Estatística II: IV – Ensaio de hipóteses não paramétrico
4.1- Teste de ajustamento ou de aderência
4.2- Breve referência ao teste de ajustamento Kolmogorov-Smirnov
4.3- Teste de independência
4.4- Teste Mann-Whitney
4.5- Teste Kruskall-Wallis

IV- Testes não-paramétricos

Um método estatístico diz-se não-paramétrico se não se cumprirem os requisitos enunciados no capítulo anterior. Se os dados estiverem expressos numa escala nominal, ou numa escala ordinal, ou, ainda numa escala de intervalos ou rácios mas a função de distribuição da v. a. não esteja especificada, deve-se proceder a um ensaio de hipóteses não paramétrico.

Por exemplo, sabendo-se que a altura média de um português era, há 20 anos, de 1,60 m, pretende-se saber se a estatura média se alterou ou não. Está-se perante um teste paramétrico do tipo

$$H_o : \mu = 1,60$$
$$H_a : \mu \neq 1,60$$

A v. a é quantitativa contínua, mas se a amostra for de pequena dimensão, é necessário testar se a amostra é proveniente de uma população normal.

4.1- Teste de ajustamento ou de aderência
4.2- Breve referência ao teste de ajustamento Kolmogorov-Smirnov
4.3- Teste de independência
4.4- Teste Mann-Whitney
4.5- Teste Kruskall-Wallis

4.1- Teste de ajustamento ou de aderência

Para responder a essa questão ou a qualquer outra do mesmo tipo é necessário proceder-se a um teste de ajustamento (ou Bondade do Ajustamento).

Generalizando, pretende-se saber se a amostra pode ser considerada como proveniente de uma população com uma determinada distribuição.

H_o : A função (densidade) de probabilidade de X é $f_o(x)$
H_a : A função (densidade) de probabilidade de X não é $f_o(x)$

A estatística do teste é:

$$T = \sum_{i=1}^{m} \frac{(f_{o,i} - f_{e,i})^2}{f_{e,i}} \overset{\circ}{\cap} \chi^2_{(m-k-1)}$$

onde:- $f_{o,i}$ é a frequência observada na amostra.

- $f_{e,i}$ é a frequência esperada caso H_o verdadeira.
- m é o número de valores (caso população discreta) ou número de classes (caso população contínua) depois de corrigidas.
- k é o número de parâmetros a estimar.

Se H_o for verdadeira, espera-se que sejam pequenas as diferenças entre as frequências observadas e as esperadas. Logo, valores baixos de T são favoráveis a H_o. Pelo contrário, valores altos de T correspondem a diferenças grandes entre as frequências observada e esperada e tendem a rejeitar H_o.

$$\Downarrow$$

RCUD

4.2- **Teste de ajustamento K-S** (breve referência)

Enquanto o teste de ajustamento do Qui-Quadrado é adequado a dados nominais, o teste Kolmogorov-Smirnov é adequado a dados ordinais.

O teste K-S de ajustamento para uma amostra permite atender à ordem inerente dos dados. Veja-se o seguinte exemplo:

"O peso dos pacotes de esparguete é normalmente distribuído, com média 500 gr e desvio-padrão 5,1 gr. Perante a amostra obtida, será que se pode afirmar que estas normas estão a ser cumpridas?"

Retirou-se uma amostra aleatória de 10 pacotes, por forma a verificar o processo de embalagem. Obteve-se a seguinte amostra:

$$(507;\ 490;\ 497;\ 489{,}5;\ 499;\ 501{,}5;\ 502{,}5;\ 498{,}5;\ 510;\ 510{,}5)$$

$$H_o : X \cap N(500;5{,}1)$$
$$H_a : X \cap N(500;5{,}1)$$

A estatística do teste será

$$d_n = \underbrace{\max}_{i=1,\cdots,n} \left| F(x_{i;n}) - S_n(x_{i-1;n}) \right|$$

onde $S_n(x)$ é a função de distribuição empírica e $F_n(x)$ é a função de distribuição da amostra.

