Tabelas de Contingência

Hildete Prisco Pinheiro, Rafael Alves e Eduardo Vargas

2013

Introdução

Introdução

Muitas vezes, a informação da amostra coletada tem a estrutura de dados categorizados, ou seja, o conjunto de dados consiste em frequências de contagens para essas categorias. O que ocorre com frequência nas áreas sociais e biomédicas. O objetivo aqui é estudar dados agrupados em categorias múltiplas.

Exemplo 1

Exemplo

Uma determinada marca de geladeira é vendida em cinco cores diferentes e uma pesquisa de mercado quer avaliar a popularidade das várias cores. As frequências abaixo são observadas para uma amostra de 300 vendas feitas num semestre. Suponha que seja de interesse testar a hipótese das cinco cores serem igualmente populares.

Vendas das cinco cores das geladeiras da marca W

marrom	*creme*	*vermelho*	*azul*	*branco*	*total*
88	65	52	40	55	300

O Modelo Multinomial

Para acomodar dados como no Exemplo 1, precisamos estender o modelo Bernoulli de forma que os resultados possam ser classificados em mais de duas categorias. Esse modelo é chamado de distribuição multinomial.

Distribuição Multinomial

a) O resultado de cada amostra pode ser classificado em uma de k respostas denotadoas por 1, 2, ..., k.

b) A probabilidade da amostra ser i é p_i, $i = 1, 2, ..., k$, com $\sum_{i=1}^{k} p_i = 1$.

c) As observações são independentes.

O Modelo Multinomial

Quando a amostragem é de uma população que consiste de elementos em diversas categorias, teremos k valores possíveis, denotaremos por $n_1, n_2, \ldots, n_k$, com $\sum_{i=1}^{k} n_i = n$ suas frequências e $p_1, p_2, \ldots, p_k$ suas probabilidades. A distribuição conjunta de $n_1, n_2, \ldots, n_k$ é chamada de distribuição multinomial e tem função densidade de probabilidade dada por:

$$f(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

em que $\sum_{i=1}^{k} n_i = n$ e com $\sum_{i=1}^{k} p_i = 1$.

O Modelo Multinomial

Se designarmos a componente n_1 como "sucesso" e juntarmos as demais numa mesma que designamos "fracasso", a variável aleatória n_1 é o número de sucessos em n observações de bernoulli cuja distribuição é $B(n, p_1)$. Portanto, $E(n_1) = np_1$, $Var(n_1) = np_1(1 - p_1))$, analogamente aplicando o mesmo argumento a cada n_i temos: $E(n_i) = np_i$ e $Var(n_i) = np_i(1 - p_i))$. Além disso como a soma de n_i é fixa $Cov(n_i, n_j) = -np_ip_j$, para $i \neq j$.

O Teste χ^2 de Pearson de Aderência

O Teste χ^2 de Pearson de Aderência

Voltando aos dados do Exemplo 1, cujas componentes tem frequências multinomiais, a hipótese nula especifica uma estrutura de probabilidades das componentes. O primeiro passo é testar se o modelo dado na hipótese nula se ajusta aos dados, esse procedimento é o chamado teste de bondade de ajuste ou teste de aderência.

Caso A: As probabilidades das componentes são especificadas completamente por H_0

Caso A: As probabilidades das componentes são especificadas completamente por H_0.

$$H_0 : p_1 = p_{10}, \ldots, p_k = p_{k0}$$

onde $p_{10}, p_{20}, \ldots, p_{k0}$ são valores tais que $p_{10} + p_{20} + \cdots + p_{k0} = 1$.

Caso A: As probabilidades das componentes são especificadas completamente por H_0

No Exemplo 1, a hipótese nula de que as cinco cores são iqualmente populares pode ser escrita como $H_0 : p_1 = p_2 = \cdots = p_k = \frac{1}{5}$. Uma vez especificadas as propabilidades das componentes, as frequências esperadas podem ser calculadas multiplicando essas probabilidades pelo tamanho da amostra n.

