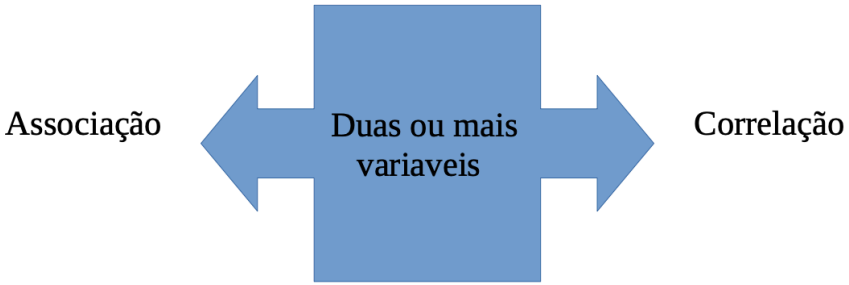
Existe…

- **associação** entre temperatura e umidade do ar?
- **correlação** entre índices de glicemia e colesterol alto?
- **associação** entre nota no ENEM e tipo de escola (pública/particular)?
- **associação** entre atividade física e a ocorrência de diabetes melittus, em homens?
- **correlação** entre área construída e preço de imóveis?
- **correlação** entre renda e consumo?

- Estamos interessados em verificar se existe “influência” - associação/correlação - de uma variável sob outra variável.

- Se existe, qual o grau da dependência?

  - que tipo de variáveis serão analisadas (qual a escala de medidas)?
  - quantas variáveis serão analisadas?
  - os resultados ”valem” em qualquer contexto?
  - associadas ou relacionadas: mesmo sentido?
  - as variáveis são relacionadas ou associadas?



E agora? O que usar?

As medidas de associação e correlação são usadas para avaliar relações entre variáveis,
com objetivos e interpretações diferentes.

- **Correlação**:
  - formalmente: correlação significa relação mútua entre dois termos ou estabelecer relação ou correlação entre os termos
  - interesse: avaliar “força” ou ‘grau” da dependência linear entre duas **variáveis quantitativas**
  - não indica **causalidade**, apenas a intensidade e direção da relação
  - coeficientes mais utilizados:
    - Pearson: relação linear entre duas variáveis quantitativas; usa os valores registrados
    - Spearman: mede a intensidade da relação entre variáveis ordinais; usa os postos
    - Kendall: mede a intensidade da relação entre duas variáveis ordinais, pareadas: concordância dos pares.

- **Associação**:
  - “força” e “direção” da relação entre **variáveis**, geralmente, **categóricas**
  - coeficientes mais usados:
    - razão de chances ou Odds Ratio: usada para verificar a chance de um evento ocorrer em um grupo em relação a outro.
    - coeficientes de contingência : usado para verificar a associação entre variáveis categóricas em tabelas de contingência.

- Também existem suposições para o uso destas medidas.

Duas variáveis qualitativas

- Os dados representam contagens das classes de respostas de duas **variáveis qualitativas**, geralmente expressos no formato de tabelas de contingência.
- Objetivo: testar a **associação** entre duas variáveis qualitativas.

Grupos	Variável		Total
	Não	Sim	
Grupo 1	n_{11}	n_{12}	$n_{1.}$
Grupo 2	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	n

- Aplicado nas situações em que os tamanhos das duas amostras são pequenos.
- Alternativa ao teste de Qui-quadrado: situações em que as frequências observadas/esperadas são menores do que 5.
- Alguns coeficientes e testes:
  - coeficiente de Pearson e de Cramér
  - teste qui-quadrado de Pearson, exato de Fisher.

Teste exato de Fisher

- **Amostra:**
  - **d**uas variáveis qualitativas e duas classes de resposta (em cada variável)
  - cada observação é classificada em uma única célula
  - os totais das linhas e colunas são fixos

- **Hipóteses:**
  - H_0: não existe diferença entre as proporções observadas nos dois grupos (os grupos são independentes; não existe associação entre os dois grupos)
  - H_1: as proporções são diferentes nos dois grupos.

  (também possível testar hipóteses unilaterais)

- **Estatística:**

$$T = n_{11}$$

- **Regra de decisão:**

  - Encontre o p-valor usando a distribuição da estatística T (distribuição hipergeométrica).
  - Usual: encontrar o p-valor:
    - some todas as probabilidades $p(k) = P(X = k)$ para todos os k tais que $p(k) \leq p(x_0)$, sendo x_0 o valor observado na tabela, isto é, o p-valor é tal que $p = P(p(k) \leq p(x_o))$.

