Estatística e Informática

Aula 03 - Somatória e Análise Combinatória

Alan Rodrigo Panosso alan.panosso@unesp.br

Departamento de Engenharia e Ciências Exatas FCAV/UNESP

(12-05-2022)

SOMATÓRIA

Definições

Conjunto de dados: Nessa notação, a variável numérica de interesse (altura, idade ou peso, por exemplo) será representada pelas letras maiúsculas X, Y, Z.

O conjunto de dados terá a o tamanho n, que representa o número de elementos que ele contém. Em outras palavras, n representa o tamanho da amostra, o **número de observações** ou de **realizações** da variável.

os valores específicos assumidos por tais variáveis (suas realizações) serão representados pelas letras minúsculas x, y e z, respectivamente, seguidas de um índice i que representa a posição daquele valor específico dentro do conjunto de dados.

Assim, para distinguir um valor do outro, utilizamos esse índice i, que pode ser entendido como uma variável auxiliar, utilizada para contage, que se inicia na posição 1 e termina na posição n, abrangendo todo o seu conjunto de dados.

Assim temos:

$$Altura : X = \{x_1, x_2, \ldots, x_n\},\ \text{com}\ i = 1, 2, .., n.$$

$$Idade : Y = \{y_1, y_2, \ldots, y_n\}, (i = 1, 2, .., n).$$

$$Peso : Z = \{z_1, z_2, \ldots, z_n\}, (i = 1, 2, .., n).$$

Nessa notação um valor típico da variável *Altura*, será designado por x_i e o valor final por x_n.

Somatória

Ao realizar a análise de uma variável quantitativa (numérica) é necessário somarmos todos os seus valores valores.

Essa operação é frenquetemente utilizada na estatística, assim, utilizaremos uma notação compacta para representar a soma de todos os valores de uma variável de interesse.

Portanto, dado a variável X a soma de todos seus valores será notada pela letra grega sigma maiúscula Σ:

Dados $X = \{x_1, x_2, x_3, x_4, x_5\}$, a soma desses 5 valores:

$x_1 + x_2 + x_3 + x_4 + x_5$

será representada pela notação:

$\displaystyle\sum_{i=1}^{n} x_i$ ou $\sum_{i=1}^{n} x_i$

onde i atua como o índice, ou seja, a cada *iteração* ele muda e representa um dos 5 valores de X.

Dado duas variáveis $X = \{3, 0, 5, 9, 7\}$ e $Y = \{2, 3, 9, 1, 2\}$, calcular:

$$\sum_{i=1}^{n} x_i = 3 + 0 + 5 + 9 + 7 = 24$$

```r
X <- c(3,0,5,9,7)
sum(X)
```

Ao invés da soma ser com os índices i de 1 a n, podemos ter, por exemplo:

$$\sum_{i=2}^{4} y_i = 3 + 9 + 1 = 13$$

```r
Y <- c(2,3,9,1,2)
sum(Y[2:4])
```

Observe que:

$$\sum_{i=1}^{n} {x_i}^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$$

```r
sum(X^2)
sum(X)^2
```

uma vez que:

$$3^2 + 0^2 + 5^2 + 9^2 + 7^2 \neq (3 + 0 + 5 + 9 + 7)^2$$

$$9 + 0 + 25 + 81 + 49 \neq (24)^2$$

$$164 \neq 576$$

Qual a soma do produto entre as variáveis X e Y:

Dado $X = \{3, 0, 5, 9, 7\}$ e $Y = \{2, 3, 9, 1, 2\}$, calcular:

$$\sum_{i=1}^{n} x_i y_i$$

$$\sum_{i=1}^{n} x_i y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3 + x_4 \cdot y_4 + x_5 \cdot y_5$$

$$\sum_{i=1}^{n} x_i y_i = 3 \cdot (2) + 0 \cdot (3) + 5 \cdot (9) + 9 \cdot (1) + 7 \cdot (2)$$

$$\sum_{i=1}^{n} x_i y_i = 6 + 0 + 45 + 9 + 14 = 74$$

```
sum(X*Y)
```

No Excel

	A	B	C	D	E	F	G	H
1	X	Y		OU			XY	
2	3	2					6	=A2*B2
3	0	3					0	=A3*B3
4	5	9					45	=A4*B4
5	9	1					9	=A5*B5
6	7	2					14	=A6*B6
7		**74**					**74**	
8		=SOMARPRODUTO(A2:A6;B2:B6)					=SOMA(G2:G6)	

Propriedades da Somatória

i) A somatória de uma constante (k) é igual ao produto $n \cdot k$.

