Bioestatística

Isolde Previdelli
itsprevidelli@uem.br
isoldeprevidelli@gmail.com

AULA 4 - Análise Bivariada

09 de Março de 2017

Sumário

Coeficientes - Estimativas pontuais

- Covariância
- Coeficiente de Contigência C
- Correlação linear de *Pearson*
- Correlação por postos de *Spearman*
- Correlação por postos de *Kendall*
- Coeficiente de concordância de *Kappa*

Objetivos

Objetivo Geral

- Descrever o comportamento de uma variável resposta em função de outra (variável independente), isto é, resumir a variável resposta na presença de outra.

Objetivos específicos

- Construir tabelas de dupla entrada
- Construção de gráficos
  - gráfico de dupla barras
  - diagrama de dispersão
  - diagrama de caixas ou boxplot

Tabelas de duas entradas ou contingência

- A ocorrência simultânea de duas variáveis qualitativas são dispostas em tabela.
- Coeficiente de Contingência C.

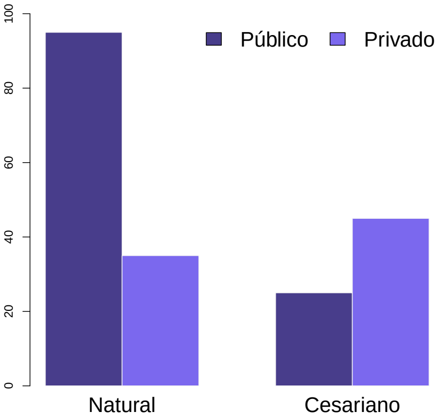
Exemplo: Deseja-se analisar o comportamento conjunto das variáveis tipo de parto e tipo de hospital, usando uma amostra de 200 pacientes de Hospitais públicos e privados.

Hospital	Parto		Total
	Natural	Cesariano	
Público	95	25	120
Privado	35	45	80
Total	130	70	200

Gráfico de barras duplas

Exemplo

Utilizado quando a variável resposta e a variável independente são qualitativas. A altura da barra representa freqüência da ocorrência simultânea das duas categorias, das variáveis qualitativas.



Coeficiente de contingência C

- É uma medida de associação entre duas variáveis em escala nomimal
- É um coeficiente de correlação não-paramétrico
- O cálculo do coeficiente de contingência é feito pela expressão abaixo:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

em que $\chi^2 = \sum_i \sum_j = \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

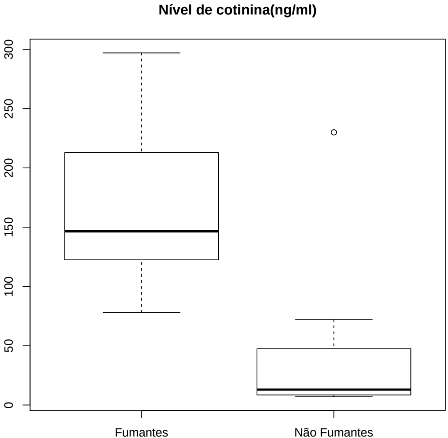
- O_{ij} = frequência onservada
- E_{ij} = frequência esperada, $E_{ij} = \frac{n_{i.} * n_{.j}}{n}$

em que

- $n_{.j}$ = totais marginais(colunas)
- $n_{i.}$ = totais marginais(linhas)

Boxplot

Em uma investigação dos fatores de risco para as doenças cardiovasculares, os níveis séricos de cotinina - produto metabólico da nicotina - foram registrados para um grupo de fumantes e um grupo de não-fumantes.



Covariância e Coeficientes de correlação linear

- Primeiro calcula-se a covariância
- Coeficiente de correlação linear de *Pearson*
  - Paramétrico, dados emparelhados (x,y) independentes
- Coeficiente de correlação por postos de *Spearman*
  - Não-Paramétrico, dados emparelhados (x,y) independentes, mas geralmente ordinais
- Coeficiente de correlação por postos de *Kendall*
  - Não-Paramétrico, útil para o mesmo tipo de dados ao qual se aplica correlação de Spearman

Covariância

- A covariância mede a relação linear entre duas variáveis
- A covariância dá uma ideia da dispersão dos valores da variável bidimensional
- A covariância não é padronizada

- O cálculo da covariância é feito pela expressão abaixo:

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Coeficiente de correlação linear de Pearson

Suposições

- A amostra de Dados Emparelhados (x,y) é independente.
- Os pares de dados (x,y) têm distribuição Normal bivariada

Para que serve?

- Mede o grau de relacionamento linear entre os valores emparelhados x e y em uma amostra.
- Utilizaremos o coeficiente de correlação linear para detectar padrões lineares.

Coeficiente de correlação linear de Pearson

- Para verificar se exite correlação linear positiva, negativa ou inexistente entre duas variáveis quantitativas aplicam-se:
  - Gráfico: Diagrama de dispersão.
- Coeficiente de correlação linear:
  - Correlação amostral é denotada pela letra "r";
  - r - estimador;
  - ρ - parâmetro.

