Estatística Descritiva e Exploratória

Gledson Luiz Picharski
e
Wanderson Rodrigo Rocha

Universidade Federal do Paraná

9 de Maio de 2008

Estatística Descritiva e exploratória

Introdução

Podemos associar números a eventos aleatórios, como no caso do lançamento de uma moeda duas vezes.

$$\Omega = \{(c,k)(k,c)(c,c)(k,k)\}$$

Sendo X a quantidade de caras, teremos: $\mathbb{X} = \{0,1,2\}$

$$X(k,k) = 0$$
$$X(c,k) = X(k,c) = 1$$
$$X(c,c) = 2$$

Função Discreta de Probabilidade

A função que atribui a cada valor da variável aleatória sua probabilidade é denominada de função discreta de probabilidade ou, simplesmente, função de probabilidade.

$$P(X = x_i) = p(x_i) = p_i, i = 1, 2, \ldots$$

Ou ainda:

X	x_1	x_2	x_3	$\ldots$
p_i	p_1	p_2	p_3	$\ldots$

Onde seja satisfeito: $0 \leq p_i \leq 1$ e $\sum p_i = 1$

Função Distribuição de Probabilidade

A função distribuição ou função acumulada, refere-se a probabilidade até um certo valor da variável.

$$F(X) = P(X \leq x)$$

Sabendo este conceito fica fácil obter a acumulada apartir de uma densidade.

Medidas de Posição para v.a. Discretas

No caso de conhecermos a distribuição de uma variável, podemos obter as medidas de tendencia central com o uso das probabilidades.

X	2	5	8	15	20
p_i	0,1	0,3	0,2	0,2	0,2

$$E(X) = 2 \times 0,1 + 5 \times 0,3 + 8 \times 0,2 + 15 \times 0,2 + 20 \times 0,2$$

Neste exemplo a esperança de X é calculada da mesma forma que uma média ponderada, que neste caso está sendo ponderada pelas probabilidades de ocorrencia do evento.

Medidas de Posição para v.a. Discretas

Para obter a mediana, devemos verificar que 50% do conjunto para cada lado.

x	$P(X \leq x)$	$P(X \geq x)$
2	0,1	1,0
5	0,4	0,9
8*	0,6*	0,6*
15	0,8	0,4
20	1,0	0,2

$Md = 8$, pois:

$P(X \leq 8) \geq 0,5$ e
$P(X \geq 8) \geq 0,5$

Medidas de Posição para v.a. Discretas

$$E(X) = \mu = \sum_{i=1}^{k} x_i p_i$$

$$E(X) = \mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$P(X \geq Md) \geq 1/2 \quad \text{e} \quad P(X \leq Md) \geq 1/2$$

$$P(X = Mo) = max(p_1, p_2, \ldots, p_n)$$

Medidas de Dispersão para v.a Discretas

- Variância

$$\sigma^2 = Var(X) = \sum_{i=1}^{k}(x_i - \mu)^2 p_i$$

$$\sigma^2 = Var(X) = \sum_{i=1}^{n} x_i^2 p_i - (\sum_{i=1}^{n} x_i p_i)^2 = \sum_{i=1}^{n} x_i^2 p_i - \mu^2$$

$$Var(X) = E(X^2) - [E(X)]^2$$

- Desvio Padrão

$$Sd(X) = Dp(X) = \sigma = \sqrt{\sigma^2}$$

Freqüência Esperada x Freqüência Observada

Caso haja conhecimento sobre o modelo probabilístico, pode-se avaliar a aderência de dados amostrais à este modelo.

Exemplo:
Num estudo sobre a incidência de câncer foi registrado, para cada paciente com esse diagnóstico, o número de casos de câncer em parentes próximos (pais, irmãos, filhos, primos e sobrinhos). Os dados de 26 pacientes são os seguintes:

Freqüência Esperada x Freqüência Observada

Estudos anteriores assumem que a incidência de câncer e, parentes próximos pode ser teoricamente modelada pela seguinte função discreta de probabilidade:

