ADC_1 – Três Tabelas de Contingência para Três Estratégias de Amostragem[1]

Quais as escalas utilizadas para as variáveis categóricas / categorizadas?

Que tipos de dados são apresentados nas tabelas abaixo (de acordo com a classificação de Nishisato)?

Tabela 1:

Opinião	Faixa Etária		Total
	< 40 anos	≥ 40 anos	
Favorável	43	25	68
Desfavorável	41	70	111
Total	84	95	169

Tabela 2:

Opinião	Faixa Etária		Total
	< 40 anos	≥ 40 anos	
Favorável	50	26	76
Desfavorável	48	76	124
Total	98	102	200

Tabela 3:

Opinião	Faixa Etária		Total
	< 40 anos	≥ 40 anos	
Favorável	54	30	84
Desfavorável	46	70	116
Total	100	100	200

[1] Exemplos de Paulino e Singer (2006) – Ver referências bibliográficas

Modelos Probabilísticos para Dados em Tabelas de Contingência

1. Revisão dos modelos

Como podemos imaginar, há modelos amostrais diferentes para descrever o processo de geração dos dados, representados por meio de tabelas de contingência *I × J.*

I) **Poisson (Tabela 1)**

As distribuições das frequências em cada célula são distribuídas de forma mutuamente independente como Poisson com parâmetros $\boldsymbol{\mu} = E(f_{ij}), \mu_{ij} \in \mathbb{R}^{+}, i = 1, \dots, \mathrm{I};\ j = 1, \dots, \mathrm{J}.$

A distribuição conjunta é dada por:

$$f(f \mid \mu) = \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{e^{-\mu_{ij}} {\mu_{ij}}^{n_{ij}}}{n_{ij!}},$$

satisfeitas as condições para uma V.A. Poisson em cada célula, onde $\mu_{ij} = m \times \lambda_{ij}$, sendo m o intervalo considerado e λ uma constante de proporcionalidade.

O tamanho da amostra *n* também apresenta uma distribuição Poisson com parâmetro $\mu_{++} = E(n)$.

A hipótese de interesse para a relação entre as variáveis é (***multiplicabilidade*** de médias):

$$\mathrm{H}_0: \ \mu_{ij} = \frac{\mu_{i+} \times \mu_{+j}}{\mu_{++}}$$

II) **Multinomial (Tabela 2)**

Para *n* fixado, selecionado de uma população infinita (ou muito grande, na prática). As *k* observações bivariadas (*k* = 1, 2, ..., n) são então classificadas em uma das *IJ* células da tabela. A distribuição conjunta para as frequências nas células é:

$$f(f \mid n, \pi) = \frac{n!}{\prod_{i=1}^{I} \prod_{j=1}^{J} n_{ij!}} \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{n_{ij}}, \qquad \text{com } 1'\, f = n;\ 1'\pi = 1$$

Um caso especial da multinomial é a binomial, onde *I* = 1 e *J* =2. Nesse caso só há duas células.

A hipótese de interesse para a relação entre as variáveis é (***independência***):

$$\mathrm{H}_0: \pi_{ij} = \pi_{i+} \times \pi_{+j}$$

III) **Produto de Multinomiais (Tabela 3)**

Quando os totais marginais de linha ou de coluna são fixados, o que acontece quando temos a variável representada pela linha ou coluna como explicativa (amostragem estratificada). No caso de totais marginais de coluna fixados, a distribuição das frequências em cada coluna é multinomial. A distribuição conjunta para todas as colunas é dada pelo produto das distribuições de cada coluna, daí o termo **produto de multinomiais**. Temos então:

$$f(f \mid n, \pi) = \prod_{j=1}^{J} \left[\frac{n_{+j}!}{\prod_{i=1}^{I} n_{ij}!} \prod_{i=1}^{I} \pi_{i \mid j}^{n_{ij}} \right]$$

A hipótese de interesse para a relação entre as variáveis é (***homogeneidade de proporções***):

$$H_0: \pi_{i \mid 1} = \pi_{i \mid 2} = \cdots = \pi_{i \mid J}$$

IV) **Hipergeométrica**

Se a população é finita (N) com frequências populacionais nas células conhecidas, uma amostragem sem reposição nos leva a um modelo de distribuição hipergeométrica:

$$f(f \mid n) = \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{N_{ij}!}{n_{ij}!\,(N_{ij} - n_{ij})!} \Big/ \frac{N!}{n!\,(N-n)!}$$

No caso de populações finitas, mas suficientemente grandes, podemos aproximar a hipergeométrica pela multinomial, desde que cada N_{ij} sejam grandes.

