Variables in Statistics

Variáveis em Estatística são informações/características que variam de observação para observação. Existem as **variáveis quantitativas**, quando estamos medindo tamanhos e quantidades; e **variáveis qualitativas**, são informações categóricas que não podem ser calculadas, normalmente contém texto, podendo conter números, mas que representem categorias.

Explorando dados do WNBA

O dataset pode ser acessado através deste link.

E o glossário dos termos neste link.

```
wnba <- read_csv("/home/marcella/Downloads/WNBA_Stats.csv")
```

```
## Rows: 143 Columns: 32
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): Name, Team, Pos, Birth_Place, Birthdate, College, Experience
## dbl (25): Height, Weight, BMI, Age, Games Played, MIN, FGM, FGA, FG%, 15:00,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(wnba)
```

```
## Rows: 143
## Columns: 32
## $ Name           <chr> "Aerial Powers", "Alana Beard", "Alex Bentley", "Alex M~
## $ Team           <chr> "DAL", "LA", "CON", "SAN", "MIN", "SEA", "PHO", "CHI", ~
## $ Pos            <chr> "F", "G/F", "G", "G/F", "G", "G", "G", "G", "G", "G", "~
## $ Height         <dbl> 183, 185, 170, 185, 175, 170, 188, 178, 185, 178, 180, ~
## $ Weight         <dbl> 71, 73, 69, 84, 78, 63, 81, 64, 76, 77, 76, 84, 113, 88~
## $ BMI            <dbl> 21.20099, 21.32944, 23.87543, 24.54346, 25.46939, 21.79~
```

```
## $ Birth_Place    <chr> "US", "US", "US", "US", "US", "US", "US", "US", "US", "~
## $ Birthdate      <chr> "January 17, 1994", "May 14, 1982", "October 27, 1990",~
## $ Age            <dbl> 23, 35, 26, 28, 23, 22, 23, 31, 24, 29, 30, 24, 24, 29,~
## $ College        <chr> "Michigan State", "Duke", "Penn State", "Georgia Tech",~
## $ Experience     <chr> "2", "12", "4", "6", "R", "R", "R", "8", "2", "5", "6",~
## $ `Games Played` <dbl> 8, 30, 26, 31, 24, 14, 16, 26, 30, 7, 30, 28, 25, 22, 4~
## $ MIN            <dbl> 173, 947, 617, 721, 137, 90, 112, 847, 834, 103, 843, 8~
## $ FGM            <dbl> 30, 90, 82, 75, 16, 9, 9, 166, 131, 14, 93, 154, 20, 18~
## $ FGA            <dbl> 85, 177, 218, 195, 50, 34, 34, 319, 346, 38, 183, 303, ~
## $ `FG%`          <dbl> 35.3, 50.8, 37.6, 38.5, 32.0, 26.5, 26.5, 52.0, 37.9, 3~
## $ `15:00`        <dbl> 12, 5, 19, 21, 7, 2, 4, 70, 29, 2, 20, 0, 2, 0, 0, 1, 0~
## $ `3PA`          <dbl> 32, 18, 64, 68, 20, 9, 15, 150, 103, 11, 62, 3, 8, 10, ~
## $ `3P%`          <dbl> 37.5, 27.8, 29.7, 30.9, 35.0, 22.2, 26.7, 46.7, 28.2, 1~
## $ FTM            <dbl> 21, 32, 35, 17, 11, 6, 2, 40, 104, 6, 38, 91, 9, 5, 0, ~
## $ FTA            <dbl> 26, 41, 42, 21, 12, 6, 2, 46, 129, 6, 51, 158, 12, 8, 0~
## $ `FT%`          <dbl> 80.8, 78.0, 83.3, 81.0, 91.7, 100.0, 100.0, 87.0, 80.6,~
## $ OREB           <dbl> 6, 19, 4, 35, 3, 3, 1, 9, 52, 3, 29, 34, 5, 12, 0, 16, ~
## $ DREB           <dbl> 22, 82, 36, 134, 9, 13, 14, 83, 75, 7, 97, 158, 18, 28,~
## $ REB            <dbl> 28, 101, 40, 169, 12, 16, 15, 92, 127, 10, 126, 192, 23~
## $ AST            <dbl> 12, 72, 78, 65, 12, 11, 5, 95, 40, 10, 50, 136, 7, 5, 1~
## $ STL            <dbl> 3, 63, 22, 20, 7, 5, 4, 20, 47, 5, 22, 48, 4, 3, 2, 1, ~
## $ BLK            <dbl> 6, 13, 3, 10, 0, 0, 3, 13, 19, 0, 4, 11, 5, 9, 0, 11, 2~
## $ TO             <dbl> 12, 40, 24, 38, 14, 11, 3, 59, 37, 2, 32, 87, 12, 6, 3,~
## $ PTS            <dbl> 93, 217, 218, 188, 50, 26, 24, 442, 395, 36, 244, 399, ~
## $ DD2            <dbl> 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 4, 0, 0, 0, 0, 0, 0, 0~
## $ TD3            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

Exemplo de variáveis qualitativas

