Análise de Dados Categóricos – Aula Inicial

Primeira parte: Introdução

1. Ilustração: O que você vê na foto?

Classificação como uma das mais antigas operações realizadas pela humanidade, talvez antes da mensuração por uma escala.

2. Mensuração:

Atribuição de numerais (por exemplo: 3, III, 11) a objetos de acordo com certas regras.

Escalas de Mensuração[1]

Escala de Mensuração	Regras Matemáticas permitidas
1. Nominal	- correspondência um a um
2. Ordinal	- correspondência um a um - relações de ordem (postos ou "rankings") com transformação monotônica
3. Intervalar	- correspondência um a um - atribuição de postos ("ranking") - igualdade de diferenças
4. Razão	- correspondência um a um - atribuição de postos ("ranking") - igualdade de diferenças - divisão e multiplicação

1. e 2. são usualmente chamados de **dados categóricos** ou dados qualitativos.

Dados nas escalas 3 e 4 são às vezes **categorizados.**

Escalonamento: Tem como objetivo melhorar o nível da escala de mensuração dos dados e capturar toda a informação contida nos mesmos[2]. Ex: dados em escala nominal transformados em ordinal, intervalar ou razão. Dados em escala intervalar ou razão em geral não necessitam de escalonamento. Dados ordinais podem ter um escalonamento ótimo.

