Análise Descritiva

Descrição x Inferência

- A análise descritiva tem por objetivo sumarizar os dados, tornando a sua visualização mais fácil
- Ela não produz inferências, ou seja, serve apenas para organizar e apresentar informações
  - No entanto, muitas vezes apenas a descrição já pode ser suficiente para resolver um problema de pesquisa

Descrição x Inferência

- Através da análise descritiva podemos resumir um conjunto grande de informações em poucas linhas
- Na aula de hoje vamos usar a base sen2018.csv
  - Para acessá-la abram o Sigaa
  - Carreguem a base e o pacote “tydeverse”

Tabelas de Frequência

- Apresenta a frequência contagem de uma variável
  - No tidy, podemos usar a função count()
  - Ou a função table(), do Rbase

Tabela de Referência Cruzada (Contingência)

- Tabelas de contingência ou de referência cruzada trazem a frequência (contagem) a partir do cruzamento de duas variáveis
  - Qual o partido que apresentou mais candidaturas de mulheres ao senado em 2018?

Proporções

- Na tabela anterior, vimos que o PSOL apresentou o maior número de candidaturas de mulheres ao senado
- No entanto, os dados absolutos não são comparáveis entre si, porque cada partido lançou um número distinto de candidaturas
- Podemos tornar os casos comparáveis usando proporções

Tabelas com 3 variáveis

- Qual o partido que apresentou o maior número de mulheres negras ao senado em 2018?

Tabelas em HTML

- Uma forma de exportar as tribbles geradas em tidyverse para outros documentos é usando a função flextable() do pacote "flextable"
  - Ela gera uma tabela em HTML que pode ser aberta no navegador e copiada para outros documentos (planilhas ou texto)

Medidas de Tendência Central

- São estatísticas descritivas que representam um valor central em uma distribuição. As medidas mais comuns são a média, a mediana e a moda.
  - Média: retorna o valor médio dos casos -> $\bar{X} = \frac{\sum X}{N}$
  - Mediana: retorna a localização da posição média de uma distribuição -> $\text{Mediana} = \frac{N+1}{2}$
  - Moda: retorna o valor mais comum em uma distribuição. Não é a frequência e sim o valor!

Problemas da Média

- A média é uma medida imprecisa quando trabalhamos com dados de cauda pesada
  - Cauda pesada = concentração de casos em um dos lados do histograma
  - Isso acontece quando há muita dispersão nos nossos dados (desigualdade)
- Temperatura média em um deserto ou renda em Belém são exemplos dito.

Medidas de tendência central

- Vimos que a média, a mediana e a moda da idade coincidem no mesmo valor (55), mas podemos afirmar que esta variável apresenta uma distribuição normal?
- Para fazer esta verificação, vamos empregar o teste de Shapiro-Wilk
  - H1 = a distribuição não é normal
  - H0 = a distribuição é normal

Medidas de tendência central

Teste de normalidade de Shapiro-Wilk, usado para verificar se os dados seguem uma distribuição normal (gaussiana).

variável que está sendo testada para normalidade

```
        Shapiro-Wilk normality test

data:  sen2018$IDADE_DATA_POSSE
W = 0.98955, p-value = 0.02514
```

Valor da estatística de teste W calculada pelo teste de Shapiro-Wilk para os seus dados. Quanto mais próximo esse valor estiver de 1, mais os dados se assemelham a uma distribuição normal. Neste caso, o valor de W é aproximadamente 0,98955.

Valor-p é igual a 0,02514, o que é menor do que 0,05. Portanto, com um nível de significância de 0,05, você rejeitaria a hipótese nula e concluiria que os dados da variável "IDADE_DATA_POSSE" não seguem uma distribuição normal.

Medidas de Dispersão

- A partir da média, podemos calcular o desvio de cada caso. Esta medida indica a distância de um caso da média, logo é dada por:
  - $\mathrm{X} - \bar{X}$
- Para calcular o desvio de uma distribuição, o mais lógico seria somar todos os desvios e dividir pelo N. No entanto $\sum(X - \bar{X})$ será sempre igual a 0.
- Isso ocorre porque, em uma distribuição, os desvios positivos (acima da média) compensam os desvios negativos (abaixo da média).

Medidas de Dispersão

- Uma forma de contornar este problema é elevar ao quadrado todos os desvios. Esta medida é a variância ou s^2 ou σ^2 (sigma ao quadrado).
  - $s^2 = \frac{\sum(X-\bar{X})^2}{N}$

s^2 representa a variância amostral.

$\sum$ denota a soma.

