Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Probabilidade e Estatística 2
1^o semestre de 2020
Gustavo L. Gilardoni

Testes Chi-quadrado

ou

Testes de Aderência e de Associação

ou

Introdução à Inferência para Dados de Contagem

26 de Novembro de 2020

0 Introdução

Exemplo 1. Um dado foi lançado $n = 100$ vezes obtendo-se o resultado na Tabela 1. É razoável assumir que o dado é não-viesado, no sentido que as suas seis faces têm a mesma probabilidade?

Face	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Frequência (n_i)	21	13	22	20	13	11	100
Freq. esperada	16.67	16.67	16.67	16.67	16.67	16.67	100

Tabela 1: Resultado de $n = 100$ lançamentos de um dado (Exemplo 1).

Exemplo 2. Uma companhia mineradora usa um número grande de correias transportadoras. Falhas nestes equipamentos impactam tanto o aspecto econômico quanto o da segurança da operação. A Tabela 2 descreve o número de falhas para $n = 50$ motores de correias transportadoras no período de um ano. Com base nesses dados, sería razoável assumir que a distribuição do número de falhas segue uma distribuição de Poisson?

# de Falhas	**0**	**1**	**2**	**3**	**Total**
# de Motores	24	10	12	4	50
probs. est. $(\hat{p}_i)$	0.399	0.367	0.169	0.066	1
Freq. esperada	19.926	18.332	8.433	3.309	50

Tabela 2: Número de falhas em um ano para 50 motores de correias transportadoras usadas por uma companhia mineradora. (Exemplo 2).

Exemplo 3. A Tabela 3 mostra as notas obtidas num curso de *Cálculo* por duas amostras: (i) a primeira com $n_1 = 100$ alunos e (ii) a segunda com $n_2 = 50$ alunas. Existe evidência no sentido que, na população de referência, a distribuição das notas é diferente para os alunos do que para as alunas?

Nota	II	MI	MM	MS	SS	Total
Homens	13	15	37	22	13	100
Mulheres	3	4	15	15	13	50
Total	16	19	52	37	26	150

Tabela 3: Distribuição das notas para uma amostra de $n_1 = 100$ alunos e outra de $n_2 = 50$ alunas numa disciplina de Cálculo (Exemplo 3).

Exemplo 4. A Tabela 4 mostra a distribuição das notas para uma amostra de $n = 200$ alunos de uma turma de Estatística Básica, classificada de acordo ao sexo do aluno. Existe evidência no sentido que a nota depende do sexo do aluno (ou vice-versa)?

Nota	II	MI	MM	MS	SS	Total
Homens	10	14	36	31	21	112
Mulheres	11	16	25	23	13	88
Total	21	30	61	54	34	200

Tabela 4: Distribuição das notas para uma amostra de $n = 200$ alunos classificados por sexo numa disciplina de Estatística Básica (Exemplo 4).

Cada um desses exemplos corresponde à uma técnica que será estudada nesta unidade:

- Exemplo 1: Teste de Aderência (*Goodness-of-Fit* no inglês) com probabilidades especificadas sob H_0;
- Exemplo 2: Teste de Aderência com probabilidades desconhecidas sob H_0;
- Exemplo 3: Teste de homogeneidade;
- Exemplo 4: Teste de independência.

1 A distribuição multinomial

A distribuição Binomial aparece relacionada com amostragem em populações classificadas em duas classes, usualmente denominadas de "sucesso" e "fracasso". Formalmente, consideramos n experimentos dicotômicos, também chamados de Ensaios de Bernoulli

- Os n ensaios são *independentes*, no sentido que o resultado (sucesso ou fracasso) de um deles não terá nenhum efeito na probabilidade de observarmos sucesso ou fracasso em qualquer um dos outros.
- Os n ensaios são *idênticos*, no sentido que todos eles tem a mesma probabilidade de sucesso p e de fracasso $q = (1 - p)$.

- No caso de populações finitas, essas duas propriedades aparecem quando fazemos amostragem com reposição.
- Denotando por X a variável aleatória que representa o número de sucessos nos n ensaios, segue que

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x} \tag{1}$$

para $x = 0, 1, \ldots, n$.

A **Distribuição Multinomial** é uma generalização da Binomial quando os ensaios (ou a população sob estudo) são **Politômicos** ao invés de dicotômicos. Um ensaio politômico pode ter um número finito $k \geq 2$ de resultados. Quando $k = 2$ temos o caso dicotômico e a distribuição do número de sucessos é Binomial. Quando $k > 2$, a distribuição do vetor formado pelo número de indivíduos em cada classe é chamada de Multinomial. Formalmente:

