UNIVERSIDADE DA REGIÃO DE JOINVILLE - UNIVILLE

Bacharelado em Engenharia de Software (BES)

Estatística para computação

Professora Priscila Ferraz Franczak

Engenheira Ambiental - UNIVILLE

Mestre em Ciência e Engenharia de Materiais - UDESC

Doutoranda em Ciência e Engenharia de Materiais - UDESC

priscila.franczak@gmail.com

Plano de Aula

Estatística Descritiva

1. Notações de soma e produto
2. Medidas de posição amostral
3. Medidas de dispersão amostral
4. Covariância e correlação
5. Exercícios

Estatística Descritiva

- É a parte da Estatística que descreve e avalia certo grupo de dados, seja ele população, seja amostra.

- No caso de estarmos trabalhando com amostras, o simples uso de estatísticas descritivas não nos permite tirar quaisquer conclusões ou inferências sobre um grupo maior.

- Para estabelecimento de inferências ou conclusões sobre um grupo maior (a população) precisaríamos usar métodos estatísticos, que caracterizam a área da Estatística conhecida como Estatística Indutiva ou Inferência Estatística.

Há na Estatística Descritiva dois métodos que podem ser usados para a apresentação dos dados:

- Métodos **gráficos** (envolvendo apresentação gráfica e tabular);
- Métodos **numéricos** (envolvendo apresentações de medidas de posição e dispersão, entre outras).

1. Notações de soma e produto

Somatório

- Muitos dos processos estatísticos exigem o cálculo da soma.

- Para simplificar a representação da operação de adição nas expressões algébricas, utiliza-se a notação Σ, que é o *sigma* maiúsculo do alfabeto grego.

- Lê-se $\sum_{i=1}^{n} x_i$ como somatório de x índice i,

com *i* variando de 1 a *n*, em que **n** é a ordem da última parcela ou limite superior do somatório.

- Na verdade, o somatório nada mais é que uma notação simplificada da adição de elementos de um conjunto.

- Exemplos:

```
> x<-c(10,20,30,40)   #vetor
> sum(x)              #somatório do vetor
[1] 100
```

- Encontre a soma $\sum_{\substack{i=2 \\ i \neq 5}}^{6} {Y_i}^2$

```
> Y<-c(65,75,85,65,95,80)
> sum(Y^2)-Y[1]^2-Y[5]^2
[1] 23475
>
> sum(Y[-c(1,5)]^2)
[1] 23475
```

Produtório

- O símbolo produtório é utilizado para facilitar a representação dos produtos.

- Emprega-se a notação Π, que é o *pi* maiúsculo do alfabeto grego.

- Lê-se $\prod_{i=1}^{n} x_i$ como produtório de x índice i,

com *i* variando de 1 a *n*, em que **n** é a ordem da última parcela ou limite superior do produtório.

- Exemplos:

- Encontre o produtório $\prod_{i=1}^{4} z_i$

```
> z<-c(15,25,35,45)   #vetor
> prod(z)             #produtório do vetor
[1] 590625
```

- Encontre o produtório $\prod_{i=1}^{2} z_i$

```
> prod(z[-c(3,4)])   #produtório do vetor menos a exceção
[1] 375
```

2. Medidas de posição amostral

Média aritmética

- **Dados não agrupados**

Sejam os elementos x_1, x_2, x_3,...,x_n de uma amostra, portanto "n" valores da variável x. A média aritmética da variável aleatória de x é definida por:

$$\bar{x} = \frac{soma\ dos\ valores\ de\ x}{número\ de\ observações} = \frac{\sum x}{n}$$

Exemplo:

Suponha o conjunto de tempo de serviço de cinco funcionários: 3, 7, 8, 10 e 11. Determinar a média aritmética simples deste conjunto de dados.

$$\bar{x} = \frac{3 + 7 + 8 + 10 + 11}{5} = \frac{39}{5} = 7{,}8$$

Interpretação: o tempo médio de serviço deste grupo de funcionários é de 7,8 anos.

