Passo-a-passo

ETAPA 5. Medidas de Disperção

Prof. Pedro Feliú

Instituto de Relações Internacionais
Universidade de São Paulo

INTRODUÇÃO

Nesta etapa executaremos comandos referentes às medidas de tendência central e dispersão. Carregue no Rstudio o banco de dados "mundo", da pasta de banco de dados no moodle. É o mesmo banco de dados utilizado na etapa anterior. Caso já esteja carregado, vá direto aos passos abaixo. Não se esqueça que os comandos abaixo devem ser precedidos do comando **attach(mundo)**, como na etapa anterior.

PASSO 1: Medidas de dispersão

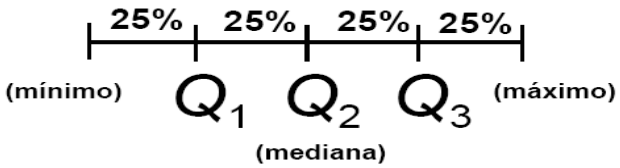
Vamos apresentar abaixo diversos comandos referentes às medidas de dispersão. Utilize os comandos no console do Rstudio, todos referentes à expectativa de vida dos países, variável que compõem o IDH. Como o banco de dados tem variáveis *missing*, utilizamos na.rm=TRUE depois do nome da variável, precedido de vírgula.

min(expectativa_vida, na.rm=TRUE) # menor valor da variável
max(expectativa_vida, na.rm=TRUE) # maior valor da variável
range(expectativa_vida, na.rm=TRUE) #intervalo da variável, mínimo ao máximo
quantile(expectativa_vida, na.rm=TRUE) #quartis da distribuição da variável
var(expectativa_vida, na.rm=TRUE) # variância da distribuição da variável
sd(expectativa_vida, na.rm=TRUE) #desvio padrão da distribuição da variável

```
> min(expectativa_vida, na.rm=TRUE)          # menor valor da variável
[1] 50.881
> max(expectativa_vida, na.rm=TRUE)         # maior valor da variável
[1] 84.27805
> range(expectativa_vida, na.rm=TRUE)      #intervalo da variável, mínimo ao máximo
[1] 50.88100 84.27805
> quantile(expectativa_vida, na.rm=TRUE)  #quartis da distribuição da variável
      0%      25%      50%      75%     100%
50.88100 66.43075 73.34600 77.30600 84.27805
> var(expectativa_vida, na.rm=TRUE)           # variância da distribuição da variável
[1] 63.05003
> sd(expectativa_vida, na.rm=TRUE)            #desvio padrão da distribuição da variável
[1] 7.940405
>
```

As medidas de dispersão descrevem o grau de heterogeneidade dos dados. O primeiro mais evidente são os valores mínimos e máximos. A expectativa de vida mais baixa de um país é 50,8 anos, enquanto a mais alta 84,2. Temos 33,4 anos de diferença entre o país em que mais se vive e aquele em que menos se vive, mostrando grande desigualdade no mundo. Note que a expectativa de vida pode ser afetada por nutrição, acesso à saúde, conflitos, violência, entre outros. A distribuição por quartis, acrescenta informação aos valores extremos ao "cortar" a distribuição em 4 partes iguais. Assim, os 25% de países com a menor expectativa de vida apresentam valores entre 50,8 anos e 66,4 anos. É como selecionar os 25% da porção inferior dos países em termos de expectativa de vida, ou quartil inferior, e medisse o mínimo e máximo dessa porção, podendo observar a heterogeneidade dentro desse grupo de países. No quartil superior, de 75% a 100%, os países variam de 77,3 anos a 84,2 anos. Em outras palavras, nos 25% dos países com maior expectativa de vida, há uma diferença de 6,9 anos entre o valor mínimo e máximo dentro desse grupo de países. Se voltarmos aos valores do quartil inferior, ou os 25% dos países com menor expectativa de vida, percebemos que a diferença entre o maior e menor valor intragrupo (66,4-50,8) retorna o valor 13,6. Isso significa que a desigualdade entre os países do quartil inferior é maior do que a desigualdade entre os países do quartil superior. Dado que a mediana é 73.3 anos (50%)

e o terceiro quartil (75%) 77,3, esse quarto de países possuem menor heterogeneidade, com uma diferença de 4 anos dentro desse grupo. A figura abaixo, presente no slide da aula 3, representa graficamente a descrição acima:



No nosso exemplo, o primeiro quartil (Q1) é igual a 66,4; o segundo quartil ou mediana (Q2) é 73,3; o terceiro quartil (Q3) é 77,3.