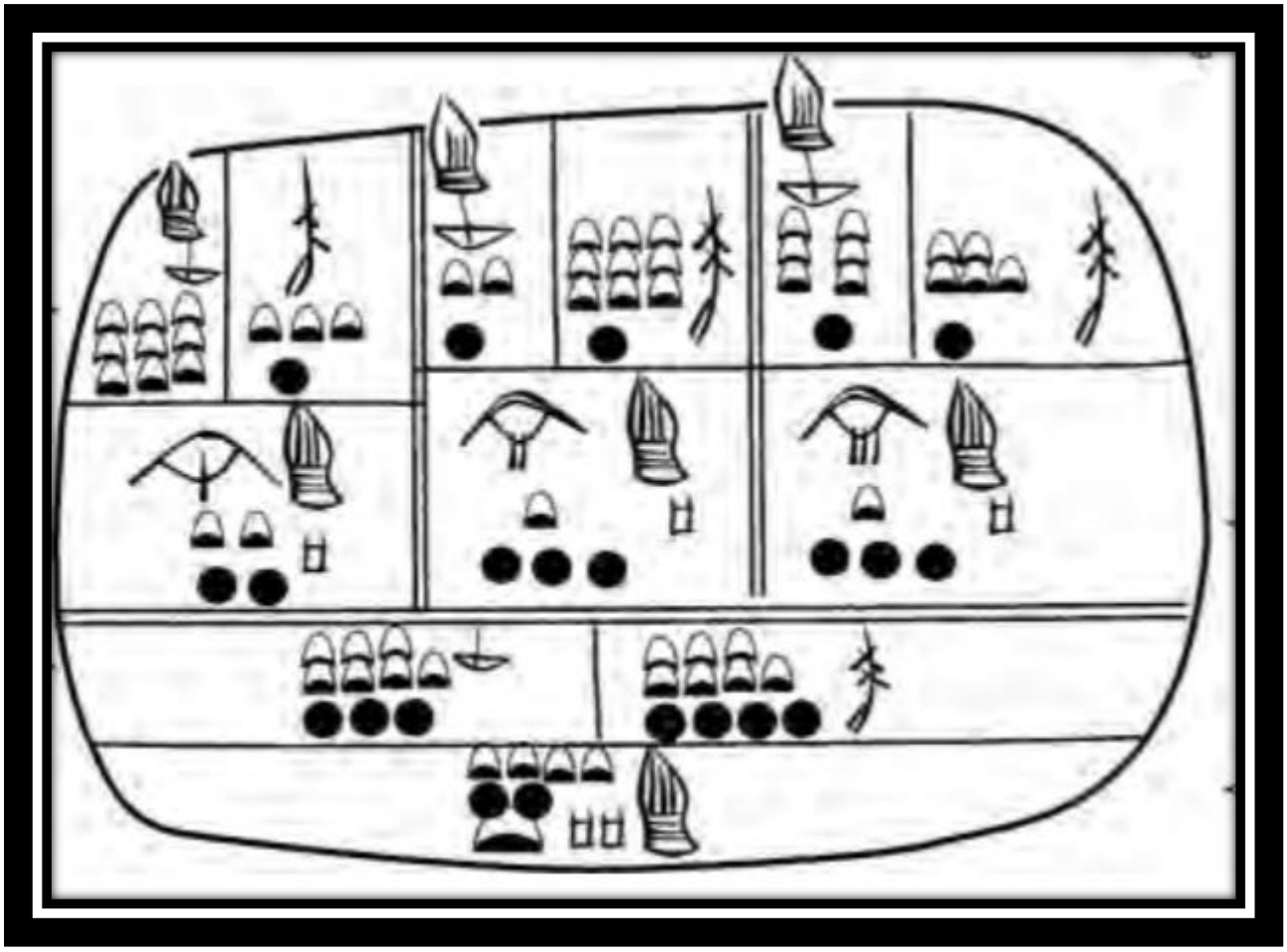
---

[1] Stevens, S.S. 1951. Mathematics, measurement and psychophysics. In Stevens, S.S. (ed.), *Handbook of Experimental Psychology*. New York: Wiley.

[2] Nishisato, S. 2007. Multidimensional nonlinear descriptive analysis. Boca Raton: CRC Press.

3. Classificação e contagem (frequências)

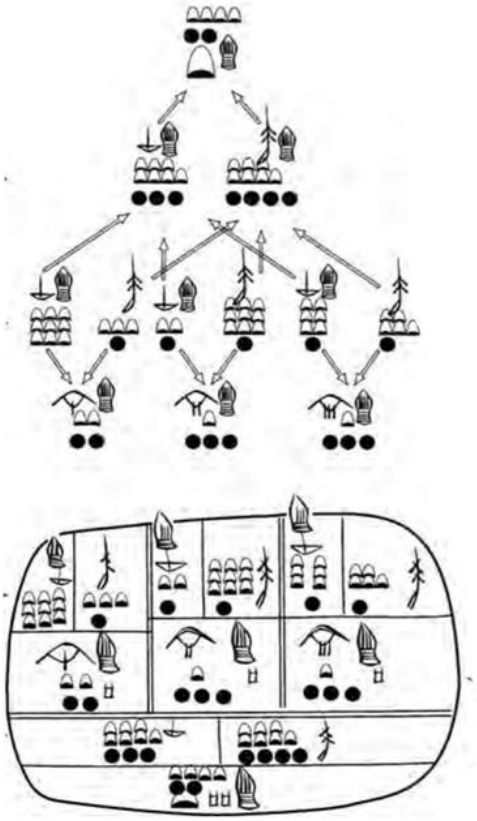
Veja uma simplificação de uma das placas de argila da foto



O que os símbolos representam?

Atribuir números a cada classificação. Escreva na sua cópia da placa. Você vê a indicação de algum padrão conhecido na placa de argila simplificada?

Como poderiam ser representadas, modernamente, as classificações e contagens feitas há mais de 5000 anos atrás?



Construir a tabela que você acredita poder representar melhor as informações da placa de argila.

		Ano da colheita			
		1	2	3	Total
Tipo de cultivar	1	9	12	16	37
	2	13	19	15	47
Total		22	31	31	84

4. Tabelas de contingência:

- Podemos adotar a seguinte notação para os dados de uma tabela de contingência bidimensional:
  $\rightarrow$ n_{ij} representando os ***valores observados*** na célula (i,j)
  $\rightarrow$ o símbolo (+), ou o símbolo (●) representando o somatório na respectiva dimensão (*i* ou *j*, *i*= 1, 2, ..., *I* e *j* = 1, 2, ..., *c*).

  Veja a tabela a seguir para *i*= 1, 2, ..., *I* e *j* = 1, 2, ..., *c:*

		Variável B						Totais marginais das linhas $\sum_J n_{ij} = n_{i+}$
		B_1	B_2	...	B_j	...	B_c	
Variável A	A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	n_{1+}
	A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	n_{2+}
	...	...	...	...	...	...	...	...
	A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	n_{i+}
	...	...	...	...	...	...	...	...
	A_I	n_{I1}	n_{I2}	...	n_{Ij}	...	n_{Ic}	n_{I+}
Totais marginais das colunas $\sum_I n_{ij} = n_{+j}$		n_{+1}	n_{+2}	...	n_{+j}	...	n_{+c}	n_{++}= n

- Notar que as categorias atribuídas a ambas as variáveis devem ser **mutuamente exclusivas** e **exaustivas**.
- Uma indagação motivadora de pesquisa poderia ser:
  "Há independência entre a variável de linha (A) e a variável de coluna (B)?"