- **Dados agrupados em uma distribuição de frequência por valores simples**

Quando os dados estiverem agrupados numa distribuição de frequência usaremos a média aritmética dos valores x_1, x_2, x_3,...,x_n, ponderados pelas respectivas frequências absolutas: f_1, f_2, f_3, ... , f_n. Assim:

$$\bar{x} = \frac{\sum x_i f_i}{n}$$

Exemplo:

Em um **determinado dia** foi registrado o número de veículos negociados por uma amostra de 10 vendedores de uma agência de automóveis obtendo a seguinte tabela:

Veículos negociados (x_i)	**Número de vendedores (f_i)**	**$x_i.f_i$**
1	1	1
2	3	6
3	5	15
4	1	4
Total	10	26

$$\bar{x} = \frac{\sum x_i f_i}{n}$$

$$\bar{x} = \frac{26}{10} = 2,6$$

Interpretação: em média, cada vendedor negociou 2,6 veículos.

- **Dados agrupados em uma distribuição de frequência por classes**

Usaremos a média aritmética **dos pontos médios** $\dot{x}_1$, $\dot{x}_2$, $\dot{x}_3$,...,$\dot{x}_n$ de cada classe, ponderados pelas respectivas frequências absolutas: f_1, f_2, f_3, ... , f_n.

Desta forma, o cálculo da média passa a ser igual ao da situação anterior. Assim:

$$\bar{x} = \frac{\sum \dot{x}_i f_i}{n}$$

- A média é a medida de posição mais conhecida e pode ser obtida facilmente no R com o comando mean( ):

```
> k<-c(1,2,3,4,5)
> mean(k)
[1] 3
```

- Em algumas situações é possível haver um ou mais dados ausentes (NA).

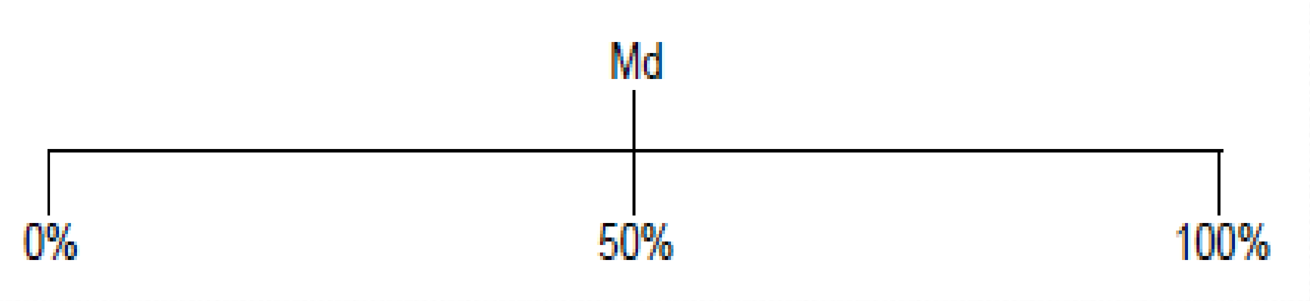
- Neste caso basta usar o argumento na.rm=T para que o R desconsidere os elementos NA no cálculo da média.

```
> w<-1:5
> w
[1] 1 2 3 4 5
>
> w[2]<-NA  #o valor 2 foi perdido
> w
[1]  1 NA  3  4  5
>
> mean(w)
[1] NA
>
> mean(w,na.rm = T)
[1] 3.25
```

Mediana

- A mediana é definida como o número que se apresenta no centro de uma série de números dispostos **segundo uma ordem**.

- Construído o ROL, o valor da mediana é o elemento que ocupa a posição central, ou seja, é o elemento que divide a distribuição em 50% de cada lado:



- A mediana é uma medida de posição indicada quando o conjunto de dados possui valores extremos discrepantes dos demais, o que pode comprometer a discussão dos dados baseados simplesmente na média.

- Embora os valores precisem estar ordenados para se calcular a mediana, o R já realiza automaticamente a ordenação:

```
> a<-c(1,2,18,7,6)
> median(a)
[1] 6
```

Moda

- Dentre as principais medidas de posição, destaca-se a Moda. **É o valor mais frequente da distribuição**.

- **Amodal**: quando nenhum valor do conjunto pode ser considerado moda.

- **Unimodal**: quando possui apenas um valor modal.

- **Bimodal**: quando tem dois valores de moda.

- **Multimodal**: para um conjunto de dados com mais de dois valores modais.

