Medidas de associação I

Tabelas de Contingência e Testes de Independência

Felipe Figueiredo

Instituto Nacional de Traumatologia e Ortopedia

Sumário

Dados categóricos

- Vamos analisar contagens de dados categóricos (ou nominais)
- Para estas variáveis qualitativas não existe ordenação interente
- Observamos apenas as contagens e frequências destes dados em uma amostra.

Example

doente/sadio, fumante/não fumante, masculino/feminino, olhos castanhos/olhos azuis/olhos verdes, etc.

Eventos independentes

Conforme vimos na aula de Probabilidades:

- Dois eventos são independentes se a ocorrência do primeiro não afeta a ocorrência do segundo
- Isto significa que a probabilidade da ocorrência do segundo não é condicional em relação ao primeiro
- Em relação aos dados de uma amostra: a frequência observada para cada categoria indica que estas são independentes?

Objetivo

Considere a seguinte tabela:

Example

Resultado	Alongou-se	Não se alongou
Lesão	18	22
Não tem lesão	211	189

(Fonte: Larson & Farber 2013)

Pergunta: a partir destes dados é possível determinar se existe alguma relação entre as variáveis? Isto é: as variáveis são independentes?

Exemplo

Medidas de associação I

Felipe Figueiredo

Example

Considere que 10% dos pacientes morrem após uma operação arriscada. Em uma amostra de 75 pacientes, observou-se que 16 pacientes morreram após a operação. Como comparar o número de óbitos osbervado e o número esperado?
Fonte: Motulsky, 1995

- O número observado de óbitos em 75 pacientes foi 16.
- O número esperado seria $75 \times 10\% = 7.5$
- A discrepância nos óbitos foi $16 - 7.5 = 8.5$

Questões

Medidas de associação I

Felipe Figueiredo

- Esse aumento reflete uma mudança real na mortalidade?
- Em uma amostra qualquer com 75 pacientes esperaríamos observar 7.5 óbitos
- Em uma amostra específica poderíamos observar mais ou menos que isso
- Provavelmente algo próximo de 7.5

Pergunta

Se a mortalidade for 10%, qual é a probabilidade de se observar 16 ou mais óbitos em uma amostra de 75 pacientes?

Roteiro

Medidas de associação I

Felipe Figueiredo

- Podemos representar as contagens observadas e esperadas em uma tabela
- A hipótese H_0 é que observamos uma amostra de uma população com 10% de mortalidade.
- As diferenças entre os dados observados e os esperados tem distribuição aproximadamente χ^2 (qui-quadrado)
- Fazendo o teste χ^2 podemos testar a hipótese H_0
- Estatística de teste: $\chi^2 = \frac{\sum(\text{observado} - \text{esperado})^2}{\text{esperado}}$

Tabela de frequências

Example

	Observado	Esperado
Óbito	16	7.5
Vivo	59	67.5
Total	75	75

Estatística de teste:

