Advanced Institute for Artificial Intelligence

Introdução à Estatística 2

https://advancedinstitute.ai

Introdução à Estatística 2

Teste de Hipótese

Referências e Fontes das Imagens

- Estatística Básica (Book)
- Think Stats (Book)
- Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python (Book)
- Stats Normaltest

Associação entre Variáveis

- Tem como objetivo, descrever simultâneamente a variabilidade de duas ou mais variáveis de forma que cada conjunto seja observado para uma mesma unidade observacional (pessoas, animais, plantas, peças, etc.)

Associação entre Variáveis

- Tem como objetivo, descrever simultâneamente a variabilidade de duas ou mais variáveis de forma que cada conjunto seja observado para uma mesma unidade observacional (pessoas, animais, plantas, peças, etc.)
  - Vamos iniciar com um par de variáveis (x,y), sendo $(x_i,\ y_i)$, $i = 1, ..., n$ pares de observações de duas variáveis:

Associação entre Variáveis

- Vamos iniciar com um par de variáveis (x,y), sendo (x_i, y_i), $i = 1, ..., n$ pares de observações de duas variáveis:

Associação entre Variáveis

- Vamos iniciar com um par de variáveis (x,y), sendo $(x_i,\ y_i)$, $i = 1, ..., n$ pares de observações de duas variáveis:

x	y
Qualitativa	Qualitativa
Qualitativa	Quantitativa
Quantitativa	Qualitativa

Associação entre Variáveis

- Vamos iniciar com um par de variáveis (x,y), sendo $(x_i,\ y_i)$, $i = 1, ..., n$ pares de observações de duas variáveis:

x	y
Qualitativa	Qualitativa
Qualitativa	Quantitativa
Quantitativa	Qualitativa

Para cada combinação de pares de variáveis, um tipo de análise será realizada.

Associação entre Variáveis

- Para entendermos a associação entre x e y, precisamos de uma medidade associação, que deve avaliar se essa associação é **forte** ou **fraca**, **positiva** ou **negativa**.

Associação entre Variáveis

- Para entendermos a associação entre x e y, precisamos de uma medidade associação, que deve avaliar se essa associação é **forte** ou **fraca**, **positiva** ou **negativa**.
- Outra possibilidade é através da **representação gráfica, podendo para esse tipo de análise ser:**
  - Sentido da associação: Positiva ou Negativa;
  - Intensidade da Associação: Forte, Moderada ou Fraca

Entender os sentidos e sua intensidade é necessário e muito utilizado em processos de análise de **predições** nos dados.

Associação entre Variáveis

- x: Altura da planta
- y: Largura da folha

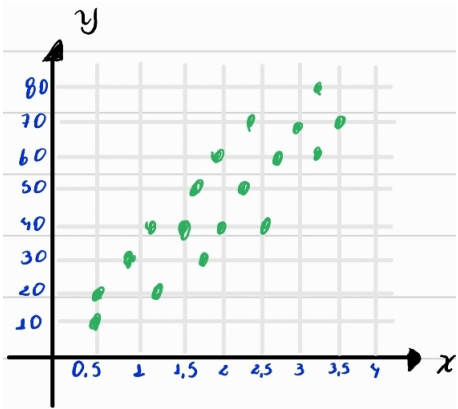


Figure: Correlação Positiva

Associação entre Variáveis

- x: Idade
- y: Nº de acidentes

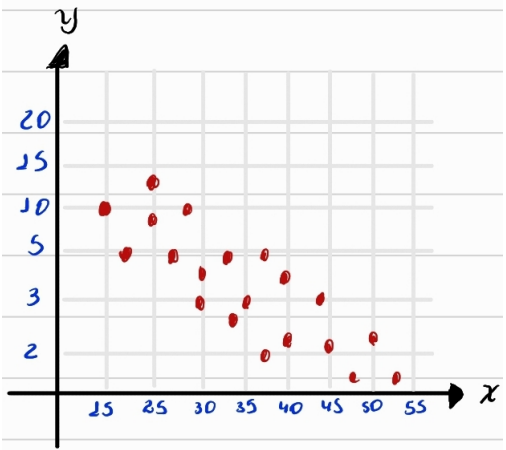


Figure: Correlação Negativa

Associação entre Variáveis

- x: Nº do sapato
- y: Nota final do semestre

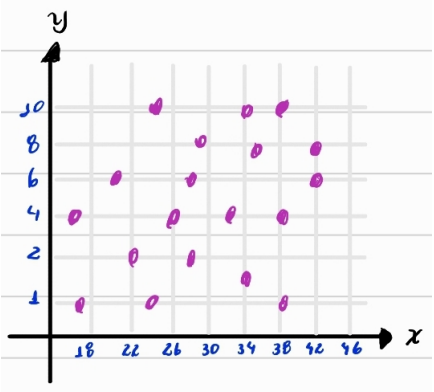


Figure: Sem Associação

Tipos de relação entre Variáveis

- Causal unilateral:
  - y depende de x ou x depende de y.

Tipos de relação entre Variáveis

- Causal unilateral:
  - y depende de x ou x depende de y.
- Exemplo:
  - Preço da venda de um produto **(y)** depende do local da venda **(x)**.

$$x \Rightarrow y$$

Tipos de relação entre Variáveis

- Causal unilateral (Preço X Local)
  - É possível perceber então que o preço do produto depende do local da venda e também pode depender de outros fatores mas, o local não depende do preço do produto.

Tipos de relação entre Variáveis

- Causal bilateral:
  - y depende de x e x depende de y.

Tipos de relação entre Variáveis

- Causal bilateral:
  - y depende de x e x depende de y.
- Exemplo:
  - Peso **(x)** e circunferência abdominal **(y)** de uma pessoa.

$$x \Leftrightarrow y$$

Tipos de relação entre Variáveis

- Causal bilateral (Peso X Circunferência abdominal)
  - Neste caso, se o peso cresce, a circunferência também aumenta porém, se a circunferência aumenta, o peso também será maior, comprovando que uma variável depende da outra.

Tipos de relação entre Variáveis

- Dependência Indireta:
  - Considerado uma condição que gera discussão em estatística devido ao fato de existir uma correlação mas não existir causa

Tipos de relação entre Variáveis

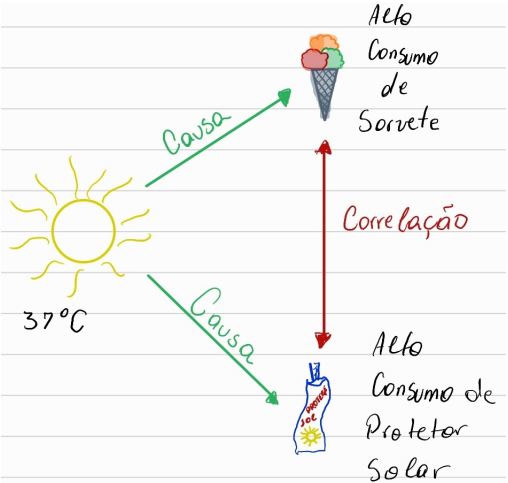
- Dependência Indireta:
  - Considerado uma condição que gera discussão em estatística devido ao fato de existir uma correlação mas não existir causa
- Exemplo:
  - Vendas de sorvete na praia **(x)**, causas de afogamento **(y)** e temperatura **(w)**.

Tipos de relação entre Variáveis

- Aqui, **(w)** possui uma relação causal unilateral com **(x)** e **(w)** também tem relação causal unilateral com **(y)**, ou seja, aumentar **(w)** irá causar aumento tanto em **(x)** como em **(y)**.

