IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

Trabalho para a Unidade Curricular Métodos Estatísticos
Licenciatura em Engenharia Informática

2° ano, 2° semestre, ano letivo 2021-2022

Testes de Ajustamento e Independência no Mercado de Venda de Videojogos

Diogo Letras - Turma SW 03, Nº 202002529

Miguel Vicente - Turma SW 03, Nº 202000563

Pedro Cunha - Turma SW 03, Nº 202000757

Resumo: O trabalho consiste na realização de alguns testes de ajustamento e independência utilizando dados do mercado de venda de videojogos a nível mundial nas suas mais variadas vertentes. Mais especificamente, foram feitos teste de ajustamento do Qui-Quadrado com uma variável, o teste de ajustamento de Kolmogorov-Smirnov com os resíduos calculados no trabalho prático 2 e o teste de independência do Qui-Quadrado para testar se existe associação entre 2 variáveis selecionadas.

Palavras-chave: Estatística, Videojogos, Testes, Ajustamento, Variáveis.

1. Introdução

Conforme requisitado pelo enunciado do 3º Trabalho de Grupo, foram feitos testes de ajustamento e independência a partir de uma base de dados fornecida pela docente responsável. A base de dados em questão é composta por 16598 videojogos, que basicamente são todos os videojogos com pelo menos 100 mil cópias vendidas entre, aproximadamente, o ano de 1980 e o ano de 2017. Ou seja, resumidamente temos como população todos os videojogos vendidos e como amostra 16598 videojogos, que são títulos com pelo menos 100 mil cópias vendidas, durante um período de tempo. Os testes não paramétricos são, tal como o nome indica, testes estatísticos onde os parâmetros não são especificados.

2. Teste de ajustamento do Qui-Quadrado

O seu objetivo é testar a adequabilidade de um modelo probabilístico a um conjunto de dados observados, ou seja, comparar a distribuição dos dados amostrais (frequências observadas) com a distribuição teórica que se associa à população de onde provém essa amostra.

Para realizar esta tarefa, foi escolhida a variável 'Year' da base de dados e, recorrendo ao teste de ajustamento do Qui-Quadrado, testou-se a distribuição uniforme discreta, considerando um nível de significância de 10%.

2.1 Distribuição a ser testada

Tal como foi referido, foi escolhida a distribuição uniforme discreta para ser testada. Neste contexto, isto significa que se pretende verificar se a variável 'Year' segue esta distribuição, ou seja, que existe um número igual de jogos lançados em cada ano.

2.2 Hipóteses testadas

Tem-se que:

H0: X segue uma distribuição uniforme discreta.

contra

H1: X não segue uma distribuição uniforme discreta.

2.3 Resultados através do valor-p

Sabendo que o nível de significância é de 10%, o valor de α é 0.10.

Calculou-se o valor-p, tendo-se obtido o seguinte valor(bruto) da Figura.1:

```
p-value < 0.00000000000000022
```

Figura.1 – Valor -p (TAQQ).

Conclui-se, utilizando a seguinte lógica:

se valor-p > α -> Não se rejeita H0

ou

se valor-p <= α-> Rejeita-se H0

Como o valor-p é claramente menor que α, rejeita-se H0.

2.4 Resultados através da região crítica

Delimitou-se a região de aceitação (RA) e a região crítica (RC), obtendo-se os intervalos, representados na Fig.2.

```
#RA=[0,48.36[ e RC=[48.36,+infinito[
```

Figura.2 – Região de aceitação e região crítica (TAQQ).

Calculou-se o valor observado, Q_{obs} , representado na Fig.3:

```
X-squared
 17632.48
```

Figura.3 – Q_{obs} (TAQQ).

Utilizando a seguinte lógica:

Q_{obs} pertence à RA -> Não se rejeita H0.

ou

Q_{obs} pertence à RC -> Rejeita-se H0.

Como Q_{obs} está fora da região de aceitação, ou dito de outra forma, Q_{obs} está dentro da região crítica, rejeita-se H0.