```
wnba %>%select(Name, College, Pos) %>% head()
```

```
## # A tibble: 6 x 3
##   Name            College        Pos
##   <chr>           <chr>          <chr>
## 1 Aerial Powers   Michigan State F
## 2 Alana Beard     Duke           G/F
## 3 Alex Bentley    Penn State     G
## 4 Alex Montgomery Georgia Tech   G/F
## 5 Alexis Jones    Baylor         G
## 6 Alexis Peterson Syracuse       G
```

Exemplo de variáveis quantitativas

```
wnba %>%select(Age, Height, PTS, FTM) %>% head()
```

```
## # A tibble: 6 x 4
##     Age Height   PTS   FTM
##   <dbl>  <dbl> <dbl> <dbl>
## 1    23    183    93    21
## 2    35    185   217    32
## 3    26    170   218    35
## 4    28    185   188    17
## 5    23    175    50    11
## 6    22    170    26     6
```

Variáveis qualitativas, quando comparamos um indivíduo com outro, conseguimos perceber se há diferença. Mas não conseguimos determinar o tamanho e nem a direção dessa diferença (quem é maior ou menor), algo que em variáveis quantitativas é possível.

Existe um sistema de regras chamado Escala de Medida que possui 4 formas de medir uma variável.

- ordinal
- nominal
- intervalo
- razão

Escala Nominal

Numa escala nominal:

- percebemos quando os indivíduos são diferentes
- não podemos quantificar essa diferença
- nem indicar a direção da diferença
- mas podemos descrever as qualidades

Ainda que os valores dessa variável use números, os números representam classificações. Se uma jogadora de basquete usa a camisa 5 e outra a camisa 8, não faz sentido analisar qual é maior ou menor, ou dizer que o tamanho da diferença entre uma e outra camisa é 3. Os números são identificadores, não quantificam nada.

```
wnba %>% select(Pos) %>% head()
```

```
## # A tibble: 6 x 1
##   Pos
##   <chr>
## 1 F
## 2 G/F
## 3 G
## 4 G/F
## 5 G
## 6 G
```

Escala Ordinal

Agora vamos pensar na seguinte situação, se transformarmos a Altura que está em valores numéricos em uma classificação de Baixo, Médio ou Alto, a nova variável está numa escala nominal?

```
wnba %>%
  mutate(Height_labels = case_when(
    Height <= 170 ~ "Baixo",
    Height > 170 & Height <= 180 ~ "Médio",
    Height > 180 ~ "Alto"
  )) %>%
  select(Height,Height_labels) %>%
  head()
```

```
## # A tibble: 6 x 2
##   Height Height_labels
##    <dbl> <chr>
## 1    183 Alto
## 2    185 Alto
## 3    170 Baixo
## 4    185 Alto
## 5    175 Médio
## 6    170 Baixo
```

Apesar de ser um texto, sabemos que Baixo é menor que Médio ou Alto, e por isso a variável não se encaixa na escala nominal.

No exemplo da Altura que convertemos para classificações foi possível:

- dizer que os indivíduos são diferentes
- qual a direção da diferença (Baixo é menor que Alto por exemplo)

- mas ainda não é possível quantificar o tamanho dessa diferença

E é com essas características que temos a escala Ordinal.

Se temos uma competição, uma corrida, onde o Atleta A ficou na posição 1 do ranking, e o Atleta B na posição 2, sabemos que o 1 foi melhor, mais rápido que o primeiro, mas só usando este dado não é possível dizer quantos segundos ou minutos mais rápido.

Ou na hora de avaliar um produto numa escala de 1 a 5, onde 1 a pessoa odiou o produto, e 5 a pessoa amou. Sabemos que a nota vai gradativamente aumentando de acordo com o grau de satisfação da pessoa, mas ainda assim não conseguimos quantificar a diferença entre essas notas.

Intervalos e razão

Com a Escala de Intervalos ou Razão é possível mensurar o tamanho da diferença entre dois indivíduos. A diferença entre Intervalo ou Razão está na natureza do ponto zero.

Na Escala de Razão, zero (0) significa sem quantidade, por exemplo Peso == 0 significa ausência de peso.

Já a mesma variável peso numa escala de Intervalo, quando está zerada indica a presença de peso. Isso se dá, pois o cálculo é feito a partir da média da população. Se o peso do indivíduo é igual ao da média, o resultado será zero. Então o cálculo indica o quanto o indivíduo está diferente da média.

