)w]^{-2} \frac{\sum_{i=1}^{k} [\sigma^2_i - (\sigma^2_\bullet)w]^2}{\sum_{i=1}^{k} n_i}$	

No caso em que $n_1 = n_2 = n_3 = n_k = n$, temos,

a) $(\sigma^2_\bullet)u = (\sigma^2_\bullet)w$

$$(\sigma^2_\bullet)w = \frac{\sum_{i=1}^{k} n_i \sigma^2_i}{\sum_{i=1}^{k} n_i} = \sum_{i=1}^{k} \frac{n\sigma^2_i}{nk} = n \sum_{i=1}^{k} \frac{\sigma^2_i}{nk} = (\sigma^2_\bullet)u$$

e portanto,

$$E(F') = \{\frac{1}{k-1}\ k-1\} = 1 = E(F)$$

b) Se $k = 2$

i) $Vu = Vw = 0$

ii) $(\sigma_{\cdot}^2)u = (\sigma_{\cdot}^2)w \Longrightarrow \dfrac{(\sigma_{\cdot}^2)u}{(\sigma_{\cdot}^2)w} = 1$

e portanto,

$$Var(F) = Var(F') = \frac{2}{k-1}$$

c) em outros casos,

$$Var(F') = \frac{2}{k-1}\left(1 + Vu\ \frac{k-2}{k-1}\right)$$

Vemos que para amostras de tamanho iguais, quando $k \neq 2$,

$$Var(F') = \frac{2}{k-1}\left(1 + Vu\ \frac{k-2}{k-1}\right),$$

e segue-se que a verdadeira *Probabilidade do Erro Tipo I é maior que o nível de significância fixado* α.

Concluímos pelo desenvolvimento acima que no caso das variâncias diferentes e populações não normais não existe uma solução simples mesmo quando os n_i, $i=1,2,\ldots,k$, são grandes. Nas situações práticas deste tipo, aconselhamos o uso de transformações.

4.3 - COMPARAÇÕES MÚLTIPLAS

Se a hipótese $H_0: \mu_1 = \mu_2 = \ldots = \mu_k = \mu$ é rejeitada, nós podemos concluir que nem todas as k médias populacionais são iguais. Quando isto acontece nós poderemos estar interessados em saber que médias diferem, ou ordenar os grupos de acordo com a ordem crescente ou decrescente baseado nas médias amostrais. Indagações como estas nos leva ao estudo das Comparações Múltiplas.

Entre os vários métodos de comparações Múltiplas citados na literatura estudaremos aqui apenas os Métodos, de Scheffé e Contrastes Ortogonais, pelo fato de apresentarmos soluções relativamente simples no caso em que as variâncias dos grupos são desiguais.

4.3.1 - Método de Scheffé

As comparações que envolvem um conjunto de médias, quase sempre são expressas por meio de contrastes que constituem um conjunto de funções estimáveis nos modelos de Análise de Variância.

Definimos contrastes por,

$$C = \sum_{i=1}^{k} a_i \mu_i \qquad (3.1.1)$$

onde, $\sum_{i=1}^{k} a_i = 0$.

Um estimador não viesado de um constraste C é dado por,

$$\hat{C} = \sum_{i=1}^{k} a_i \bar{x}_i \qquad (3.1.2)$$

Suponhamos as amostras como na Tabela 1.1.1 extraídas de populações caracterizadas por variáveis aleatórias que seguem a distribuição Normal e as observações independentes. Para testar a hipótese,

$$H_0: C = 0 \text{ versus } H_a \neq 0,$$

distinguiremos dois casos:

a) As variâncias são iguais.

Quando $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$ a variância de $\hat{C}$ e um estimador não viesado desta variância são respectivamente:

$$\mathrm{Var}(\hat{C}) = \sigma^2 \cdot \frac{\sum_{i=1}^{k} a_i^2}{n_i} \tag{3.1.3}$$

e

$$\widehat{\mathrm{Var}(\hat{C})} = \mathrm{QMD} \, \frac{\sum_{i=1}^{k} a_i^2}{n_i} \tag{3.1.4}$$

onde σ^2 é a variância comum e QMD é o quadrado médio do resíduo da Análise de Variância correspondente.