Mais formalmente[3], temos a tabela de contingência definida assim:

- As frequências observadas n_{ij} formam um conjunto *O* de $N=n_{++}$ entidades, classificadas ao mesmo tempo em uma partição $I = \{i_1, ..., i_l\}$ e uma partição $J = \{j_1, ..., j_c\}$ de acordo com as categorias de cada uma das duas variáveis.
- Uma das variáveis (linha) apresenta l categorias, $i \in I$, e a outra variável (coluna) *c* categorias, $j \in J$.
- A tabela formada pelos números de co-ocorrência, n_{ij}, ou seja, as cardinalidades das interseções par a par $i \cap j$ ($i \in I; j \in J$), é chamada de tabela de contingência ou de classificação cruzada.
- As cardinalidades das classes $i \in I$ e $j \in J$ são usualmente denominadas marginais, representadas por n_{i+} e n_{+j}, respectivamente.
- As proporções observadas, $p_{ij} = n_{ij}/N$, $p_{i+} = n_{i+}/N$ e $p_{+j} = n_{+j}/N$ são frequentemente utilizadas, constituindo a matriz $P = (p_{ij})$ de proporções.
- Quando o conjunto de entidades *O* é uma amostra aleatória de uma população, a tabela de contingência equivalente $N = (n_{ij})$ ou, mais corretamente, $P = (p_{ij})$, é considerada como uma estimativa da distribuição bivariada associada às duas variáveis categóricas.
- Considere que os símbolos gregos π_{ij}, π_{i+} e π_{+j} representam os parâmetros da distribuição a partir da qual os dados foram observados e obtidos os estimadores p_{ij}, p_{i+} e p_{+j}, respectivamente.

[3] Ver, por exemplo, Mirkin, B. 2001. Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables, *The American Statistician*, Vol. 55, No. 2 (May, 2001), pp. 111-120

- Logo, π_{ij} representa a probabilidade de que uma observação retirada da população seja da i-ésima categoria da variável de linha e da j-ésima categoria da variável de coluna.
- Considere então F_{ij} como sendo a frequência esperada na célula *ij* da tabela ao observarmos *N* casos. Ou seja, $F_{ij} = N \times \pi_{ij}$. Logo, $F_{ij} = E(n_{ij})$.
- Independência estatística pode ser entendida pela hipótese H_0: $\pi_{ij} = \pi_{i+} \times \pi_{+j}$
- Logo, $F_{ij} = N \times \pi_{i+} \times \pi_{+j}$, onde os parâmetros marginais são geralmente desconhecidos.
- As estimativas por MV para as proporções marginais podem ser obtidas empiricamente: $p_{i+} = n_{i+} / N$ e $p_{+j} = n_{+j} / N$. (Veremos nas próximas aulas)
- Então $E_{ij} = N \times p_{i+} \times p_{+j}$. Alternativamente, $E_{ij} = (n_{i+} \times n_{+j})/N$. E_{ij} é o valor esperado (estimado) na célula *ij* **se as duas variáveis são independentes**, estimativa de F_{ij}.

- Podemos realizar então um ***teste de hipóteses não-paramétrico***, no qual as hipóteses são:

  H_0: as variáveis de linha e de coluna que formam a tabela são ***independentes***

  H_1: as variáveis de linha e de coluna que formam a tabela ***não*** são independentes (isto é, as variáveis de linha e de coluna são ***associadas***)

  Resumindo, de uma forma mais intuitiva:

- Observe que (n_{i+}/n_{++}) é uma estimativa da probabilidade de obtermos um item que pertença à categoria A_i e que (n_{+j}/n_{++}) é uma estimativa da probabilidade de obtermos um item que pertença à categoria B_j .
- Portanto, se consideramos H_0 (as variáveis A e B são *independentes*), a probabilidade de obtermos um item em A_i **e** B_j simultaneamente será igual a:

$$P(A_i \text{ e } B_j) = P(A_i) \times P(B_j) = (n_{i+}/n_{++}) \times (n_{+j}/n_{++})$$

- Logo, podemos definir um novo valor na situação H_0 (independência entre as variáveis), o *valor esperado estimado*(e_{ij}) em cada célula da tabela, que tem um total de observações $n_{++}= N$, que seria:

$$e_{ij} = (n_{i+}/n_{++}) \times (n_{+j}/n_{++}) \times n_{++} = (n_{i+} \times n_{+j})/n_{++}$$

Então, podemos utilizar uma métrica que relacione o que foi observado ao que teria sido observado na situação H_0: a *estatística qui-quadrado de Pearson*, que podemos representar por