Incidência	0	1	2	3	4	5
p_i	0,1	0,1	0,3	0,3	0,1	0,1

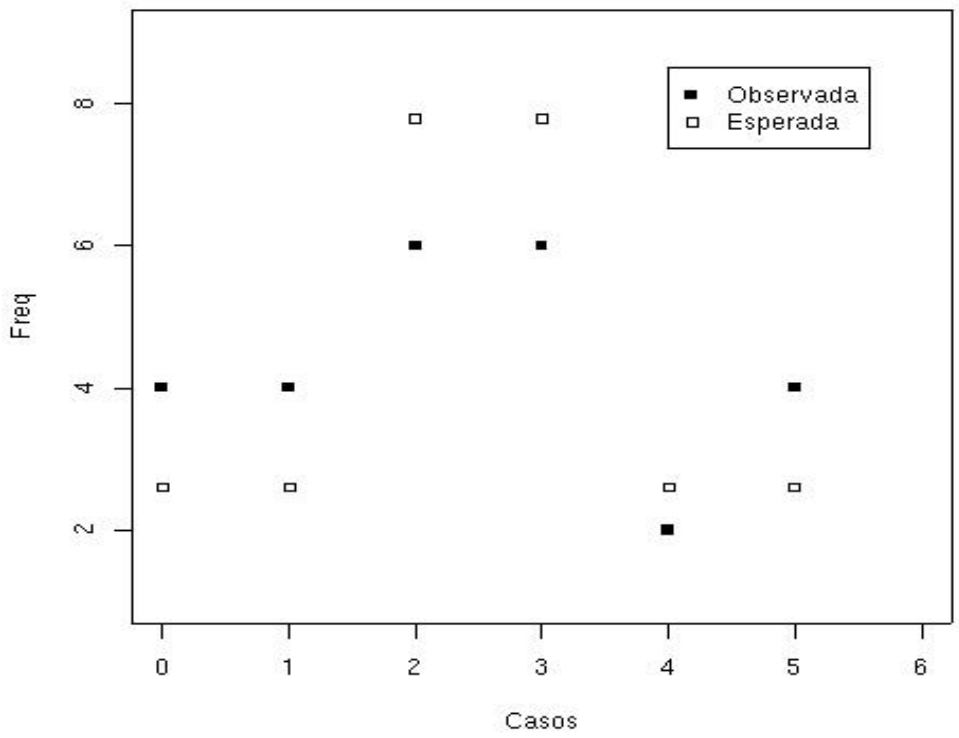
Freqüência Esperada x Freqüência Observada

Fazendo comparação dos dados obtidos com o modelo teórico podemos observar a tendencia dos dados.

Incidência	n_i	e_i
0	4	2,6
1	4	2,6
2	6	7,8
3	6	7,8
4	2	2,6
5	4	2,6
Total	26	26

Freqüência Esperada x Freqüência Observada

O gráfico mostra uma comparação entre os dados observados e esperados.



Modelo Uniforme Discreto

Seja X uma variável aleatória discreta cujos possíveis valores são representados por $x_1, x_2, x_3, \cdots, x_k$.
Dizemos que X segue o modelo Uniforme Discreto se sua função de probabilidade é dada por:

$$P(X = x_j) = \frac{1}{k}, \forall j = 1, 2, 3, \cdots, k.$$

Modelo Bernoulli

Em muitas situações práticas a variável de interesse assume somente dois valores:

- uma peça é classificada como boa ou defeituosa;
- o entrevistado concorda ou não com a afirmação feita;
- a vacina imunizou ou não a criança.

Estas situações têm alternativas dicotômicas, que genericamente podem ser representadas por respostas do tipo sucesso-fracasso. Experimentos deste tipo recebem o nome de Ensaios de Bernoulli e dão origem a uma variável aleatória com o mesmo nome

Modelo Bernoulli

Com p representando a probabilidade de sucesso, $0 \leq p \leq 1$, sua função discreta de probabilidade é dada por:

$$P(X = x) = p^x(1-p)^{1-x}, x = 0, 1.$$

OBS: A repetição de ensaios de Bernoulli independentes dá origem à mais importante variável aleatória discreta cujo modelo é denominado Modelo Binomial.