$$\chi^2 = \frac{(16-7.5)^2}{7.5} + \frac{(59-67.5)^2}{67.5} \approx 10.70$$

A tabela Qui-Quadrado

χ^2	P	χ^2	P	χ^2	P	χ^2	P
0.0	1.000	4.1	0.0429	8.1	0.0044	12.1	0.0005
0.1	0.7518	4.2	0.0404	8.2	0.0042	12.2	0.0005
0.2	0.6547	4.3	0.0381	8.3	0.0040	12.3	0.0005
0.3	0.5839	4.4	0.0359	8.4	0.0038	12.4	0.0004
0.4	0.5271	4.5	0.0339	8.5	0.0036	12.5	0.0004
0.5	0.4795	4.6	0.0320	8.6	0.0034	12.6	0.0004
0.6	0.4386	4.7	0.0302	8.7	0.0032	12.7	0.0004
0.7	0.4028	4.8	0.0285	8.8	0.0030	12.8	0.0003
0.8	0.3711	4.9	0.0269	8.9	0.0029	12.9	0.0003
0.9	0.3428	5.0	0.0253	9.0	0.0027	13.0	0.0003
1.0	0.3173	5.1	0.0239	9.1	0.0026	13.1	0.0003
1.1	0.2943	5.2	0.0226	9.2	0.0024	13.2	0.0003
1.2	0.2733	5.3	0.0213	9.3	0.0023	13.3	0.0003
1.3	0.2542	5.4	0.0201	9.4	0.0022	13.4	0.0003
1.4	0.2367	5.5	0.0190	9.5	0.0021	13.5	0.0002
1.5	0.2207	5.6	0.0180	9.6	0.0019	13.6	0.0002
1.6	0.2059	5.7	0.0170	9.7	0.0018	13.7	0.0002
1.7	0.1923	5.8	0.0160	9.8	0.0017	13.8	0.0002
1.8	0.1797	5.9	0.0151	9.9	0.0017	13.9	0.0002
1.9	0.1681	6.0	0.0143	10.0	0.0016	14.0	0.0002
2.0	0.1573	6.1	0.0135	10.1	0.0015	14.1	0.0002
2.1	0.1473	6.2	0.0128	10.2	0.0014	14.2	0.0002
2.2	0.1380	6.3	0.0121	10.3	0.0013	14.3	0.0002
2.3	0.1294	6.4	0.0114	10.4	0.0013	14.4	0.0001
2.4	0.1213	6.5	0.0108	10.5	0.0012	14.5	0.0001
2.5	0.1138	6.6	0.0102	10.6	0.0011	14.6	0.0001
2.6	0.1069	6.7	0.0096	10.7	0.0011	14.7	0.0001
2.7	0.1003	6.8	0.0091	10.8	0.0010	14.8	0.0001
2.8	0.0943	6.9	0.0086	10.9	0.0010	14.9	0.0001
2.9	0.0886	7.0	0.0082	11.0	0.0009		
3.0	0.0833	7.1	0.0077	11.1	0.0009		

Comparando as frequências

- Assumos a hipótese H_0 de que não houve aumento da mortalidade do procedimento.
- Encontramos a estatística de teste $\chi^2 = 10.7$ para a amostra.
- Fazendo o teste χ^2, encontramos $p = 0.0011$.
- Como $p = 0.0011 < 0.05$, decidimos rejeitar H_0.
- Conclusão: rejeitamos a hipótese de que não houve aumento na mortalidade, ao nível de significância de 5%.

Tabelas de Contingência

Definition

Uma tabela de contingência mostra as frequências observadas para duas ou mais variáveis complementares.

- Podemos calcular as frequências esperadas, baseado no tamanho das amostras
- Comparamos assim a frequência observada com a frequência esperada
- Obs: a tabela do exemplo anterior (óbitos) não é uma tabela de contingência! (Por que?)

Tabelas de contingência 2x2

Example

Frequências observadas:

	doença progrediu	doença não progrediu
AZT	76	399
Placebo	129	332

- Existe relação entre o uso do AZT e a progressão da doença?
- Ou: nessa amostra o AZT foi mais eficiente que o placebo (rejeitar H_0)?

Tabelas de contingência 2x2

- H_0: o AZT não é mais eficaz que o placebo
- Pergunta: assumindo a H_0, qual seria a frequência esperada para a progressão da doença?
- Em outras palavras: quantos pacientes tiveram progressão na doença, em relação ao total?

Tabelas de contingência 2x2

Example

Frequências observadas:

	progrediu	não progrediu	total
AZT	76	399	475
Placebo	129	332	461
total	205	731	936

- Frequência esperada $E = \dfrac{205}{936} \approx 0.2190 = 21.90\%$
- Número esperado: $475 \times 0.2190 = 104.025 \approx 104.0$ pacientes

Tabelas de contingência 2x2

- Se a H_0 fosse verdadeira, esperaríamos que 104.0 tivessem a progressão da doença, usando o AZT.
- Mas observamos 76.
- Discrepância $|104.0 - 76| = 28$ pacientes
- Procedendo de maneira análoga, podemos descobrir todos os valores esperados, para cada categoria da tabela
- Para simplificar a interpretação, podemos usar a seguinte fórmula:

$$E = \frac{\text{total por linha} \times \text{total por coluna}}{\text{total da tabela}}$$

Tabelas de contingência 2x2

Example

Frequências observadas:

	progrediu	não progrediu	total
AZT	76	399	475
Placebo	129	332	461
total	205	731	936

- AZT + Progressão = $\frac{205 \times 475}{936} = 104.0$
- AZT + Não progressão = $\frac{731 \times 475}{936} = 371.0$
- Placebo + Progressão = $\frac{205 \times 461}{936} = 101.0$
- Placebo + Não progressão = $\frac{731 \times 461}{936} = 360.0$

Tabelas de contingência 2x2

Medidas de associação I

Felipe Figueiredo

Colocando os valores em uma tabela semelhante:

Example

Frequências esperadas:

	progrediu	não progrediu	total
AZT	104.0	371.0	475.0
Placebo	101.0	360.0	461.0
total	205.0	731.0	936.0

Observe que os totais esperados devem ser iguais aos observados!

