Introdução

Dados Categorizados

- Uma variável categorizada é aquela na qual a escala de medida consiste em um conjunto de categorias

- Exemplos:
  - Opinião sobre horário eleitoral –a favor ou contra
  - Grau de instrução: analfabeto, fundamental, médio, superior
  - Diagnóstico de uma doença - presente ou ausente
  - Faixa de Renda - 0 a 2 S.M, 2 a 4 S.M, > 4 S.M

Variável Resposta/Explicativa

- Muitas análises estatísticas fazem distinção entre variáveis resposta e explicativas.

- Exemplos:
  - Modelos de Regressão descrevem como a distribuição de uma variável resposta contínua, renda anual, mudam de acordo com os níveis de algumas variáveis explicativas, tais como: número de anos de estudo e número de anos de trabalho

Variável Resposta/Explicativa

- **Variável Resposta** ⇒ Variável Dependente
- **Variável Explicativa** ⇒ Variável Independente ou preditora

- Modelos Estatísticos para variáveis respostas categorizadas analisam como as respostas são influenciadas pelas variáveis explicativas. As variáveis explicativas podem ser categorizadas ou contínuas

- ***Exemplo*** -Estudo sobre a associação entre opinião sobre maioridade penal (a favor, contra) e os seguintes fatores: sexo, raça (branca, negra, outra), religião (católica, evangélica, outra), grau de instrução (fundamental, médio e superior)

Escala de Mensuração das Variáveis Categorizadas

- Existem dois tipos de escalas para variáveis categorizadas
  - **Nominal** - na qual as categorias da variável não apresentam uma ordenação natural
    - Religião: Católica, Budista, Espírita, Protestante
    - Tipo de Música Favorita: Clássica, Sertaneja, Rock, Jazz
  - Para variáveis nominais, a ordem das categorias é irrelevante e a análise estatística não deveria depender da ordenação. Os métodos de análise fornecem os mesmos resultados independente da ordem em que as categorias são apresentadas.

Escala de Mensuração das Variáveis Categorizadas

- **Ordinal** - na qual as categorias da variável apresentam uma ordenação natural
  - Opinião sobre a legalização do aborto: desaprovação total, aprovação parcial e aprovação total
  - Diagnóstico se um paciente é doente mental: certo, provável, improvável e definitivamente não
- Para variáveis ordinais os métodos estatísticos levam em consideração a ordem das categorias. Se listarmos as categorias de forma crescente ou decrescente os resultados não se alteram. No entanto, se reordenamos de uma outra forma os resultados podem ser mudados.

Escala de Mensuração das Variáveis Categorizadas

- Métodos designados para variáveis ordinais não podem ser usados com variáveis nominais.
- Métodos para variáveis nominais podem ser usados para variáveis ordinais. No entanto, quando usamos estes métodos com variáveis ordinais, devemos saber que eles não levam em consideração a ordenação das categorias. Isto pode resultar em uma perda de poder.
- Variáveis Categorizadas são frequentemente referidas como *variáveis qualitativas*, para distinguir daquelas que apresentam valores numéricos, ou *variáveis quantitativas*. Entretanto, é normalmente vantajoso tratar dados ordinais de maneira quantitativa, designando-se escores ordenados as categorias.

Distribuições de Probabilidade para Dados Categorizados

- Na análise de Dados Categorizados, ou de qualquer outro tipo de dado, são fixadas suposições sobre o processo aleatório que gera os dados.
- Nos modelos para dados contínuos de regressão e de análise de variância, a distribuição normal desempenha um papel central.
- Nos modelos para dados categorizados três distribuições desempenham papel importante no mecanismo aleatório gerador dos dados. São elas:
  - Distribuição de Poisson
  - Distribuição Binomial
  - Distribuição Multinomial

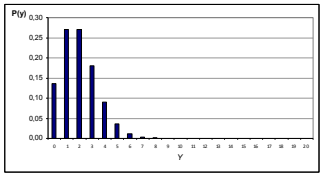
Distribuição de Poisson

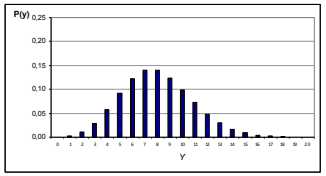
- Y: número de sucessos por intervalo de tempo, área ou volume

$$P(Y=y)=\frac{\mu^y e^{-\mu}}{y!} \qquad y=0,1,2,...$$

onde $\mu=\lambda t$ e λ é a taxa de ocorrências por unidade de tempo ou área ou volume

Tem-se que: $E(Y)=\mu$ e $V(Y)=\mu$





- Uma característica da distribuição é que sua variância cresce a medida que a média aumenta.