Coeficiente de correlação linear de Pearson

Para verificar o grau de relacionamento linear entre os valores emparelhados x e y em uma amostra e verificar se existe correlação linear positiva, negativa ou inexistência de correlação calcula-se:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i}(y_i - \bar{y})^2}}$$

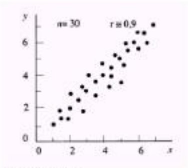
Ou ainda,

$$r = \frac{Cov(x, y)}{s_x s_y}$$

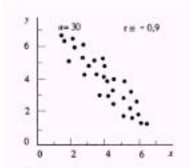
Estatística r:

- é adimensional
- indicao grau de correlação entre duas amostras independentes
- varia de -1 a 1
- se "r"for zero, não há relação linear entre as variáveis

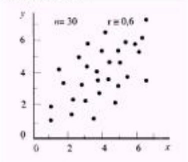
Diagrama de dispersão



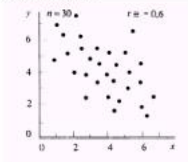
.1 - Correlação Positiva



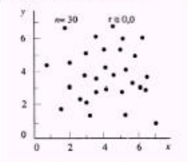
.2 - Correlação Negativa



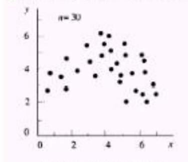
.3 - Pode haver Correlação Positiva



.4 - Pode haver Correlação Negativa



.5 - Não Há Correlação



.6 - Não Há Correlação


UM BOM EXEMPLO DE CORRELAÇÃO NEGATIVA ($r_{XY} < 0$) É DADO POR X = IDADE DE UMA PESSOA E Y = EXPECTATIVA DE VIDA DESSA MESMA PESSOA.
A RAZÃO É SIMPLES: QUANTO MAIS UMA PESSOA VIVE, MENOS ANOS TEM PELA FRENTE — JÁ QUE A VIDA É LIMITADA.
GLUP! ISTO QUER DIZER QUE FALTA POUCO!!...

Coeficiente de correlação por postos de Spearman

É utilizado quando as duas variáveis, x e y:

- São números com uma ordem de uma classificação
- Não apresentam distribuição conjunta normal bivariada

Ou ainda

- Nos casos dos dados não formarem uma nuvem comportada com pontos distantes dos demais, parece existir uma relação crescente ou decrescente num formato de curva
- Desvantagens: Não leva em consideração os pesos amostrais

Coeficiente de correlação por postos de Spearman

Passos para calcular o coeficiente:

- Ordenar todos os valores em cada amostra (X_i, Y_i)
- Atribuir números de ordem (atribuindo o número de ordem médio quando há empate) para cada amostra $Posto_{X_i}$ e $Posto_{Y_i}$

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{(n^3 - n)}$$

em que $d_i = X_i - Y_i$

Coeficiente de correlação por postos de Kendall

Passos para calcular o coeficiente:

- Primeiro reordenamos os valores de modo que aparecam a ordem natural(1,2,3,..,n) para X_i
- Em seguida calculamos $S = \sum_{i=1}^{n} S_i$

Veja o exemplo abaixo sobre notas atribuídas por dois médicos a 4 imagens de ressonância magnética:

Imagens	d	c	a	b
Médico X	1	2	3	4
Médico Y	2	4	3	1

Coeficiente de correlação por postos de Kendall

- 1^0 escore do médico Y é 2 e os pares que podemos formar são:(2,4);(2,3) e (2,1). Para o par que possui 1^0 valor $< 2^0$ valor atribuímos o valor 1 e em caso contrário -1. Nesse caso temos, $S_1 = 1 + 1 - 1 = 1$
- 2^0 escore do médico Y é 4 e então $S_2 = -3$
- 3^0 escore do médico Y é 3 e então $S_3 = -1$

Assim $S = 1 - 3 - 1 = -3$
Então calculamos o grau de correlação:

$$\tau = \frac{S}{\frac{1}{2}(n(n-1))}$$

Portanto, $\tau = -0,5$

Útil para avaliar concordância entre observadores, avaliadores, etc.

- Observador é uma fonte de erro
- Por exemplo, examinar um raio X ou realizar um exame físico
- Em geral os dados são apresentados em uma tabela de contigência *SxS*
- *Kappa* não é indicado para tabelas $2x2$ (Neste caso usa-se McNemar)

Coeficiente de *Kappa*:

$$\hat{k} = \frac{\Pi_0 - \Pi_e}{1 - \Pi_e}$$

Sendo $\Pi_0 = \sum_{i=1}^{s} p_{ii} = \sum_{i=1}^{s} = \frac{n_{ii}}{n}$, a probabilidade de concordância com, p_{ii} a probabilidade de um indivíduo ser classificado na categoria i por ambos os observadores e ,

$\Pi_e = \sum_{i=1}^{s} (p_{i+})(p_{+i}) = \sum_{i=1}^{s} \frac{n_{i+}n_{+i}}{nn}$

($\hat{k} = 1$, concordância perfeita, $\hat{k} = 0$, discordância total)

Exemplo(Giolo S.R., 2004): Concordância entre diagnóstico de dois neurologistas. Os dados abaixo se referem a classificação de pacientes com esclerose múltipla, em 4 classes de diagnóstico, por dois neurologistas.

Neurologista 2	Neurologista 1				Totais
	1	2	3	4	
1	38	5	0	1	44
2	33	11	3	0	47
3	10	14	5	6	35
4	3	7	3	10	23
Totais	84	37	11	17	149

Obteve-se:

$\Pi_0 = (38 + 11 + 5 + 10)/149 = 0,4295$

$\Pi_e = ((44 * 84) + (47 * 37) + (35 * 11) + (23 * 17))/149^2 = 0,2798$

$$\hat{k} = \frac{0,4295 - 0,2798}{1 - 0,2798} = 0,2079$$

Tal resultado indica uma fraca concordância entre os neurologistas


Statistics
organization
descriptive
scientific
collection
interpretation
data
statistical
Inference
theory
experiments
science
mathematical
discipline
probability
design
parameters
mathematics
provides
called
population
prediction
samples
analysis
word
total
statistic

Obrigada!