2.5 Conclusão

Conclui-se que a hipótese “H1: X não segue uma distribuição uniforme discreta.” é verdadeira.

Os resultados do teste podem parecer à primeira vista, ser um pouco absurdos, tendo em conta que Q_{obs} está claramente na região crítica, e o valor-p é bastante menor que α.

No entanto, estes resultados eram previsíveis, tendo em conta que olhando para a evolução da indústria de videojogos, é de conhecimento geral que nos anos atuais há muito mais jogos a serem lançados do que há 30 ou 40 anos atrás. Portanto, a variável não tem uniformidade.

3. Teste de ajustamento de Kolmogorov-Smirnov

O seu objetivo é testar a adequabilidade de um modelo probabilístico a um conjunto de dados observados, ou seja, comparar a função de distribuição teórica (referente à população) com a função de distribuição amostral (referente à amostra).

2.1 Transformação dos resíduos

Para realizar esta tarefa, testou-se os resíduos obtidos no trabalho 2, de modo a perceber se os mesmos seguem um modelo normal. Para tal considerou-se os resíduos padronizados, a distribuição normal padrão e um nível de significância de 5%.

Nota: Os resíduos padronizados (e'_i) foram obtidos a partir da seguinte transformação:

$$e'_i = \frac{e_i}{s_e} \quad (1)$$

onde e_i são os resíduos obtidos na regressão linear simples (trabalho 2) e s_e é o desvio padrão dos resíduos.

2.2 Hipóteses testadas

Tem-se que:

H0: X segue uma distribuição normal com média 0 e desvio padrão 1.

contra

H1: X não segue uma distribuição normal com média 0 e desvio padrão 1.

2.3 Resultados através do valor-p

Sabendo que o nível de significância é de 5%, o valor de α é 0.05.

Calculou-se o valor-p, tendo-se obtido o seguinte valor(bruto) da Figura.4:

```
p-value < 0.00000000000000022
```

Figura.4 – Valor -p (TAKS).

Conclui-se, utilizando a seguinte lógica:

se valor-p > α -> Não se rejeita H0

ou

se valor-p <= α-> Rejeita-se H0

Como o valor-p é claramente menor que α, rejeita-se H0.

2.4 Resultados através da região crítica

Delimitou-se a região de aceitação (RA) e a região crítica (RC), obtendo-se os intervalos, representados na Fig.5.

```
RA=[0,0.023[ e RC=[0.023,+infinito[
```

Figura.5 – Região de aceitação e região crítica (TAKS).

Observação: Para calcular o valor em que a região crítica (RC) começa (0.023), teve-se que recorrer à "Tabela dos Valores Críticos da Distribuição da Estatística (Kolmogorov-Smirnov)", disponibilizada pelos docentes.

Neste caso, como α = 0.05 e o n é 3475, ou seja, n>40, utilizou-se a fórmula assinalada a vermelho na figura.6.

Para n>40 os valores críticos de D_n podem ser aproximados pelas seguintes expressões:

α				
0.20	**0.10**	**0.05**	**0.02**	**0.01**
$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Figura.6 – Cálculo do valor em que a região crítica começa (TAKS).

Calculou-se o valor observado, Q_{obs} , representado na Fig.7:

```
          D
0.09282074
```

Figura.7 – Q_{obs} (TAKS).

Utilizando a seguinte lógica:

Q_{obs} pertence à RA -> Não se rejeita H0.

ou

Q_{obs} pertence à RC -> Rejeita-se H0.

Como Q_{obs} está fora da região de aceitação, ou dito de outra forma, Q_{obs} está dentro da região crítica, rejeita-se H0.

2.5 Conclusão

Conclui-se que os resíduos padronizados não parecem seguir uma distribuição normal padrão, no entanto é importante mencionar que devido às características técnicas do software R, tais como arredondamento ou o cálculo impreciso dos graus de liberdade, não se pode excluir totalmente que esta conclusão não seja verdadeira.