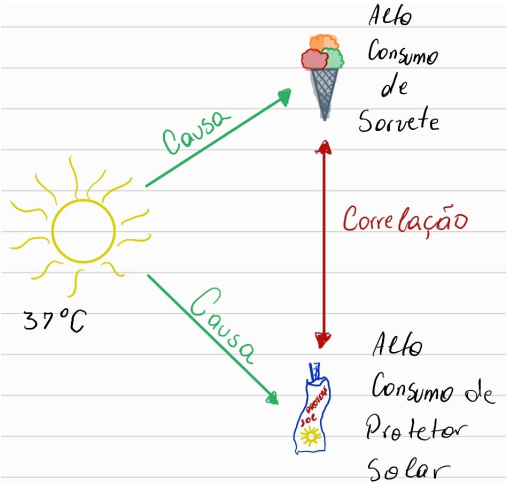
Correlação não é Causalidade

- **Altas temperaturas:** Variável independente (x)
- **Aumento das vendas:** Variável dependente (y)



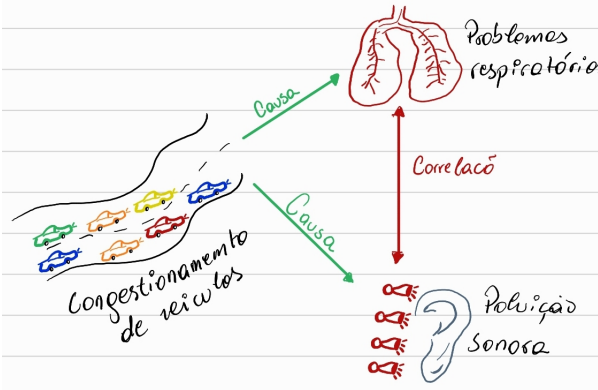
Correlação não é Causalidade

- Existe uma correlação entre o consumo de sorvetes e o consumo de protetor solar porém, um não causa o outro (Associação não causal)



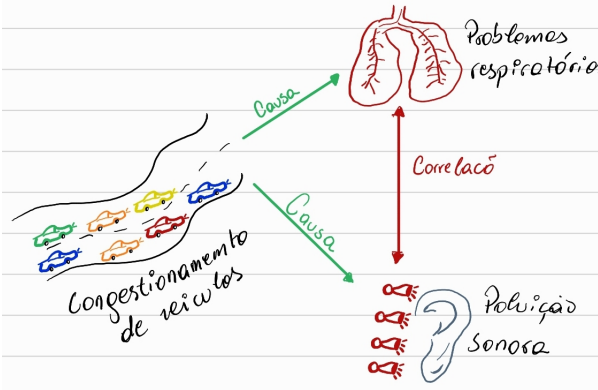
Correlação não é Causalidade

- **Congestionamento de veículos:** Variável independente (x)
- **Nível de Poluição:** Variável dependente (y)



Correlação não é Causalidade

- Existe uma correlação entre os problemas respiratórios e a poluição sonora porém, um não causa o outro (Associação não causal)



Testes para Comparação

- A primeira coisa a ser feita é encontrar qual é a **variável dependente**

Testes para Comparação

- A primeira coisa a ser feita é encontrar qual é a **variável dependente**
- Vamos dividir nosso exemplo em 3 partes:

Testes para Comparação

- A primeira coisa a ser feita é encontrar qual é a **variável dependente**
- Vamos dividir nosso exemplo em 3 partes:
  - Qualitativa Nominal;
  - Qualitativa Ordinal;
  - Quantitativa

Testes para Comparação

- Vamos iniciar com as variáveis Qualitativas Nominais

Testes para Comparação

- Vamos iniciar com as variáveis Qualitativas Nominais
  - Qualitativa Nominal: Como já sabemos, para quando possuímos uma variável que tem uma característica que é uma qualidade e não uma ordem.

Testes para Comparação (Qualitativas Nominais)

- Teste Hipotético I
  - Vamos imaginar que um psicólogo quer **comparar** o nível de ansiedade de seu grupo de pacientes, para isso, aplicou um questionário contendo 20 questões.

Testes para Comparação (Qualitativas Nominais)

Teste Hipotético I

- Vamos imaginar que um psicólogo quer **comparar** o nível de ansiedade de seu grupo de pacientes, para isso, aplicou um questionário contendo 20 questões.
- De acordo com os valores das questões no questionário será dado o diagnóstico inicial sendo, caso obtenha 10 pontos ou mais, será considerado como **estressado**, caso obtenha menos de 10 pontos, seu diagnóstico terá um quadro **normal**.

Testes para Comparação (Qualitativas Nominais)

Teste Hipotético I

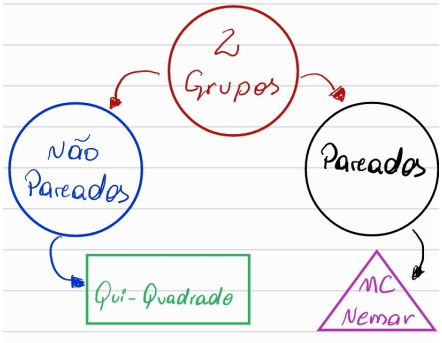
- Vamos imaginar que um psicólogo quer **comparar** o nível de ansiedade de seu grupo de pacientes, para isso, aplicou um questionário contendo 20 questões.
- De acordo com os valores das questões no questionário será dado o diagnóstico inicial sendo, caso obtenha 10 pontos ou mais, será considerado como **estressado**, caso obtenha menos de 10 pontos, seu diagnóstico terá um quadro **normal**.
- Podemos verificar então que ao final, a resposta da nossa pesquisa será se o indivíduo está estressado ou não.

Testes para Comparação (Qualitativas Nominais)

- Mas... qual seria o melhor teste comparativo?

Testes para Comparação (Qualitativas Nominais)
Não Pareados

- Diagrama para variáveis qualitativas nominais (2 grupos)



Testes para Comparação (Qualitativas Nominais)
Não Pareados e Pareados

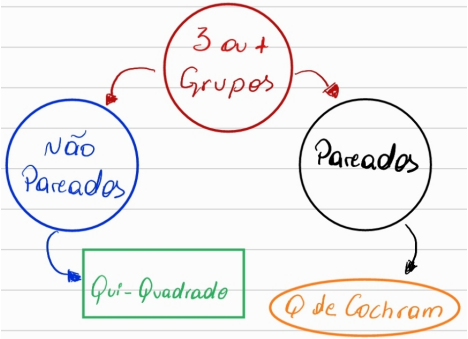
- Diagrama para variáveis qualitativas nominais (2 grupos)
  - Qui-quadrado: trata-se de um teste de hipóteses que se destina a encontrar um valor da dispersão para duas variáveis **categóricas nominais** e avaliar a associação existente entre essas variáveis.

Testes para Comparação (Qualitativas Nominais)
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (2 grupos)
  - Qui-quadrado: trata-se de um teste de hipóteses que se destina a encontrar um valor da dispersão para duas variáveis **categóricas nominais** e avaliar a associação existente entre essas variáveis.
  - MC Nemar: é um teste não paramétrico que se baseia em dados ordinais e nominais e não requerem os pressupostos dos testes paramétricos.

