Noções de Estatística

Aula 2

- Medidas Resumo
- Boxplots

2.1 Medidas Resumo

- O que é uma medida resumo?
- Para que serve uma medida resumo?

Para entendermos a função das medidas resumo, vamos retomar como exemplo parte da Tabela 1.1.

Tabela 1.1 Tabela de um estudo hipotético

Nº	Estado Civil	Grau de escolaridade	Nº de filhos	Salário (x sal min)	Idade	Região de procedência
1	Solteiro	E.f.	-	4,00	26	Interior
2	casado	E.M.	1	1,56	23	capital
3	Solteiro	E.S.	2	2,33	25	Interior
4	casado	E.f.	-	2,50	23	capital
5	Solteiro	E.M.	3	4,00	26	Interior
6	casado	E.S.	4	5,12	27	capital
7	Solteiro	E.f.	2	2,21	22	Interior
8	casado	E.M.	1	2,12	24	capital
9	Solteiro	E.S.	-	3,21	25	Interior
10	casado	E.f.	2	1,00	26	capital
11	Solteiro	E.M.	1	1,21	27	Interior
12	Solteiro	E.S.	3	1,00	32	capital
13	casado	E.f.	1	1,21	21	Interior
14	Solteiro	E.M.	2	1,21	31	capital
15	casado	E.S.	-	3,21	41	capital
16	Solteiro	E.f.	3	4,21	52	Interior
17	Solteiro	E.M.	2	3,21	53	capital
18	casado	E.S.	1	1,21	54	Interior
19	Solteiro	E.f.	-	1,86	34	capital
20	casado	E.M.	1	1,23	32	Interior
21	Solteiro	E.S.	2	1,32	42	capital

Vemos que, da forma como os dados se apresentam, fica difícil formar uma intuição generalizada sobre a amostra. Como discutimos na última aula, a tabela de distribuição de frequências, gráficos de barras e histogramas são formas de resumir informações de modo a nos auxiliar a ter uma visão mais geral sobre os meus dados. Podemos falar o mesmo sobre as medidas resumo.

2.2 Medidas de Posição:

As medidas de posição são valores utilizados para representar toda a séria de observações.

Média:

A média é uma medida resumo muito importante. Quando calculamos a média para uma variável quantitativa, estamos buscando resumir todas as nossas observações em um único valor.

Para calcular a média de uma variável, utilizamos a fórmula:

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

Vamos entender como funciona esse operador. O "símbolo" $\sum$ é o símbolo de somatório, ele indica que será necessário fazer somas sucessivamente. Observe também que abaixo e acima do somatório temos duas letras. Convencionalmente, quando falamos em uma amostra, falamos que uma amostra tem tamanho $\boldsymbol{n}$. Também dizemos que cada observação da nossa variável X na nossa amostra, ou seja, cada indivíduo observado é o indivíduo $\boldsymbol{x_i}$. A letra abaixo do somatório

indica o início do somatório, ou seja, a partir de que i devemos começar a somar, e a letra acima indica quando devemos parar de somar. Lemos essa fórmula da seguinte maneira: Para i começando em 1, até que o i se torne n, some todas as observações x_i. Ao finalizar, divida o resultado dessa soma por n.

Resumindo...

$$\frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{n}$$

Mediana

A mediana é outra medida resumo interessante e muito utilizada. A mediana ocupa o valor central de um conjunto de n observações ordenadas. Isso significa que, se eu sei a mediana de um conjunto de dados, eu sei que 50% das demais observações se encontram abaixo desse valor e os outros 50%, acima desse valor.

Exemplo:

Suponha que eu tenho a seguinte tabela:

Indivíduo i	Número de filhos
1	2
2	3
3	0
4	1
5	2

Primeiro passo para calcularmos a mediana:

Ordenar as observações da variável número de filhos:

0 1 2 2 3

A mediana é o valor que ocupa a posição $\frac{n+1}{2}$.

Nesse pequeno exemplo, temos que n = 5. Portanto a mediana é o 3º valor dessa sequência. $\frac{n+1}{2} = \frac{5+1}{2} = 3$

O 3º valor é o número 2.