4. Teste de independência do Qui-Quadrado

O seu objetivo é estudar a relação entre duas variáveis qualitativas, tentando perceber se existe associação entre elas.

Para realizar esta tarefa, escolheu-se duas variáveis da base de dados e, recorrendo ao teste de independência do Qui-Quadrado, testou-se se existe associação entre as variáveis selecionadas, considerando um nível de significância de 1%.

2.1 Construção da tabela de contingência

Escolheu-se 2 variáveis qualitativas, 'Genre' (tipo de jogo) e 'Publisher' (editora do jogo), para tentar perceber se existe uma associação entre as duas variáveis. Fez-se então a tabela de contingência, representada na tabela.1, para descrever a frequência dos níveis de uma das variáveis relativamente aos níveis da outra variável.

Tabela.1 – Tabela de contingência (TIQQ).

	Publisher		
Genre	Activision	Electronic Arts	Nintendo
Action	310	183	79
Adventure	25	13	35
Fighting	7	39	18
Misc	103	46	100
Platform	60	16	112
Puzzle	7	7	74
Racing	74	159	37
Role-Playing	41	35	106
Shooter	159	139	26
Simulation	23	116	29
Sports	144	561	55
Strategy	22	37	32

2.2 Hipóteses testadas

Tem-se que:

H0: o tipo de jogo é independente da editora do jogo.

contra

H1: o tipo de jogo não é independente da editora do jogo.

2.3 Resultados através do valor-p

Sabendo que o nível de significância é de 1%, o valor de α é 0.01.

Calculou-se o valor-p, tendo-se obtido o seguinte valor(bruto) da Figura.8:

```
[1] 0.000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
0001015245
```

Figura.8 – Valor -p (TIQQ).

Conclui-se, utilizando a seguinte lógica:

se valor-p > α -> Não se rejeita H0

ou

se valor-p <= α-> Rejeita-se H0

Como o valor-p é claramente menor que α, rejeita-se H0.

2.4 Resultados através da região crítica

Delimitou-se a região de aceitação (RA) e a região crítica (RC), obtendo-se os intervalos, representados na Fig.9.

```
RA=[0,40.28[ e RC=[40.28,+infinito[
```

Figura.9 – Região de aceitação e região crítica (TIQQ).

Calculou-se o valor observado, Q_{obs} , representado na Fig.10:

```
X-squared
 1128.057
```

Figura.10 – Q_{obs} (TIQQ).

Utilizando a seguinte lógica:

Q_{obs} pertence à RA -> Não se rejeita H0.

ou

Q_{obs} pertence à RC -> Rejeita-se H0.

Como Q_{obs} está fora da região de aceitação, ou dito de outra forma, Q_{obs} está dentro da região crítica, rejeita-se H0.

2.5 Conclusão

Conclui-se que a hipótese "H1: o tipo de jogo não é independente da editora do jogo" é verdadeira.

À semelhança do que já aconteceu anteriormente, os resultados do teste podem parecer à primeira vista, ser um pouco absurdos, tendo em conta que Q_{obs} está claramente na região crítica, e o valor-p é bastante menor que α.

No entanto, estes resultados eram previsíveis, devido ao facto de na indústria dos videojogos, normalmente, as editoras tendem a especializar-se num determinado tipo de jogo, criando assim uma associação significativa entre tipo de jogo e editora.

5. Conclusões

Com este trabalho, foi possível realizar 3 testes de ajustamento distintos aos nossos dados.

Através do seu desenvolvimento, aprofundou-se e consolidou-se os conhecimentos sobre o tópico.

Portanto, este trabalho permitiu ao grupo adquirir um leque de competências importante na área da Estatística, mais especificamente na realização de testes de ajustamento. Sem dúvidas, que estes conhecimentos serão úteis no futuro, caso haja interesse de algum membro do grupo em estudar ou trabalhar nessa área.

Referências Bibliográficas

- Fichas 6, 7 e 8, das aulas práticas.
- Slides disponibilizados pelos professores no moodle, ”Capítulo 7 – Testes de Hipóteses Não Paramétricos”.