Seja C um contraste, como definido acima, então Scheffé (1955) provou que,

$$\Pr(\hat{C} - S\sqrt{\widehat{\mathrm{Var}(\hat{C})}} < C < \hat{C} + S\sqrt{\widehat{\mathrm{Var}(\hat{C})}}) = 1 - \alpha, \tag{3.1.5}$$

onde $s^2 = (k-1) \cdot F_{1,(N-k)(1-\alpha)}^{(1)}$ e a expressão do primeiro membro

(1) - $F_{a,b,1-\alpha} = F_c$ é o ponto tal que $P(|F|>F_c) = \alpha$, onde F tem distribuição F Snedecor com a,b graus de liberdade.

de (3.1.5) significa "Probabilidade de que simultaneamente todos os contrastes satisfaçam a desigualdade entre parenteses".

Para cada contraste a hipótese H_0: $C = 0$ é rejeitada se o intervalo (3.1.5) correspondente não contém o (0) zero e o nível de significância conjunto é α.

b) As variâncias são diferentes.

Aqui a variância do estimador de C, $\hat{C}$ é dada por,

$$Var(\hat{C}) = \sum_{i=1}^{k} \frac{a_i^2 \sigma_i^2}{n_i} \qquad (3.1.6)$$

onde σ_i^2 é a variância do i-ésimo grupo. Um estimador não viesado de $Var(\hat{C})$ é da forma,

$$\widehat{Var}\ (\hat{C}) = \sum_{i=1}^{k} \frac{a_i^2 s_i^2}{n_i} \qquad (3.1.7)$$

onde

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}{(n_i - 1)}, \qquad i=1,2,\ldots,k.$$

Neste caso, (variâncias heterogêneas) se substituirmos na expressão (3.1.5) $\widehat{Var}(\hat{C})$ por (3.1.7), teremos a mesma alteração nos graus de liberdade da distribuição F considerada na Análise de Variância quando as variâncias são heterogêneas (Vide a seção 2.2).

Portanto temos,

$$\Pr\left(\hat{C} - S\sqrt{\sum_{i=1}^{k}\frac{a_i^2 s_i^2}{n_i}} < C < \hat{C} + S\sqrt{\sum_{i=1}^{k}\frac{a_i^2 s_i^2}{n_i}}\right) = 1 - \alpha$$

onde $S^2 = (k-1)F_{1,f,1-\alpha}$.

e

$$\frac{1}{f} = \sum_{i=1}^{k}\frac{d_i^2}{(n_i-1)} \qquad d_i = \frac{\frac{s_i^2}{n_i}}{\sum_{i=1}^{k}\frac{s_i^2}{n_i}}$$

4.3.2 - Contrastes Ortogonais

DEFINIÇÃO - Sejam

$$C_\ell = \sum_{i=1}^{k} a_{\ell_i}\mu_i \quad \text{e} \quad C_m = \sum_{i=1}^{k} a_{m_i}\mu_i$$

dois contrastes de um conjunto de $(k-1)$ contrastes. Dizemos que o conjunto é ortogonal se, g.l. do contraste (?)

$$\frac{\sum_{i=1}^{k} a_{\ell_i} a_{m_i}}{n_i} = 0$$

para todo ℓ e m tal que $\ell \neq m = 1,2,\ldots,(k-1)$.

Suponhamos as amostras como na Tabela 1.1.1, extraídas de populações caracterizadas por variáveis aleatórias que seguem a distribuição normal e as observações independentes. Seja C_e um contraste que pertence a um conjunto de (k-1) contrastes ortogonais. Para testar a hipótese,

$$H_0: C_\ell = 0 \text{ contra } H_a: C_\ell \neq 0,$$

consideremos duas situações,

a) As variâncias são iguais.

Seja $\hat{C}_e$ o estimador não viesado de C_ℓ definido em (3.1.2). A variância de $\hat{C}_\ell$ é dada por

$$\mathrm{Var}\ (C_\ell) = \sigma^2 \sum_{i=1}^{k} \frac{a^2_{\ell_i}}{n_i} \tag{3.2.1}$$

Como a distribuição das observações é normal podemos dizer que

$$\frac{(\hat{C}_\ell - C_\ell)^2}{\dfrac{\sum\limits_{i=1}^{k} a_i^2 \sigma^2}{n_i}},$$

tem distribuição χ^2_1.