- Temos n ensaios politômicos, onde o resultado de cada ensaio pode assumir k valores que denotamos por $C_1, C_2, \ldots, C_k$ (ou simplesmente por $1, \ldots, k$);
- Os n ensaios são independentes, no sentido que o resultado de um (ou um grupo) deles não afeta as probabilidades relativas aos outros ensaios;
- Os n ensaios são idênticos, no sentido que as probabilidades das classes $C_1, C_2, \ldots, C_k$, que denotamos por $p_1, p_2, \ldots, p_k$ são as mesmas em todos os ensaios
- É claro que

$$p_i \geq 0 \;\; (i = 1, 2, \ldots, k) \;\text{ e }\; \sum_{i=1}^{k} p_i = 1 \tag{2}$$

- Denote por $X = (X_1, X_2, \ldots, X_k)$ o evento que nos n ensaios foram observados X_1 vezes a classe C_1, X_2 vezes a classe C_2 e assim por diante.
- É claro da descrição do problema que X pode assumir valores $(x_1, x_2, \ldots, x_k)$ tais que os x_is são inteiros e

$$x_i \geq 0 \;\; (i = 1, 2, \ldots, k) \;\text{ e }\; \sum_{i=1}^{k} x_i = n \tag{3}$$

(compare com a equação (2)!) Portanto, embora tanto o vetor dos p_is quanto o das quantidades aleatórias X_is tem k componentes, a sua dimensão efetiva é $(k-1)$.

- A distribuição do vetor $(X_1, \ldots, X_k)$ é chamada Multinomial com parâmetros n e $p = (p_1, \ldots, p_k)$ e é dada por

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} p_i^{x_i} = \frac{n!}{x_1!\, x_2!\, \ldots\, x_k!} p_1^{x_1}\, p_2^{x_2}\, \ldots\, p_k^{x_k}, \tag{4}$$

onde os x_is satisfazem a equação (3).

- Se $X = (X_1, \ldots, X_k) \sim$ Multinomial(n,$p = (p_1, \ldots, p_k)$), então cada um dos X_i segue uma distribuição Binomial(n,p_i). Portanto

$$\mathbb{E}(X_i) = n\,p_i \;\text{ e }\; \text{Var}(X_i) = n\,p_i\,(1 - p_i)\,.$$

  Ainda, para $i \neq j$,

$$\text{cov}(X_i, X_j) = \mathbb{E}[(X_i - n\,p_i)\,(X_j - n\,p_j)] = -n\,p_i\,p_j\,.$$

Exemplo 1 (Continuação). Se o dado do exemplo for efetivamente não-viesado, no sentido que as seis faces tem a mesma probabilidade, então a distribuição dos números de cada face nos 100 lançamentos é Multinomial com $n = 100$ e $p = (1/6, 1/6, \ldots, 1/6)$. Ao realizar os $n = 100$ lançamentos o valor esperado do número de vezes que aparece a face "1" é $\mathbb{E}(X_1) = n\,p_1 = (100)(1/6) \doteq 16.67$.

2 Testes de Aderência

Seja $X = (X_1, \ldots, X_k) \sim$ Multinomial(n,$p = (p_1, \ldots, p_k)$). Queremos testar a H_0 que $p = p_0 = (p_{0,1}, p_{0,2}, \ldots, p_{0,k})$ contra a H_a que $p \neq (p_{0,1}, p_{0,2}, \ldots, p_{0,k})$ (i.é. pelo menos um dos p_i é diferente do que o correspondente $p_{0,i}$). Está é a situação do Exemplo 1 onde $k = 6$, $n = 100$ e $p_0 = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6)$. No seguinte exemplo temos uma situação onde as probabilidades $p_{i,0}$ não são todas iguais.

Exemplo 5. A coloração das flores de certas plantas (snapdragon, ou boca-de-lobo) mostra um fenômeno de *dominância parcial*. Sucintamente, temos dois gens A e a de forma que a combinação AA produz flores roxas, as combinações Aa e aA produzem flores rosas e a combinação aa produz flores brancas. De acordo a esta teoria, numa população em equilibrio deveríamos ter 25% de flores roxas e brancas e 50% de flores rosas. A Tabela 5 mostra a distribuição das cores para uma amostra de $n = 1000$ mudas. Queremos saber se essa distribuição é consistente com a hipótese da população estar em equilíbrio. Portanto, a nossa hipótese nula é que $p_{1,0} = p_{3,0} = 0.25$ e $p_{2,0} = 0.50$.

Cor	**Roxa**	**Rosa**	**Branca**	**Total**
Freq. obs.	262	535	203	1000
Freq. esp.	250	500	250	1000

Tabela 5: Cores das flores para uma amostra de $n = 1000$ mudas de *snapdragon* (Exemplo 5).