$\sum_{i=1}^{n} k = k + k + \cdots + k = n \cdot k$

ii) A somatória dos produtos de uma constante k e uma variável X é igual ao produto da constante pela soma dos valores da variável.

$$\sum_{i=1}^{n} k \cdot x_i = k \cdot x_1 + k \cdot x_2 + \cdots + k \cdot x_n = k \cdot (x_1 + x_2 + \cdots + x_n) = k \sum_{i=1}^{n} x_i$$

iii) A somatória da soma de duas variáveis é igual à adição das somatórias individuais dessas duas variáveis:

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1 + x_2 + y_2 + \cdots + x_n + y_n)$$

$$\sum_{i=1}^{n} (x_i + y_i) = (x_1 + y_1 + x_2 + y_2 + \cdots + x_n + y_n) = (x_1 + x_2 + \cdots + x_n) + (y_1 + y_2 + \cdots + y_n)$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Exercícios

1) Calcular: $\sum_{i=1}^{n} (k \cdot x_i + k) = ?$

2) Dado a média sendo:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Provar que:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Resposta

1) Calcular, portanto $\sum_{i=1}^{n}(k \cdot x_i + k) =?$

$$\sum_{i=1}^{n}(k \cdot x_i + k) = \sum_{i=1}^{n} k \cdot x_i + \sum_{i=1}^{n} k$$

$$\sum_{i=1}^{n}(k \cdot x_i + k) = \sum_{i=1}^{n} k \cdot x_i + \sum_{i=1}^{n} k$$

$$\sum_{i=1}^{n}(k \cdot x_i + k) = k \cdot \sum_{i=1}^{n} x_i + n \cdot k$$

Resposta

2) Dado: $\bar{x} = \frac{\sum\limits_{i=1}^{n} x_i}{n}$,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n \cdot \bar{x}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n \cdot \frac{\sum\limits_{i=1}^{n} x_i}{n}$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

ANÁLISE COMBINATÓRIA E MÉTODOS DE NUMERAÇÃO

Exemplo Prático

Imaginemos uma situação prática onde temos um rebanho com 100 animais contendo 20 indivíduos doentes e, consequentemente, 80 não doentes (sadios).

1) Quais as probabilidade de, em uma amostra aleatória simples sem reposição de 3 animais, selecionarmos:
a) nenhum doente, consequentemente, 3 animais sadios?
b) 1 doente e 2 sadios?
c) 2 doentes e 1 sadios?
d) 3 doentes

2) Construir o gráfico dessas probabilidade de cada evento eixos (Y) pelo nº animais doentes na amostra eixo (X).

Probabilidade e Análise Combinatória

A Probabilidade permite analisar ou calcular as chances de obter determinado resultado diante de um experimento que dizemos aleatório. Por exemplo, qual a chance do número 5 sair no lançamento de um dado. Ou no lançamento de 4 moedas, sair pelo menos uma face **Cara**.

A partir disso, a probabilidade é determinada pela razão entre o número de eventos possíveis e o número de eventos favoráveis, sendo apresentada pela seguinte expressão:

$$P(E) = \frac{\text{nº de resultados favoráveis}}{\text{nº total de resultados possíveis}}$$

onde $P(E)$ é a probabilidade de ocorrência do evento E.

Assim, para respondermos a quetão da amostragens de animais doentes, precisaremos calcular:

nº de resultados favoráveis

Ou seja, de quantas maneiras poderemos escolher, por exemplo, 2 animais do total de 20 doentes e 1 do total de $80 $animais sadios.

nº total de resultados possíveis

De quantas maneiras poderemos escolher 3 animais (independentemente de serem sadios ou doentes) em um total de 100 animais.