Modelo Binomial

Considere a repetição de n ensaios de Bernoulli independentes e todos com a mesma probabilidade de sucesso p.
A variável aleatória X que conta o número total de sucessos é denominada Binomial com parâmetros n e p e a denotaremos por $X \sim b(n,p)$.
Sua função de probabilidade é dada por:

$$P(X=k) = \left(\begin{array}{c} n \\ k \end{array} \right) \times p^k \times (1-p)^{n-k}, k = 0, 1, 2, \cdots, n.$$

Modelo Geométrico

Dizemos que uma variável aleatória X tem distribuição Geométrica de parâmetro p, ie $X \sim G(p)$, se sua função de probabilidade tem a forma

$$P(X = k) = p(1-p)^k, 0 \leq p \leq 1, k = 0, 1, 2, \cdots .$$

Interpretando p como a probabilidade de sucesso, a distribuição Geométrica pode ser pensada como o número de ensaios de Bernoulli até o primeiro sucesso.

Modelo de Poisson

Uma variável aleatória X tem distribuição de Poisson com parâmetro $\lambda > 0$, ie $X \sim Po(\lambda)$, se sua função de probabilidade é dada por

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0, 1, 2, \cdots,$$

com o parâmetro λ sendo usualmente referido como a taxa de ocorrência ou também a frequência média ou esperada de ocorrências num determinado intervalo de tempo.

Estatística Descritiva e exploratória

Introdução

Em diversas análises são comuns o estudo de muitas variáveis, ao aplicarmos um questionário por exemplo, o interesse pode estar em registrar: sexo, idade, renda, time de preferência, etc. Neste caso, cada respondente tem associado a si um vetor de informações que representa uma observação multidimensional. A partir disso podemos estudar conjuntamente as diversas variáveis aplicando ferramentas estatísticas adequadas.

Função de probabilidade conjunta

Sejam X e Y duas variáveis aleatórias discretas originárias do mesmo fenômeno aleatório e valores atribuídos a partir do mesmo espaço amostral teremos:

$$p(x,y) = P[(X = x) \cap (Y = y)] = P(X = x, Y = y)$$

Propriedades:

$$1) \sum_x \sum_y p(x,y) = 1$$

$$2) \sum_x p(x,y) = p(y)$$

$$3) \sum_y p(x,y) = p(x)$$

Exemplo

Uma região foi subdividida em 10 sub-regiões. Em cada uma delas foram observadas duas variáveis: O número de poços artesianos X e o número de riachos ou rios Y presentes na sub-regiao. Os resultados encontrados foram:

sub-região	X Número de poços	Y Número de rios
1	0	1
2	0	2
3	0	1
4	0	0
5	1	1
6	2	0
7	1	0
8	2	1
9	2	2
10	0	2

Espaço Amostral para a variável conjunta

A partir dos resultados encontrados anteriormente e considerando que a probabilidade de selecionar alguma região seja de 1/10, teremos que os pares (x,y) apresentam as seguintes probabilidades:

(x,y)	P(X = x, Y = y)
(0,0)	$1/10$
(0,1)	$2/10$
(0,2)	$2/10$
(1,0)	$1/10$
(1,1)	$1/10$
(2,0)	$1/10$
(2,1)	$1/10$
(2,2)	$1/10$

Tabela de Dupla Entrada e Marginais para X e Y

Construindo a tabela de dupla entrada para a variável conjunta (X,Y) obtemos:

$X \backslash Y$	0	1	2	$P(X = x)$
0	$1/10$	$2/10$	$2/10$	$5/10$
1	$1/10$	$1/10$	0	$2/10$
2	$1/10$	$1/10$	$1/10$	$3/10$
$P(Y = y)$	$3/10$	$4/10$	$3/10$	1

Pode-se notar que as marginais na tabela de dupla entrada correspondem aos valores que as variáveis assumem em suas tabelas de frequência individuais. Por exemplo: $P(X = 0) = P(X = 0, Y = 0) + P(X = 0, Y = 1) + P(X = 0, Y = 2) = 5/10$

Probabilidade Condicional para variáveis aleatórias discretas

Sejam as variáveis aleatórias discretas X e Y, temos que a probabilidade de $X = x$ dado a ocorrência de um $Y = y$ é dada pela expressão:

$$P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$$

Exemplo utilizando o exercício anterior:

$$P(X = 0|Y = 1) = \frac{P(X = 0 \cap Y = 1)}{P(Y = 1)} = \frac{2/10}{4/10} = 1/2$$

.

Independência entre variáveis aleatórias discretas

Sejam X e Y variáveis aleatórias discretas, teremos independência entre as variáveis quando:

$$P(X = x|Y = y) = P(X = x)$$

Ou de forma alternativa:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

É de fundamental importância entender que X e Y serão independentes se as relações acima forem válidas para todos os pares x e y.