X representa cada valor individual nos dados.

$\bar{X}$ é a média dos dados.

N é o número de observações na amostra.

A variância é uma medida importante, mas os valores estão em unidades ao quadrado, o que pode ser difícil de interpretar diretamente. Portanto, frequentemente usamos a raiz quadrada da variância, que é chamada de desvio padrão (s ou σ), para obter uma medida de dispersão na mesma escala que os dados originais.

Medidas de Dispersão

- Apesar de ser uma medida de dispersão, a variância nos diz muito pouco sobre a dispersão da distribuição, uma vez que ela não representa o valor exato da média dos desvios. Para isto, simplesmente podemos obter uma raiz quadrada da variância, ou o desvio padrão:
  - $s = \sqrt{\frac{\sum(X-\bar{X})^2}{N}}$
- No R, o desvio padrão pode ser obtido através da função sd()

Medida de Dispersão

- Como interpretar o Desvio Padrão?
  - É uma medida de dispersão, ou seja, quanto menor o seu valor, mais próximo da média se encontram os valores da distribuição.
  - Ele pode indicar a homogeneidade ou heterogeneidade dos dados: valores muito alto indicam que os dados estão muito dispersos, ou seja, variam muito em relação à média.
  - Imagine uma sala medindo 5m x 5m. No meio dela, você coloca uma mesa de centro e passa a distribuir o restante dos móveis a partir dela. A ± 1 metro da mesa você coloca os sofás, a 2 metros uma estante com livros, a -2m um vaso de flor... O desvio padrão seria o metro, ou a medida que você usa para distribuir os móveis nesta sala. Quanto maior for o desvio padrão, mais longe da mesa de centro os restantes dos móveis estarão posicionados.

Medidas de dispersão

- O valor da variância é mostrado em forma de notação científica.
  - Podemos alterar isso com a função options(scipen = 999)
  - Isso faz com que o R exiba números em seu formato decimal normal, sem usar a notação científica.

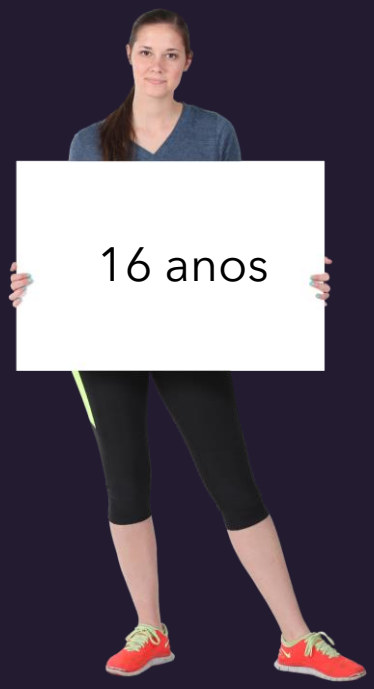
Medidas de dispersão

Exemplo genérico para calcular no quadro:

Todo mês acontece na UFPA a Feira de Agricultura Familiar.

Em uma das barriquinhas tem 3 crianças.

Queremos:
*Identificamos as idades das crianças
*Saber a média das idades
*Identificar a variância e o desvio padrão



$$\text{Média}(\bar{x}) = \frac{\sum \text{valores}}{\text{número de observações}}$$

$$\text{Variância}(s^2) = \frac{\sum(\text{valor} - \bar{x})^2}{\text{número de observações}}$$