Teste de Hipótese

Medidas de associação I

Felipe Figueiredo

- H_0 não há relação entre o uso do AZT e a progressão da doença.
- Determinamos as diferenças quadráticas entre o valor observado e o esperado como fizemos anteriormente

$$\chi^2 = \frac{\sum(\text{observado} - \text{esperado})^2}{\text{esperado}}$$

- Fazemos o teste χ^2 e julgamos a hipótese H_0

Teste de Hipótese

Medidas de associação I

Felipe Figueiredo

Example

- AZT + P = $\frac{(76 - 104.0)^2}{104.0} = \frac{28^2}{104.0} \approx 7.54$
- AZT + NP = $\frac{(399 - 371.0)^2}{371.0} = \frac{28^2}{371.0} \approx 2.11$
- Placebo + P = $\frac{(129 - 101.0)^2}{101.0} = \frac{28^2}{101.0} \approx 7.76$
- Placebo + NP = $\frac{(332 - 360.0)^2}{360.0} = \frac{28^2}{360.0} \approx 2.18$
- $\chi^2 = 7.54 + 2.11 + 7.76 + 2.18 = 19.59$

O teste Qui-Quadrado

- Quanto maior for o valor da estatística de teste, menor será o valor-p.
- Calculamos a estatística de teste para a amostra e encontramos $\chi^2 = 19.59$
- Qual é o p-valor desta análise?

A tabela Qui-Quadrado

Medidas de associação I

Felipe Figueiredo

χ^2	P	χ^2	P	χ^2	P	χ^2	P
0.0	1.000	4.1	0.0429	8.1	0.0044	12.1	0.0005
0.1	0.7518	4.2	0.0404	8.2	0.0042	12.2	0.0005
0.2	0.6547	4.3	0.0381	8.3	0.0040	12.3	0.0005
0.3	0.5839	4.4	0.0359	8.4	0.0038	12.4	0.0004
0.4	0.5271	4.5	0.0339	8.5	0.0036	12.5	0.0004
0.5	0.4795	4.6	0.0320	8.6	0.0034	12.6	0.0004
0.6	0.4386	4.7	0.0302	8.7	0.0032	12.7	0.0004
0.7	0.4028	4.8	0.0285	8.8	0.0030	12.8	0.0003
0.8	0.3711	4.9	0.0269	8.9	0.0029	12.9	0.0003
0.9	0.3428	5.0	0.0253	9.0	0.0027	13.0	0.0003
1.0	0.3173	5.1	0.0239	9.1	0.0026	13.1	0.0003
1.1	0.2943	5.2	0.0226	9.2	0.0024	13.2	0.0003
1.2	0.2733	5.3	0.0213	9.3	0.0023	13.3	0.0003
1.3	0.2542	5.4	0.0201	9.4	0.0022	13.4	0.0003
1.4	0.2367	5.5	0.0190	9.5	0.0021	13.5	0.0002
1.5	0.2207	5.6	0.0180	9.6	0.0019	13.6	0.0002
1.6	0.2059	5.7	0.0170	9.7	0.0018	13.7	0.0002
1.7	0.1923	5.8	0.0160	9.8	0.0017	13.8	0.0002
1.8	0.1797	5.9	0.0151	9.9	0.0017	13.9	0.0002
1.9	0.1681	6.0	0.0143	10.0	0.0016	14.0	0.0002
2.0	0.1573	6.1	0.0135	10.1	0.0015	14.1	0.0002
2.1	0.1473	6.2	0.0128	10.2	0.0014	14.2	0.0002
2.2	0.1380	6.3	0.0121	10.3	0.0013	14.3	0.0002
2.3	0.1294	6.4	0.0114	10.4	0.0013	14.4	0.0001
2.4	0.1213	6.5	0.0108	10.5	0.0012	14.5	0.0001
2.5	0.1138	6.6	0.0102	10.6	0.0011	14.6	0.0001
2.6	0.1069	6.7	0.0096	10.7	0.0011	14.7	0.0001
2.7	0.1003	6.8	0.0091	10.8	0.0010	14.8	0.0001
2.8	0.0943	6.9	0.0086	10.9	0.0010	14.9	0.0001
2.9	0.0886	7.0	0.0082	11.0	0.0009		
3.0	0.0833	7.1	0.0077	11.1	0.0009		