Considere as variáveis aleatórias X e Y com a tabela de dupla entrada e as marginais representadas abaixo:

$X \backslash Y$	2	3	4	5	$P(X = x)$
2	$2/25$	$2/25$	$1/25$	0	$5/25$
3	$2/25$	$5/25$	$2/25$	$2/25$	$11/25$
4	$1/25$	$2/25$	$2/25$	$4/25$	$9/25$
$P(Y = y)$	$5/25$	$9/25$	$5/25$	$6/25$	1

Considerando a distribuição conjunta acima observamos que X e Y não são independentes pois:

$$P(X = 3, Y = 4) = 2/25 \neq P(X = 3)P(Y = 4) = 11/125$$

.

Covariância

Uma medida de dependência linear entre X e Y pode ser dada pela Covariância:

$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)] = E(XY) - E(X)E(Y)$$

Se as variáveis X e Y forem independentes teremos:

$$E(XY) = E(X)(Y)$$

o que implica: $Cov(X,Y) = 0$.

Obs: Se X e Y forem váriaveis aleatórias Independentes $Cov(X,Y) = 0$, mas se obtermos $Cov(X,Y) = 0$ não necessariamente as variáveis serão Independentes.

Coeficiente de correlação e suas características

O coeficiente de correlação entre duas variáveis X e Y é calculado pela seguinte expressão:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

A divisão pelos desvios-padrão tem como objetivo padronizar a medida para posteriores comparações. Propriedades:
$\rho_{X,Y}$ é adimensional;
$-1 \leq \rho_{X,Y} \leq 1$;
valores próximos de -1 ou de 1 indicam forte correlação.

Outras propriedades Importantes

Outras propriedades importante na análise de variáveis bidimensionais são dadas a seguir:

$$E(X+Y) = E(X) + E(Y)$$

A Esperança da soma de duas variáveis é igual a soma de suas esperanças.

$$Var(X+Y) = VAR(X) + VAR(Y) + 2Cov(X,Y)$$

Temos que se X e Y forem independentes a $Cov(X,Y) = 0$ e a variância torna-se a soma das variâncias de X e Y.

Estatística Descritiva e exploratória

Introdução

Discutiremos agora a caracterização de variáveis cujos possíveis valores ocorrem aleatoriamente e pertencem a um intervalo dos números reais reais: *variáveis aleatórias contínuas.*

Exemplos de variáveis aleatórias contínuas

Exemplos de variáveis aleatórias contínuas:
-Renda;
-Salário;
-Tempo de uso de um equipamento;
-Comprimento de uma peça;
-Área atingida por certa praga agrícola.

Podemos caracterizar completamente a atribuição de probabilidades para o caso contínuo. Ela será definida pela área abaixo de uma função positiva, denominada densidade de probabilidade.

Função Densidade de Probabilidade

Dizemos que $f(x)$ é uma função contínua de probabilidade ou função densidade de probabilidade para uma variável aleatória contínua X, se satisfaz duas condições:

1. $f(x) \geq 0$, para todo $x \in (-\infty, \infty)$.

2. A área definida por $f(x)$ é igual a 1, ou seja: $\int_{-\infty}^{\infty} f(x)dx = 1$.

Para calcularmos probabilidades por exemplo para $a \leq b$.
Utilizaremos:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

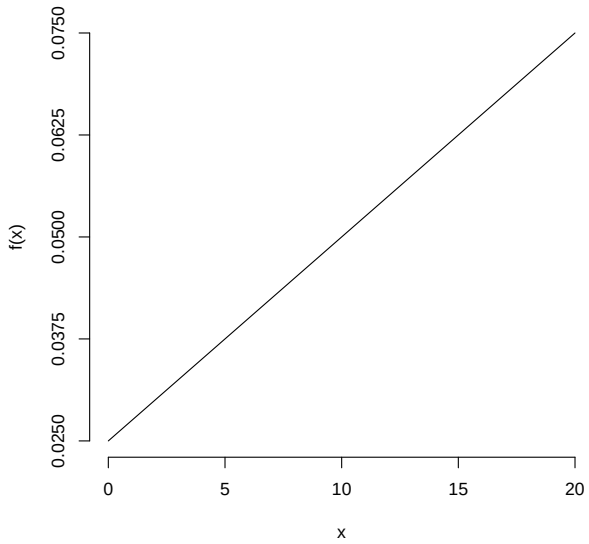
A integral acima determina a área abaixo da função no intervalo $[a, b]$.
A probabilidade de um evento estar entre os valores $[a, b]$ será definida pela área compreendida entre esses valores.
OBS: Teremos área zero sob qualquer valor individual, ou seja, $P(X = k) = 0$ para qualquer k. Com isso intervalos abertos ou fechados não modificarão o valor das probabilidades.