Testes para Comparação (Qualitativas Nominais)
Pareados

- Diagrama para variáveis qualitativas nominais (3 grupos ou mais)



Testes para Comparação (Qualitativas Nominais)
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (3 grupos ou mais)
  - Qui-quadrado: trata-se de um teste de hipóteses que se destina a encontrar um valor da dispersão para duas variáveis **categóricas nominais** e avaliar a associação existente entre essas variáveis.

Testes para Comparação (Qualitativas Nominais)
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (3 grupos ou mais)
  - Qui-quadrado: trata-se de um teste de hipóteses que se destina a encontrar um valor da dispersão para duas variáveis **categóricas nominais** e avaliar a associação existente entre essas variáveis.
  - Q de Cochram: é um teste estatístico não paramétrico para verificar se k tratamentos têm efeitos similares ao realizado nas análises de delineamentos em blocos aleatorizados, onde a variável de resposta pode assumir apenas dois valores possíveis ($\mathbf{0}$ ou $\mathbf{1}$).

Testes para Comparação

- Vamos agora com as variáveis Qualitativas Ordinal

Testes para Comparação

- Vamos agora com as variáveis Qualitativas Ordinal
  - Qualitativa Ordinal: Como próprio nome diz, existe uma ordem (hierarquia) na classificação dos dados na qual precisas ser respeitada

Testes para Comparação (Qualitativas Ordinais)

- Teste Hipotético II
  - Vamos imaginar que um psicólogo quer **comparar** o nível de ansiedade de seu grupo de pacientes, para esse novo teste, o psicólogo aplicou agora um questionário contendo 50 questões.

Testes para Comparação (Qualitativas Ordinais)

Teste Hipotético II

- Vamos imaginar que um psicólogo quer **comparar** o nível de ansiedade de seu grupo de pacientes, para esse novo teste, o psicólogo aplicou agora um questionário contendo 50 questões.
- Nesse novo experimento, a análise realizada ao final será aplicada dependendo das respostas do indivíduo, classificando como: **ansiedade**, **ansiedade moderada**, **ansiedade leve**, **muito ansioso** e **sem ansiedade**.

Testes para Comparação (Qualitativas Ordinais)

Teste Hipotético II

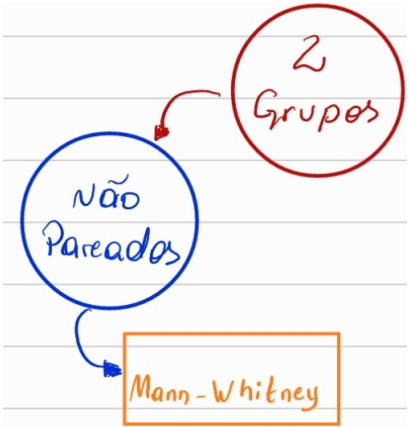
- Vamos imaginar que um psicólogo quer **comparar** o nível de ansiedade de seu grupo de pacientes, para esse novo teste, o psicólogo aplicou agora um questionário contendo 50 questões.
- Nesse novo experimento, a análise realizada ao final será aplicada dependendo das respostas do indivíduo, classificando como: **ansiedade**, **ansiedade moderada**, **ansiedade leve**, **muito ansioso** e **sem ansiedade**.
- Podemos verificar então que ao final, a resposta da nossa pesquisa irá apresentar em que nível de estresse se encontra um determinado indivíduo.

Testes para Comparação (Qualitativas Ordinais)

- Mas... qual seria o melhor teste comparativo?

Testes para Comparação (Qualitativas Ordinais)

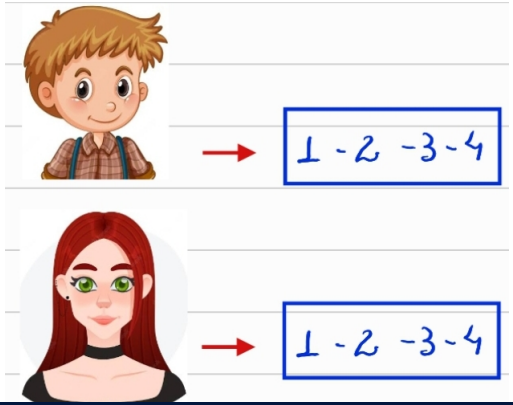
- Diagrama para variáveis qualitativas nominais (2 grupos)



Testes para Comparação (Qualitativas Ordinais)

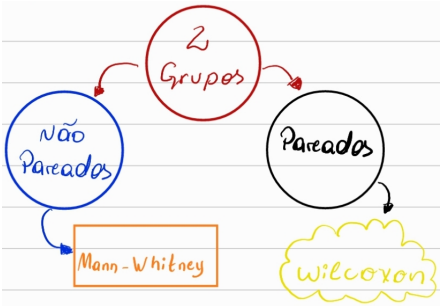
Não Pareados

- Diagrama para variáveis qualitativas nominais (2 grupos)



Testes para Comparação (Qualitativas Ordinais)
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (2 grupos)



Testes para Comparação (Qualitativas Ordinais)
Não Pareados e Pareados

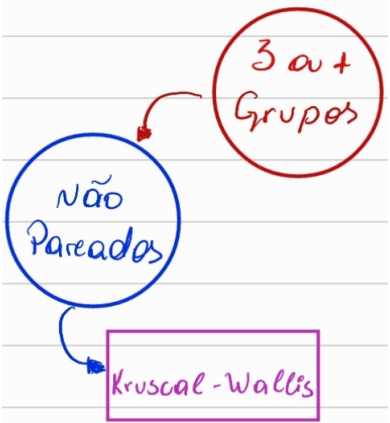
- Diagrama para variáveis qualitativas nominais (2 grupos ou mais)
  - Mann-Whitney: indicado para comparação de dois grupos não pareados para se verificar se pertencem ou não à mesma população e cujos requisitos para aplicação do **teste t de Student** não foram cumpridos

Testes para Comparação (Qualitativas Ordinais)
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (2 grupos ou mais)
  - Mann-Whitney: indicado para comparação de dois grupos não pareados para se verificar se pertencem ou não à mesma população e cujos requisitos para aplicação do **teste t de Student** não foram cumpridos
  - Wilcoxon: é um método não-paramétrico para comparação de duas amostras pareadas ou não pareadas.

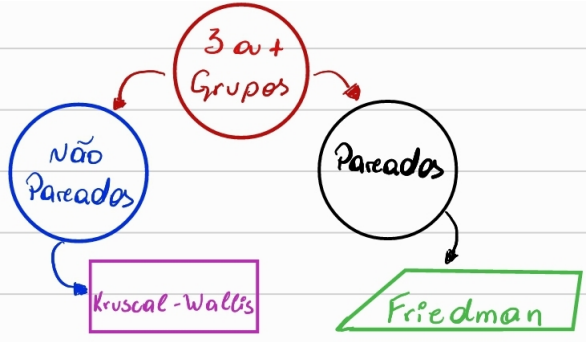
Testes para Comparação (Qualitativas Ordinais)

- Diagrama para variáveis qualitativas nominais (3 grupos)



Testes para Comparação (Qualitativas Ordinais)

- Diagrama para variáveis qualitativas nominais (3 grupos)



Testes para Comparação (Qualitativas Ordinais)
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (3 grupos ou mais)
  - Kruskal-Wallis: é um método não paramétrico para testar se amostras se originam ou não da mesma distribuição. Muito utilizado para comparar duas ou mais amostras independentes de tamanhos iguais ou diferentes.