Primeiro, ordenem-se os dados e depois construa-se a seguinte tabela:

	$S_n(x)$
$x < 489{,}5$	0
$489{,}5 \leq x < 490$	0,1
$490 \leq x < 497$	0,2
$497 \leq x < 498{,}5$	0,3
$498{,}5 \leq x < 499$	0,4
$499 \leq x < 501{,}5$	0,5
$501{,}5 \leq x < 502{,}5$	0,6
$502{,}5 \leq x < 507$	0,7
$507 \leq x < 510$	0,8
$510 \leq x < 510{,}5$	0,9
$x \geq 510{,}5$	1

x_k	z_k	$F(x_k)$	$S_n(x_{k-1})$	$F(x_k) - S_n(x_{k-1})$
489,5	-2,06	0,0197	0,0	0,0197
490,0	-1,96	0,0250	0,1	-0,075
497,0	-0,58	0,2810	0,2	0,081
498,5	-0,29	0,3859	0,3	0,0859
499,0	-0,2	0,4207	0,4	0,0207
501,5	0,29	0,6141	0,5	0,1141
502,5	0,49	0,6879	0,6	0,0879
507,0	1,37	0,9147	0,7	0,2147 max
510,0	1,96	0,9750	0,8	0,175
510,5	2,06	0,9803	1,0	0,0803

$$\boxed{\frac{490-500}{5.1} = -1{,}96}$$

$$\boxed{P[Z \leq -1{,}96] = 0{,}025}$$

$d_{10} = 0{,}2147$.

O valor tabelado na tabela "Quantis da estatística de Kolmogorov-Smirnov para uma amostra" de dimensão $n = 10$ $(\alpha = 0{,}05)$ é 0,409. Como 0,2147<0,409, não se rejeita H_o, ou seja, não há evidência estatística de que a máquina não esteja a funcionar de acordo com o especificado. Logo, pode-se concluir que a distribuição do peso dos pacotes segue distribuição normal com média 500 gr e desvio-padrão de 5,1 gr.

Note-se que se a distribuição normal não estiver completamente especificada no que diz respeito aos seus parâmetros, torna-se necessário recorrer à estimação desses parâmetros, o que torna o teste conservativo já que tende a não rejeitar H_o quando ela é falsa. Por isso Lilliefors apresentou em 1967 tabelas modificadas para o caso do ajustamento à normal sem parâmetros especificados, tendo por base a mesma estatística de teste.

Conclusão

Em resumo, o teste K-S de ajustamento envolve uma série de passos conducentes à tomada de decisão. Eles são, respectivamente:

1- Ordenem-se as observações (ordem crescente) e identifique-se a função de distribuição empírica.

2- Estandardizem-se os limites superiores de cada classe.

3- Calculem-se as probabilidades acumuladas até àqueles limites já estandardizados e obtenha-se a função de distribuição amostral.

4- Calculem-se as diferenças entre as funções de distribuição amostral e empírica.

5- Seleccione-se o maior valor em termos absolutos para aquelas diferenças.

6- Tome-se a decisão comparando aquele valor máximo com o valor tabelado na tabela "Quantis da estatística de Kolmogorov-Smirnov" para essa dimensão amostral e para um determinado nível de significância atendendo que a RC é unilateral à direita.

4.3- Teste de independência

Questões como "homens e mulheres têm preferências diferentes, no que respeita a programas de TV" e "independentemente da idade, o português gosta de fado" podem ser testadas com recurso a este tipo de teste onde se pretende averiguar se as características em estudo podem ser independentes ou não.

H_o : as características são independentes (não estão relacionadas)
H_a : as características não são independentes (estão relacionadas)

Como se sabe, em caso de independência a probabilidade conjunta é igual ao produto das marginais. Ou seja,

$$H_o : p_{ij} = p_i * p_j$$
$$H_a : p_{ij} \neq p_i * p_j \text{ para pelo menos um } p_{ij}$$

A estatística do teste a utilizar é:

$$T = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(f_{oi,j} - f_{ei,j})^2}{f_{ei,j}} \cap \chi^2_{(r-1)(k-1)}$$

onde:- $f_{oi,j}$ é a frequência observada na amostra.

- $f_{ei,j}$ é a frequência esperada caso H_o verdadeira.
- r e k são o número de modalidades diferentes em que cada característica é classificada.

$\Downarrow$

RCUD: se as diferenças entre as frequências observadas e as esperadas forem grandes e, portanto, valores altos de T, então é porque H_o deve ser falsa; pelo contrário, valores baixos de T, que resultam de pequenas diferenças entre as frequências, indicam que H_o deve ser verdadeira.