Análise Combinatória

Portanto, para calcularmos a probabilidade do evento (escolher 2 animais doentes e 1 sadio), faz-se necessário relembrarmos os conceitos de métodos de numeração e análise combinatória (*Permutações*, *Arranjos* e *Combinações*).

Princípio Fundamental da Contagem

O princípio fundamental da contagem, também chamado de princípio multiplicativo, postula que:

> "quando um evento é composto por n etapas sucessivas e **independentes**, de tal modo que as possibilidades da primeira etapa é x e as possibilidades da segunda etapa é y, o número total de possibilidades do evento ocorrer é dado pelo produto $x \cdot y$".

Em resumo, no princípio fundamental da contagem, multiplica-se o número de opções independentes entre as escolhas que lhe são apresentadas.

Exemplos

1. Em um restaurante, há 5 aperitivos no *menu*, 10 entradas principais e 4 sobremesas. De quantas maneiras uma refeição pode ser solicitada se um aperitivo, um prato principal e uma sobremesa forem escolhidos?

2. Quantos códigos numéricos de 5 caractéres são possíveis se:
   (a) não há restrições?
   (b) o primeiro dígito não pode ser 0?
   (c) nenhuma repetição é permitida?
   (d) além de números você pode utilizar todas as letras do teclado (maiúsculas e minúsculas - 52 ao todo) e mais 33 símbolos especiais e sem qualquer restrição quanto à repetição.

3. Um dos métodos utilizados para roubo de senhas e invasão de contas é o método de Ataque de Força-bruta (Brute-force attack). Nessa técnica são testadas combinações de números, letras e caracteres especiais até que a senha seja descoberta. Levando-se em consideração que um programa espião pode testar em média 435 milhões senhas por segundo:
   a. qual o tempo máximo em segundos ele levaria para decodificar uma senha composta na condição (d) do exercício anterior? b. qual o tempo máximo em dias ele levaria para decodificar uma senha composta na condição (d) do exercício anterior, mas com 9 dígitos?

Permutações

Quando estamos preocupados em responder "de quantas maneiras um evento pode ocorrer", utilizamos o conceito de permutação para essa tarefa. Digamos que temos n objetos diferentes, de quantas maneiras podemos **dispor** (**permutar**) esses objetos?

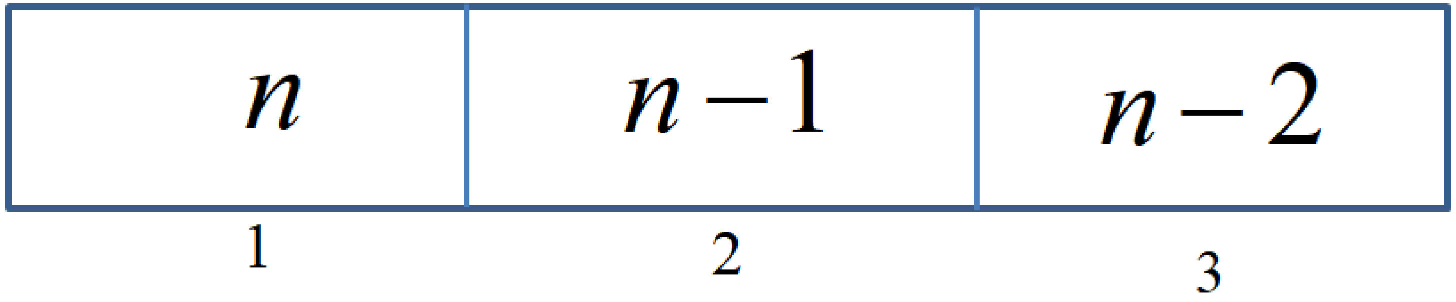
Dado o conjunto de letras $L = \{a, b, c\}$, de quantas maneiras podemos permutá-las?

a, b, c
a, c, b
b, a, c
b, c, a
c, a, b
c, b, a

Equivale em colocar os elementos dentro de caixas com n compartimentos em alguma ordenação.