Testes para Comparação (Qualitativas Ordinais)
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (3 grupos ou mais)
  - Kruskal-Wallis: é um método não paramétrico para testar se amostras se originam ou não da mesma distribuição. Muito utilizado para comparar duas ou mais amostras independentes de tamanhos iguais ou diferentes.
  - Friedman: trata-se de um teste não-paramétrico utilizado para comparar dados amostrais vinculados, ou seja, quando o mesmo indivíduo é avaliado mais de uma vez. Esse testa não utiliza os dados numéricos diretamente, mas sim os postos ocupados por eles após a ordenação feita para cada grupo separadamente.

Testes para Comparação (Quantitativas)

Teste Hipotético III

- Vamos imaginar que um psicólogo quer **comparar** o nível de ansiedade de seu grupo de pacientes, só que agora, o psicólogo aplicou um teste psicológico que classificava o nível de ansiedade do indivíduo através de uma escala numérica.

Testes para Comparação (Quantitativas)

Teste Hipotético III

- Vamos imaginar que um psicólogo quer **comparar** o nível de ansiedade de seu grupo de pacientes, só que agora, o psicólogo aplicou um teste psicológico que classificava o nível de ansiedade do indivíduo através de uma escala numérica.
- Nessa escala que vai de 0 à 100, quanto maior o *score* alcançado, maior as chances do indivíduo estar em um quadro de estresse.

Testes para Comparação (Quantitativas)

Teste Hipotético III

- Vamos imaginar que um psicólogo quer **comparar** o nível de ansiedade de seu grupo de pacientes, só que agora, o psicólogo aplicou um teste psicológico que classificava o nível de ansiedade do indivíduo através de uma escala numérica.
- Nessa escala que vai de 0 à 100, quanto maior o *score* alcançado, maior as chances do indivíduo estar em um quadro de estresse.
- Agora, nesse tipo de análise nossa variável resposta independente se trata de um número, sendo esse responsável por impactar o resultado.

Testes para Comparação (Quantitativas)

- Mas... qual seria o melhor teste comparativo?

Testes para Comparação (Quantitativas

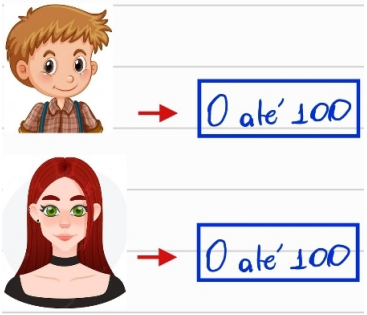
Não Pareados

- Diagrama para variáveis quantitativas (2 grupos)

Testes para Comparação (Quantitativas

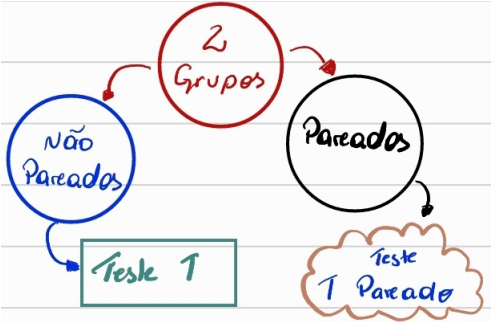
Não Pareados

- Diagrama para variáveis quantitativas (2 grupos)



Testes para Comparação (Quantitativas
Não Pareados e Pareados

- Diagrama para variáveis quantitativas (2 grupos)



Testes para Comparação (Quantitativas
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (3 grupos ou mais)
  - Teste T: é um teste de hipóteses utilizado quando queremos tirar conclusões de um grupo inteiro de indivíduos com base em apenas uma pequena amostra coletada.

Testes para Comparação (Quantitativas
Não Pareados e Pareados

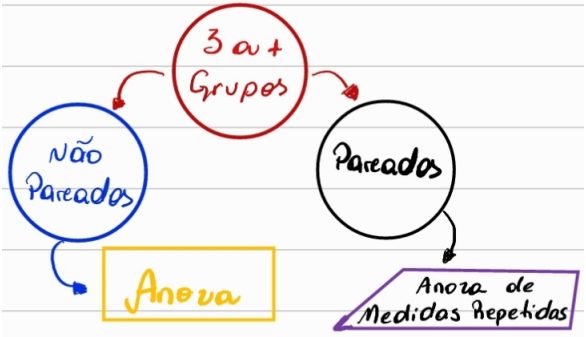
- Diagrama para variáveis qualitativas nominais (3 grupos ou mais)
  - Teste T: é um teste de hipóteses utilizado quando queremos tirar conclusões de um grupo inteiro de indivíduos com base em apenas uma pequena amostra coletada.
  - Teste-t pareado: simplesmente calcula a diferença entre observações emparelhadas (por exemplo, antes e depois) e, em seguida, realiza um teste-t para 1 amostra sobre as diferenças.

Testes para Comparação (Quantitativas
Não Pareados

- Diagrama para variáveis quantitativas (3 grupos)

Testes para Comparação (Quantitativas
Não Pareados e Pareados

- Diagrama para variáveis quantitativas (3 grupos)



Testes para Comparação (Quantitativas
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (3 grupos ou mais)
  - Anova: é uma fórmula estatística usada para comparar as variâncias entre as medianas (ou médias) de grupos diferentes.

Testes para Comparação (Quantitativas
Não Pareados e Pareados

- Diagrama para variáveis qualitativas nominais (3 grupos ou mais)
  - Anova: é uma fórmula estatística usada para comparar as variâncias entre as medianas (ou médias) de grupos diferentes.
  - Anova de Medidas Repetidas: compara as médias de uma ou mais variáveis que se baseiam em observações repetidas. Um modelo ANOVA de medidas repetidas também pode incluir zero ou mais variáveis independentes.

Iniciando os Testes

- Basicamente, sempre que iremos utilizar uma amostra de dados, calculamos algumas métricas estatísticas conhecidas (como média, mediana, desvio padrão e outras) e generalizamos esse valor para todo uma população, esse processo é conhecido como **inferência estatística**.
  - Mas essa generalização pode realmente ser feita?

Iniciando os Testes

- Basicamente, sempre que iremos utilizar uma amostra de dados, calculamos algumas métricas estatísticas conhecidas (como média, mediana, desvio padrão e outras) e generalizamos esse valor para todo uma população, esse processo é conhecido como **inferência estatística**.
  - Mas essa generalização pode realmente ser feita?
  - Será que nossa amostra é realmente uma boa representação da população?

Iniciando os Testes

- Basicamente, sempre que iremos utilizar uma amostra de dados, calculamos algumas métricas estatísticas conhecidas (como média, mediana, desvio padrão e outras) e generalizamos esse valor para todo uma população, esse processo é conhecido como **inferência estatística**.
  - Mas essa generalização pode realmente ser feita?
  - Será que nossa amostra é realmente uma boa representação da população?
  - Como provar isso estatisticamente?

Iniciando os Testes

- Basicamente, sempre que iremos utilizar uma amostra de dados, calculamos algumas métricas estatísticas conhecidas (como média, mediana, desvio padrão e outras) e generalizamos esse valor para todo uma população, esse processo é conhecido como **inferência estatística**.
  - Mas essa generalização pode realmente ser feita?
  - Será que nossa amostra é realmente uma boa representação da população?
  - Como provar isso estatisticamente?
  - **Simples, usando teste de hipóteses!**

Iniciando os Testes

- O teste de hipóteses é um ótima ferramenta para validar as nossas inferências, mas mesmo sendo ótima, muitas vezes ela é utilizada de forma errada ou simplesmente é esquecido o que realmente o teste representa.