Uma primeira aproximação para testar H_0 é comparar o número de vezes que ocorre cada classe (x_i) com o seu valor esperado sob H_0, $\mathbb{E}(X_i \,|\, H_0) = n\,p_{i,0}$ (a terceira linha nas Tabelas 1 e 5. Por exemplo, poderíamos considerar a soma do quadrado da diferença entre a frequência observada e a esperada. No exemplo 1 teríamos $(21 - 16.67)^2 + (13 - 16.67)^2 + \cdots + (11 - 16.67)^2 \doteq 117.33$, que na literatura é usualmente denotado por

$\sum(n_{obs} - n_{esp})^2$ (leia "n" como "frequência"). Ainda, é claro que sob a H_a esperaríamos que está soma fosse maior do que sob a H_0. No entanto, proceder dessa forma tem um problema que é mais fácil de visualizar no Exemplo 5. Suponha nesse caso que tivéssemos observado na amostra de $n = 1000$ mudas $x_1 = 260$ flores roxas e $x_2 = 510$ flores rosas. A contribuição desses dos valores para a soma acima seria a mesma, $(260 - 250)^2 = (510 - 500)^2 = 100$, embora parece claro intuitivamente que uma diferença de 10 flores roxas é mais importante do que uma diferença de 10 flores rosas, pois, por exemplo, 260 é 4% a mais do que 250, enquanto 510 é somente 2% a mais do que 500. Por esse motivo, é usual na soma referida acima dividir cada quadrado pela frequência esperada.

- Problema: Dada a observação $x_1, \ldots, x_k$ de uma distribuição Multinomial (n,$p = (p_1, \ldots, p_k)$). Queremos testar a H_0 que $p = p_0 = (p_{0,1}, p_{0,2}, \ldots, p_{0,k})$ contra a H_a que $p \neq (p_{0,1}, p_{0,2}, \ldots, p_{0,k})$
- Estatístico do teste:

$$\chi^2_{obs} = \sum_{i=1}^{k} \frac{(x_i - n\,p_{i,0})^2}{n\,p_{i,0}} \tag{5}$$

  ou

$$\chi^2_{obs} = \sum \frac{(n_{obs} - n_{esp})^2}{n_{esp}}\,. \tag{6}$$

- Sob H_0, mostra-se que quando n tende a ∞, a distribuição do estatístico tende a uma χ^2_{k-1}. A explicação intuitiva para termos $(k-1)$ graus de liberdade é que temos somente $(k-1)$ probabilidades independentes, pois $\sum_{i=1}^{k} p_i = 1$.
- Sob a hipótese alternativa, esperamos que χ^2_{obs} seja grande. Portanto, ao nível de significância aproximado $100\alpha\%$, a H_0 é rejeitada se $\chi^2_{obs} > \chi^2_{(k-1);\alpha}$. O p-valor do teste é $\mathbb{P}(\chi^2_{(k-1)} > \chi^2_{obs})$.
- Embora o resultado anterior faz referência a $n \to \infty$, é claro que n será finito em qualquer aplicação e pode ocorrer que a convergência à distribuição limite seja muito devagar. Usualmente recomenda-se que esse teste seja aplicado somente se $n\,p_{i,0} > 5$ para todo i. Quando isso não ocorre, precisa (i) aumentar o tamanho amostral n ou (ii) colapsar classes com frequências esperadas pequenas até que todas as frequências esperadas sejam maiores do que 5, embora é preciso mencionar que nesse caso vamos estar testando uma hipóteses (ligeiramente?) diferente.

Exemplo 1 (Continuação). O estatístico observado é $\chi^2_{obs} = (21 - 16.67)^2/16.67 + (13 - 16.67)^2/16.67 + \cdots + (11 - 16.67)^2/16.67 \doteq 7.04$. Usando $\alpha = 0.05$, o valor crítico é $\chi^2_{5,0.05} \doteq 11.07$ e portanto não rejeitamos a H_0 que o dado é não-viesado. O p-valor do teste é $\mathbb{P}(\chi^2_5 > 7.04) \doteq 0.22$.

Exemplo 5 (Continuação). O estatístico observado é $\chi^2_{obs} = (262 - 259)^2/250 + (535 - 500)^2/500 + (203 - 250)^2/250 \doteq 11.86$. Usando $\alpha = 0.05$, o valor crítico é $\chi^2_{2,0.05} \doteq 5.99$ e portanto rejeitamos a H_0 que a população está em equilibrio. O p-valor do teste é $\mathbb{P}(\chi^2_2 > 11.86) \doteq 0.003$ ou 0.3%.

Quando as probabilidades das classes dependem de um ou mais parâmetros desconhecidos $(\theta_1, \theta_2 \ldots, \theta_r)$, como no Exemplo 2, é necessário estimar primeiro os parâmetros para depois calcular as probabilidades das classes. Nesse caso o cálculo do estatístico χ^2 é semelhante ao discutido acima, com a diferença que para calcular as frequências esperadas precisamos primeiro calcular primeiro estimativas $(\hat\theta_1, \hat\theta_2 \ldots, \hat\theta_r)$, depois as probabilidades estimadas das classes $\hat p_i = p_i(\hat\theta_1, \ldots, \hat\theta_k)$ e finalmente as frequências esperadas $n_i\,\hat p_i$. Quando os parâmetros são estimados pelo método de Máxima Verossimilhança (ou algum outro método que é equivalente a ele quando $n \to \infty$), mostra-se que a distribuição limite do estatístico χ^2 é χ^2 com $(k-1-r)$ graus de liberdade. Isto é, por cada parâmetro que é necessário estimar para calcular as frequências esperadas, perde-se um grau de liberdade na distribuição limite.