0 1 2 2 3

Observe que metade das demais observações se encontram abaixo e a outra metade, acima desse valor.

No caso de termos n observações pares, por exemplo...

0 1 2 3 4 5

A mediana se localizaria entre dois valores. Quando isso acontecer, faça uma média simples entre esses valores.

Portanto a mediana para o exemplo acima seria 2.5.

Podemos também utilizar a mesma fórmula.

$$Posição = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$$

Ou seja, a fórmula também nos indica que o valor estaria na posição 3.5, e, portanto, seria a média entre os valores que estão na posição 3 e 4, resultando também em 2.5.

Percentis

O percentil de ordem p x 100 ($0 < p < 1$), em um conjunto de dados de tamanho n, é o valor da variável que ocupa a posição p x (n + 1) do conjunto de dados ordenados.

Ou seja, a mediana é um percentil de ordem 50. (50% pois ela é o centro das observações ordenadas, como discutimos acima.

Além da mediana, temos outros percentis muito úteis para a metodologia estatística:

25 → Q1 (Primeiro quartil)
50 → Q2 (Segundo quartil corresponde à mediana)
75 → Q3 (Terceiro quartil)

Assim como a mediana é o centro das observações ordenadas, o que você acha que acontece com os demais percentis apresentados?

Cálculo dos percentis:

Para encontrar a posição do nosso percentil, utilize a fórmula:

Posição = p x (n +1)

Por exemplo:

Número	Posição
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

A posição do percentil de ordem 25º da nossa amostra é obtida (utilizado a fórmula dada acima) ao multiplicarmos 0.25 por 9 = 2.25. Porém, a posição 2.25 não existe. Como resolvemos esse problema?

É muito importante ficar claro que nem todo mundo concorda com a definição de um percentil. Diferentes pessoas podem trabalhar com diferentes definições.

Veja abaixo possíveis definições sobre o que é um percentil. (Vou exemplificar utilizando a ordem de 25%)

Definição 1: O menor valor que é maior que 25% dos valores.

Definição 2: O menor valor que é maior ou igual a 25% dos valores.

Infelizmente essas duas definições podem levar a resultados diferentes. No nosso exemplo, a definição 1 aponta para a posição número 3 e a definição 2 aponta

para a posição número 2. Há uma terceira definição que procura corrigir esse problema.

Calcule a diferença entre as duas possibilidades. No nosso exemplo (5 e 7), a diferença é 2. Multiplique essa diferença pela parte decimal da nossa posição. No nosso exemplo, 0.25.

2 x 0.25 = 0.5

Por último, some 0.5 ao valor com a menor posição candidata.

0.5 + 5 = 5.5

Assim, com essa definição, o nosso Q1 seria 5.5.

Se você está fazendo uma análise estatística, pode escolher trabalhar com a definição que desejar. Porém eu recomendo que você discuta com o professor da sua turma a definição que ele está utilizando para evitar futuras confusões!

Outras medidas resumo importantes são:

Max e Min: Simplesmente as observações de valores máximo e mínimo da amostra.

Moda: O valor mais frequente da amostra.

Outras medidas muito importantes para a estatística são as medidas de dispersão.

As principais medidas de dispersão são a variância e o desvio-padrão. O desvio-padrão é simplesmente a raiz quadrada da variância, então vamos nos concentrar apenas na fórmula do cálculo da variância:

Assim como na fórmula do cálculo da média, novamente teremos que lidar com o somatório para o cálculo da variância.

$$\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = S^2$$

Como se calcula a variância?

A variância é uma medida resumo muito importante, pois ela nos fornece uma informação muito útil sobre a dispersão das observações em torno na média.

Também pode ser interessante calcular o Coeficiente de variação: É uma medida de dispersão relativa, elimina o efeito da magnitude dos dados e exprime a variabilidade em relação à média.

O CV é muito utilizado quando queremos comparar a variação entre dados com magnitudes diferentes.