Exemplo

Arqueólogos estudaram uma certa região e estabeleceram um modelo teórico para a variável X, comprimento de fósseis da região (em cm). Suponha que X é uma variável aleatória contínua coma a seguinte função densidade de probabilidade:

$$f(x) = \begin{cases} \frac{1}{40}\left(\frac{x}{10}+1\right) & 0 \leq x \leq 20; \\ 0, & \text{caso contrário.} \end{cases}$$

O gráfico da distribuição anterior teria o seguinte comportamento:



Medidas de posição para variáveis aleatórias contínuas

Definição: Valor esperado também conhecido por média, expectância ou esperança de uma variável aleatória contínua X é dado pela expressão :

$$E(X) = \mu = \int_{-\infty}^{\infty} xf(x)dx$$

Definição: A mediana de uma variável aleatória contínua é um valor Md que satisfaz a seguinte propriedade :

$P(X \geq Md) \geq 0,5$ e $P(X \leq Md) \geq 0,5$

Moda de uma distribuição contínua

Definição: A moda de uma variável aleatória X é o valor Mo tal que:

$$f(Mo) = max_x f(x)$$

ou seja, Mo é o valor de máximo da função $f(x)$.

Variância de uma variável aleatória contínua

Para uma variável aleatória X com densidade $f(x)$, a variância é dada por:

$$\sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$$

Alternativamente, ela pode ser calculada por:

$$\sigma^2 = E(X^2) - \mu^2$$

Onde
$E(X^2) = \mu = \int_{-\infty}^{\infty} x^2 f(x)dx$

O desvio padrão σ é calculado através da raiz da variância.

Modelo Uniforme Contínuo

O primeiro modelo a ser apresentado traz uma situação análoga ao modelo uniforme discreto. Uma variável que assume valores no intervalo $[a, b]$ com $a < b$, é dita ter distribuição uniforme contínua se sua função densidade de probabilidade é dada por :

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \leq x \leq b; \\ 0, & \textit{caso contrário}. \end{cases}$$

Modelo Uniforme Contínuo

Neste caso, a função $f(x)$ é constante no intervalo $[a, b]$ e sua área total pode ser calculada através da área de um retângulo de base $b - a$ e altura $1/(b - a)$. Pelo produto da base pela altura do retângulo, verificamos que esta é uma função densidade de probabilidade pois sua área é igual a 1.
No modelo uniforme contínuo, a média e variância são dados respectivamente por :

$$E[X] = \frac{a + b}{2}$$

$$V[X] = \frac{(b - a)^2}{12}$$

Modelo Exponencial

O modelo exponencial é muito aplicado quando o interesse é descrever em termos probabilísticos o tempo (espaço) até a ocorrência de um evento de interesse. Alguns exemplos de variáveis modeladas por esta distribuição são :

- Tempo de espera na linha telefônica até o serviço de atendimento
- Tempo até a ativação de um neurônio
- Tempo de vida de um paciente com câncer
- Distância até encontrar uma deformidade em uma rodovia.

Modelo Exponencial

De forma mais geral, uma variável aleatória contínua é modelada pela distribuição exponencial se sua função densidade de probabilidade é descrita por:

$$f(x) = \alpha e^{-\alpha x}, x \geq 0$$

Modelo Exponencial

Para o modelo exponencial, a média e variância são inversamente proporcionais ao parâmetro α:

$$E[X] = \frac{1/}{\alpha}$$

$$V[X] = \frac{1}{\alpha^2}.$$

Na distribuição exponencial, a probabilidade da variável aleatória pertencer ao intervalo (a,b) é obtida através de :

$$P(a \leq X \leq b) = \int_a^b \alpha e^{-\alpha x} dx = e^{-\alpha a} - e^{-\alpha b}$$