Distribuição de Poisson

- **Exemplo:**
- Considere uma rodovia com alto tráfego de veículos pesados e de passeio. Suponha que pesquisadores desejam estudar a taxa de acidentes fatais nesta rodovia. O estudo irá catalogar, no próximo ano todos os acidentes fatais ocorridos na rodovia. Os dados consistirão de contagens semanais.
- A distribuição de Poisson é um potencial modelo probabilístico para o número de acidentes fatais em uma dada semana. Ela é indexada pelo parâmetro μ, sua média.

Distribuição de Poisson

Seja Y o número de acidentes fatais por semana com distribuição de Poisson.

A probabilidade é dada por:

$$P(y)=\frac{e^{-\mu}\mu^y}{y!} \qquad y=0,1,2,\cdots$$

Suponha que os acidentes fatais ocorram a uma taxa de 2 por semana. A probabilidade de que não ocorra acidentes fatais é dada por:

$$P(0)=\frac{e^{-2}2^0}{0!}=0,135$$

Ao longo do período de t semanas mantendo-se constante a taxa média de acidentes fatais em dois por semana, o modelo de Poisson para o número total de acidentes fatais tem média igual a 2t.

Distribuição Binomial

No exemplo dos acidentes fatais, o número de acidentes semanais é aleatório e não fixo. O número de acidentes semanais não fatais também é aleatório. Antes de uma particular semana, não sabemos quantos acidentes irão ocorrer.

Todavia, muitas aplicações tem fixados o tamanho da amostra. Por exemplo, suponha que os pesquisadores planejem classificar o resultado de todos os acidentes até que *n* ocorram, visando estimar a proporção π de acidentes fatais. O tamanho da amostra é então fixo e d igual a *n*. Neste delineamento o número de acidentes fatais não tem distribuição de Poisson.

Distribuição Binomial

- Considere *n* ensaios independentes e identicamente distribuídos, com dois resultados possíveis (sucesso e falha).
- Estes ensaios são frequentemente chamados de ensaios de Bernoulli.
- Seja π a probabilidade de sucesso em cada ensaio.
- Se Y o número de sucessos nos *n* ensaios. A variável Y tem **distribuição Binomial** comparâmetros *n* e π.

Distribuição Binomial

- *Se a probabilidade de sucesso é* π *em cada um dos* n *ensaios idênticos independentes, então :*

  Y : número de sucessos nos n ensaios

  tem distribuição Binomial com parâmetros n *e* π

$$Y \sim Bin(n,\pi)$$

- A função de probabilidade de *Y* é:

$$P(Y=y)=\binom{n}{y}\pi^{y}(1-\pi)^{n-y} \qquad y=0,1,2,\ldots,n$$

$$onde: \quad \binom{n}{y}=\frac{n!}{y!(n-y)!}$$

Tem-se que:

$$E(Y)=n\pi \qquad e \qquad V(Y)=n\pi(1-\pi)$$

Distribuição Binomial

- **Exemplo:**

Para ilustrar, seja Y o número de acidentes fatais em 10 acidentes.

A probabilidade de sucesso (acidentes fatais) é 0,2 para cada acidente.

Então *n* = 10 e π = 0,2.

A probabilidade de Y = 0 é dada por:

$$P(0)=\frac{10!}{(0!10!)}(0,2)^{0}(0,8)^{10}=0,107$$

O número médio de acidentes fatais será :

$$E(Y)=n\pi=10x0,2=2 \qquad com \qquad V(Y)=n\pi(1-\pi)=10x0,2x0,8=1,6$$

- Diferente da Poisson, a variância da Binomial é menor do que a média.