Passemos agora para a análise da variância, que pode ser definida com a fórmula abaixo:

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

A variância é a média dos quadrados de todos os desvios em relação à média. No caso que estamos analisando, a média da expectativa de vida dos países, [**mean(expectativa_vida, na.rm = TRUE)**], é 71,7 anos. Assim, pegamos cada valor da expectativa de vida de um país e subtraímos por 71,7. A soma de todos os valores resultados desta operação por país, dividido pelo total de países (média) retorna o valor desejado, 63 no caso da expectativa de vida mundial. A finalidade de se elevar ao quadrado é para evitar o sinal negativo que alguns desvios possuem. Uma distribuição com maior dispersão tenderá a apresentar desvios maiores em relação à média e, por conseguinte, uma variância mais elevada. Em outras palavras, quanto maior a variância, maior a heterogeneidade da distribuição dos dados.

O desvio padrão é basicamente a raiz quadrada da variância, também medindo o grau de heterogeneidade dos valores analisados, indicando quantas unidades as observações desviam da média:

$$Dp = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

No exemplo acima, na última linha de comando, observamos que o desvio padrão da expectativa de vida dos países do mundo é 7,9 anos. Enquanto a variância, por ser elevada ao quadrado, nos retorna uma unidade anos o quadrado (63), o desvio padrão nos retorna o valor na mesma unidade da variável, em anos. Deste modo, um desvio padrão de 7,9 anos significa o quão afastada da média é a distribuição da expectativa de vida no mundo. Faça o mesmo teste com outras variáveis explorando os comandos relatados acima.

PASSO 2: Funções que retornam várias estatísticas descritivas

Vamos utilizar uma função e um pacote que resumem várias medidas de dispersão e tendência central com apenas um comando. Comecemos com a função **summary()** para a variável crescimento do PIB de cada país, contabilizada em porcentagem ("pib_crescimento").

summary(pib_crescimento)

```
> summary(pib_crescimento)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
-20.599   1.333   3.084   2.958   4.754  25.163       7
> |
```

A função exposta na figura acima retorna o valor mínimo, o primeiro quartil, a mediana, média, terceiro quartil, valor máximo e quantidade de dados ausentes, respectivamente. Reparem que há países que decrescem -20,5% e crescem 25,1% seus PIB. A média é parecida com a mediana, indicando que são poucos os valores extremos como os mencionados. A distância inter-quartil, medida por Q3 - Q1 (4,75-1,33), é 3,4 do crescimento do PIB no mundo. Vejamos mais algumas medidas dessa mesma variável.

Utilizaremos agora o pacote **pastecs**. Instale esse pacote e depois o carregue com a função **library()**, em seguida, digite os comandos abaixo para a variável "pib_ crescimento":

install.packages("pastecs")
library(pastecs)
stat.desc(pib_crescimento)

```
> stat.desc(pib_crescimento)
     nbr.val     nbr.null       nbr.na          min          max        range          sum       median
 188.0000000    1.0000000    7.0000000  -20.5987707   25.1625331   45.7613038  556.1931583    3.0842984
        mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
   2.9584742    0.3271660    0.6454105   20.1230660    4.4858740    1.5162796
> |
```

A função **stat.desc** do pacote **pastecs** retorna várias medidas:

nbr.val – o número de observações (188 países)
nbr.null – o número de células vazias (1)
nbr.na – o número de dados ausentes (7)
min – o valor mínimo (-20,5%)
max – o valor máximo (25,1%)
range – o valor máximo menos o valor mínimo
sum – soma dos valores
median –mediana
mean – média
SE.mean – erro padrão da média
CI.mean.0.95 – intervalo de confiança de 95% da estimação da média
var – variância
std.dev – desvio padrão
coef.var – coeficiente de variação

Das medidas que ainda não vimos retornadas acima, destacarei apenas o coeficiente de variação, as demais medidas que dizem respeito à estatística inferencial (CI e SE) serão abordadas na segunda fase do curso. O coeficiente de variação nada mais é do que o desvio padrão dividido pela média. Ele é muito útil para comparar o grau de dispersão de duas distribuições que não estejam na mesma unidade. Por exemplo, comparar a dispersão do crescimento do PIB (%) no mundo com a dispersão das florestas no mundo (Km). Façamos o comando stat.desc() para realizar a comparação

stat.desc(floresta)

```
> stat.desc(floresta)
     nbr.val     nbr.null       nbr.na          min          max        range          sum       median
1.920000e+02 2.000000e+00 3.000000e+00 0.000000e+00 8.149305e+06 8.149305e+06 3.988518e+07 2.693500e+04
        mean      SE.mean CI.mean.0.95          var      std.dev     coef.var
2.077353e+05 5.667234e+04 1.117840e+05 6.166569e+11 7.852750e+05 3.780171e+00
> |
```