Teorema Central do Limite

- Antes de apresentar as características de um teste de hipótese, é importante entendermos o que é o **Teorema Central do Limite** - **TCL**, pois trata-se de um dos principais conceitos por trás da inferência estatística.

Teorema Central do Limite

- Antes de apresentar as características de um teste de hipótese, é importante entendermos o que é o **Teorema Central do Limite - TCL**, pois trata-se de um dos principais conceitos por trás da inferência estatística.
- Basicamente, esse teorema nos diz que conforme aumentamos o tamanho de uma amostra, a distribuição amostral da sua média aproxima-se cada vez mais de uma distribuição normal, independentemente da distribuição da população.

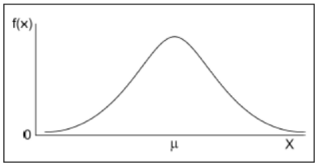


Figure: Distribuição normal: simétrica e com média

Teorema Central do Limite

- Mas o que isso significa???

Teorema Central do Limite

- Mas o que isso significa???
- Se utilizarmos a *feature* **Renda** do nosso conjunto de dados para analisar o salário médio dos brasileiros. Pela definição de média, teríamos que somar o salário de todos os brasileiros e dividir pelo tamanho da população do Brasil.

Teorema Central do Limite

- Mas o que isso significa???
- Se utilizarmos a *feature* **Renda** do nosso conjunto de dados para analisar o salário médio dos brasileiros. Pela definição de média, teríamos que somar o salário de todos os brasileiros e dividir pelo tamanho da população do Brasil.
- Isso é um pouco inviável de ser feito não acham?

Teorema Central do Limite

- Nesse caso, como não é possível ter informações da população toda, teríamos que utilizar amostras. Com isso, decidimos perguntar o salário de **10 pessoas** que encontramos aleatoriamente na rua (vamos ignorar o fato de que praticamente ninguém responderia a uma pergunta dessa rss)

Teorema Central do Limite

- Nesse caso, como não é possível ter informações da população toda, teríamos que utilizar amostras. Com isso, decidimos perguntar o salário de **10 pessoas** que encontramos aleatoriamente na rua (vamos ignorar o fato de que praticamente ninguém responderia a uma pergunta dessa rss)
- Então, pegaríamos a média desses 10 salários, o que nos retornaria um valor X_1.

Teorema Central do Limite

- Nesse caso, como não é possível ter informações da população toda, teríamos que utilizar amostras. Com isso, decidimos perguntar o salário de **10 pessoas** que encontramos aleatoriamente na rua (vamos ignorar o fato de que praticamente ninguém responderia a uma pergunta dessa rss)
- Então, pegaríamos a média desses 10 salários, o que nos retornaria um valor X_1.
- Em seguida, perguntamos o salário de mais 10 pessoas para obtermos mais uma média, X_2.

Teorema Central do Limite

- Nesse caso, como não é possível ter informações da população toda, teríamos que utilizar amostras. Com isso, decidimos perguntar o salário de **10 pessoas** que encontramos aleatoriamente na rua (vamos ignorar o fato de que praticamente ninguém responderia a uma pergunta dessa rss)
- Então, pegaríamos a média desses 10 salários, o que nos retornaria um valor X_1.
- Em seguida, perguntamos o salário de mais 10 pessoas para obtermos mais uma média, X_2.
- Repetimos esse processo com mais 10 pessoas, obtemos então outra média, X_3.

Teorema Central do Limite

- Nesse caso, como não é possível ter informações da população toda, teríamos que utilizar amostras. Com isso, decidimos perguntar o salário de **10 pessoas** que encontramos aleatoriamente na rua (vamos ignorar o fato de que praticamente ninguém responderia a uma pergunta dessa rss)
- Então, pegaríamos a média desses 10 salários, o que nos retornaria um valor X_1.
- Em seguida, perguntamos o salário de mais 10 pessoas para obtermos mais uma média, X_2.
- Repetimos esse processo com mais 10 pessoas, obtemos então outra média, X_3.
- Vamos supor agora que repetimos esse processo até obter mil médias amostrais, (X_1, X_2, ... X_{1000}).

Teorema Central do Limite

- Pelo **Teorema Central do Limite**, a distribuição dessas **mil médias** tende a ser normal.

Teorema Central do Limite

- Pelo **Teorema Central do Limite**, a distribuição dessas **mil médias** tende a ser normal.
- Se ao invés de perguntarmos a 10 pessoas, perguntássemos a 30, a distribuição dessas médias obtidas por meio de amostras de 30 pessoas, se aproximaria mais ainda de uma normal!

Teorema Central do Limite

- Pelo **Teorema Central do Limite**, a distribuição dessas **mil médias** tende a ser normal.
- Se ao invés de perguntarmos a 10 pessoas, perguntássemos a 30, a distribuição dessas médias obtidas por meio de amostras de 30 pessoas, se aproximaria mais ainda de uma normal!
- Conforme aumentássemos o número de pessoas questionadas sobre o seu salário, mais a distribuição dessas médias se aproximaria de uma normal.

Teorema Central do Limite

- E o melhor, a média dessa distribuição normal é uma ótima aproximação da média dos salários de toda a população.

Teorema Central do Limite

- E o melhor, a média dessa distribuição normal é uma ótima aproximação da média dos salários de toda a população.
- Com isso teríamos uma estimativa, ou seja, teríamos uma inferência da média dos salários de toda a população brasileira sem precisar perguntar isso para cada pessoa que vive no país. :-)

Teorema Central do Limite

- E o melhor, a média dessa distribuição normal é uma ótima aproximação da média dos salários de toda a população.
- Com isso teríamos uma estimativa, ou seja, teríamos uma inferência da média dos salários de toda a população brasileira sem precisar perguntar isso para cada pessoa que vive no país. :-)
- Podemos observar também que em nenhum momento há a definição de que os salários são normalmente distribuídos.

Teorema Central do Limite

- E o melhor, a média dessa distribuição normal é uma ótima aproximação da média dos salários de toda a população.
- Com isso teríamos uma estimativa, ou seja, teríamos uma inferência da média dos salários de toda a população brasileira sem precisar perguntar isso para cada pessoa que vive no país. :-)
- Podemos observar também que em nenhum momento há a definição de que os salários são normalmente distribuídos.
- Pelo TCL essa característica não é necessária. A distribuição das médias amostrais dos salários seguirão uma normal, mesmo que os salários sigam qualquer outra distribuição.

Teorema Central do Limite

- Vamos supor que os salários seguem uma distribuição normal.

Teorema Central do Limite

- Vamos supor que os salários seguem uma distribuição normal.
- Muitas pessoas recebem um salário médio e o número de pessoas que recebem um salário menor que a média é o mesmo número de pessoas que recebem um salário maior que a média.

Teorema Central do Limite

- Vamos supor que os salários seguem uma distribuição normal.
- Muitas pessoas recebem um salário médio e o número de pessoas que recebem um salário menor que a média é o mesmo número de pessoas que recebem um salário maior que a média.
- Por outro lado, acreditamos que o salário das pessoas seguem uma **distribuição exponencial**.

Teorema Central do Limite

- Vamos supor que os salários seguem uma distribuição normal.
- Muitas pessoas recebem um salário médio e o número de pessoas que recebem um salário menor que a média é o mesmo número de pessoas que recebem um salário maior que a média.
- Por outro lado, acreditamos que o salário das pessoas seguem uma **distribuição exponencial**.
- Muitas pessoas recebem um salário mais baixo, enquanto poucas pessoas recebem um salário mais alto.