O comando table( ) cria uma tabela de frequência de cada elemento de determinado objeto:

```
> b<-c(0,12,3,4,5,5,5,5,6,6,7,7,8)
> table(b)
b
 0  3  4  5  6  7  8 12
 1  1  1  4  2  2  1  1
```

- Para achar a moda, podemos usar o comando mfv, do pacote modeest (que precisa ser instalado no R), digitando no console:

```
install.packages("modeest")
```

```
> b<-c(0,12,3,4,5,5,5,5,6,6,7,7,8)
> mfv(b)
[1] 5
```

- Outro comando que pode ser usado , sem a necessidade de instalação de pacote é:

```
> b<-c(0,12,3,4,5,5,5,5,6,6,7,7,8)
> table(b)
b
 0  3  4  5  6  7  8 12
 1  1  1  4  2  2  1  1
> subset(table(b),table(b)==max(table(b)))
5
4
```

- Resumindo as medidas de posição:

```
> d<-c(20,7,5,9,6,21,24,10,12,22,21,16,13,
+       6,6,2,19,3,10,7,2,18,4,6,18,12,4,13,9,3)
>
> mean(d)
[1] 10.93333
>
> median(d)
[1] 9.5
>
> table(d)
d
 2  3  4  5  6  7  9 10 12 13 16 18 19 20 21 22 24
 2  2  2  1  4  2  2  2  2  2  1  2  1  1  2  1  1
>
> mfv(d)
[1] 6
>
> subset(table(d),table(d)==max(table(d)))
6
4
```

3. Medidas de dispersão amostral

- Essas medidas descrevem a variabilidade que ocorre no conjunto de dados analisado e são úteis para complementar as informações fornecidas pelas medidas de posição.

- Assim, as medidas de dispersão são elementos fundamentais na caracterização de uma amostra.

Variância (s^2)

- A variância é uma das medidas que fornecem informações complementares à informação contida na média aritmética.

- Ela apenas indica se há dispersão em relação à média.

- É definida como sendo a média aritmética dos quadrados dos desvios em relação à média da população (ou amostra).

- A variância é expressa pela fórmula:

$$s^2 = \frac{\sum f_i\,(d_i)^2}{n}$$

Quanto maior a variância, maior a dispersão dos dados amostrais

Observação: Alguns autores usam s^2 para indicar a variância e s para indicar o desvio padrão da amostra. Quando temos amostra com número de elementos menor que 30, dividimos por (n-1).

Exemplo: Nascimentos diários na maternidade M no período X.

x_i	f_i	$\bar{x}$	$d_i = x_i - \bar{x}$	$(d_i)^2$	$f_i(d_i)^2$
1	2	4	-3	9	18
2	3	4	-2	4	12
3	1	4	-1	1	1
4	3	4	0	0	0
5	3	4	1	1	3
7	2	4	3	9	18
8	1	4	4	16	16
Total	$\sum fi = 15$				$\sum f_i(d_i)^2 = 68$

$$s^2 = \frac{\sum f_i\,(d_i)^2}{n-1}$$

$$s^2 = \frac{68}{14}$$

$$s^2 = 4{,}86$$

- Com apenas um comando podemos obter a variância amostral:

```
> nascimentos<-c(1,1,2,2,2,3,4,4,4,5,5,5,7,7,8)
> var(nascimentos)
[1] 4.857143
> signif(var(nascimentos),3)
[1] 4.86
```

Desvio padrão (s)

- O desvio padrão é uma medida de variação em relação a média largamente usada nos testes estatísticos.
- Indica por meio de uma medida padronizada o quanto um dado está afastado da média.
- É útil, por exemplo, para verificar se há melhoria de qualidade na produção de determinado elemento através de novo processo de fabricação

O desvio padrão é calculado extraindo a raiz quadrada da variância:

$$s = \sqrt{\frac{\sum f_i\,(d_i)^2}{n}}$$

Exemplo: A média de idade de funcionários de uma empresa é:

38,44 ±11,58 anos

$$\bar{x} \pm s$$

- Com apenas um comando podemos obter o desvio padrão:

```
> nascimentos
 [1] 1 1 2 2 2 3 4 4 4 5 5 5 7 7 8
> sd(nascimentos)
[1] 2.203893
> signif(sd(nascimentos),3)
[1] 2.2
>    Ou raiz quadrada da variância
> sqrt(var(nascimentos))
[1] 2.203893
> signif(sqrt(var(nascimentos)),3)
[1] 2.2
```

Amplitude total

- A amplitude total é dada pela diferença entre a variável de maior valor e a variável de menor valor da amostra.