Observe que devemos ter cuidado para que as probabilidades estimadas das classes somem um ou, em outras palavras, as classes devem ser definidas de forma que sejam *exaustivas* (veja o exemplo abaixo).

Exemplo 2 (Continuação). Como o problema não específica o valor de θ da distribuição de Poisson, precisamos estima-lo. A estimativa de máxima verossimilhança é a média das observações $\hat\theta = \bar x = [(0)\,(24) + (1)\,(10) + (2)\,(12) + (3)\,(4)]/50 \doteq 0.92$ Calculamos depois as probabilidades das classes

$$\hat p_0 = \mathbb{P}(X = 0 \,|\, \hat\theta) = \frac{e^{-\hat\theta}\,\hat\theta^0}{0!} \doteq e^{-0.92} \doteq 0.398\,,$$

$$\hat p_1 = \mathbb{P}(X = 1 \,|\, \hat\theta) = \frac{e^{-\hat\theta}\,\hat\theta^1}{1!} \doteq (0.92)\,e^{-0.92} \doteq 0.367\,,$$

$$\hat p_2 = \mathbb{P}(X = 2 \,|\, \hat\theta) = \frac{e^{-\hat\theta}\,\hat\theta^2}{2!} \doteq \frac{(0.92)^2\,e^{-0.92}}{2} \doteq 0.169$$

e finalmente

$$\hat p_3 = \mathbb{P}(X \geq 3 \,|\, \hat\theta) = 1 - \hat p_0 - \hat p_1 - \hat p_2 \doteq 0.066\,.$$

Veja que a última classe foi definida como "$X \geq 3$" (ao invés de "$X = 3$"), de forma que as probabilidades estimadas somem um.

Com as probabilidades estimadas calculamos as frequências esperadas $n\,\hat p_i$ (veja a Tabela 2) e o estatístico $\chi^2_{obs} = (24 - 19.926)^2/19.926 + \cdots + (4 - 3.309)^2/3.309 \doteq 6.273$ Esse estatístico é comparado com a cauda à direita da distribuição χ^2 com $(k-1-r) = (4-1-1) = 2$ graus de liberdade. Por exemplo, ao nível de significância $\alpha = 0.10$, o valor da tabela é $\chi^2_{2;0.90} \doteq 4.605$, de forma que rejeitamos a H_0 da distribuição ser Poisson. O p-valor do teste é $\mathbb{P}(\chi^2_2 > 6.273) \doteq 0.043$ ou 4.3%.

Observação 1. Como explicamos acima, a convergência da distribuição do estatístico χ^2 pode ter problemas quando alguma das frequências esperadas é menor do que 5, que é o caso do exemplo para a classe "$X \geq 3$". Portanto, poderia ser indicado nesse caso agrupar as duas últimas classes numa outra que chamaríamos "$X \geq 2$". Nesse caso teríamos somente um grau de liberdade na distribuição limite.

Quando a distribuição da variável observada é continua, é necessário agrupar os dados por intervalos para definir as respectivas classes.

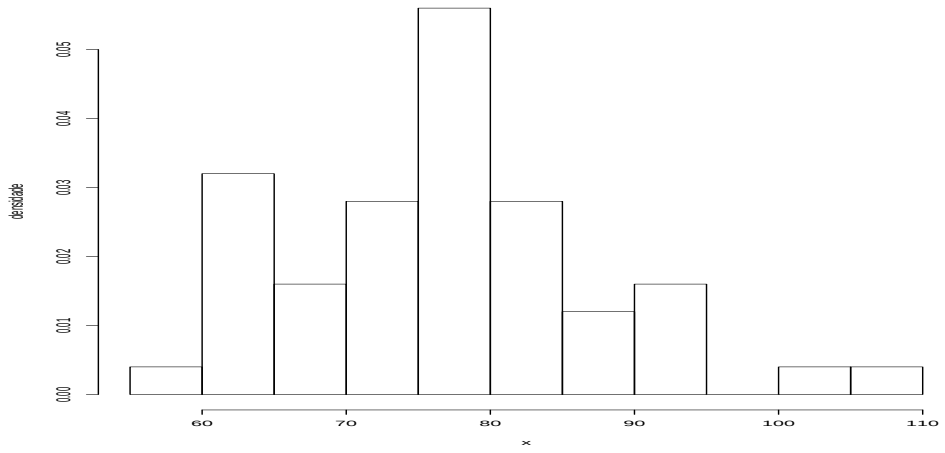

Figura 1: Histograma para os dados da Tabela 6.

Exemplo 6. A Tabela 6 mostra o tempo (em minutos) que cada aluno de uma turma de uma disciplina introdutória de estatística levou para terminar uma prova (os dados estão ordenados para facilitar os cálculos). A Figura 1 mostra um histograma desses dados produzido com a linguagem **R**. O objetivo do estudo é avaliar se é razoável supor que o tempo para completar a prova segue uma distribuição Normal.