O teste Qui-Quadrado

Medidas de associação I

Felipe Figueiredo

- Consultando a tabela χ^2, encontramos um $p < 0.0001$
- Interpretação: Se a H_0 for verdadeira, temos uma chance menor que 0.01% de observar uma discrepância tão grande entre os valores observados e os esperados.
- Conclusão: devemos rejeitar a H_0

Interpretação

Rejeitamos a hipótese de que o AZT não é mais eficiente que o placebo.

O teste Qui-Quadrado

Medidas de associação I

Felipe Figueiredo

- O teste χ^2 é apenas uma aproximação da distribuição dos dados, que pode ser usado para amostras grandes.
- Vantagem: simples de consultar na tabela
- Desvantagem: a aproximação é ruim para amostras pequenas
- Nunca usar se alguma célula da tabela tiver valor < 5
- O teste indicado para este tipo de cenário é o teste exato de Fisher

O teste de Fisher (teste F)

- Para as seguintes situações deve-se usar o teste exato de Fisher:
  1. Quando se tem amostras pequenas
  2. Quanto se tem amostras de tamanho moderado, e se tiver uma ferramenta computacional disponível
- Se sua amostra for enorme (milhares de dados), prefira o teste χ^2, pois:
  1. o cálculo do teste F pode ser lento
  2. a aproximação será boa

Tabelas de Contingência maiores

- E quando temos mais do que duas categorias complementares?
- Resposta: procedemos como no caso anterior, mas precisamos considerar os graus de liberdade do teste χ^2

$$gl = (l-1)(c-1) = (\text{linhas}-1) \times (\text{colunas}-1)$$

- Obs: no caso 2×2 temos $gl = (2-1) \times (2-1) = 1 \times 1 = 1$

Tabelas de Contingência maiores

Example

Em dois hospitais, os resultados de 575 autópsias foram comparados com as causas de morte listadas nos atestados. Um dos hospitais que participou do estudo era comunitário (A); o outro era universitário (B).

Hospital	Precisão confirmada	Falta de informações	Recodificação incorreta
A	157	18	54
B	268	44	34

Os resultados sugerem práticas diferentes no preenchimento de atestados de óbito nos dois hospitais?

Fonte: Aula Hacker & Simões (2008 - Fiocruz)

Tabelas de Contingência maiores

- H_0: Dentro de cada categoria do status do atestado, as proporções de atestados de óbitos no hospital A são idênticas ao hospital B.
- H_1: As proporções não são idênticas
- Graus de liberdade:

$$(l-1) \times (c-1) = (2-1) \times (3-1) = 1 \times 2 = 2$$

Tabelas de contingência maiores

Preenchendo os totais por linha e coluna:

Example

Hospital	Confirmada	Incompleta	Incorreta	total
A	157	18	54	229
B	268	44	34	346
total	425	62	88	575

Tabelas de contingência maiores

Incluindo os valores esperados em parênteses temos:

Example

Hospital	Confirmada	Incompleta	Incorreta	total
A	157 (169.3)	18 (24.7)	54 (35.0)	229
B	268 (255.7)	44 (37.3)	34 (53.0)	346
total	425	62	88	575

Tabelas de Contingência maiores

Calculando a estatística de teste χ^2:

- $\chi^2 = \dfrac{(157 - 169.3)^2}{169.3} + \dfrac{(18 - 24.7)^2}{24.7} + \dots$
- $\chi^2 = 21.62$