```
wnba <- wnba %>%
  mutate(Weight_deviation = Weight - mean(Weight, na.rm = TRUE))

wnba %>% select(Weight, Weight_deviation) %>% sample_n(5)
```

```
## # A tibble: 5 x 2
##   Weight Weight_deviation
##    <dbl>            <dbl>
## 1     84             5.02
## 2     88             9.02
## 3     65           -14.0
## 4    104            25.0
## 5     82             3.02
```

A variávei Peso (Weight) está numa escala de razão, enquanto a Weight_deviation está numa escala de intervalo.

Na escala de razão há duas formas de estabelecer a diferença entre 2 indivíduos:

- subtrair o valor do Peso de um indivíduo para o outro (Ex: 90 kg - 75 kg = 15 kg de diferença)
- calcular a proporção de diferença entre um Peso para outro (Ex: o indivíduo de 90 kg está 1,2 acima comparado ao indivíduo com peso de 75 kg. Basta dividir 90 por 75)

Agora na escala de intervalo, até é possível calcular a diferença de peso entre 2 indivíduos, olhando o exemplo abaixo, tanto pela escala de intervalo como de razão notamos 5kg de diferença. No entanto na escala de intervalo, se formos calcular proporção como fizemos na escala de razão, dividindo 10 por 5 vai resultar em 2 e induzir ao erro, fazendo entender que um indivíduo tem o peso 2 vezes maior que o outro.

```
wnba %>% select(Name, Weight,Weight_deviation) %>%
  filter(Name %in% c("Clarissa dos Santos","Alex Montgomery"))
```

```
## # A tibble: 2 x 3
##   Name                Weight Weight_deviation
##   <chr>                <dbl>            <dbl>
## 1 Alex Montgomery         84             5.02
## 2 Clarissa dos Santos     89            10.0
```

Com esse exemplo, a escala de intervalo parece não ser tão útil e portanto rara. No entanto existe utilidade para essa escala dependendo da informação, um exemplo é o tempo. Afinal não temos uma data com o valor zero, que defina com precisão o início do tempo, ou a ausência dele. Em outras palavras não conseguimos definir numa data como o “agora” há quanto tempo estamos desde o tempo zero.

Outro exemplo é a temperatura, quando dizemos que está 0ºC não significa que há ausência de temperatura. E que inclusive os graus podem ir abaixo de zero. Também é incorreto dizer que se ontem foi 15ºC e hoje está 30ºC, hoje foi o dobro da temperatura de ontem, só podemos afirmar que foi 15ºC mais quente. Dessa forma não podemos analisar uma informação de usa escala de intervalo com o raciocínio de uma escala de razão. Exceto a temperatura na escala Kelvin, onde 0k de fato significa ausência de temperatura.

Abaixo um resumo do comportamento das escalas.

checar/medir	Nominal	Ordinal	Intervalo	Proporção
se é diferente	S	S	S	S
a direção da diferença	N	S	S	S
o tamanho da diferença	N	N	S	S
variáveis quantitativas	N	S	S	S
variáveis qualitativas	S	S	N	N

Variáveis Discreta e Contínua

Uma lição que aprendemos até então é que nas escalas de intervalo e razão só é possível utilizar valores numéricos. Vamos dar uma olhada em duas informações da base, o Peso e os Pontos acumulados na temporada.

Se um jogador tem 92 pontos e outro tem 93, eles estão a 1 ponto de diferença, não tem como nessa informação ter meio ponto, ou 0.2 ponto. Os valores são inteiros, e dessa forma podemos chamar a variável de **discreta**.

Agora se olhamos o peso, entre 92 e 93 kg existe uma infinidade de valores possíveis nas casas decimais, e portanto chamamos essas variáveis de **contínuas**.

```
wnba %>% select(Name, Weight,PTS) %>%
  head()
```

```
## # A tibble: 6 x 3
##   Name            Weight   PTS
##   <chr>            <dbl> <dbl>
## 1 Aerial Powers       71    93
## 2 Alana Beard         73   217
## 3 Alex Bentley        69   218
## 4 Alex Montgomery     84   188
## 5 Alexis Jones        78    50
## 6 Alexis Peterson     63    26
```

Para determinar se a variável é contínua ou discreta, precisamos entender a natureza da informação, e não somente se limitar as valores disponíveis no momento. Por exemplo o próprio Peso que vemos acima aparenta ser discreta, mas sabemos que isso se dá apenas pelo fato da informação estar arredondada e existe uma infinidade de pesos nas casas decimais que representam as gramas com precisão.

Agora vamos olhar o Peso de algumas jogadoras.

```r
wnba %>%
  filter(Weight == 77) %>%
  pull(Weight) %>% head()
```

```
## [1] 77 77 77 77 77 77
```

Apesar de várias jogadoras terem na base o Peso de 77kg, sabemos que o peso não é igual. Provavelmente o valor foi arredondado e se tivesse uma casa decimal poderia variar entre 76,5kg até 77,5kg. Portanto, em geral, toda variável contínua possui um intervalo entre os valores, e o limite desses valores são chamados de **limites reais**, sendo o menor valor possível chamado de limite inferior (Ex 76,5kg), e o maior valor de limite superior (Ex 77,5kg).