58	61	63	64	64	65	65	65	65	67
67	68	69	72	72	73	73	73	73	75
76	76	76	77	77	78	78	78	79	79
79	80	80	80	81	81	82	83	84	84
85	87	88	88	92	93	93	95	102	107

Tabela 6: Tempo (em minutos) de 50 alunos para completar uma prova de um curso de estatística básica.

Para definir os intervalos e agrupar os dados, usamos inicialmente a opção padrão da função `hist` do **R**. O resultado está mostrado na Tabela 7. Para o cálculo das frequências esperadas calculamos primeiro estimadores para a média e a variância da distribuição Normal; $\hat{\mu} = \bar{x} \doteq 77.4$ e $\hat{\sigma}^2 = s^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \doteq (10.51)^2$. As frequências esperadas são então as probabilidades de cada intervalo numa distribuição Normal com essa média e variância. Por exemplo, para o segundo intervalo calculamos primeiro $\mathbb{P}(60 < X \leq 65) \doteq 0.0701$ e então $n_{esp} \doteq n\,(0.0701) \doteq (50)\,(0.0701) \doteq 3.51$.

Como pode ser visto da tabela, existem vários intervalos (6 no total) nos extremos cujas frequências esperadas são menores que 5. Dessa forma, para o cálculo do estatístico χ^2 agrupamos primeiro os dois primeiros intervalos e os últimos quatro, de forma que todas as classes tivessem frequências esperadas acima de 5. O resultado desse agrupamento está mostrado na Tabela 8.

Int.	$(-\infty,60]$	(60,65]	(65,70]	(70,75]	(75,80]	(80,85]
n_{obs}	1	8	4	7	14	7
n_{esp}	2.44	3.51	6.08	8.45	9.40	8.38

Int.	(85,90]	(90,95]	(95,100]	(100,105]	$(105,\infty)$
n_{obs}	3	4	0	1	1
n_{esp}	5.98	3.41	1.56	0.57	0.22

Tabela 7: Frequências observadas e esperadas para os intervalos obtidos com a opção padrão da função `hist` e as estimativas $\hat{\mu} \doteq 77.4$ e $\hat{\sigma}^2 \doteq (10.51)^2$.

Finalmente, calculamos o estatístico do teste de acordo à equação (6), obtendo $\chi^2_{obs} \doteq (9-5.95)^2/5.95 + \cdots + (6-5.76)^2/5.76 \doteq 6.493$. Como temos 7 intervalos e estimamos dois parâmetros (μ e σ^2), devemos usar a distribuição chi^2 com $7-1-2=4$ graus de liberdade. Ao nível de significância 5%, o valor crítico é $\chi^2_{4;0.05} \doteq 9.488$, de forma que não rejeitamos a hipótese nula de normalidade. O p-valor do teste é $\mathbb{P}(\chi^2_4 > 6.493) \doteq 0.165$.

Int.	$(-\infty,65]$	(65,70]	(70,75]	(75,80]	(80,85]	(85,90]	(90, ∞]
n_{obs}	9	4	7	14	7	3	6
n_{esp}	5.95	6.08	8.45	9.40	8.38	5.98	5.76

Tabela 8: Frequências observadas e esperadas para os intervalos finais e as estimativas $\hat{\mu} \doteq 77.4$ e $\hat{\sigma}^2 \doteq (10.51)^2$. Note que todas as frequências esperadas são maiores do que 5.

Observação 2. Como pode ser visto do exemplo, a definição dos intervalos que definem as classes é arbitrária e seria possível que a nossa decisão mude dependendo de como são definidas as classes. Dessa forma, a metodologia deve ser considerada somente como uma forma de explorar os dados e não um procedimento formal para testar normalidade.

Observação 3. Na verdade, estamos somente testando se as probabilidades dos intervalos por nós definidos coincidem com as de uma distribuição Normal. Existem testes mais específicos e apropriados para normalidade, tais como o critério de Kolmogorov e von Mises ou o teste de Shapiro e Wilk, mas o tratamento deles fica fora do escopo da nossa disciplina.

3 Teste de homogeneidade

Suponha que temos I populações politômicas tais que os seus indivíduos podem ser classificados nas categorias $A_1, A_2, \ldots, A_J$. Seja p_{ij} a probabilidade da categoria A_j na $i$ésima população, como indicado Tabela 9. Chama-se *teste de homogeneidade* o teste da hipótese nula que as I distribuições são iguais. Mais precisamente, a H_0 específica que $p_{1j} = p_{2j} = \cdot = p_{Ij}$ para todo $j = 1, \ldots, J$ (i.é., que os elementos das J colunas da Tabela 9, são todos iguais).

	Classe						
População	A_1	A_2	$\cdots$	A_j	$\cdots$	A_J	Total
1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1J}	1
2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	p_{2J}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{iJ}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	p_{I1}	p_{I2}	$\cdots$	p_{Ij}	$\cdots$	p_{IJ}	1

Tabela 9: I populações classificadas em J classes ou eventos.