Como podemos observar, o coeficiente de variação (CV) do crescimento do PIB (figura anterior) é 1,5, enquanto o CV da distribuição de florestas no mundo é 3,7. Como o CV permite a comparação, podemos dizer que a distribuição das florestas no mundo é mais dispersa do que o crescimento econômico. Florestas são mais desigualmente distribuídas no mundo do que o crescimento econômico.

Há mais pacotes que realizam função semelhante à recém-executada, pesquise algumas e poste os comandos no mural de avisos no moodle.

PASSO 3: Medidas de dispersão e tendência central agrupadas

Neste último passo vamos realizar as estatísticas descritivas por um grupo de variáveis específicas. Vamos mudar o banco de dados **SCM_Brazil_stata12**. Notem que esse banco está no formato STATA, assim, quando for carregado no Rstudio, deve ser escolhida a opção From Stata. Lembrem-se das etapas anteriores para carregar o banco de dados. Esse banco tem alguns países do mundo e variáveis selecionadas no formato série histórica, ou seja, para cada país, temos o dado desde 1990 a 2017. Assim, o que queremos fazer são as estatísticas descritivas por país. Vamos usar a variável PKO (Peace Keeping Operations), que mede quanto soldados cada país cedeu para as missões de paz da ONU em cada ano da nossa série histórica. Assim podemos comparar a variabilidade entre os países. Para tanto, recorreremos ao pacote *psych*. Utilizem os seguintes comandos:

install.packages("psych")
library(psych)
attach(SCM_Brazil_stata12)
describeBy(PKO, cname, na.rm=TRUE)

Console Terminal Jobs

~/

```
 Descriptive statistics by group
group: Argentina
   vars  n   mean     sd median trimmed    mad min  max range  skew kurtosis    se
X1    1 28 738.71 341.64    779  738.79 348.41  39 1436  1397 -0.05    -0.39 64.56
------------------------------------------------------------------------------------
group: Brazil
   vars  n   mean     sd median trimmed     mad min  max range skew kurtosis     se
X1    1 28 898.39 787.98   1137  848.42 1386.23  14 2493  2479  0.3    -1.21 148.91
------------------------------------------------------------------------------------
group: Canada
   vars  n   mean     sd median trimmed   mad min  max range skew kurtosis     se
X1    1 28 633.29 825.79    279  478.33 217.2  43 3285  3242 1.95     2.94 156.06
------------------------------------------------------------------------------------
group: Colombia
   vars  n  mean    sd median trimmed  mad min max range skew kurtosis   se
X1    1 28 18.86 40.31    4.5   10.75 6.67   0 208   208 3.69    14.39 7.62
------------------------------------------------------------------------------------
group: Democratic Republic of Congo
   vars  n  mean     sd median trimmed   mad min max range skew kurtosis    se
X1    1 28 78.75 248.22     10   12.17 14.08   0 977   977 3.14     8.21 46.91
------------------------------------------------------------------------------------
group: Egypt
   vars  n    mean      sd median trimmed    mad min  max range skew kurtosis    se
X1    1 28 1573.25 1658.35    838 1394.29 1150.5   0 5409  5409  0.8    -0.61 313.4
------------------------------------------------------------------------------------
group: Germany
   vars  n   mean     sd median trimmed    mad min  max range skew kurtosis    se
X1    1 28 366.61 337.42    285     323 164.57   6 1351  1345 1.56     1.71 63.77
------------------------------------------------------------------------------------
group: India
   vars  n    mean      sd median trimmed     mad min  max range  skew kurtosis     se
X1    1 28 4871.46 3388.58   4908 4895.25 4521.19  35 9483  9448 -0.07    -1.74 640.38
------------------------------------------------------------------------------------
group: Indonesia
   vars  n   mean     sd median trimmed    mad min  max range skew kurtosis     se
X1    1 28 956.11 968.95  517.5  881.58 756.87   5 2854  2849 0.54    -1.22 183.11
------------------------------------------------------------------------------------
```