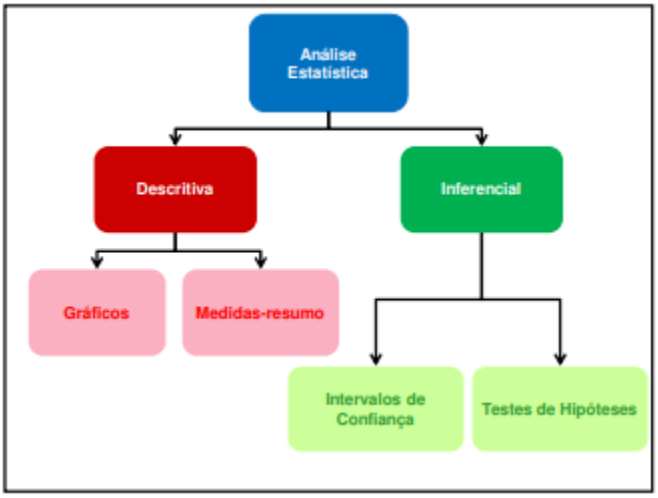
Estatística e Probabilidade
Aula 1 - Introdução

Professora: Josceli Tenório

Conceitos básicos

Estatística descritiva e inferencial



População e amostra

Para iniciar uma investigação é necessário estabelecer quais são os objetos em análise. A diferenciação entre os conceitos de população e amostra é importante para conduzir o tratamento dos dados e seu potencial para previsão.

Quando, em um estudo, todos os objetos são envolvidos na investigação teremos uma população. Por exemplo, todos os indivíduos que receberam um diploma de engenharia durante o ano acadêmico mais recente. Quando as informações desejadas estiverem disponíveis para todos os objetos da população, temos o que é denominado **censo**.

Restrições de tempo, dinheiro e outros recursos escassos normalmente tomam um censo impraticável ou inviável. Em vez disso, um subconjunto da população - uma amostra - é selecionado de uma forma prescrita. Dessa forma, podemos selecionar uma amostra dos formandos em engenharia do ano anterior para obter um retorno sobre a qualidade dos currículos.

Uma amostra deve representar significativamente a população da qual foi extraída. A amostra deve apresentar todas as características qualitativas e quantitativas do universo reproduzido.

O que nos induz a pensar a Figura 1?

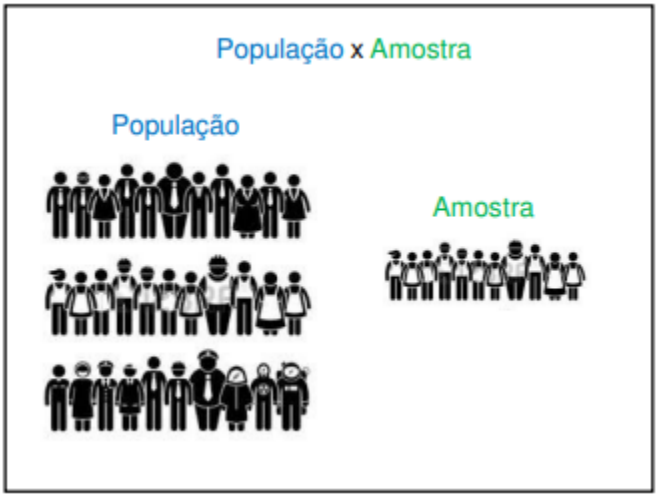


Figura 1. População e amostra.

Antes de podermos entender o que uma determinada amostra pode nos dizer sobre a população, devemos entender a incerteza associada à tomada da amostra de uma dada população. A Figura 2 mostra a relação entre esses elementos: a probabilidade faz suas considerações da população para a amostra (raciocínio dedutivo) e a inferência estatística faz considerações da amostra para a população (raciocínio indutivo).

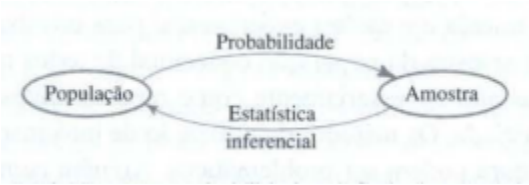


Figura 2. Relação entre probabilidade e estatística.

Normalmente, estamos interessados apenas em certas características dos objetos de uma população. Uma característica pode ser categorizada, como tipo de defeito, ou pode ter natureza numérica. No primeiro caso, o valor da característica é uma categoria enquanto, no último caso, o valor é um número. Uma variável é qualquer característica cujo valor pode mudar de um objeto para outro na população. A Figura 3 mostra os tipos de variáveis.