4.4- Teste à igualdade de distribuições em duas amostras independentes: Teste Mann-Whitney

É a alternativa não-paramétrica ao teste *t-student* para a diferença de médias. Assenta nos postos dos valores observados da variável em estudo. O posto de uma observação é o número de ordem que lhe corresponde no conjunto das duas amostras. Caso não haja empates, a observação de valor mais baixo recebe o posto 1, a segunda mais baixa recebe o posto 2, e assim sucessivamente. Em caso de empates, i. e., observações com o mesmo valor, atribua-se a estas observações o posto médio dos que lhes corresponderiam caso tais empates não existissem.

As hipóteses a testar são:

H_o : as duas amostras são provenientes de populações com a mesma distribuição

H_a : as duas amostras não são provenientes de populações com a mesma distribuição

Considerem-se duas amostras independentes e suponha-se que $n_1 < n_2$. O modo como o teste é construído torna-o particularmente sensível às diferenças nas medianas das distribuições. Se θ_1 for a mediana da primeira população e θ_2 a mediana da segunda população, as hipóteses podem ser definidas do seguinte modo:

$$\underbrace{\begin{array}{l} H_o : \theta_1 = \theta_2 \\ H_a : \theta_1 \neq \theta_2 \end{array}}_{\text{RCBILATERAL}} \qquad \underbrace{\begin{array}{l} H_o : \theta_1 \leq \theta_2 \\ H_a : \theta_1 > \theta_2 \end{array}}_{\text{RCUD}} \qquad \underbrace{\begin{array}{l} H_o : \theta_1 \geq \theta_2 \\ H_a : \theta_1 < \theta_2 \end{array}}_{\text{RCUE}}$$

Se os valores da primeira população estiverem tendencialmente abaixo dos da segunda, deve-se proceder a um teste unilateral à esquerda; se, pelo contrário, os valores da primeira população estiverem tendencialmente acima dos da segunda, deve-se proceder a um teste unilateral à direita.

Suponha-se que se está perante um teste bilateral. A estatística do teste é igual à soma dos postos atribuídos aos valores da primeira amostra.

$$R_1 = \sum_{i=1}^{n_1} R_1(i)$$

A função de probabilidade de R_1 é simétrica em relação a $\frac{n_1(n+1)}{2}$. Como o teste é bilateral, se $\alpha = 0,05$ e se $n_1, n_2 \leq 20$, o quantil de probabilidade $\alpha/2$ para estes valores de n_1 e n_2 pode ser consultado na tabela "Quantis da estatística de Mann-Whitney".

A distribuição normal pode ser utilizada como aproximação se as dimensões dos grupos forem grandes. Se os grupos não forem "especialmente" grandes, é conveniente fazer-se a correcção de continuidade.

4.5- Teste à igualdade de distribuições em mais de duas amostras independentes: Teste Kruskall-Wallis

É a alternativa não-paramétrica à análise da variância simples. Pretende-se verificar se as k amostras aleatórias independentes podem ser provenientes de populações com a mesma distribuição. Assim, as hipóteses são:

H_o : as k populações partilham a mesma distribuição
H_a : pelo menos uma das populações tem distribuição diferente das restantes

Tal como no teste anterior, também este é particularmente sensível às diferenças entre as medianas das k populações já que assenta na atribuição de postos das observações. Assim, as hipóteses podem ser redefinidas como

$$H_o : \theta_1 = \theta_2 = \cdots = \theta_k$$
$$H_a : \ \exists\ i, j: \quad \theta_i \neq \theta_j$$

Se: - $R(X_{ij})$ for o posto atribuído a X_{ij}

- $R_i = \sum_{j=1}^{n_i} R(X_{ij})$ for a soma dos postos das observações da *i-ésima* amostra

- $n = \sum_{i=1}^{k} n_i$

A estatística do teste é dada por:

$$T = \frac{1}{s^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{n(n+1)^2}{4}\right)$$

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} R(X_{ij})^2 - \frac{n(n+1)^2}{4}\right)$$

Caso não existam empates, a estatística reduz-se a

$$\boxed{T' = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1)}$$

- ♦ Se existirem apenas três grupos, com dimensões inferiores ou iguais a 5, e não haja empates entre os grupos, os valores desta estatística encontram-se na tabela "Quantis da estatística de Kruskall-Wallis para pequenas amostras".
- ♦ Nas outras situações, utiliza-se como distribuição aproximada a Qui-Quadrado com $(k-1)$.