Por outro lado

$$\frac{QMD}{\sigma^2},$$

onde QMD é o quadrado médio do Resíduo da ANOVA, tem distribuição $\chi^2_{(N-k)}$ e é independente do primeiro. Portanto, para testarmos a hipótese,

$$H_0: C_\ell = 0,$$

a estatística para o teste é dada por,

$$F_\ell = \frac{\dfrac{(\hat{C}_\ell)^2}{\sum\limits_{i=1}^{k} \dfrac{a_i^2\sigma^2}{n_i}}}{\dfrac{QMD}{\sigma^2}} \tag{3.2.2}$$

o qual sob H_0 tem distribuição F Snedecor Central com 1 (um) grau de liberdade.

A hipótese H_0 é rejeitada para valores de F_0 tais que $|F_0| > F_{1,(N-k),1-\alpha}$ e neste caso dizemos que o contraste C_e difere de (0) zero.

Observemos que embora cada contraste de conjunto de (k-1) contrastes ortogonais é testado separadamente o nível de significância se mantêm.

b) As variâncias são diferentes.

Consideremos um conjunto de k-1 contrastes ortogonais como definido em (3.2.a). Aqui $\hat{C}_\ell$ tem distribuição normal com média C_e e variância dada por,

$$\mathrm{Var}\ (\hat{C}_\ell) = \sum_{i=1}^{k} a_{\ell_i}^2 \frac{\sigma_i^2}{n_i} \tag{3.2.3}$$

Baseado no artigo de Brown e Fosrythe (1974) e Satterthwait (1941), poderemos testar a hipótese

$$H_0 : C_\ell = 0$$

usando a estatística,

$$F_\ell^* = \frac{\left(\sum\limits_{i=1}^{k} a_{\ell_i} \bar{x}_{i\cdot}\right)^2}{\sum\limits_{i=1}^{k} a_{\ell_i}^2 \frac{s_i^2}{n_i}} \tag{3.2.4}$$

a qual sob H_0, segue aproximadamente uma distribuição F Snedecor Central com 1 e f graus de liberdade, onde,

$$\frac{1}{f} = \sum_{i=1}^{k} \frac{d_i^2}{(n_i - 1)} \quad \text{e} \quad d_i = \frac{\left(\sum\limits_{\ell=1}^{k-1} a_{\ell_i}^2\right) s_i^2/n_i}{\sum\limits_{i=1}^{k} \left(\sum\limits_{\ell=1}^{k-1} a_{\ell_i}^2\right)^2 s_i^2/n_i}$$

Para o nível de significância α, rejeitaremos H_0 se $|F_0^*| > F_{1,f,1-\alpha}$ onde F_ℓ^* segue a distribuição F Snedecor com 1 e f graus de liberdade.

Aqui vake a mesma observação feita para o caso de variâncias iguais.

EXEMPLO 4.2.3

Os dados abaixo representam medidas de resistência de telhas, quando sujeitos a pressão de baixo para cima. Foram utilizadas no experimento de 5 diferentes vãos de fixação, no sentido do comprimento da telha. Os vãos têm as seguintes medidas (1080, 1690, 2300, 2910, 3520). Testaremos a hipótese,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5,$$

ao nível de significância $\alpha = 0{,}05$.

V	Ã	O		
1080	1690	2300	2910	3520
219	244	138	118	75
292	205	141	92	78
301	197	129	91	86
278	215	133	103	80
286	180	140	98	79
263	164	145	103	83
270	225	112	114	83
312	220	133	98	82
219	187	141	94	88
257	209	121	101	88
281	191	121	102	86
284	196	112	85	83
264	192	144	100	82
232	195	117	85	81
280	178	138	90	78
$\bar{x}_{1.} = 268,4$	$\bar{x}_{2.} = 199,9$	$\bar{x}_{3.} = 131,0$	$\bar{x}_{4.} = 98,3$	$\bar{x}_{5.} = 82,13$
$s_1^2 = 859,40$	$s_2^2 = 416,84$	$s_3^2 = 133,86$	$s_4^2 = 88,38$	$s_5^2 = 14,41$

Baseados em grandes amostras sabemos que as variâncias são heterogêneas e para este caso a estatística para o teste é $F_0^* = 297,11$. A estatística F^*, segue a distribuição F Snedecor com 4 e $f = 34$ graus de liberdade onde o valor tabelado $F_{4,34,0,35} = 2,66$. Como $F_0^* > 2,66$, rejeitaremos H_0 e verificamos que a resistência média das telhas diferem nos diferentes vãos.

A aplicação de comparações Múltiplas, está diretamente ligada com a rejeição da hipótese H_0.