Frequentemente, para realizar o teste, observa-se uma amostra de cada população considerada e contam-se quantos indivíduos pertencem a cada uma das classes. Denotamos n_{ij} o número de indivíduos da i-ésima amostra que estão na classe A_j e por $n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$ o tamanho da i-ésima amostra. De forma semelhante podemos definir o total de indivíduos na classe A_j, $n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$, e o total de indivíduos observados $n = n_{\bullet\bullet}$ (Veja a Tabela 10).

	Classe						
População	A_1	A_2	$\cdots$	A_j	$\cdots$	A_J	Totais
1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Totais	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	$n = n_{\bullet\bullet}$

Tabela 10: .

Para construir o estatístico do teste é útil pensar como estimaríamos as probabilidades da Tabela 9 no caso da H_0 ser ou não ser verdadeira:

- Começamos pelo caso que H_0 é falsa, isto é, que não temos nenhuma informação sobre a relação entre as probabilidades da Tabela. Nesse caso, a melhor forma de estimar p_{ij}, a probabilidade da classe A_j na i-ésima população, é com a proporção de indivíduos dessa classe na respectiva amostra, $\hat{p}_{ij} = n_{ij}/n_{i\bullet}$.

- Veja que, nesse caso, temos $I\,J$ probabilidades mas vamos precisar estimar "somente" $I\,(J-1)$ delas, pois a soma de cada linha da Tabela 9 é igual a um.

- Quando H_0 é verdadeira, as I populações tem a mesma distribuição de probabilidade. Em outras palavras, as I linhas da Tabela 9 são idênticas. Se as populações

tem a mesma distribuição de probabilidade, podemos agrupá-las e considerar o conjunto de todos $n = n_{\bullet\bullet}$ como se fosse uma única amostra dessa grande população. Logo, as estimativas das probabilidades das classe são $\tilde{p}_{ij} = n_{\bullet j}/n_{\bullet\bullet}$.

- Veja que, quando H_0 é verdadeira, temos inicialmente J probabilidades mas precisamos estimar "somente" $(J-1)$, pois a soma de todas elas tem que ser um.

A construção da estatística do teste de homogeneidade passa então pela construção de uma medida da "distância" (ou, na terminologia da Teoría da Informação, da "divergência") entre as distribuições de probabilidade definidas pelas estimativas $\hat{p}_{ij}$ e $\tilde{p}_{ij}$. Se a "distância" for "grande", isso significa que as estimativas das probabilidades da Tabela 9 são muito diferentes segundo assumirmos que H_0 é verdadeira ou não, e portanto vamos rejeitar a H_0 de homogeneidade.

Existem muitas medidas de distância ou divergência que podem ser usadas. A mais tradicional, devida ao estatístico Karl Pearson no final do século XIX, é chamada χ^2 de Pearson e é definida por

$$\chi^2(\hat{p};\tilde{p}) = \sum_{i,j} n_{i\bullet} \frac{(\hat{p}_{ij} - \tilde{p}_{ij})^2}{\tilde{p}_{ij}}$$

[cuidado: essa medida não é simétrica, assim que em geral $\chi^2(\hat{p};\tilde{p}) \neq \chi^2(\tilde{p};\hat{p})$].

Uma vez que decidimos usar a divergência χ^2, precisamos definir o significado de "grande" ou, em outras palavras, achar uma distribuição de referência para fazer o teste. Nesse sentido, o resultado fundamental é que quando todos os tamanhos amostrais tendem a infinito, a distribuição de $\chi^2(\hat{p};\tilde{p})$ se aproxima de uma distribuição χ^2 com $(I-1)(J-1)$ graus de liberdade. Duas observações são importantes aqui:

- Como em todos os problemas que temos visto anteriormente, os graus de liberdade do estatístico χ^2 é a diferença entre o número de parâmetros livres no caso geral $[I(J-1)]$ e o número de de parâmetros livres sob H_0 $(J-1)$.
- Defina $n_{ij}^{obs} = n_{ij}$ e $n_{ij}^{esp} = n_{i\bullet}\,\tilde{p}_{ij} = n_{i\bullet}\,n_{\bullet j}/n$. Então

$$\begin{aligned}\chi^2_{obs} &= \chi^2(\hat{p};\tilde{p}) \\ &= \sum_{i,j} n_{i\bullet} \frac{(\hat{p}_{ij} - \tilde{p}_{ij})^2}{\tilde{p}_{ij}} = \sum_{i,j} n_{i\bullet} \frac{(n_{ij}/n_{i\bullet} - n_{\bullet j}/n)^2}{n_{\bullet j}/n} = \sum_{i,j} \frac{(n_{ij}^{obs} - n_{ij}^{esp})^2}{n_{ij}^{esp}} \qquad (7)\end{aligned}$$

[compare com a equação (6)]