O Teste χ^2 de Pearson de Aderência

Um teste de aderência tenta determinar se existe uma discrepância entre as frequências observadas e as esperadas, sob a hipótese nula. Uma medida útil para tal discrepância é:

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

onde O_i e E_i simbolizam a frequência observada e a respectiva frequência esperada e o valor de χ^2 pode ser comparado com um quantil α da distribuição χ^2 com $k-1$ graus de liberdade.

O Teste de χ^2 de Pearson

Procedimento

- $H_0 : p_1 = p_{10}, \ldots, p_k = p_{k0}$
- $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i0})^2}{np_{i0}} = \sum \frac{(O-E)^2}{E}$
- Rejeita-se H_0 quando $\chi^2 \geq \chi^2_{\alpha;k-1}$, para um nível α de significância.

Já que o teste é aproximado como regra geral é adequado usarmos para n suficientemente grande, de modo que cada componente tenha frequência esperada de pelo menos 5.

Exemplo

Exemplo

Usando os dados do Exemplo 1, teste a hipótese que as cinco cores das geladeiras são iqualmente populares. Use $\alpha = 0,05$.

Teste de χ^2 de Pearson para o Exemplo 1

Componente	*marrom*	*creme*	*vermelho*	*azul*	*branco*	*total*
Frequencia Observada	88	65	52	40	55	300
Prob. sob H_0	0,2	0,2	0,2	0,2	0,2	1,0
Frequencia Esperada	60	60	60	60	60	300
$\frac{(O-E)^2}{E}$	13,07	0,42	1,07	6,67	0,42	21,63

Exemplo

Como $\chi^2_{0,05;4} = 9,487 \leq 21,63 = \chi^2$, a hipótese nula é rejeitada. Concluimos que há discrepância significativa entre o que foi observado e a hipótese nula.

Caso B: As probabilidades das componentes não são especificadas completamente por H_0

Caso B: As probabilidades das componentes não são especificadas completamente por H_0

Algumas vezes é necessário o conhecimento de alguns parâmetros para testar um modelo específico. Por exemplo, se quizéssemos testar se os dados são provenientes de uma distribuição normal precisaríamos saber μ e σ^2.

Caso B: As probabilidades das componentes não são especificadas completamente por H_0

O procedimento nesse caso é estimar os parâmetros desconhecidos a partir dos dados observados e então usar os valores estimados como se fossem os valores dos parâmetros, para determinar as probabilidades das componentes. As frequências esperadas e a estatística de teste são calculadas como antes, mas o número de graus de liberdade é reduzido de acordo com os parâmetros estimados, agora teremos número de componentes-1-número de parâmetros estimados.

Exemplo

Exemplo

Um estatístico amostra 200 famílias e registra a distribuição de frequência do número de vezes que essas famílias usaram o seguro de saúde num período de quatro anos.

Distribuição de frequência do número de vezes que o seguro foi usado

No. de utilização do seg.	0	1	2	3	4	5	6	7	total
Frequência	22	53	58	39	20	5	2	1	200

Teste se o modelo de Poisson descreve adequadamente esses dados.

Exemplo

Podemos estimar λ, a média da distribuição de Poisson, usando a média amostral

$$\bar{x} = \frac{\sum(valor \times frequencia)}{n} = \frac{410}{200} = 2,05 \sim 2,0$$

Consultando a tabela para o valor estimado 2 para λ e multiplicando por 200 obtemos a frequência esperada dos dados.

Exemplo

Teste χ^2 de Pearson para os dados de seguro de saúde

Utilizacao do seguro	0	1	2	3	4	5	6	7	*total*
Frequencia Observada	22	53	58	39	20	5	2	1	200
Prob. Poisson$(\lambda = 2)$	0, 135	0, 271	0, 271	0, 180	0, 090	0, 036	0, 012	0, 05	1
Frequencia Esperada	27	54, 2	54, 2	36	18, 2	7, 2	2, 4	1	200
$\frac{(O-E)^2}{E}$	0, 926	0, 027	0, 266	0, 250	0, 222	0, 672	0, 067	0	2, 33

Exemplo

Para $\alpha = 0,05$ o valor de $\chi^2_{0,05;8-1-1} = \chi^2_{0,05;6} = 12,592$, que é maior do que o observado $\chi^2 = 2,33$ e portanto a hipótese nula não é rejeitada, isto é, o modelo de Poisson não contradiz os dados.
Nesse tipo de teste é importante verificar que concordância com a hipótese nula não significa que o modelo proposto está correto mas sim que ele é um dos modelos de acordo com os dados.