Para exemplificar a teoria faremos o intervalo de confiança de Scheffé usando um só contraste C e em seguida testaremos um conjunto constrastes Ortogonais.

EXEMPLO 4.3.1 b)

Com base nas médias amostrais, consideremos um contraste C tal que o estimador

$$\hat{C} = -268,4 - 199,9 + 98,3 + 82,13 = -287,83$$

nosso objetivo é testar

$$H_0: C = 0.$$

Então temos,

$$Pr\{-287,83 - 46,02 < C < -287,83 + 46,02\} = 0,95.$$

$$Pr(-329,06 < C < -237,02) = 0,95.$$

O intervalo acima não contém o (0) zero e portanto a hipótese H_0: $C = 0$ é rejeitada.

Observe que

$$\frac{1}{f} = 0,0352 \Longrightarrow f = 28,93 \simeq 29.$$

e

$$F_{4,29,0,95} = 5,73$$

EXEMPLO 4.3.2 b)

Veremos agora uma aplicação de teste para um conjunto de Constrastes Ortogonais, quando as variâncias nos grupos diferem.

Consideremos o exemplo (4.2.3) e seja C_1, C_2, C_3 e C_4 constrastes cujos coeficientes são:

C_1:	-2	-1	0	1	2	(Testa o efeito linear)
C_2:	2	-1	-2	-1	2	(Testa o efeito Quadrático)
C_3:	1	2	0	-2	1	(Testa o efeito Cúbico)
C_4:	1	-4	6	-4	1	(Testa o efeito de 4ª Ordem)

Queremos verificar se,

a) $H_{0_1}: C_1 = 0$,
isto é, não existe efeito linear.

b) $H_{0_2}: C_2 = 0$,
isto é, não existe efeito quadrático.

c) $H_{0_3}: C_3 = 0$,
isto é, não existe efeito cúbico.

d) $H_{0_4}: C_4 = 0$,
Isto é, não existe efeito de 4ª ordem.

Os valores observados da estatística

$$F'_\ell = \frac{\hat{C}^2_\ell}{\dfrac{\sum_{i=1}^{5} a^2_{\ell_i} s^2_i}{15}} \qquad \ell=1,2,3,4,$$

assim como o valor tabelado $F_{1,f,0,95}$ estão colocados na Tabela a seguir e f é da forma,

$$\frac{1}{f} = \sum_{i=1}^{k} \frac{d_i^2}{(n_i - 1)} \quad \text{e} \quad d_i = \frac{\left(\sum_{\ell=1}^{4} a_{\ell_i}^2\right) \frac{s_i^2}{n_i}}{\sum_{i=1}^{5} \left(\sum_{\ell=1}^{4} a_{\ell_i}^2\right) \frac{s_i^2}{n_i}}$$

EFEITOS	F_i' OBSERVADO	F TABELADO $F_{1,f,0,95}$	CONCLUSÃO
Linear	1.082,89	4,01	Significante
Quadrático	61,83	4,01	Significante
Cúbico	1,386	4,01	Não signific.
4ª Ordem	3,218	4,01	Não signific.

Pela tabela acima rejeitamos as hipóteses $H_{0_1}: C_1 = 0$ e H_{0_2} $C_2 = 0$ o que significa dizer que as médias apresentam efeito linear, e Quadrático. Aceitamos porém $H_{0_3}: C_3 = 0$ e $H_{0_4}: C_4 = 0$ e verificamos assim que não existe efeito e de 4a. Ordem.

- o -

SUMÁRIO

Vamos apresentar para fins práticos, um resumo das soluções de testes sobre médias e variâncias quando as suposições usuais não são satisfeitas.

1. Testes de Hipóteses envolvendo uma População não normal.

$H_0: \sigma^2 = \sigma_0^2$

1. Amostras pequenas. Deve-se usar transformações da variável (Bartlett, 1947) ou testes não paramétricos (Conover 1971).
2. Amostras grandes. A estatística para o teste é

$$\left(\frac{n-1}{2}\right)^{1/2} \cdot \left(\frac{s^2}{\sigma_0^2} - 1\right)$$

Sob H_0 a estatística tem distribuição normal com média zero e variância $(1 + \frac{1}{2}\gamma_2)$.

γ_2 = Coeficiente de Curtose -3.