Fórmula:

$$CV = \frac{S}{\bar{x}}\ 100\%$$

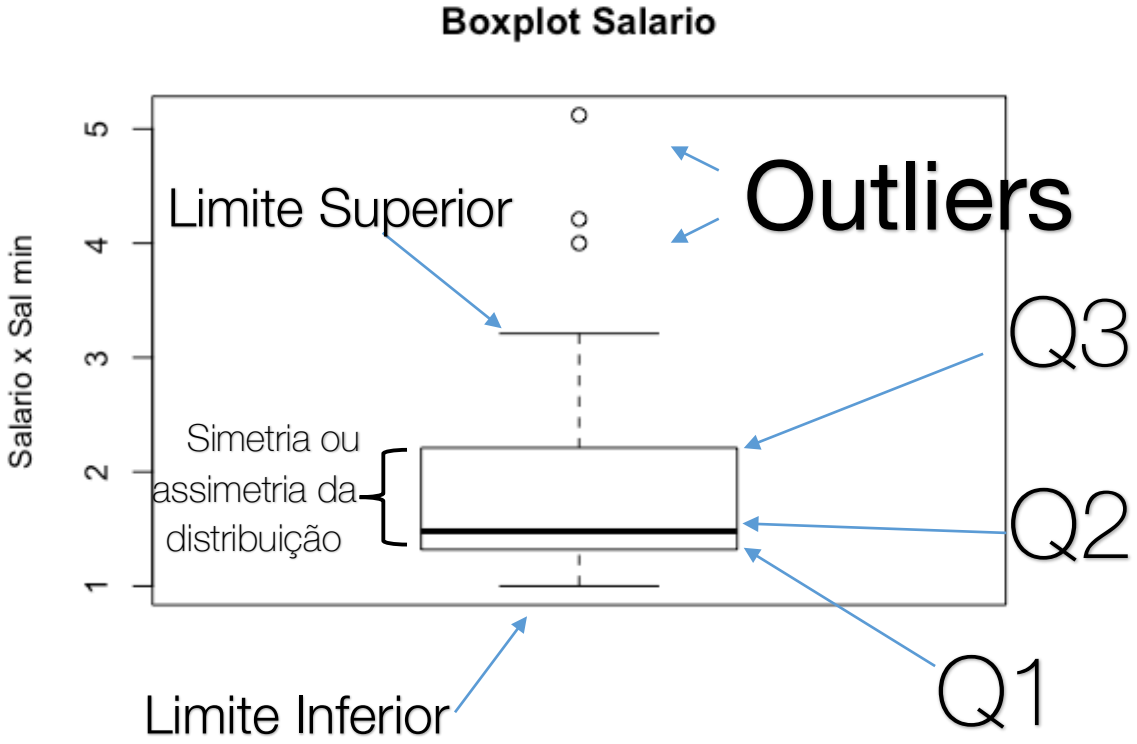
Interpretação: Quanto maior o CV, mais dispersos estão os dados ao redor da média.

Quanto menor o CV, mais concentrados os dados ao redor da média.