- Para variáveis com mais do que duas classes de respostas,:
  - obtemos a estatística de Pearson:

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

em que: $$E_{ij} = \frac{(\text{total de observações na linha } i) \times (\text{total de observações na coluna } j)}{\text{total de observações}}$$

- Alguns coeficientes baseados na estatística de Pearson:

- **Coeficiente de Cramér:**

$$T_1 = \sqrt{\frac{\chi^2}{n \times (q-1)}},$$

  - *q* é o menor valor entre o número de linhas e o número de colunas
  - $0 < T_1 < 1$.
  - apropriado para tabelas maiores que 2 × 2.

- **Coeficiente de McNemar e Siegel:**

$$T_2 = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

  - $0 < T2 < 1$.

- Para ambas as medidas: 0 => não associação e, 1 => forte associação entre as duas variáveis.

Variáveis quantitativas

Correlação linear: coeficiente de Pearson

- Considere um par de variáveis contínuas (X, Y) e uma amostra de observações (pareadas) - independentes - destas variáveis: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

- Suposição: os valores observados tem distribuição normal (principalmente se n < 40).

- Usando um diagrama de dispersão é possível verificar a existência de um relação linear entre as variáveis

- **Coeficiente**:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

• **Observações**:

- assume valores no intervalo [-1,1]

- o sinal indica direção - positiva ou negativa - da relação:
  - se “+” => relação positiva entre as duas variáveis
  - se “-‘ => relação negativa entre as duas variáveis

- o valor sugere a força da relação entre as variáveis

- o **coeficiente não diferencia entre variáveis independentes e variáveis dependentes** (correlação entre X e Y é igual à correlação entre Y e X)

- o valor da correlação não muda se a unidade de mensuração das variáveis for alterada

- o coeficiente é **adimensional** (não tem unidade física)

- o coeficiente de **correlação é fortemente afetado** por valores discrepantes.

- **Mais observações**:

  Esta medida:

  Deve ser usada para:
  - explorar relações: “ajudar” a determinar a força e a direção de uma relação linear entre variáveis, através de uma medida de coeficiente numérico
  - auxiliar na seleção de variáveis em modelos preditos.

  Não deve ser usada:
  - para entender a causalidade: o coeficiente da análise de correlação não deve ser interpretado como uma evidência de **causalidade** entre as variáveis
  - para descrever relações não-lineares
  - no caso de existirem valores extremos no conjunto de dados.

- **Correlação espúria**

  - é uma relação entre variáveis que parecem estar correlacionadas, mas na verdade não têm relação de causa e efeito.

  - essa relação pode surgir devido a fatores externos ou **coincidências** estatísticas, levando a conclusões equivocadas se não forem analisadas corretamente.

  - alguns exemplo: *site* ***Spurious correlations***:

  a) correlação alta (66%) entre o número de pessoas afogadas por cair em uma piscina com o número de filmes em que Nicolas Cage aparece => quanto maior o número de filmes em que Nicolas Cage aparece, maior a chance de pessoas de se afogarem?

  b) relação entre a taxa de divórcio nos EUA e o consumo per capita de margarina => consumir mais margarina implica em aumento de problemas conjugais?

Correlação ρ de Spearman

- Este coeficiente mede a intensidade da relação, entre variáveis **ordinais**.
- Atribuição de postos
- Este coeficiente não é sensível a assimetrias na distribuição, nem a presença de valores discrepantes. Por isso, não é necessário que os dados provenham de duas populações normais.
- Pode ser aplicado em variáveis quantitativas - como alternativa ao coeficiente de Pearson (pela suposição de normalidade dos dados)
- Nos casos em que os dados não formam uma nuvem “bem comportada” (observar no diagrama de dispersão), com alguns pontos muito afastados dos restantes, ou nos casos em que parece existir uma relação crescente ou decrescente em formato de curva, o coeficiente de Spearman é mais apropriado.

- Considere um par de variáveis, pelo menos ordinais, (X, Y) e uma amostra de observações pareadas - independentes - destas variáveis: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

- Atribua postos separadamente para as observações de X e para as observações de Y. Se ocorrerem empates, atribua o posto médio.