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

$$\bar{x} = 12$$

$$S^2 = \frac{24}{3} = 8$$

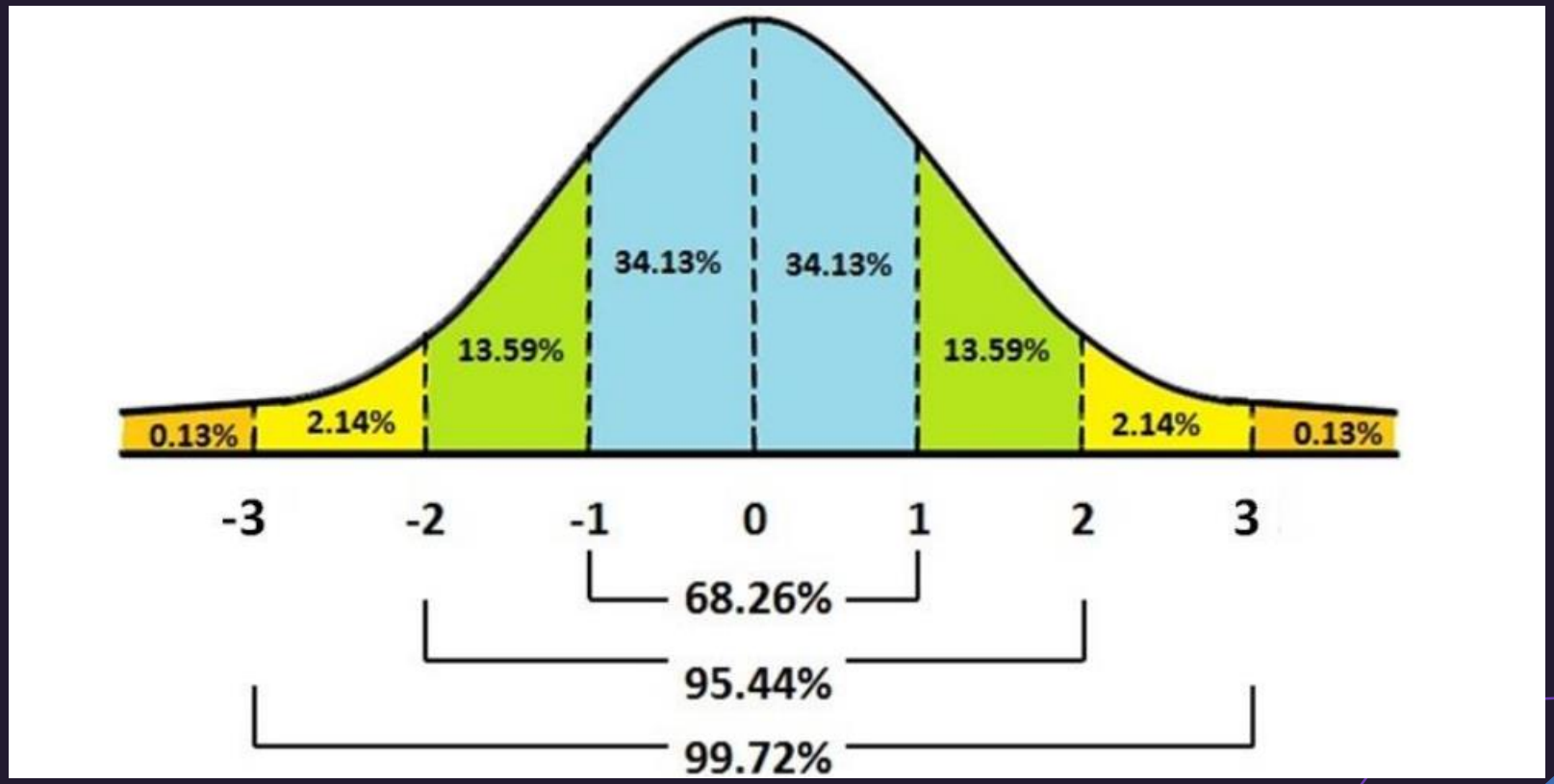
$$S = \sqrt{8} \approx 2{,}8$$

Escore Z (também conhecido como escore padrão)

- É uma medida de distância de um valor em relação à média do conjunto
  - Escore aponta a distância em desvios padrão
- Valores de escore Z negativos indicam que o valor produto é inferior à média, já valores positivos significa que estão situados acima da média.

Escore Z (também conhecido como escore padrão)

• Seus valores oscilam entre -3 < Z < +3 e isto corresponde a 99,72% da área sob a curva da Distribuição Normal.



Escore Z

Comumente utilizado em estudos de desempenho acadêmico, análises financeiras, controle de qualidade e etc.

- Calcular o escore z é bem simples:

- $Z = \frac{X - \bar{X}}{\sigma}$

- Onde:
  - X = valor bruto
  - $\bar{X}$ = média da distribuição
  - σ = desvio padrão

A média das idades é 55 anos

Escore Z

- Vamos calcular o escore z da idade do senador Jader Barbalho

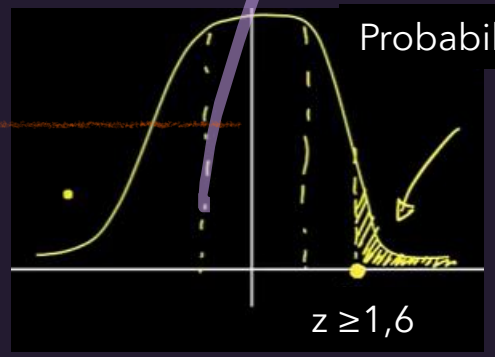
```
#####Escore Z#####
######Escore Z do Jader Barbabalho######
jader_z ← (74-mean(sen2018$IDADE_DATA_POSSE))/
  sd(sen2018$IDADE_DATA_POSSE)

jader_z #[1] 1.693606
```

1.indica que a idade de Jader está cerca de 1.693606 desvios padrão acima da média das idades no conjunto de dados. Ou seja? a idade de Jader é relativamente maior em comparação com a maioria das idades no conjunto de dados.