E como consequência do princípio da multiplicação, as possibilidades de inserção dentro de uma caixa são multiplicadas pelas possibilidades da caixa seguinte, e assim sucessivamente:

Permutações



$$n! = n \cdot (n-1) \cdot (n-2) \cdots (2) \cdot (1)$$

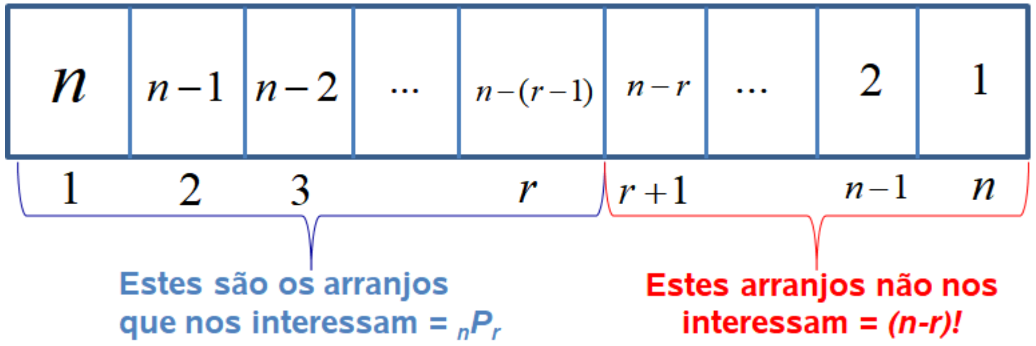
$$3! = 3 \cdot (3-1) \cdot (3-2) = 3 \cdot 2 \cdot 1 = 6$$

Exemplos:

1. Quantos números de 7 algarismos **distintos** podem ser formados com os algarismos 1,2,3,4,5,6 e 7?
2. Quantos números de 7 algarismos **distintos** podem ser formados com os algarismos 1,2,3,4,5,6 e 7 de modo que em todos os números formados, o algarismo 6 seja imediatamente seguido pelo algarismo 7?
3. Três (03) chineses, 04 americanos e 05 italianos serão dispostos em fila de modo que as pessoas da mesma nacionalidade estejam sempre juntas. De quantas maneiras a fila poderá ser formada, de modo que o primeiro da fila sempre seja um italiano?

Arranjos

O arranjo é uma extensão do conceito de fatorial de modo que, se temos n elemntos e vamos tomar r amostras desses n (sempre com $0 < r < n$), de quantas maneiras poderemos **dispor** os r elementos amostrados? Nesse caso, **a ordem** com a qual os elementos serão dispostos **IMPORTA** para a contagem.



$$A_r^n = \frac{n!}{(n-r)!}$$

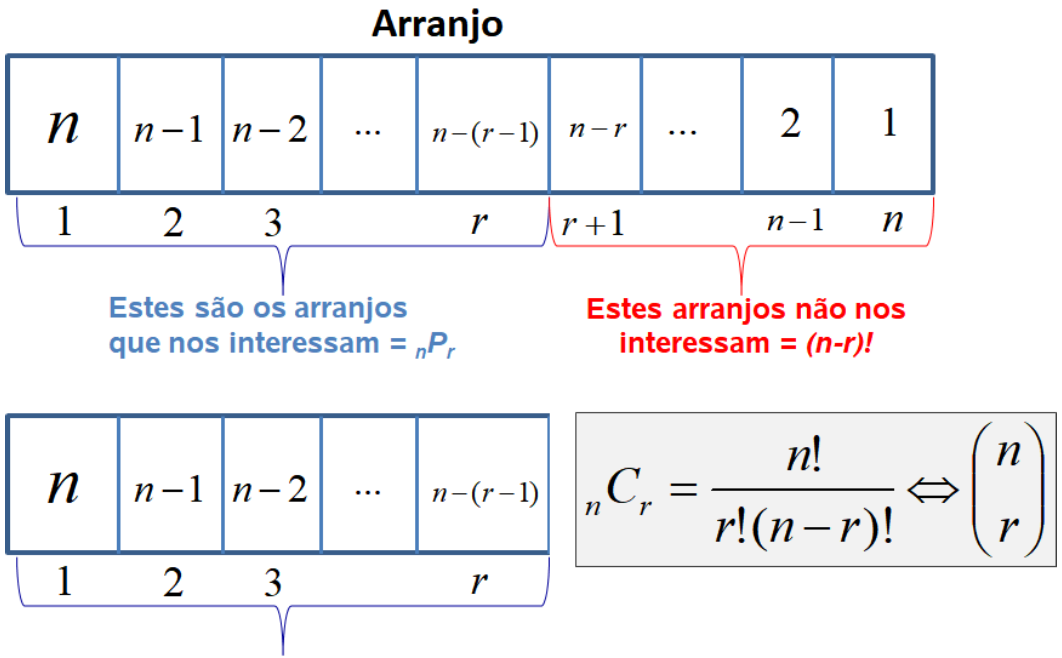
1) A senha de uma conta bancária é formada de 6 números ou de 4 números dentre os números de 0 a 9, dependendo do banco. Os gerentes sempre aconselham utilizar todos os números distintos (sem repetição),