- Leva em conta os valores extremos da série em prejuízo dos valores intermediários.

- Usa-se a amplitude total quando se quer determinar, por exemplo, a variação de temperatura de um dia do ano ou quando a compreensão popular é mais importante que a exatidão e a estabilidade dos resultados.

PREVISÃO DO TEMPO PARA OS PRÓXIMOS DIAS

SÁB	DOM	SEG	TER	QUA	QUI	SEX
01/04	02/04	03/04	04/04	05/04	06/04	07/04
↑ 25°	↑ 26°	↑ 27°	↑ 28°	↑ 28°	↑ 29°	↑ 28°
↓ 18°	↓ 17°	↓ 19°	↓ 20°	↓ 20°	↓ 22°	↓ 23°
3mm	0mm	0mm	2mm	0mm	0mm	2mm
60%	0%	0%	60%	0%	0%	40%

Fonte: https://www.climatempo.com.br/previsao-do-tempo/cidade/381/joinville-sc

Exemplo:

Consideremos os seguintes conjuntos de valores que representam o número de pacientes atendidos em postos de saúde de três bairros A, B e C, num período de 5 dias. Temos os resultados, para cada posto, dados por:

bairro A: 60, 60, 60, 60, 60

$\overline{x_A} = 60$

$H_A = 0$

bairro B: 58, 62, 59, 61, 60

$\overline{x_A} = 60$

$H_A = 4$

bairro C: 5, 15, 115, 105, 60

$\overline{x_A} = 60$

$H_A = 110$

- No R obtemos a amplitude dos dados através do comando:

```
> range(nascimentos)
[1] 1 8
```

Coeficiente de variação (C.V.)

- Desvio padrão é limitado
- Podemos caracterizar a dispersão ou variabilidade dos dados em termos relativos a seu valor médio:

$$C.V. = \frac{s}{\bar{x}}.100$$

- Mede a dispersão relativa do conjunto de dados
- Assim, podemos comparar duas ou mais séries de valores

Interpretações do coeficiente de variação:

Se:

- C.V. < 15% há baixa dispersão
- 15% ≤ C.V. < 30% há média dispersão
- C. V. ≥ 30% há elevada dispersão

```
> cv<-(sd(nascimentos)/mean(nascimentos))*100
> cv
[1] 55.09732
```

4. Covariância e correlação

- Quando existirem duas séries de dados, existirão várias medidas estatísticas que podem ser usadas para capturar como as duas séries se movem juntas através do tempo.

- As duas mais largamente usadas são a <u>correlação</u> e a <u>covariância</u>.

- A covariância fornece uma medida não padronizada do grau no qual elas se movem juntas, e é estimada tomando o produto dos desvios da média para cada variável em cada período.

- O sinal na covariância indica o tipo de relação que as duas variáveis tem.

- Um sinal positivo indica que elas movem juntas e um negativo que elas movem em direções opostas.

- Enquanto a covariância cresce com o poder do relacionamento, ainda é relativamente difícil fazer julgamentos sobre o poder do relacionamento entre as duas variáveis observando a covariância, pois ela não é padronizada.

- A correlação é a medida padronizada da relação entre duas variáveis.

- A correlação nunca pode ser maior do que 1 ou menor do que -1.

- Uma correlação próxima a zero indica que as duas variáveis não estão relacionadas.

- Uma correlação positiva indica que as duas variáveis movem juntas, e a relação é forte quanto mais a correlação se aproxima de 1.

- No R, a covariância e a correlação entre dois conjuntos de dados podem ser obtidas pelos comandos cov(x,y) e cor(x,y):

```
> x<-c(1,2,3,4,5)
> y<-c(6,7,8,9,10)
> cov(x,y)
[1] 2.5
> cor(x,y)
[1] 1
```

5. Exercícios