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}$$

- Pode-se mostrar que se A e B são variáveis independentes e os valores e_{ij} são suficientemente grandes (na prática $e_{ij} \geq 5$), a estatística acima é uma variável aleatória que pode ser representada aproximadamente por uma *distribuição teórica qui-quadrado* com *(l -1) × (c-1) graus de liberdade*.
- Esta distribuição teórica tem como valor esperado (média) o número de graus de liberdade, e é assimétrica positiva (veja tabela anexa, ao final do exemplo)
- A hipótese nula H_0 será rejeitada se a estatística de teste calculada for maior do que o valor crítico de X2:

$$\chi^2_{calc.} \geq \chi^2_{\alpha,(l-1)\times(c-1)GL}$$

- Os valores esperados e_{ij} (valores teóricos) podem ser calculados para a tabela do exemplo:

		Ano da colheita			Total
		1	2	3	
Tipo de cultivar	1	$\frac{(37)(22)}{84} = 9{,}69$	$\frac{(37)(31)}{84} = 13{,}65$	$\frac{(37)(31)}{84} = 13{,}65$	37
	2	$\frac{(47)(22)}{84} = 12{,}31$	$\frac{(47)(31)}{84} = 17{,}35$	$\frac{(47)(31)}{84} = 17{,}35$	47
Total		22	31	31	84

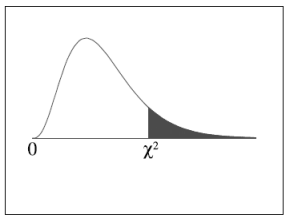
- Podemos agora calcular o valor da estatística de teste qui-quadrado (χ^2):

$$\chi^2 = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} = \frac{(9-9{,}69)^2}{9{,}69} + \frac{(12-13{,}65)^2}{13{,}65}$$
$$+\frac{(13-12{,}31)^2}{12{,}31} + \frac{(19-17{,}35)^2}{17{,}35} + \frac{(15-17{,}35)^2}{17{,}35}$$
$$\chi^2 = 1{,}1662$$

Na ***tabela***, veremos o valor crítico para a regra de decisão:

$$\chi^2_{\alpha=0{,}05,(2-1)\times(3-1)GL} = 5{,}991$$

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188

- Qual a decisão, no nível de significância α = 0,05? Interprete esta decisão.

2ª Parte: Classificação de dados categóricos (Nishisato, 2007)

Como são as estruturas de dados (tabelas ou matrizes) para os diferentes tipos de dados categóricos?

1. Dados de incidência
2. Dados de dominância

1. Dados de Incidência

Seus *elementos* são a ausência (0) ou presença (1) de um atributo, que nos dão as frequências de tais atributos.

1.1 Tabelas de Contingência (bidimensional)

Os dados representam as frequências conjuntas de **dois** conjuntos de categorias.

Exemplo: Tipos de laxantes e efeitos

	Efeito				
Laxante	Nenhum	Leve	Adequado	Demasiado	TOTAL
A	0	3	6	21	30
B	5	15	9	1	30
C	2	18	10	0	30
TOTAL	7	36	25	22	90

1.2 Dados de Múltipla Escolha

Extensão da tabela de contingência bidimensional, usada para mais de duas variáveis categóricas. Difícil representação em tabelas com mais de três variáveis categóricas e maior possibilidade de células em branco. É preferível utilizar a forma de matriz indicadora (*respondentes ou sujeitos* por categorias de

perguntas de múltipla escolha – *objetos,* mutuamente exclusivas e exaustivas).

Ex: **Pergunta1**: Faixa Etária [20-30; 31-40; 41 +]

Pergunta2: Você concorda com nova lei de porte de armas? [sim; não]

Pergunta3: Em que região você mora? [A; B; C; D]

Exemplo de matriz indicadora com os dados:

Obj.	20-30	31-40	41+	Sim	Não	A	B	C	D
1	0	1	0	1	0	1	0	0	0
2	0	0	1	1	0	0	0	1	0
....									
n	1	0	0	0	1	0	1	0	0

1.3 Dados de Separação ("Sorting Data")

Não são largamente utilizados, mas há situações em que são necessários.

Exemplo: Sete disciplinas: A= Inglês; B= História; C= Matemática; D= Física; E=Psicologia; F= Biologia; G= Educação. A um número de alunos ou *sujeitos* (n = 8, por exemplo) é solicitado que comecem com o numeral 1 em qualquer disciplina e então continuar com este numeral para todas as disciplinas similares. Passar então para o numeral 2 e repetir a operação até que todas as disciplinas estejam agrupadas em "pilhas" de similaridade. A decisão sobre o número de "pilhas" e sobre o tamanho das mesmas é arbitrário (não há restrições quanto ao julgamento).