Tabelas de Contingência

Quando dois ou mais atributos são observados para cada elemento amostrado, os dados podem ser simultaneamente classificados com respeito aos níveis de ocorrência para cada um dos atributos. Por exemplo, empregados podem ser classificados de acordo com os anos de escolaridade e tipo de ocupação, flores podem ser classificadas com respeito ao tipo de folhagem e tamanho da flor. Dados de frequência aparecem da classificação simultânea de duas ou mais características e são chamados de tabelas de contingência.

Tabelas de Contingência

Na maioria das tabelas de contingência o objetivo é estudar se certa característica parece se manifestar independentemente da outra ou se níveis de uma característica tendem a estar associados com níveis da outra.

Exemplo

Exemplo

Uma amostra aleatória de 500 pessoas responde um questionário sobre filiação partidária e atitude mediante um programa de racionamento de energia. As frequências observadas são apresentadas a seguir.

Tabela de contingência para filiação partidária e opinião sobre o racionamento de energia

	favorável	indiferente	contrário	total
PT	138	83	64	285
PSDB	64	67	84	215
Total	202	150	148	500

Exemplo

Responda as sequintes perguntas:

a) Os dados indicam que a opinião sobre racionamento de energia é independente da filiação partidária?

b) Podemos medir quantitativamente a associação entre as duas características?

Exemplo

Antes de apresentar uma análise formal estatística consideramos a tabela de um ponto de vista descritivo, transformando as contagens em proporções, primeiramente por linhas, depois por componente.

a) Proporções por linhas

	favorável	indiferente	contrário	total
PT	0,48	0,30	0,22	1
PSDB	0,30	0,31	0,39	1

b) Proporções por componente

	favorável	indiferente	contrário	total
PT	0,276	0,166	0,128	0,570
PSDB	0,128	0,134	0,168	0,430
Total	0,404	0,300	0,296	1

Exemplo

Inspeção visual dessas tabelas revelam diferenças aparentes nas distribuições ao longo das linhas, colunas ou das proporções dos totais de observações em cada componente. Por exemplo, na tabela *a*) a distribuição parece variar com as linhas, daí a suspeita da presença de associação. A primeira linha diminui, enquanto a segunda aumenta. Na tabela *b*) o interesse está nas componentes com maior e menor concentração de observações. Todas as componentes na tabela *b*) tem proporções moderadas com a componente PT/favorável sendo a maior. Uma possível associação deve ser confirmada por um teste estatístico.

Teste de Independência

Para um tratamento geral de teste de independência em tabelas de contingência, considere duas características designadas por A e B e suponha que existem r categorias $A_1, A_2, \ldots, A_r$ para A e c categorias $B_1, B_2, \ldots, B_c$ para B. Suponha que uma amostra de tamanho n é classificada e distribuída nas componentes da tabela produzindo uma tabela de frequência em que:
n_{ij} = frequência de $A_i B_j$.
n_{i0} = total da i-ésima linha, ou frequência de A_i.
n_{0j} = total da j-ésima coluna, ou frequência de B_j.

Teste de Independência

Tabelas de contingência $r \times c$

	B_1	B_2	$\cdots$	B_c	total da linha
A_1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{10}
A_2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r0}
total da coluna	n_{01}	n_{02}	$\cdots$	n_{0c}	n

Teste de Independência

Podemos usar a população classificada em termos de proporções populacionais e a tabela anterior fica:

Probabilidade das componentes

	B_1	B_2	$\cdots$	B_c	total da linha
A_1	p_{11}	p_{12}	$\cdots$	p_{1c}	p_{10}
A_2	p_{21}	p_{22}	$\cdots$	p_{2c}	p_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	p_{r1}	p_{r2}	$\cdots$	p_{rc}	p_{r0}
total da coluna	p_{01}	p_{02}	$\cdots$	p_{0c}	1