Exemplo 3 (Continuação). As tabelas 11-13 mostram respectivamente (i) as probabilidades estimadas no caso geral, (ii) as probabilidades estimadas sob a restrição de homogeneidade e (iii) as frequências esperadas sob a restrição de homogeneidade. Usando por exemplo as frequências observadas da Tabela 3 e as esperadas da Tabela 13, calculamos $\chi^2_{obs} \doteq (13-10.667)^2/10.667 + \cdots + (13-8.667)^2/8.667 \doteq 7.407$. Como temos $I(J-1)-(J-1) = (I-1)(J-1) = 4$ graus de liberdade, ao nível de significância de 5%, o valor da tabela é $\chi^2_{4,0.05} \doteq 9.488$, de forma que não rejeitamos a H_0 de homogeneidade. Em outras palavras, ao nível de significância de 5%, não existe evidência que a distribuição das notas é diferentes segundo o sexo. O p-valor do teste é $\mathbb{P}(\chi^2_4 > 7.407) \doteq 0.116$.

Nota	II	MI	MM	MS	SS	Total
Homens	0.13	0.15	0.37	0.22	0.13	1
Mulheres	0.06	0.08	0.30	0.30	0.26	1

Tabela 11: Probabilidades estimadas sem restrições para o Exemplo 3: $\hat{p}_{ij} = n_{ij}/n_{i\bullet}$.

Nota	II	MI	MM	MS	SS	Total
Homens	0.107	0.127	0.347	0.247	0.173	1
Mulheres	0.107	0.127	0.347	0.247	0.173	1

Tabela 12: Probabilidades estimadas sob restrição de homogeneidade para o Exemplo 3: $\tilde{p}_{ij} = n_{\bullet j}/n_{\bullet\bullet}$.

4 Testes de independência

Suponha uma única população classificada de acordo a dois fatores ou características, digamos $A_1, A_2, \ldots, A_I$ e $B_1, B_2, \ldots, B_J$. Seja p_{ij} a probabilidade de um indivíduo escolhido ao acaso dessa população possuir as características A_i e B_j simultaneamente, $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ a probabilidade (marginal) dele possuir a caracaterística A_i e $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$ a probabilidade (marginal) dele possuir a caracaterística B_j (veja a Tabela 14). Por outro lado, suponha que foi observada uma única amostra de tamanho n dessa população e os indivíduos foram classificados de acordo aos dois fatores A e B, de forma que a frequência observada da classe $A_i \cap B_j$ foi n_{ij}. Defina ainda a frequência associada com a classe A_i, $n_{i\bullet} = \sum_j n_{ij}$ e com a classe B_j, $n_{\bullet j} = \sum_i n_{ij}$. A situação aparece descrita na Tabela 15

Queremos testar a H_0 que os dois fatores são (probabilisticamente) independentes. Lembrando que dois eventos são independentes se a probabilidade da sua interseção é igual ao produto das probabilidades dos eventos, no caso de independência deveriamos ter que $p_{ij} = \mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)\,\mathbb{P}(B_j) = p_{i\bullet}\,p_{\bullet j}$. Portanto as nossas hipóteses são

$$H_0 : p_{ij} = p_{i\bullet}\,p_{\bullet j} \text{ para todo par } (i,j)$$

e

$$H_a : p_{ij} \neq p_{i\bullet}\,p_{\bullet j} \text{ para pelo menos um par } (i,j)\,.$$

O procedimento do teste é muito parecido com o recém discutido teste de homogeneidade. Isto é, começamos por ver como estimaríamos as probabilidades p_{ij} dependendo da H_0 ser verdadeira ou não:

- Quando H_0 é falsa, não temos nenhuma informação sobre a relação entre as probabilidades da Tabela 14. Nesse caso, a melhor forma de estimar p_{ij}, a probabilidade da interseção de A_i e B_j, é com a proporção de indivíduos que possuem essa característica na respectiva amostra. Em outras palavras, $\hat{p}_{ij} = n_{ij}/n_{i\bullet}$.
- Veja que, nesse caso, temos (IJ) probabilidades mas vamos precisar estimar "somente" $(IJ-1)$ delas, pois a soma de todas as probabilidades da Tabela 14 é igual

Nota	**II**	**MI**	**MM**	**MS**	**SS**	**Total**
Homens	10.667	12.667	34.667	24.667	17.333	100
Mulheres	5.333	6.333	17.333	12.333	8.667	50

Tabela 13: Frequências esperadas no caso de homogeneidade no Exemplo 3: $n_{ij}^{esp} = n_{i\bullet}\,\tilde{p}_{ij} = n_{i\bullet}\,n_{\bullet j}/n_{\bullet\bullet}$

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	Total
A_1	p_{11}	p_{12}	$\cdots$	p_{1j}	$\cdots$	p_{1J}	$p_{1\bullet}$
A_2	p_{21}	p_{22}	$\cdots$	p_{2j}	$\cdots$	p_{2J}	$p_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	p_{i1}	p_{i2}	$\cdots$	p_{ij}	$\cdots$	p_{iJ}	$p_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	p_{I1}	p_{I2}	$\cdots$	p_{Ij}	$\cdots$	p_{IJ}	$p_{I\bullet}$
Total	$p_{\bullet 1}$	$p_{\bullet 2}$	$\cdots$	$p_{\bullet j}$	$\cdots$	$p_{\bullet J}$	1

Tabela 14: Uma única população classificada de acordo a dois fatores, A com I níveis e B com J níveis.

a um.