Quantas senhas podem ser formadas no banco do Banco do Brasil, que utiliza 6 números, e no banco Santander que utiliza 4 números, seguindo o conselho dos gerentes?

2) Quantos números de 4 algarismos podemos formar com os dígitos 1, 2, 3, 4 e 5 de modo que haja pelo menos dois dígitos iguais?

Combinações

Semelhante aos arranjo, nas **Combinações** temos n elementos e vamos amostrar r elementos $(0 < r < n)$, e estamos interessados em contar de quantas maneiras poderemos **combinar** os r elementos amostrados, nesse caso **a ordem** com a qual os elementos são amostrados **NÃO IMPORTA** para a contagem.



Assim, teremos que descontar dos Arranjos de r elementos as permutações desses r elementos, para isso, basta dividirmos os Arranjos por $r!$ ou seja $C_r^n = \binom{r}{x}$.

Combinações

Voltando para o exemplo do rebanho composto por 100 animais, onde temos 20 indivíduos doentes e 80 sadios, vamos calcular a probabilidade de em uma amostra de 3 animais selecionarmos 2 animais doentes e 1 animais sadios:

$\text{nº de resultados favoráveis}$

Ou seja, de quantas maneiras poderemos escolher 2 animais do total de 20 doentes e 1 dos 80 sadios?

$C_2^{20} \cdot C_1^{80} = \binom{20}{2} \cdot \binom{80}{1}$

$C_2^{20} \cdot C_1^{80} = \frac{20!}{2!(20-2)!} \cdot \frac{80!}{1!(80-1)!}$

$C_2^{20} \cdot C_3^{80} = 190 \cdot 80$

$C_2^{20} \cdot C_3^{80} = 15200$

nº total de resultados possíveis

Devemos calcular agora de quantas maneiras poderemos escolher 3 animais (independente de serem sadios ou doentes) em um total de 100 animais.

$C_3^{100} = \binom{100}{3}$

$C_3^{100} = \frac{100!}{3!(100-3)!}$

$C_3^{100} = 161700$

Finalmente, a probabilidade do evento, amostrar 2 indivíduos doentes e 1 sadio será:

$$P(E) = \frac{\text{nº de resultados favoráveis}}{\text{nº total de resultados possíveis}} = \frac{15200}{161700} = 0,094 \text{ ou } 9,40\%$$

Agora que temos uma fórmula, podemos calcular a probabilidade de na amostragem de 5 animais tomarmos:

a) nenhum animal doente (todos sadios).
b) 01 doente e 02 sadios.
c) 02 doentes e 01 sadios.
d) todos doentes.

Distribuição de Probabilidade

Essa fórmula que acbamos de encontrar decreve uma ditribuição discreta de probabilidade, conhecida na estatística como **Distribuição Hipergeométrica**.