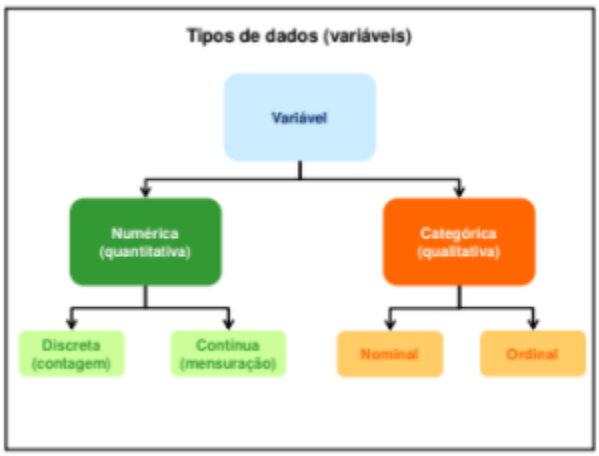


Figura 3. Tipos de variáveis

O que fazer com essas variáveis?

- **numéricas:** medidas-resumo:média, quartis, valores mínimo e máximo, desvio-padrão gráficos: diagrama de dispersão unidimensional, histograma, de pontos (scatter plot), boxplot
- **categóricas:** medidas-resumo: frequências absoluta e relativa (porcentagem) gráficos: setor circular (pizza), barra

Um conjunto de dados univariado consiste em observações sobre uma única variável. Temos dados bivariados quando as observações são feitas em cada uma de duas variáveis. Dados multivariados surgem quando são feitas observações sobre mais de duas variáveis. Em muitos conjuntos de dados multivariados, algumas variáveis são numéricas e outras são categorizadas.

O que usar?

1. Softwares: SPSS, Statistica
2. Linguagens de programação
    a. Python - pacote Pandas - (Google Colab)
    b. R programming
    c. NodeJS: https://www.npmjs.com/package/simple-statistics

R programming - Aula 0

```
#Vetores - tipos de dados
numerico <- c(100, 10, 49)
caractere <- c("a", "b", "c")
string<-c("segunda","terça")
boolean <- c(TRUE, FALSE, TRUE)
numerico[1]
boolean[c(2,3)]

#Lista
a <- list(
  a = 1:3,
  b = "sextou",
  c = pi,
  d = list(-1, -5)
)
a
```

```
vet <- c(-24, -50, 100, -350, 10)
names(vet) <- c("segunda-feira", "terça-feira", "quarta-feira",
"quinta-feira", "sexta-feira")
vet[1]

#Converter uma variável para factor()
status <- c("estudante", "não estudante", "estudante", "não
estudante")
estudante <- factor(status)
estudante

temperatura <- c("alta", "baixa", "alta","baixa", "media")
factor_temperatura <- factor(temperatura, order = TRUE, levels =
c("baixa", "media", "alta"))
temperatura

#Estrutura condicional
x <- 30
y <- 50
if (x > y) {
  print("x é maior")
} else {
  print("y é maior")
}

#Estrutura de repetição
valor <- c(16, 9, 13, 5, 2, 17, 14)
for(i in 3:length(valor)) {
  print(valor[i])
}


#DATAFRAMES

#O efeito da vitamina C no crescimento dos dentes em
porquinhos-da-índia
data(ToothGrowth)
str(ToothGrowth)
View(ToothGrowth)
#calculando a media
mean(ToothGrowth$len)
sd(ToothGrowth$len)
head(ToothGrowth)
```

```
head(ToothGrowth, 10)
str(ToothGrowth)


#GRAFICO
library(ggplot2)

# scatter plot
ggplot(ToothGrowth, aes(x = len, y = dose, color = supp)) +
  geom_point(size = 4)

#histograma
ggplot(ToothGrowth, aes(x = len)) +
  geom_histogram()

#densidade
ggplot(ToothGrowth, aes(x = len)) +
  geom_density()

#violin
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
ggplot(ToothGrowth, aes(x=dose, y=len)) +
  geom_violin() +
coord_flip()

#violin separado por cores
ggplot(ToothGrowth, aes(x=dose, y=len, color=dose)) +
  geom_violin() +
  coord_flip()

#violin com pontos
ggplot(ToothGrowth, aes(x=dose, y=len, color=dose)) +
  geom_violin() +
  geom_dotplot(binaxis='y', stackdir='center', dotsize=1)
  coord_flip()
```