- Quando H_0 é verdadeira sería suficiente estimar somente as probabilidades *marginais* $p_{i\bullet}$ e $p_{\bullet j}$, pois depois as outras probabilidades da Tabela 14 seriam obtidas como o produto das duas marginais correspondentes. Portanto, podemos definir as estimativas $\tilde{p}_{i\bullet} = n_{i\bullet}/n$ e $\tilde{p}_{\bullet j} = n_{\bullet j}/n$ e depois teríamos simplesmente que $\tilde{p}_{ij} = \tilde{p}_{i\bullet}\,\tilde{p}_{\bullet j}$.

- Veja que, quando H_0 é verdadeira, temos $(I + J)$ probabilidades marginais mas, dessas, somente precisamos estimar $(I + J - 2)$, pois $\sum_i p_{i\bullet} = \sum_j p_{\bullet j} = 1$.

	B_1	B_2	$\cdots$	B_j	$\cdots$	B_J	Total
A_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1J}	$n_{1\bullet}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2J}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iJ}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	$\cdots$	n_{Ij}	$\cdots$	n_{IJ}	$n_{I\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet J}$	$n = n_{\bullet\bullet}$

Tabela 15: Uma única amostra classificada de acordo a dois fatores, A com I níveis e B com J níveis.

O estatístico do teste é

$$\chi^2_{obs} = \sum_{i,j} n \, \frac{(\hat{p}_{ij} - \tilde{p}_{ij})^2}{\tilde{p}_{ij}} = \sum_{i,j} \frac{(n^{obs}_{ij} - n^{esp}_{ij})^2}{n^{esp}_{ij}} \,,$$

onde $n^{obs}_{ij} = n\,\hat{p}_{ij} = n_{ij}$ e $n^{obs}_{ij} = n\,\tilde{p}_{ij} = n_{i\bullet}\,n_{\bullet j}/n$ [novamente, compare com as equações (6) e (7)]. Esse valor é comparado com valores críticos da distribuição χ^2 com $(IJ-1)-(I+J-2) = (I-1)(J-1)$ graus de liberdade, de forma que a H_0 de independência é rejeitada se $\chi^2_{obs} > \chi^2_{(I-1)(J-1);\alpha}$.

Exemplo 4 (Continuação). A Tabela 4 mostra as frequências observadas n_{ij}. Comm base nela calculamos as probabilidades estimadas sem restrição ($\hat{p}_{ij} = n_{ij}/n$) e sob a restrição de independencia ($\tilde{p}_{ij} = n_{i\bullet}n_{\bullet j}/n^2$) e as frequências esperadas sob H_0 ($n^{esp}_{ij} = n_{i\bullet}n_{\bullet j}/n$), mostradas respectivamente nas tabelas 16-18.

Nota	**II**	**MI**	**MM**	**MS**	**SS**	**Total**
Homens	0.050	0.070	0.180	0.155	0.105	0.56
Mulheres	0.055	0.080	0.125	0.115	0.065	0.44
Total	0.105	0.150	0.305	0.270	0.170	1

Tabela 16: Probabilidades estimadas sem restrição no Exemplo 4 ($\hat{p}_{ij} = n_{ij}/n$).

Nota	**II**	**MI**	**MM**	**MS**	**SS**	**Total**
Homens	0.0588	0.0840	0.1708	0.1512	0.0952	0.56
Mulheres	0.0462	0.0660	0.1342	0.1188	0.0748	0.44
Total	0.105	0.150	0.305	0.270	0.170	1

Tabela 17: Probabilidades estimadas sob restrição de independência para o Exemplo 4 ($\tilde{p}_{ij} = n_{i\bullet}n_{\bullet j}/n^2$).

Nota	**II**	**MI**	**MM**	**MS**	**SS**	**Total**
Homens	11.76	16.80	34.16	30.24	19.04	112
Mulheres	9.24	13.20	26.84	23.76	14.96	88
Total	21	30	61	54	34	200

Tabela 18: Frequências esperadas sob restrição de independência para o Exemplo 4 ($n^{esp}_{ij} = n\,\tilde{p}_{ij} = n_{i\bullet}n_{\bullet j}/n$).

O estatístico do teste é $\chi^2_{obs} = (10-11.76)^2/11.76 + \cdot + (13-14.96)^2/14.96 \doteq 2.386$. Ao nível de significância de 5%, o valor crítico é $\chi^2_{4;0.05} \doteq 9.488$. Portanto, não rejeitamos a H_0 de independência. O p-valor aproximado do teste é $\mathbb{P}(\chi^2_4 > 2.386) \doteq 0.665$ ou 66.5%.