$H_0: \mu = \mu_0$

1. Amostras pequenas. Usa-se transformações de variáveis (Bartlett, 1947) ou teste não paramétrico (Conover, 1971).
2. Amostras grandes. A estatística para o teste é

$$\frac{(n)^{1/2}(\bar{x}-\mu_0)}{\sigma}$$ quando a variância é conhecida,

$$\frac{(n)^{1/2}(\bar{x}-\mu_0)}{s}$$ quando a variância é desconhecida.

Sob H_0 ambas as estatísticas, para amostras grandes, têm distribuição Normal com média zero e variância um.

2. Testes de Hipóteses envolvendo duas populações não Normais.

$H_0: \sigma_1^2 = \sigma_2^2$

1. As médias μ_1 e μ_2 são conhecidas e iguais a zero. A estatística,

$$\frac{n_2}{n_1} \frac{\sum_{j=1}^{n_1} x_{1j}^2}{\sum_{j=1}^{n_2} x_{2j}^2}$$

segue sob H_0, distribuição F Snedecor com $n_1 d$ e $n_2 d$ graus de liberdade e

$$d = \left[1 + \frac{1}{2}\left\{\frac{N+2}{N-1-(b_2-3)}\right\}(b_2-3)\right]^{-1}$$

para $N = n_1 + n_2$ grande

$$d = [1 + \frac{1}{2}(b_2-3)]^{-1}$$

e

$$b_2 = \frac{(N+2)\left\{\sum_{j=1}^{n_1} x_{1j}^4 + \sum_{j=1}^{n_2} x_{2j}^4\right\}}{\left(\sum_{j=1}^{n_1} x_{1j}^2 + \sum_{j=1}^{n_2} x_{2j}^2\right)^2}$$

2. As médias μ_1 e μ_2 são desconhecidas. A estatística para o teste é:

$$\frac{\chi_1^{*2}}{1+\frac{1}{2}c_2},$$

segue sob H_0 distribuição χ_1^2, para $n_1 = n_2$,

$$c_2 = \frac{2\sum_{i=1}^{2} K_{4i}}{\left(\sum_{i=1}^{2} k_{2i}\right)^2}$$

$$K_{4i} = \frac{n_i(n_i+1)S_{4i}-3(n_i-1)S_{2i}^2}{(n_i-1)(n_i-2)(n_i-3)} \qquad i=1,2$$

$$K_{2i} = \frac{S_{2i}}{(n_i-1)} \qquad i=1,2$$

$$S_{ri} = \sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^r \qquad i=1,2$$

$H_0: \mu_1=\mu_2$

1. Amostras independentes e $\sigma_1^2 = \sigma_2^2$. A estatística para o teste, é:

$$\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\bar{x}_{i.}-\bar{x}_{..})^2}{\frac{1}{(N-2)}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i.})^2},$$

a qual sob H_0 segue uma distribuição F Snedecor com d e (N-2)d graus de liberdade

i) se $n_1 = n_2$,

$$d = 1 + \left\{\frac{(N+1)}{(N-1)} \cdot \frac{c_2}{(N-c_2)}\right\}$$

ii) se $n_1 \neq n_2$,

$$d = \left\{1 + \frac{(N+1)}{(N-1)} \cdot \frac{c_2}{(N^{-1}+A)^{-1}-c_2}\right\}$$

$$c_2 = \frac{k_4}{k_2^2} \qquad k_2 = \frac{S_2}{N-1}$$

$$k_4 = \frac{N(N+1)S_4-3(N-1)S_2^2}{(N-1)(N-2)(N-3)}$$

$$A = \frac{N+1}{2(K-1)(N-K)}\left(\frac{K^2}{N} - \sum_{i=1}^{K} \frac{1}{n_i}\right)$$

$$S_r = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^r$$

iii) se $N = n_1 + n_2$ grande d aproxima-se de 1.

2. Amostras independentes e $\sigma_1^2 \neq \sigma_2^2$. Amostras grandes. A estatística para o teste,

$$\frac{(\bar{x}_1 - \bar{x}_2)}{S_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{1/2}},$$

segue sob H_0 distribuição normal com média zero e variância

$$\frac{\theta + R}{R\theta + 1}, \text{ onde } \theta = \frac{\sigma_1^2}{\sigma_2^2} \text{ e } R = \frac{n_1}{n_2}.$$

Porém se $n_1 = n_2$, $\frac{\theta + R}{R\theta+1} = 1$, independente da distribuição da população ser normal ou não.