2.Como o escore Z é positivo, isso indica que a idade de Jader está acima da média das idades no conjunto de dados. Quanto maior o valor do escore Z, maior a diferença entre a idade de Jader e a média.

1,6 desvios padrões da média
P (z ≥1,6)

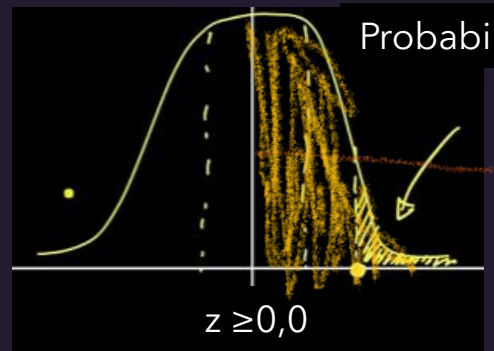
*mostrar o gráfico e ir para a tabela



```
> summary(sen2018$IDADE_DATA_POSSE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   27.0    47.0    55.0    55.1    63.0    83.0
>
```

Escore Z

```
######Calcular todos os Escores Z da base#####
sen2018$idade_z ← (sen2018$IDADE_DATA_POSSE -
                     mean(sen2018$IDADE_DATA_POSSE))/
                   sd(sen2018$IDADE_DATA_POSSE)

summary(sen2018$idade_z)
mean(sen2018$idade_z)

# > summary(sen2018$idade_z)
# Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
# -2.518651 -0.726202 -0.009222  0.000000  0.707758  2.500208
# > mean(sen2018$idade_z) #Explicar a leitura do valor em notação
# [1] 9.594753e-17
```



Qual a probabilidade abaixo da curva ?

*tabela

Escore Z

TABELA DA ÁREA SOB A CURVA DA DISTRIBUIÇÃO NORMAL

JADER

1,6 equivale a 0,94 ou seja 94%

OU SEJA

Até chegar a 1,6 eu tenho 94% das observações

Para calcular a probabilidade usamos o percentual

100% = 1

1-94= 0,07

Então a probabilidade de ter uma pessoa na base de dados com idade maior que a Jader é de 7%

BASE

0,0 equivale a 0,5 ou seja 50%

...

Até chegar a 0,0 eu tenho 50% das observações

O cálculo seria o mesmo, porém estamos olhando para as idades que apresenta uma curva de distribuição normal.
O que isso significa? A probabilidade de ter uma pessoa com até maior que 55 anos é de 50%

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Risco Relativo

- O Risco Relativo (RR) é uma medida derivada da epidemiologia que ajuda a entender as chances as chances de um evento ocorrer.
- Permite identificar a chance de um evento ocorrer em detrimento de outro (ou todos os outros) eventos
- Através dele podemos identificar a magnitude do fator de risco de um evento.

Risco Relativo

- O cálculo do risco relativo é feito a partir de tabelas 2x2 (duas linhas e duas colunas) e é dado por:

$$RR = \frac{Probabilidade\ de\ Ocorrência\ no\ Fator\ de\ Risco}{Probabilidade\ de\ Ocorrência\ fora\ do\ Fator\ de\ Risco}$$

Risco Relativo

- Ilustração:

	Ocorrência do Evento	
	SIM	Não
Presença do Fator de Risco	A	B
Ausência do Fator de Risco	C	D

$$RR = \frac{\frac{A}{A+B}}{\frac{C}{C+D}}$$

Risco Relativo

- Qual o risco relativo de uma mulher ser eleita senadora?

Risco Relativo

- Como interpretar?
  - Valores negativos indicam uma menor probabilidade de uma evento acontecer.
  - No exemplo, as mulheres tem menos 0,69 vezes a chance de serem eleitas, ou
  - 1-0,69 = -31% de chances de serem eleitas

Risco Relativo

- Limite para o Risco Relativo
  - Cervi (2014) aponta que fatores de risco inferiores a ±50% (risco relativo entre 0,5 e 1,5) são considerados não práticos, ou seja, devem ser desconsiderados.
  - No exemplo anterior, o Risco Relativo é de -31%, o que apontaria um fator não relevante.