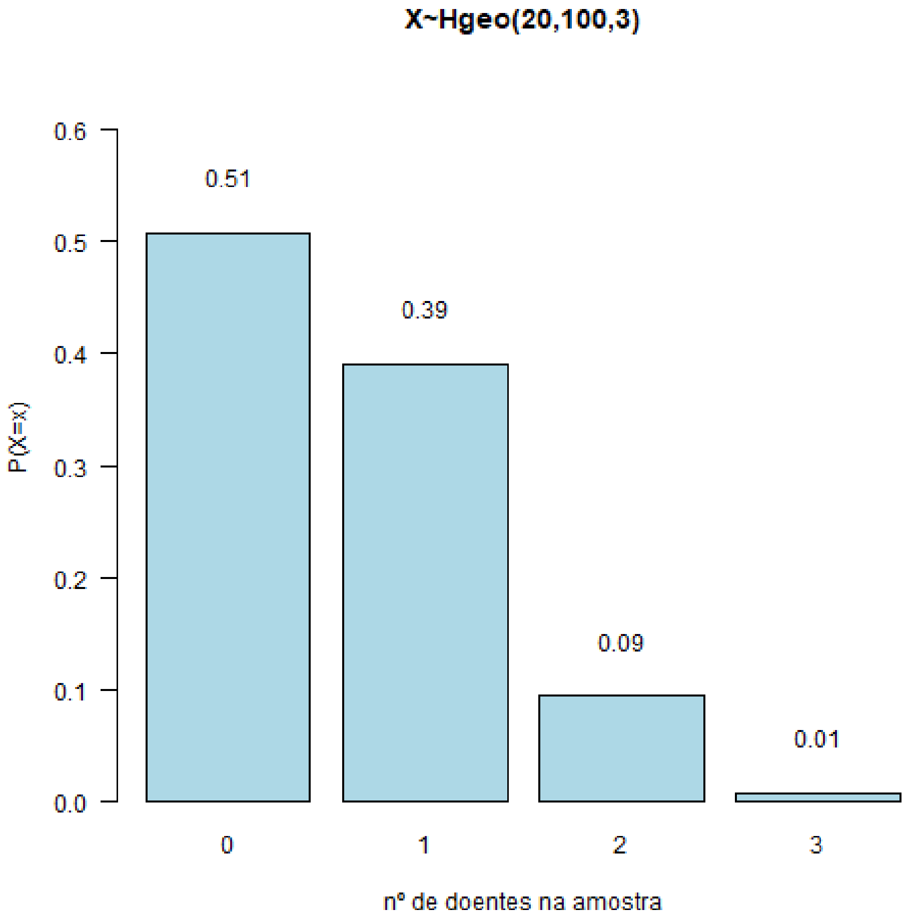
Considere uma população com N objetos nos quais M são classificados como do tipo A e $N-M$ são classificados como do tipo B. Tomamos uma amostra ao acaso, **sem reposição** e não ordenada de r objetos. Seja X a variável que conta o número de objetos classificados como do tipo A na amostra. Então a distribuição de probabilidade de X será dada por:

$$P(X=x) = \frac{\binom{M}{x} \cdot \binom{N-M}{r-x}}{\binom{N}{r}}$$

Assim, vamos construir essa distribuição no R para termos todas as probabilidades.

```
N = 100 # Número total de elementos na população
M = 20 # número total de elementos de interesse (doentes) na população
r = 3 # tamanho da amostra.
x=0:r # os possiveis valores de doentes na amostra.
px<-dhyper(x,M,N-M,r) # função que faz a distribuição hipergeométrica
barplot(px,names=x,col="lightblue",las=1,ylim=c(0,0.65),
        xlab="nº de doentes na amostra",ylab="P(X=x)",main="X~Hgeo(20,100,3)")
text(1:length(x),px+.05,round(px*100,2))
```

Distribuição Hipergeométrica




```
## # A tibble: 4 x 2
##       x      px
##   <int>   <dbl>
## 1     0 0.508
## 2     1 0.391
## 3     2 0.0940
## 4     3 0.00705
```

OBS: O número de doentes da amostra de 3 animais pode variar de 0 a 3, ao conjunto dessas possibilidades denominamos espaço amostral (letras Ω ou S), com a soma de suas probabilidades é sempre igual a 1:

$$\sum_{x=0}^{r} P(x_i) = 1$$