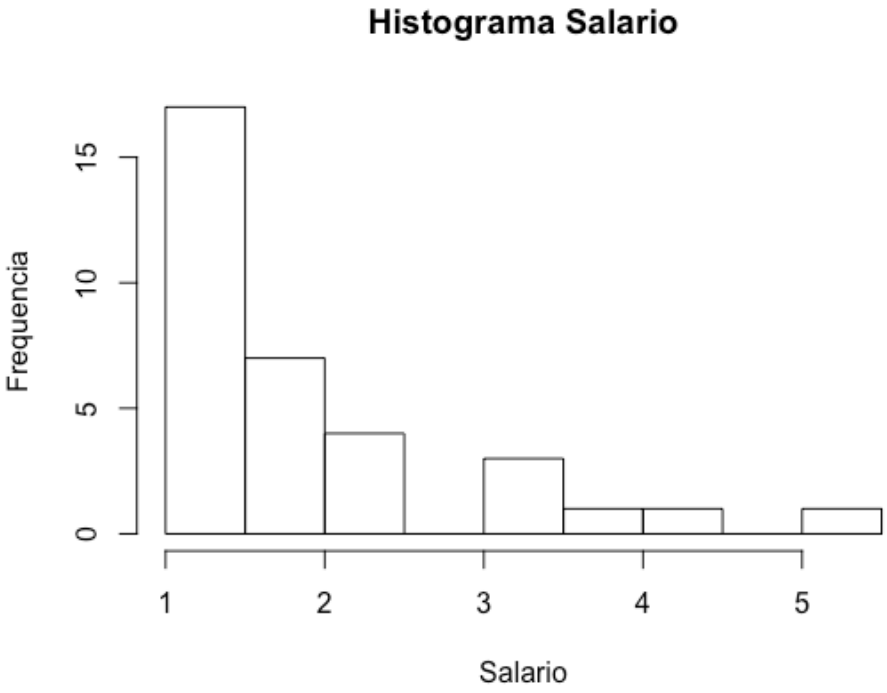
Boxplot

Agora que já estudamos os percentis, podemos falar sobre outro método de representação gráfica muito comum: o Boxplot.

O Boxplot fornece vários tipos de informações ao mesmo tempo. É utilizado quando se quer ter uma noção geral sobre o comportamento da variável.

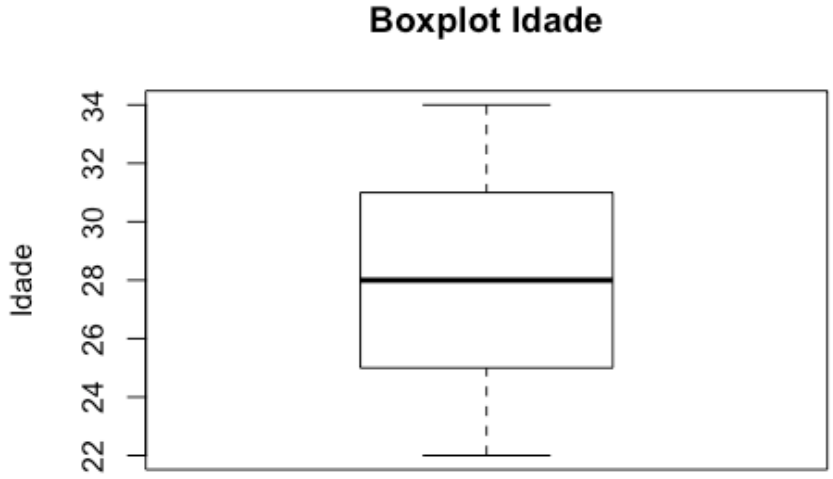


Observe um histograma para a mesma variável Salário.



Note a assimetria na distribuição à esquerda, ou seja, como há uma grande concentração de observações entre os valores de Salários de 1 a 1,5. Essa informação também é verificada no boxplot quando comparamos a espessura da faixa entre Q1 e Q2 com a espessura da faixa entre Q2 e Q3. Como? Já sabemos que 25% das observações estão abaixo de Q1, outros 25% entre Q1 e Q2, outros 25% entre Q2 e Q3 e os 25% restantes, acima de Q3. Assim, as espessuras dos retângulos do boxplot entregam os valores em que essas porcentagens de realizam. Se os valores estivessem igualmente distribuídos, esperaríamos observar um boxplot com retângulos de tamanhos iguais. Uma espessura pequena indica uma alta concentração e uma espessura grande uma alta dispersão dos dados.

Observe outro boxplot gerado a partir de uma sequência de idades com apenas uma observação para cada valor.



Observe o histograma correspondente:

Também é importante falarmos sobre os tracejados acima e abaixo da caixa: Os limites inferior e superior.

A grande maioria das nossas observações se encontra “dentro da caixa”, ou seja, entre Q1 e Q3.

Os limites inferior e superior são calculados para que possamos encontrar os outliers ou valores discrepantes, que são valores que se destacam dos demais. Isso pode ser útil também para identificarmos possíveis erros de digitação presentes no nosso banco de dados.

Fórmula para encontrar os limites superior e inferior:

$LS = Q3 + 1{,}5(Q3 - Q1)$
$LI = Q1 - 1{,}5(Q3 - Q1)$