Teorema Central do Limite

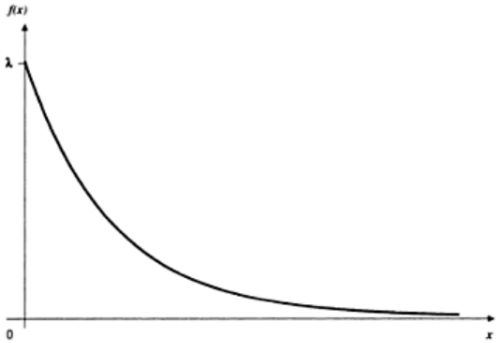


Figure: Distribuição exponencial.

Teorema Central do Limite

- Uma pessoa sonhadora acredita que vivemos em uma sociedade muito igualitária, onde os salários seguem uma **distribuição uniforme**, ou seja, as pessoas recebem salários parecidos.

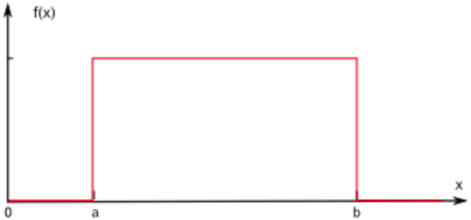


Figure: Distribuição uniforme.

Teorema Central do Limite

- Para obtermos uma melhor visualização, geramos uma distribuição contendo 100.000 valores aleatórios que seguem, respectivamente, uma distribuição **normal**, **exponencial** e **uniforme**.

Teorema Central do Limite

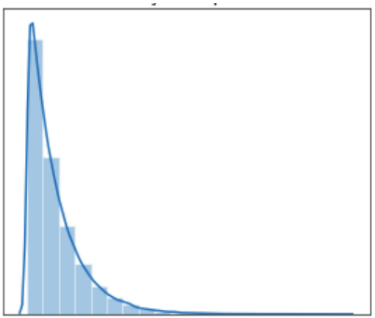
- Para obtermos uma melhor visualização, geramos uma distribuição contendo 100.000 valores aleatórios que seguem, respectivamente, uma distribuição **normal**, **exponencial** e **uniforme**.
- O valor médio das três distribuições é aproximadamente **2.000**, ou seja, a suposição é que o salário médio da população é = R$2.000.

Teorema Central do Limite

- Distribuição Normal

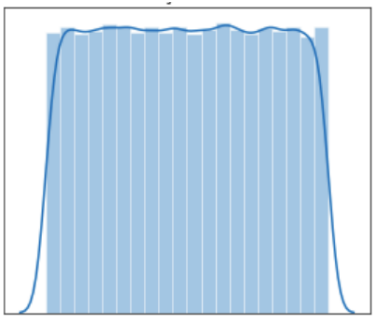
Teorema Central do Limite

- Distribuição Exponencial



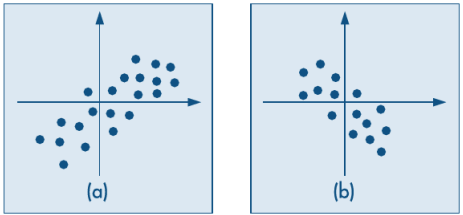
Teorema Central do Limite

- Distribuição Uniforme



Teorema Central do Limite

- Tipos de associações entre duas variáveis
  - (a) Associação linear direta (ou **positiva**)
    - Soma do produto das coordenadas será sempre positivo
  - (b) Dependência linear inversa (ou **negativa**)
    - Soma dos produtos das coordenadas será negativa



Construção das Hipóteses

- Agora que já sabemos o que é Teorema Central do Limite, vamos então entender o que realmente é um **teste de hipóteses**

Construção das Hipóteses

- Agora que já sabemos o que é Teorema Central do Limite, vamos então entender o que realmente é um **teste de hipóteses**
  - O **teste de hipóteses** é um método que nos permite verificar se certos dados amostrais trazem evidências que confirmam ou refutam uma hipótese já formulada.

Construção das Hipóteses

- Agora que já sabemos o que é Teorema Central do Limite, vamos então entender o que realmente é um **teste de hipóteses**
  - O **teste de hipóteses** é um método que nos permite verificar se certos dados amostrais trazem evidências que confirmam ou refutam uma hipótese já formulada.
- Mas o que isso quer dizer?

Construção das Hipóteses

- Agora que já sabemos o que é Teorema Central do Limite, vamos então entender o que realmente é um **teste de hipóteses**
  - O **teste de hipóteses** é um método que nos permite verificar se certos dados amostrais trazem evidências que confirmam ou refutam uma hipótese já formulada.
- Mas o que isso quer dizer?
  - Na construção das três populações que acabamos de verificar, deliberadamente definimos que as médias seriam aproximadamente 2.000.

Construção das Hipóteses

- Agora que já sabemos o que é Teorema Central do Limite, vamos então entender o que realmente é um **teste de hipóteses**
  - O **teste de hipóteses** é um método que nos permite verificar se certos dados amostrais trazem evidências que confirmam ou refutam uma hipótese já formulada.
- Mas o que isso quer dizer?
  - Na construção das três populações que acabamos de verificar, deliberadamente definimos que as médias seriam aproximadamente 2.000.
  - Então qualquer amostra que eu retirar dessa população iria confirmar a hipótese de que a média é igual a 2.000.

Construção das Hipóteses

- Porem, isso não ocorre com dados reais. Nós não sabemos qual é a média populacional de uma variável, temos apenas uma hipótese de qual seria esse valor, geralmente baseada em outras estimativas.

Construção das Hipóteses

- Porem, isso não ocorre com dados reais. Nós não sabemos qual é a média populacional de uma variável, temos apenas uma hipótese de qual seria esse valor, geralmente baseada em outras estimativas.
  - Um exemplo seria fazer uma pesquisa sobre o mercado de trabalho e encontrar matérias que dizem o seguinte sobre o salário:

Construção das Hipóteses

- Porem, isso não ocorre com dados reais. Nós não sabemos qual é a média populacional de uma variável, temos apenas uma hipótese de qual seria esse valor, geralmente baseada em outras estimativas.
  - Um exemplo seria fazer uma pesquisa sobre o mercado de trabalho e encontrar matérias que dizem o seguinte sobre o salário:
- Salário de cientista de dados está entre R$13.1 e R$26.7 mil.

Construção das Hipóteses

- Porem, isso não ocorre com dados reais. Nós não sabemos qual é a média populacional de uma variável, temos apenas uma hipótese de qual seria esse valor, geralmente baseada em outras estimativas.
  - Um exemplo seria fazer uma pesquisa sobre o mercado de trabalho e encontrar matérias que dizem o seguinte sobre o salário:
- Salário de cientista de dados está entre R$13.1 e R$26.7 mil.
  - Um excelente salário, não? Em um primeiro momento você fica animado com a perspectiva de receber esse salário mas, depois nos questionamos: (como cientista de dados, sempre faz!) "será que o salário de um cientista de dados é realmente esse?"

Construção das Hipóteses

- Porem, isso não ocorre com dados reais. Nós não sabemos qual é a média populacional de uma variável, temos apenas uma hipótese de qual seria esse valor, geralmente baseada em outras estimativas.
  - Um exemplo seria fazer uma pesquisa sobre o mercado de trabalho e encontrar matérias que dizem o seguinte sobre o salário:
- Salário de cientista de dados está entre R$13.1 e R$26.7 mil.
  - Um excelente salário, não? Em um primeiro momento você fica animado com a perspectiva de receber esse salário mas, depois nos questionamos: (como cientista de dados, sempre faz!) "será que o salário de um cientista de dados é realmente esse?"
  - O salário de um cientista de dados é R$ 19.900 (valor médio dos dois valores apresentados) ou não?