Observações importantes

- *Em dados categorizados freqüentemente olha-se par o número de sucessos e de insucessos simultaneamente.*

Se

$$Y_1 \rightarrow número\ de\ sucessos$$

$$Y_2 \rightarrow número\ de\ insucessos$$

então :

$$Y_2 = n - Y_1$$

e $\quad Y_2 \sim Bin(n, 1-\pi)$

- *Portanto:*

$$E(Y_2)=n(1-\pi)$$

$$V(Y_2)=n(1-\pi)\pi=V(Y_1)$$

$$cov(Y_1,Y_2)=-n\pi(1-\pi)$$

$$corr(Y_1,Y_2)=-1$$

- Ambas as distribuições Binomial e Poisson se aproximam da distribuição Normal quando N e a média aumentam

Distribuição Multinomial

Alguns ensaios possuem mais de dois resultados. Por exemplo, para cada motorista em cada acidente podemos categorizar em:

- sem ferimentos,
- com ferimentos sem hospitalização,
- com ferimentos com hospitalização
- com ferimento fatal.

A distribuição de probabilidade em várias categorias é **multinomial**. A distribuição Binomial é uma caso particular.

Distribuição Multinomial

- *A distribuição MULTINOMIAL é uma generalização da Binomial para mais de 2 categorias.*
- *Suponha que se tem **n** ensaios idênticos e independentes. Em cada ensaio considere que um de **k** categorias :*

$$O_1, O_2, \ldots, O_k$$

pode ocorrer, com probabilidade

$$\pi_j = P(O_j), \qquad 0 < \pi_j < 1$$

e

$$\pi_1 + \pi_2 + \ldots + \pi_k = 1$$

- *Seja*

$$Y_j \rightarrow \text{número de vezes que a categoria } O_j \text{ é observada em n ensaios}$$

- *O vetor*

$$Y' = (Y_1, Y_2, \ldots, Y_k), \quad com \quad \sum_{j=1}^{k} Y_j = n$$

- *É chamado vetor multinomial com **k** categorias e **n** ensaios.*

Distribuição Multinomial

- *A distribuição do vetor* $\mathbf{Y}$ *, isto é, a distribuição conjunta de* $Y_1, Y_2, \ldots, Y_k$ *é chamada DISTRIBUIÇÃO MULTINOMIAL com parâmetros*

$$n, \pi_1, \pi_2, \ldots, \pi_k, \quad \sum_{j=1}^{k} \pi_j = 1.$$

$$(Y_1, Y_2, \ldots, Y_k) \sim Mult(n, \pi_1, \pi_2, \ldots, \pi_k)$$

e

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_k = y_k,) = \frac{n!}{y_1! y_2! \ldots y_k!} \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_q^{y_k} (1-\pi)^{n-y}$$

$$onde: \quad \sum_{j=1}^{k} x_j = n \quad e \quad \sum_{j=1}^{k} \pi_j = 1$$

- ***Propriedades:***

$$E(Y_j) = n\pi_j \qquad e \qquad V(Y_j) = n\pi_j(1-\pi_j) \qquad \forall j = 1,2,\ldots,k$$

$$\text{cov}(Y_i, Y_j) = -n\pi_i\pi_j \qquad e \qquad corr(Y_i, Y_j) = -\sqrt{\frac{\pi_i \pi_j}{(1-\pi_i)(1-\pi_j)}} \qquad \forall i \neq j$$

Observações importantes

- ***lim*** $\mathbf{Bin(n,\pi) \sim Poisson(\mu)}$

$$\left.\begin{array}{l} \mathbf{n \rightarrow \infty} \\ \boldsymbol{\pi} \rightarrow \mathbf{o} \end{array}\right\} \textbf{de mod o que } \mathbf{n\pi \rightarrow \mu}$$

- $Y_1, Y_2, \ldots, Y_k \quad v.a.\ independentes\ com\ Y_i \sim Poisson(\mu_i)$
  $então \quad Y_1 + Y_2 + \ldots + Y_k \sim Poisson(\mu_1 + \mu_2 + \ldots + \mu_k)$