3. Amostras dependentes (dados emparelhados). A estatística para o teste,

$$t^2 = \frac{n.\bar{y}^2}{s_y^2},$$

sob H_0, tem distribuição F Snedecor com d e (n-1)d graus de liberdade, onde

$$d = \left\{1 + \frac{b_2 - 3}{n\left(1 - \frac{b_2}{n+2}\right)}\right\}$$

$$b_2 = \frac{(n+2)\sum_{i=1}^{n} y_j^2}{\left(\sum_{j=1}^{n} y_j^2\right)^2}$$

3. Mais de duas populações não Normais.

$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2 = \sigma^2$

1. As médias μ_i são conhecidas e iguais a zero. A estatística para o teste é

$$\frac{\chi^{*2}}{\{1 + \frac{1}{2}(b_2' - 3\}}$$

sob H_0 segue a distribuição χ_{k-1}^2.

$$b_2' = \frac{N \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^4}{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 \right)^2}$$

2. As médias μ_i são desconhecidas e amostras grandes. A estatística

$$\frac{\chi^{*2}}{1 + \frac{1}{2} c_2},$$

segue, sob H_0 uma distribuição χ^2_{k-1}. Se $n_i = n$

$$c_2 = \frac{K \sum_{i=1}^{K} k_{4i}}{\left(\sum_{i=1}^{K} k_{2i} \right)^2}$$

$$k_{4i} = \frac{n_i(n_i-1)S_{4i} - 3(n_i-1)S_{2i}}{(n_i-1)(n_i-2)(n_i-3)}$$

$$k_{2i} = \frac{S_{2i}}{n_i - 1}$$

$$S_{ri} = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^r \qquad i=1,2,\ldots,k.$$

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

1. As variâncias $\{\sigma_i^2\}$ são iguais. A estatística usual F da ANOVA, segue sob H_0, F Snedecor com $(k-1)$d e $(N-k)$d graus de liberdade

   i) se $n_1 = n_2 = \ldots = n_k = n$,

$$d = \left\{1 + \frac{N+1}{N-1} \cdot \frac{c_2}{(N-c_2)}\right\}$$

ii) se os n_i $(i=1,2,\ldots,k)$ diferem

$$d = \left\{1 + \frac{N+1}{N-1} \cdot \frac{c_2}{(N^{-1}+A)^{-1} - c_2}\right\}$$

Para N grande, d aproxima-se de um.

$$c_2 = \frac{k_4}{(k_2)^2} \qquad k_2 = \frac{S_2}{N-1}$$

$$k_4 = \frac{N(N+1)S_4 - 3(N-1)S_2^2}{(N-1)(N-2)(N-3)}$$

$$S_r = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^r$$

$$A = \frac{N+1}{2(K-1)(N-K)} \left(\frac{K^2}{N} - \sum_{i=1}^{K} \frac{1}{n_i} \right)$$

2. As variâncias $\{\sigma_i^2\}$ são diferentes.

Verifica-se, que mesmo para n_i iguais e grande a verdadeira probabilidade do erro Tipo I é maior que o nível de significância fixado α

Populações Normais e Variâncias diferentes

1. H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$. A estatística

$$F = \frac{\sum_{i=1}^{k} (\bar{x}_{i.} - \bar{x}_{..})^2}{\sum_{i=1}^{k} \left(1 - \frac{n_i}{N}\right) s_i^2}$$

segue sob H_0 a distribuição F Snedecor com (k-1) e f graus de liberdade

$$\frac{1}{f} = \sum_{i=1}^{k} \frac{d_i^2}{(n_i - 1)} \quad \text{e} \quad d_i = \frac{(1 - \frac{n_i}{N}) s_i^2}{\sum_{i=1}^{k} (1 - \frac{n_i}{N}) s_i^2}$$

2. H_0: $C = 0$ (onde C é um contraste).

a) Método de Scheffé

$$\Pr\left(\hat{C} - S\sqrt{\sum_{i=1}^{k} \frac{a_i^2 s_i^2}{n_i}} < C < \hat{C} + S\sqrt{\sum_{i=1}^{k} \frac{a_i^2 s_i^2}{n_i}}\right) = 1 - \alpha$$

onde $S^2 = (k-1) F_{1,f,1-\alpha}$.