Teste de Independência

em que:
$p_{ij} = P(A_i B_j)$ é a propabilidade da ocorrência conjunta de A_i e B_j.
$p_{i0} = P(A_{i0})$ é a propabilidade total da i-ésima linha.
$p_{0j} = P(B_{0j})$ é a propabilidade total da j-ésima coluna.
O interesse é testar se A e B são classificações independentes, ou seja, pretende-se observar se $P(A_i B_j) = P(A_i)P(B_j)$ para todo $i = 1, 2, \ldots, r$ e $j = 1, 2, \ldots, c$

Teste de Independência

Hipótese nula de independência:

$$H_0 : p_{ij} = p_{i0}p_{0j} \text{ para todas as componentes } (i, j).$$

O modelo de independência especifica as probabilidades das componentes em termos das probabilidades marginais que são parâmetros desconhecidos. Como $p_{i0} = P(A_i)$, um estimador natural é a frequência relativa amostra de A_i,

$$\hat{p}_{i0} = \frac{n_{i0}}{n}.$$

Da mesma forma, p_{0j} é estimado por

$$\hat{p}_{0j} = \frac{n_{0j}}{n}.$$

estimadores esses também obtidos pelo método de máxima verossimilhança e portanto podemos usar um teste χ^2.

Teste de Independência

Usando essas estimativas a probabilidade da componente (i, j) é estimada por

$$\hat{p}_{ij} = \hat{p}_{i0}\hat{p}_{0j} = \frac{n_{i0}n_{0j}}{n^2}.$$

Logo, a frequência relativa esperada sob o modelo de independência é

$$E_{ij} = n\hat{p}_{ij} = \frac{n_{i0}n_{0j}}{n}.$$

e a estatística do teste é dada por

$$\chi^2 = \sum_{rc \text{ componentes}} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}.$$

que tem distribuição χ^2 com $(r-1)(c-1)$ graus de liberdade, para n grande.

Exemplo

Exemplo

Usando os dados sobre filiação partidária, a tabela abaixo mostra as frequências observadas e as esperadas.

Tabela de contingência para filiação partidária e opinião sobre racionamento

	favorável	indiferente	contrário	total
PT	138	83	64	285
	(115,14)	(85,50)	(84,36)	
PSDB	64	67	84	215
	(86,86)	(64,50)	(63,64)	
Total	202	150	148	500

Exemplo

A estatística χ^2 tem o valor observado de
$\chi^2 = 4,539 + 0,073 + 4,914 + 6,016 + 0,097 + 6,514 = 22,153$ com $(2-1)(3-1) = 2$ g.l.
Usando o nível de significância $\alpha = 0,05$, o χ^2 tabulado é $5,991$ que é menor que o observado e daí a hipótese nula de indepêndencia é rejeitada.

Exemplo

Quando o teste leva a rejeição da hipótese nula de independência concluímos que os dados dão evidência de uma associação estatística entre as duas características. No entanto, não é o bastante para declararmos que existe relação de causa e efeito entre as características.

Tabelas de contingência com uma das margens fixas

Até aqui o esquema de amostragem utizado foi baseado numa amostra aleatória de tamanho n que é classificada com respeito a duas características simultaneamente. Nesse caso, as duas frequências marginais totais são variáveis aleatórias.

Tabelas de contingência com uma das margens fixas

Se o esquema de amostragem for de dividir a população em duas sub-populações ou estratos de acordo com as categorias de uma característica, com uma amostra de um tamanho pré-determinado para cada estrato e classificada de acordo com as categorias da outra característica então esta será uma situação de tabela de contingência com margens fixas.
Por exemplo, no caso do problema de filiação partidaria, seriam selecionadas amostras aleatórias de tamanhos 200 e 300 das populações Petistas e Psdbistas e se classificaria essas amostras de acordo com a atitude (favorável, indiferente ou contrário).

Tabelas de contingência com uma das margens fixas

O interesse então é estudar as proporções nessas categorias para determinar se elas são aproximadamente iguais para as diferentes populações. Ou seja, queremos testar se as populações são homogêneas.