Construção das Hipóteses

- No caso, a hipótese já formulada afirma que o salário de um profissional da ciência de dados é R$ 19.900, segundo as informações obtidas.

Construção das Hipóteses

- No caso, a hipótese já formulada afirma que o salário de um profissional da ciência de dados é R$ 19.900, segundo as informações obtidas.
  - Essa hipótese já formulada é o que no teste de hipótese chamamos de **Hipótese Nula**, ou H_0.

Construção das Hipóteses

- No caso, a hipótese já formulada afirma que o salário de um profissional da ciência de dados é R$ 19.900, segundo as informações obtidas.
  - Essa hipótese já formulada é o que no teste de hipótese chamamos de **Hipótese Nula**, ou H_0.
  - Por sua vez, a hipótese que não confirma essa informação é chamada de **Hipótese Alternativa**, H_1.

Construção das Hipóteses

- No caso, a hipótese já formulada afirma que o salário de um profissional da ciência de dados é R$ 19.900, segundo as informações obtidas.
  - Essa hipótese já formulada é o que no teste de hipótese chamamos de **Hipótese Nula**, ou H_0.
  - Por sua vez, a hipótese que não confirma essa informação é chamada de **Hipótese Alternativa**, H_1.
- Temos então:

Construção das Hipóteses

- No caso, a hipótese já formulada afirma que o salário de um profissional da ciência de dados é R$ 19.900, segundo as informações obtidas.
  - Essa hipótese já formulada é o que no teste de hipótese chamamos de **Hipótese Nula**, ou H_0.
  - Por sua vez, a hipótese que não confirma essa informação é chamada de **Hipótese Alternativa**, H_1.
- Temos então:
  - H_0**:** O salário é **igual** a R$ 19.900.

Construção das Hipóteses

- No caso, a hipótese já formulada afirma que o salário de um profissional da ciência de dados é R$ 19.900, segundo as informações obtidas.
  - Essa hipótese já formulada é o que no teste de hipótese chamamos de **Hipótese Nula**, ou H_0.
  - Por sua vez, a hipótese que não confirma essa informação é chamada de **Hipótese Alternativa**, H_1.
- Temos então:
  - H_0**:** O salário é **igual** a R$ 19.900.
  - H_1**:** O salário é **diferente** de R$ 19.900.

Calculando e interpretando o p-valor

- Por definição o **p-valor** é a probabilidade de observarmos na população um valor no mínimo tão extremo quanto o obtido na amostra, considerando que a hipótese nula é verdadeira.
- Se $\mathcal{X}$ e $\mathcal{Y}$ variam juntos, seus desvios tendem a ter o mesmo sinal
- Se os multiplicarmos $dx_i\ dy_i$, o produto é positivo quando os desvios têm o mesmo sinal e negativo quando têm sinais opostos;
- Somar os produtos dá uma medida da tendência de variar em conjunto;
  - Normalizar pelo tamanho da amostra

Calculando e interpretando o p-valor

- Por definição o **p-valor** é a probabilidade de observarmos na população um valor no mínimo tão extremo quanto o obtido na amostra, considerando que a hipótese nula é verdadeira.
- Se $\mathcal{X}$ e $\mathcal{Y}$ variam juntos, seus desvios tendem a ter o mesmo sinal
- Se os multiplicarmos $dx_i \; dy_i$, o produto é positivo quando os desvios têm o mesmo sinal e negativo quando têm sinais opostos;
- Somar os produtos dá uma medida da tendência de variar em conjunto;
  - Normalizar pelo tamanho da amostra

$$Cov(\mathcal{X}, \mathcal{Y}) = \frac{1}{n} \sum_{i=1}^{n} dx_i \; dy_i$$

Correlação

- Normalização da covariância pelo desvio padrão;
- Produção de medida sem unidade;
  - Comparação entre dois pares de variáveis de unidades diferentes;
- Cálculo do Z-score
  - Variação entre -1 e 1;
- **Correlação de Pearson**
  - Dependência **Linear** (!!!)

Correlação

- Normalização da covariância pelo desvio padrão;
- Produção de medida sem unidade;
  - Comparação entre dois pares de variáveis de unidades diferentes;
- Cálculo do Z-score
  - Variação entre -1 e 1;
- **Correlação de Pearson**
  - Dependência **Linear** (!!!)

$$Corr(\mathcal{X}, \mathcal{Y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{dp(\mathcal{X})} \right) \left(\frac{y_i - \overline{y}}{dp(\mathcal{Y})} \right) = \frac{Cov(\mathcal{X}, \mathcal{Y})}{dp(\mathcal{X})\ dp(\mathcal{Y})}$$

Relações não Lineares

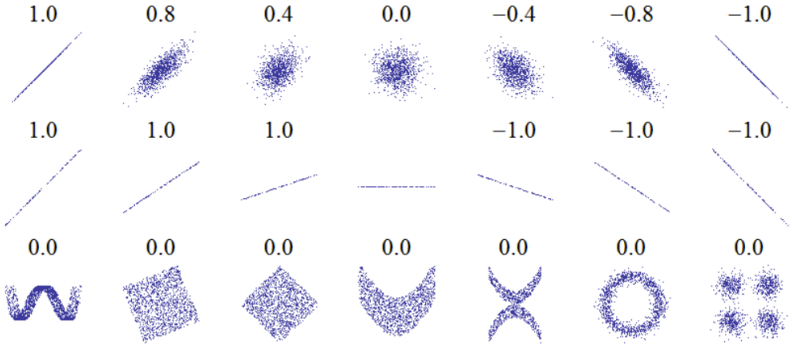


Figure: Exemplos de Correlações

Correlação e Causalidade

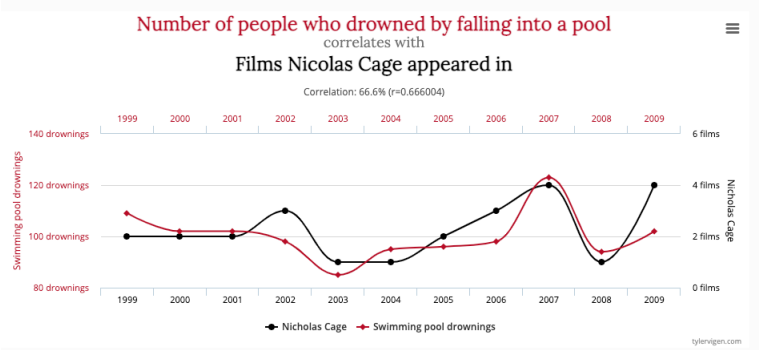
- Erro comum a ser evitado;

Correlação e Causalidade

- Erro comum a ser evitado;
- ***Correlation does not imply causation!***

Correlação e Causalidade

- Erro comum a ser evitado;
- ***Correlation does not imply causation!***



Introdução à Estatística

Distribuições de Probabilidade

Probabilidades

- Distribuição de frequências é importante para avaliarmos a variabilidade das observações de um fenômeno;
  - Medidas de posição e variabilidade;
  - **Estimativas de quantidades desconhecidas**, associadas a populações das quais os dados foram extraídos na forma de **amostras**;
- Frequências (relativas) são **estimativas de probabilidades** de ocorrências de certos eventos;
- Criar um modelo teórico que **reproduza de maneira razoável a distribuição das frequências** de quando o fenômeno é observado diretamente;

Probabilidades

- **Espaço amostral** Ω, que consiste, no caso discreto, da enumeração (finita ou infinita) de todos os resultados possíveis do experimento em questão: $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$

Probabilidades

- **Espaço amostral** Ω, que consiste, no caso discreto, da enumeração (finita ou infinita) de todos os resultados possíveis do experimento em questão: $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$
- **Probabilidade,** $P(\omega)$, para cada ponto amostral. A probabilidade do que chamaremos de um evento aleatório ou simplesmente evento.