Exemplo: A população formada por todas as declarações de imposto de renda 2019, ano base 2018, apresenta as informações sobre o estado da federação onde reside o declarante e a faixa de renda bruta, representadas por duas variáveis categóricas. Uma tabela de contingência bidimensional apresenta células N_{ij}. Com o objetivo de selecionar aleatoriamente amostras para se realizar uma análise detalhada (malha fina), podemos adotar várias estratégias de amostragem: a) multinomial ou hipergeométrica, dependendo dos tamanhos de N_{ij}. Caso sejam relativamente grandes, pode-se adotar o modelo multinomial; b) se as declarações são retiradas aleatoriamente em um período fixo de tempo (por exemplo, os dois últimos dias do prazo para entrega das declarações) não teremos o tamanho *n* da amostra fixado. Em cada célula, o número de declarações pode ser considerado Poisson, independentemente das outras células e a distribuição conjunta será produto de Poisson. No entanto a distribuição condicional das distribuições Poisson mutuamente independentes, dado o tamanho da amostra obtido, passa a ser multinomial; c) a população pode ser dividida em relação às categorias de uma das variáveis categóricas (por exemplo, o estado de residência). Amostras aleatórias de tamanhos pré-determinados são então selecionadas de cada sub-população (estado da federação). A distribuição neste caso é produto de multinomiais ou produto de hipergeométricas. Tais distribuições podem ser obtidas da multinomial ou hipergeométrica, condicionada pelos tamanhos amostrais das sub-populações (totais marginais de linha ou de coluna).

Amostragens complexas, envolvendo múltiplos estágios de estratificação e/ou aglomeração, estratégias muito utilizadas em inquéritos amostrais de larga escala, irão exigir modelos probabilísticos mais sofisticados. (Paulino e Singer, 2006; pp 35, 40-41).

2. Estimadores de MV para cada modelo

- *Poisson*: As frequências nas células (*nij)* são VA com valores esperados *E[nij]=*μ_{ij}. Os estimadores para os parâmetros das distribuições independentes Poisson são $\hat{\mu}_{ij} = \frac{n_{i+} \times n_{+j}}{n}$. Os valores verdadeiros nas células, ou seja, *E[nij],* devem satisfazer a hipótese de independência, dada por:

$$\mu_{ij} = E[n_{ij}] = \frac{E[n_{i+}]E[n_{+j}]}{E[n]E[n]} E[n] = \frac{E[n_{i+}]E[n_{+j}]}{E[n]}$$

- *Multinomial*: os estimadores para os parâmetros são $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$

- *Produto de Multinomiais*: Os estimadores para os totais marginais sem restrição, $\widehat{fi+}$ ou $\widehat{f+j}$, são dados por n_{i+}/n ou n_{+j}/n, respectivamente. As frequências esperadas sob a hipótese de independência são estimadas por ($n_{i+} \times n_{+j}$) / n, como nos casos anteriores, sendo a hipótese conhecida como homogeneidade de proporções de linha ou de coluna.

Tomando os marginais de coluna (n_{+j}) fixados, como no exemplo da Tabela 3, as proporções de coluna esperadas a serem estimadas devem ser homogêneas em todas as colunas. Estaremos, nessa situação amostrando independentemente a partir de *J* sub-populações. Logo a hipótese de independência $\pi_{ij} = \pi_{i+} \times \pi_{+j}$ deve ser escrita da forma alternativa: $\frac{\pi_{ij}}{\pi_{+j}} = \pi_{i+}$,

que afirma serem as distribuições condicionais para cada nível de *i* em cada coluna *j* equivalentes às distribuições marginais para cada nível de *i*. Os valores esperados de proporções para cada célula sob este modelo são obtidos ao reescrevermos a relação

($n_{i+} \times n_{+j}$) / n sob a forma: $n_{ij} / n_{+j} = n_{i+} / n$, o que é o mesmo que:

$\pi_{1|1} = \pi_{1|2} = \cdots = \pi_{1|J}$ para o caso de tabela 2 × *J*.

As proporções de coluna esperadas assim estimadas n_{ij} / n_{+j} para cada nível de *I* em cada coluna devem ser homogêneas para todas as *J* colunas.

Podemos concluir que todos os modelos amostrais considerados para uma tabela de contingência *I* × *J* apresentam as mesmas estimativas para as frequências observadas sob o pressuposto de independência. Logo, qualquer estatística que utilize tais estimativas, apresentará o mesmo valor para qualquer um dos modelos amostrais discutidos.

Estaremos interessados em duas estatísticas, atendidos alguns pressupostos:

a) A estatística qui-quadrado de Pearson (χ^2 *ou* G^2): $\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n}$

ou, alternativamente, b) A estatística razão de verossimilhança (*likelihood ratio*) (χ^2 *ou* H^2):
$2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} ln\left(\frac{n_{ij}n}{n_{i+}n_{+j}}\right)$.

Ambas apresentam distribuição χ^2 com (*I* – 1) (*J* – 1) graus de liberdade na hipótese de independência, atendidos alguns pressupostos.