Disciplinas	Alunos							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
A	1	1	2	3	4	3	1	2
B	1	2	2	3	3	3	1	1
C	2	3	1	2	2	2	2	3
D	2	3	1	2	2	2	2	3
E	3	4	2	1	1	3	1	4
F	4	4	2	2	5	1	2	5
G	1	1	2	1	1	3	1	1
No. de Categ.	**4**	**4**	**2**	**3**	**5**	**3**	**2**	**5**

Pergunta:

Como representar as respostas para a linha A: [11234312] em termos de matriz indicadora com as linhas sendo as disciplinas e as colunas sendo os respondentes e suas respectivas categorias?

Resposta:

[(1000), (1000), (01), (001), (00010), (001), (10), (01000)]

2. Dados de dominância

Aqui os *elementos dos dados* são as mensurações ordinais e o objetivo da quantificação é distinto daquele para os dados de incidência. Usado em pesquisas de marketing, por exemplo.

2.1 Dados de comparação par-a-par

Os respondentes (*sujeitos*) decidem para cada par apresentado qual a preferência ou qual elemento do par é mais importante. Para dois *objetos* (X_j, X_k) a resposta do *sujeito i* é codificada assim:

$_if_{jk}$= 1 se $X_j > X_k$

0 se $X_j = X_k$

-1 se $X_j < X_k$

É também comum usar a codificação 1, 2 e 0, para preferência pelo primeiro elemento, segundo elemento e ausência de preferência, respectivamente.

Exemplo: Comparação par-a-par entre ***quatro*** frutas, feita por ***cinco*** indivíduos, para os pares: A: (maçã, pera), B: (maçã, manga), C: (maçã, uva), D: (pera, manga), E: (pera, uva), F: (manga, uva), utilizando a codificação 1, 2 e 0 vista acima.

Sujeitos	Pares de frutas					
	A	B	C	D	E	F
1	1	2	2	1	0	1
2	2	2	2	1	1	1
3	2	2	1	1	0	1
4	2	2	2	2	2	2
5	1	1	1	1	2	2

2.2 Dados ordenados por postos

Respostas codificadas como 1, 2, 3, etc., onde "1" indica a primeira escolha ou a mais preferida, e o numeral maior corresponde à última escolha.

Exemplo: Cinco gerentes (A, B, C, D, E) classificaram sete candidatos a uma vaga de emprego após entrevista. No caso de empate, utiliza-se o posto médio. Por exemplo, se dois candidatos são as primeiras preferências, a eles é dados o posto 1,5. Se os três primeiros estão empatados, cada um recebe o posto 2. Desta forma, a soma de todos os postos é fixa, sendo igual a n(n+1)/2

Cand. / Ger.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
A	3	6	5	4	1	7	2
B	2	7	5	4	3	6	1
C	2	5	6	3	4	7	1
D	3	7	4	5	1	6	2
E	4	6	7	5	2	3	1

2.3 Dados com categorias sucessivas

Em essência, o mesmo que dados de incidência do tipo múltipla-escolha, só que o mesmo conjunto de categorias sucessivas

ordenadas é utilizado para o julgamento de todas as perguntas (é muito utilizado em "*surveys*" sobre atitudes, satisfação, etc.).

Exemplo: Considere o seguinte conjunto de categorias: 1 = Baixa; 2 = Média; 3 = Alta. Cinco indivíduos são questionados quanto à motivação para: fazer exercícios físicos, fazer dieta, fazer tratamento clínico contra obesidade, fazer tratamento cirúrgico contra obesidade.

Sujeitos	Atividade			
	Ex. Fís.	Dieta	Trat. Clín.	Trat. Cir.
1	1	1	3	3
2	1	1	2	3
3	2	1	3	3
4	2	2	2	2
5	1	1	2	2

Por que não nos referimos a esta tabela como contendo dados de múltipla escolha apenas?

Podemos fazer o escalonamento dos "tratamentos" (atividades), através de algum critério, mas também os limites (fronteiras) das categorias, um entre "Baixa" e "Média" e outro entre "Média" e "Alta".

Sendo assim, teremos os dados convertidos a dados *ordenados* por postos, tanto para os limites das categorias quanto para os "tratamentos". Esta é a razão para considerarmos este caso especial de dados de múltipla-escolha como dados de dominância.

Leitura complementar: *A Matemática da Escolha Social*, de Steffenon e Jabuinski (arquivo pdf anexo). Seis métodos de escolha majoritária (1.3.1 a 1.3.6) para discussão em aula.