- **Coeficiente**:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} \left(R(X_i) - R(Y_i)\right)^2}{n(n^2 - 1)}$$

- Assume valores no intervalo entre [-1, 1]:
  - quanto mais próximo de 1 => mais forte e a correlação positiva;
  - se for igual a (ou próximo de) zero => não existe correlação monotônica (as variáveis não têm relação ordenada).

- Como o coeficiente é obtido usando os postos, a escala das variáveis não afeta o valor da estatística:

- A ideia é a mesma do coeficiente de Pearson, porém usando postos.

Coeficente τ de Kendall

- Medida que avalia a força e a direção da associação entre duas variáveis ordinais.

- Útil em situações nas quais os dados não seguem uma distribuição normal, o que a torna uma alternativa robusta ao coeficiente de correlação de Pearson.

- Usado para comparar pares de observações: para todos os pares (x_i, y_i), (x_j, y_j), verifica-se se estes estão em concordância ou discordância. Se são:

  - concordantes: ambos os valores de um par estão na mesma ordem com outro par

  - discordantes: a ordem é invertida.

- O coeficiente varia no intervalo [1, 1].

- Considere duas variáveis (pareadas), pelo menos ordinais, (X, Y) e uma amostra de observações independentes destas variáveis: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

- Dada todas as possíveis combinações de 2 pares, $\binom{n}{2}$, temos
  - Nc: o número de pares concordantes
  - N_D: o número de pares discordantes

- **Coeficiente**:

$$\tau = \frac{N_C - N_D}{\frac{n(n-1)}{2}}$$

  - se todos os pares forem concordantes => $\tau = 1$

  - se todos os pares forem discordantes => $\tau = -1$

- **Observações**:

- **Coeficiente de correlação de Pearson** **não** deve ser usado para testar **independência** entre duas variáveis.
  - se for próximo de zero => ausência de correlação linear, mas as variáveis ainda podem ter relação não linear.

- **Independência estatística** significa que o conhecimento de uma variável não fornece informação sobre a outra.

- **Testes adequado para independência**
  - para testar independência entre duas variáveis:
    - variáveis qualitativas => **teste qui-quadrado de independência**
    - variáveis quantitativas => **teste de correlação de Spearman e de Kendall.**

- Com a aplicação dessas medidas (de Pearson, de Spearman e de Kendall) estimamos a influência total de uma variável aleatória sobre a outra, incluindo a influência indireta “sentida” porque a segunda variável aleatória está correlacionada não apenas com a primeira, mas talvez com uma terceira variável aleatória, que por sua vez está correlacionada com a primeira variável aleatória e, portanto, atua como portadora de influência indireta entre a primeira e a segunda variáveis aleatórias.

- Às vezes o interesse está obter uma medida da correlação entre duas variáveis, sob a condição de que a influência indireta devido às outras variáveis seja de alguma forma eliminada.

- Considerando três ou mais variáveis, os coeficientes de Pearson e de Kendall foram estendidos.

Coeficientes parciais de Pearson e de Spearman

- extensão dos coeficientes de correlação (observando a escala de medidas e suposições)

- mede a associação entre duas variáveis na situação em que a influência de uma variável sob as outras foi eliminada

- utilizado quando se deseja avaliar a relação entre duas variáveis eliminando o efeito de uma terceira.

- Caso particular: considere três variáveis, pelo menos ordinais, (X, Y, Z) e uma amostra de observações, independentes, destas variáveis: (x_1, y_1, z_1), (x_2, y_2, z_2),…, (x_n, y_n, z_n).

- **Coeficientes:**

$$\tau_{XY.Z} = \frac{\tau_{XY} - \tau_{XZ}\tau_{YZ}}{\sqrt{(1-\tau_{XZ}^2)(1-\tau_{YZ}^2)}}$$

e

$$r_{XY.Z} = \frac{r_{XY} - \tau_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}}$$

- Se $\tau_{XY.Z}$ (ou $r_{XY.Z}$) for:

- próximo de 1 => X e Y estão fortemente correlacionados, eliminado o efeito de Z

- próximo de 0 => X e Y não são correlacionados, eliminado o efeito de Z

- negativo => a relação entre X e Y é inversa, eliminado o efeito de Z.

Vamos ao R!