Tabelas de contingência com uma das margens fixas

Suponha que amostras aleatórias independentes de tamanho $n_{10}, \ldots, n_{r0}$ são selecionadas de r populações $A_1, \ldots, A_r$ respectivamente. Classificando cada amostra em $B_1, \ldots, B_c$, obtemos uma tabela de contigência $r \times c$ onde os totais das linhas são tamanhos de amostras fixos.

Tabelas de contingência $r \times c$ com totais das linhas fixos

	B_1	B_2	$\cdots$	B_c	total da linha
A_1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{10}
A_2	n_{21}	n_{22}	$\cdots$	n_{2c}	n_{20}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r0}
total da coluna	n_{01}	n_{02}	$\cdots$	n_{0c}	n

Tabelas de contingência com uma das margens fixas

As probabilidades das várias categorias de B dentro de cada sub-população de A também são apresentadas a seguir, onde cada w representa uma probabilidade condicional,

$$w_{ij} = P(B_J|A_i) = \text{ probabilidade de } B_j \text{ dentro da população} A_i.$$

Probabilidades das categorias B dentro de cada população

	B_1	B_2	$\cdots$	B_c	total da linha
A_1	w_{11}	w_{12}	$\cdots$	w_{1c}	1
A_2	w_{21}	w_{22}	$\cdots$	w_{2c}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	w_{r1}	w_{r2}	$\cdots$	w_{rc}	1

Tabelas de contingência com uma das margens fixas

A hipótese nula de iqualdade das categorias B para as r populações é

$$H_0 : w_{1j} = w_{2j} = \cdots = w_{rj}, \text{ para todo } j = 1, 2, \ldots c.$$

Sob H_0, a probabilidade comum da categoria B_j pode ser estimada do conjunto de amostras notando que de um total de n elementos amostrados, n_{0j} possuem a característica B_j, daí a probabilidade estimada fica

$$\hat{w}_{1j} = \hat{w}_{2j} = \cdots = \hat{w}_{rj} = \frac{n_{0j}}{n}.$$

Tabelas de contingência com uma das margens fixas

A frequência esperada estimada na componente (i, j) sob H_0 é E_{ij} =(Número de A_i amostrados)$\times$(Probabilidade de B_j dentro de A_i)$= n_{i0}\hat{w}_{ij} = \frac{n_{i0}n_{0j}}{n}$.
e o teste estatístico é dado por

$$\chi^2 = \sum_{\text{componentes}} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}.$$

que tem $(r-1)(c-1)$ graus de liberdade como antes.

Tabelas de contingência com uma das margens fixas

Pode-se observar que as fórmulas e os graus de liberdade dessa seção são iquais ao da seção anterior, somente o método de amostragem e a formalização da hipótese nula são diferentes.

Exemplo

Exemplo

Foi feita uma pesquisa para determinar a incidência de alcoolismo em diferentes grupos profissionais. Amostras aleatórias de religiosos, educadores, executivos e comerciantes foram entrevistados. Os dados são apresentados na tabela:

Tabela de contingência de alcolismo vs profissão

	alcoólatras	não alcoólatras	tamanho da amostra
Religiosos	$32(58,25)$	$268(241,75)$	300
Educadores	$51(48,54)$	$199(201,46)$	250
Executivos	$67(58,25)$	$233(241,75)$	300
Comerciantes	$83(67,96)$	$267(282,04)$	350
Total	233	967	1200

Exemplo

Representando por p_1, p_2, p_3 e p_4 as proporções de alcoólatras na população de religiosos, educadores, executivos e comerciantes, queremos testar a hipótese

$$H_0 : p_1 = p_2 = p_3 = p_4.$$

o valor da estatística é

$$\chi^2 = \frac{(32 - 58,25)^2}{58,25} + \cdots + \frac{(267 - 282,04)^2}{282,04} = 20,59 \text{ com } (4-2)(2-1) = 3 \text{ gl}$$

o valor tabulado para o teste χ^2 é $7,815$ para $\alpha = 0,05$ e portanto a hipótese nula é rejeitada ao nível de $\alpha = 0,05$

Exemplo

Como a hipótese nula foi rejeitada verificamos que há indícios de que a proporção de alcoólatras nas classes profissionais não é homogênea, fato que provavelmente é devido á relativa pequena proporção dos alcoólatras na classe dos religiosos.