A tabela Qui-Quadrado

	α												
df	0.250	0.200	0.150	0.100	0.070	0.060	0.050	0.040	0.030	0.020	0.010	0.005	0.001
1	1.323	1.642	2.072	2.706	3.283	3.537	3.841	4.218	4.709	5.412	6.635	7.879	10.827
2	2.773	3.219	3.794	4.605	5.319	5.627	5.991	6.438	7.013	7.824	9.210	10.597	13.815
3	4.108	4.642	5.317	6.251	7.060	7.407	7.815	8.311	8.947	9.837	11.345	12.838	16.266
4	5.385	5.989	6.745	7.779	8.666	9.044	9.488	10.026	10.712	11.668	13.277	14.860	18.466
5	6.626	7.289	8.115	9.236	10.191	10.596	11.070	11.644	12.375	13.388	15.086	16.750	20.515
6	7.841	8.558	9.446	10.645	11.660	12.090	12.592	13.198	13.968	15.033	16.812	18.548	22.457
7	9.037	9.803	10.748	12.017	13.088	13.540	14.067	14.703	15.509	16.622	18.475	20.278	24.321
8	10.219	11.030	12.027	13.362	14.484	14.956	15.507	16.171	17.011	18.168	20.090	21.955	26.124
9	11.389	12.242	13.288	14.684	15.854	16.346	16.919	17.608	18.480	19.679	21.666	23.589	27.877
10	12.549	13.442	14.534	15.987	17.203	17.713	18.307	19.021	19.922	21.161	23.209	25.188	29.588
11	13.701	14.631	15.767	17.275	18.533	19.061	19.675	20.412	21.342	22.618	24.725	26.757	31.264
12	14.845	15.812	16.989	18.549	19.849	20.393	21.026	21.785	22.742	24.054	26.217	28.300	32.909
13	15.984	16.985	18.202	19.812	21.151	21.711	22.362	23.142	24.125	25.471	27.688	29.819	34.527
14	17.117	18.151	19.406	21.064	22.441	23.017	23.685	24.485	25.493	26.873	29.141	31.319	36.124
15	18.245	19.311	20.603	22.307	23.720	24.311	24.996	25.816	26.848	28.259	30.578	32.801	37.698
16	19.369	20.465	21.793	23.542	24.990	25.595	26.296	27.136	28.191	29.633	32.000	34.267	39.252
17	20.489	21.615	22.977	24.769	26.251	26.870	27.587	28.445	29.523	30.995	33.409	35.718	40.791
18	21.605	22.760	24.155	25.989	27.505	28.137	28.869	29.745	30.845	32.346	34.805	37.156	42.312
19	22.718	23.900	25.329	27.204	28.751	29.396	30.144	31.037	32.158	33.687	36.191	38.582	43.819
20	23.828	25.038	26.498	28.412	29.991	30.649	31.410	32.321	33.462	35.020	37.566	39.997	45.314
21	24.935	26.171	27.662	29.615	31.225	31.895	32.671	33.597	34.759	36.343	38.932	41.401	46.796
22	26.039	27.301	28.822	30.813	32.453	33.135	33.924	34.867	36.049	37.659	40.289	42.796	48.268
23	27.141	28.429	29.979	32.007	33.675	34.370	35.172	36.131	37.332	38.968	41.638	44.181	49.728
24	28.241	29.553	31.132	33.196	34.893	35.599	36.415	37.389	38.609	40.270	42.980	45.558	51.179
25	29.339	30.675	32.282	34.382	36.106	36.824	37.652	38.642	39.880	41.566	44.314	46.928	52.619
26	30.435	31.795	33.429	35.563	37.315	38.044	38.885	39.889	41.146	42.856	45.642	48.290	54.051
27	31.528	32.912	34.574	36.741	38.520	39.259	40.113	41.132	42.407	44.140	46.963	49.645	55.475
28	32.620	34.027	35.715	37.916	39.721	40.471	41.337	42.370	43.662	45.419	48.278	50.994	56.892
29	33.711	35.139	36.854	39.087	40.919	41.679	42.557	43.604	44.913	46.693	49.588	52.335	58.301
30	34.800	36.250	37.990	40.256	42.113	42.883	43.773	44.834	46.160	47.962	50.892	53.672	59.702

Tabelas de Contingência maiores

- Calculamos a estatística de teste $\chi^2 = 21.62$
- Encontramos um p-valor $p < 0.001$ (valor fora da tabela)
- Rejeitamos H_0 ao nível de significância de $\alpha = 0.05$.
- Conclusão: Há associação entre o hospital e o status do atestado.
- Parece que o hospital A tem maior proporção de atestados incorretos.