C é um contraste qualquer.

$$\frac{1}{f} = \sum_{i=1}^{k} \frac{d_i^2}{(n_i - 1)} \qquad d_i = \frac{\frac{s_i^2}{n_i}}{\sum_{i=1}^{k} \frac{s_i^2}{n_i}}$$

b) Contrastes Ortogonais

A estatística para o teste

$$F_{\ell}^{*} = \frac{\left(\sum_{i=1}^{k} a_{\ell_i} \bar{x}_{i\cdot}\right)^2}{\sum_{i=1}^{k} \frac{a_{\ell_i}^2 s_i^2}{n_i}}$$

a qual sob H_0 segue F Snedecor com 1 e f graus de liberdade para cada $\ell=1,2,\ldots,(k-1)$.

Neste caso C pertence a um conjunto de contrastes ortogonais.

$$\frac{1}{f} = \sum_{i=1}^{k} \frac{d_i^2}{(n_i-1)} \qquad d_i = \frac{\left(\sum_{\ell=1}^{k-1} a_{\ell_i}^2\right) \frac{s_i^2}{n_i}}{\sum_{i=1}^{k} \left(\sum_{\ell=1}^{k-1} a_{\ell_i}^2\right) \frac{s_i^2}{n_i}}$$

- o -

BIBLIOGRAFIA

[1] - Bartlett,M.S., "Properties of sufficiency and Statistical tests", Proceedings of the *Royal Society of London*, A,160, 1937, pp. 268-282.

[2] - Bartlett,M.S., "The use of transformations", *Biometrics*, 3, 1947, pp. 39-53.

[3] - Box, G.E.P., "Non Normality and tests on variances", *Biometrika*, 40, 1953, pp.318-335.

[4] - Box, G.E.P. & Andersen,S.L., "Permutation Theory in the Derivation of Robust Criteria and the study of Departures from Assumption", *Journal of the Royal Statistical Society*, B, 17, 1955, pp. 1-34

[5] - Brown,M.B.& Forsythe,A.B., "The Small sample behavior of some Statistics which test the equality of several means", *Tecnometrics*, 16, 1974, pp. 129-132.

[6] - Brown,M.B.& Forsythe,A.B., "The ANOVA and Multiple Comparisons for data with Heterogeneous Variances", *Biometrics*, 30, 1974, pp. 719-724.

[7] - Brownlee,K.A., *Statistical Theory and Methodology in science and Engineering*, (2a. Edition), John Wiley & Sons, 1960.

[8] - Cochran, W.C., "Some Consequences when the assumptions of the Analysis of Variance, *Biometrics*, 3, 1947, pp. 22-38.

[9] - Cramér,H., *Mathematical Methods of Statistics*, Princeton University Press, 1946.

[10] - Dixon,W.J. & Massey,F.J.Jr., *Introduction to Statistical Analysis*, (3a. Edition), International Student Edition, 1969.

[11] - Eisenhat, C. "The Assumptions Underlying the Analysis of Variance", *Biometrics*, 3, 1947, pp. 1-21.

[12] - Fisher,R.A., *Comments on the Notes by Neyman, Bartleet and Welch in this Journal*, Journal of the Royal Statistics, Princeton University Press, 1946.

[13] - Guenther William G., *Concepts of Statistical Inference*, McGrow-Hill Book Company, 1965.

[14] - Kendall,M.G.,*The Advanced Theory of Statistics*, vol.II, (3a.Edition) Charles Griffin & Company Limited, 1951.

[15] - Lehaman,E.L. & Stein,C., "On the Theory of Some Non-Parametric Hypothesis, *Ann.Math. Statist.*, 20, 1949.pp.28-45.

[16] - Pitman,E.J.G., "Significance Tests which may be applied to samples from any Populations, III, The Analysis of Variance Testes, *Biometrika*, 29, 1937.pp.322-335.

[17] - Satterhawait,F.E., "Synthesis of Variance, *Psicometrika*, 6, 1941. pp. 309-316.

[18] - Scheffé,H., *The Analysis of Variance*, John Wiley & Sons, Inc.,N. York, 1959.

[19] - Welch,B.L., "The significance of the difference between two means when teh population variances are unegual", *Biometrika*, 29, 1937, pp.350-362

[20] - Welch,B.L., "On Testes for Homogeneity", *Biometrika*, 30,1938,pp. pp. 149-158.

[21] - Welch,B.L., "The generalization of "Student's Problem when several Diferent Population Variances are Involved, *Biometrika*, 34, 1947, pp.28-35.