Probabilidades

- **Espaço amostral** Ω, que consiste, no caso discreto, da enumeração (finita ou infinita) de todos os resultados possíveis do experimento em questão: $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$
- **Probabilidade,** $P(\omega)$, para cada ponto amostral. A probabilidade do que chamaremos de um evento aleatório ou simplesmente evento.
- E.g.: Lançamos uma moeda duas vezes. Se C indicar cara e R indicar coroa, então um espaço amostral será: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, sendo $\omega_1 = (C, C)$, $\omega_2 = (C, R)$, $\omega_3 = (R, C)$ e $\omega 4 = (R, R)$

Probabilidades

- **Espaço amostral** Ω, que consiste, no caso discreto, da enumeração (finita ou infinita) de todos os resultados possíveis do experimento em questão: $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$
- **Probabilidade,** $P(\omega)$, para cada ponto amostral. A probabilidade do que chamaremos de um evento aleatório ou simplesmente evento.
- E.g.: Lançamos uma moeda duas vezes. Se C indicar cara e R indicar coroa, então um espaço amostral será: $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, sendo $\omega_1 = (C, C)$, $\omega_2 = (C, R)$, $\omega_3 = (R, C)$ e $\omega 4 = (R, R)$
- No caso de querermos descobrir a probabilidade do evento $\mathcal{A}$ que consiste de termos duas faces iguais, teríamos:

$$P(\mathcal{A} = P(\{\omega_1, \omega_4\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Função Probabilidade

Exemplo

Um empresário pretende estabelecer uma firma para montagem de um produto composto de uma esfera e um cilindro. As partes são adquiridas em fábricas diferentes (A e B), e a montagem consistirá em juntar as duas partes e pintá-las. O produto acabado deve ter o comprimento (definido pelo cilindro) e a espessura (definida pela esfera) dentro de certos limites, e isso só poderá ser verificado após a montagem. Para estudar a viabilidade de seu empreendimento, o empresário quer ter uma ideia da distribuição do lucro por peça montada.

Função Probabilidade

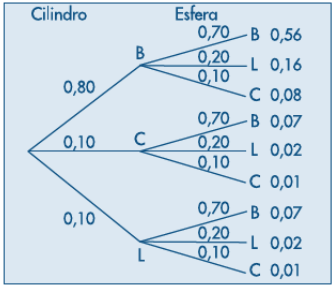
Exemplo - Cont.

Sabe-se que cada componente pode ser classificado como bom, longo ou curto, conforme sua medida esteja dentro da especificação, maior ou menor que a especificada, respectivamente. Além disso, foram obtidos dos fabricantes o preço de cada componente ($5,00) e as probabilidades de produção de cada componente com as características bom, longo e curto. Se o produto final apresentar algum componente com a característica C (curto), ele será irrecuperável, e o conjunto será vendido como sucata ao preço de $5,00. Cada componente longo poderá ser recuperado a um custo adicional de $5,00. Se o preço de venda de cada unidade for de $25,00, como seria a distribuição de frequências da variável X: lucro por conjunto montado?

Função Probabilidade

Produto		Fábrica A Cilindro	Fábrica B Esfera
Dentro das especificações	bom (B)	0,80	0,70
Maior que as especificações	longo (L)	0,10	0,20
Menor que as especificações	curto (C)	0,10	0,10

Função Probabilidade



Produto	Probabilidade	Lucro por montagem (X)
BB	0,56	15
BL	0,16	10
BC	0,08	–5
LB	0,07	10
LL	0,02	5
LC	0,01	–5
CB	0,07	–5
CL	0,02	–5
CC	0,01	–5

Função Probabilidade

- $\mathcal{X}$ pode assumir um dos seguintes valores:
  - **15**, se ocorrer o evento $A_1 = \{BB\}$;
  - **10**, se ocorrer o evento $A_2 = \{BL, LB\}$;
  - **5**, se ocorrer o evento $A_3 = \{LL\}$;
  - **–5**, se ocorrer o evento $A_4 = \{BC, LC, CB, CL, CC\}$
- Cada um desses eventos tem uma probabilidade associada:
  - $P(A_1) = 0,56,\ P(A_2) = 0,23,\ P(A_3) = 0,02,\ P(A_4) = 0,19$

Função Probabilidade

- A função $(x, p(x))$ é chamada função de probabilidade da v.a. $\mathcal{X}$:

x	$p(x)$
15	0,56
10	0,23
5	0,02
–5	0,19
Total	1,00

Função Probabilidade

Valor Médio de uma Variável Aleatória

- Qual o lucro médio por conjunto montado que o empresário espera conseguir?

$$(0,56)(15) + (0,23)(10) + (0,02)(5) + (0,19)(5) = 9,85.$$

Função Probabilidade

Valor Médio de uma Variável Aleatória

- Qual o lucro médio por conjunto montado que o empresário espera conseguir?

$$(0,56)(15) + (0,23)(10) + (0,02)(5) + (0,19)(5) = 9,85.$$

- Dada a v.a. $\mathcal{X}$ discreta, assumindo os valores $x_1, ..., x_n$, chamamos valor médio ou esperança matemática de $\mathcal{X}$ ao valor

$$E(X) = \sum_{i=1}^{n} x_i P(\mathcal{X} = x_i) = \sum_{i=1}^{n} x_i p_i$$

Função Densidade de Probabilidade

- Para o caso de variáveis contínuas
- Cálculo de probabilidade para um dado intervalo;
- Valor $= 0$ em um ponto arbitrariamente pequeno
- Teoricamente, qualquer função f, que seja não negativa e cuja área total sob a curva seja igual à unidade, **caracterizará uma v.a. contínua**;
- E.g., Considerando $f(x) = 2x$, a probabilidade de $\mathcal{X}$ assumir um valor menor que $1/2$ é:

$$P(0 \leq X \leq 1/2) = \frac{1}{2}\left(\frac{1}{2} \times 1\right) = \frac{1}{4}$$

Função Densidade de Probabilidade

Valor médio de uma v.a. contínua

- Sendo $f()$, não negativa e $\int_{-\infty}^{\infty} f(x)dx = 1$, dizemos que f define a v.a. contínua $\mathcal{X}$
- Podemos dizer também que $P(a \leq \mathcal{X} \leq b) = \int_{a}^{b} f(x)dx$
- Por completude, temos que o valor médio da v.a. $\mathcal{X}$ é $E(\mathcal{X}) = \int_{-\infty}^{\infty} xf(x)dx$
- Por extensão temos a variância para uma v.a. contínua $\mathcal{X}$ definida como:

$$Var(\mathcal{X}) = E[((X) - E(\mathcal{X}))^2] = \int_{-\infty}^{\infty} (x - E(\mathcal{X}))^2 f(x)dx.$$

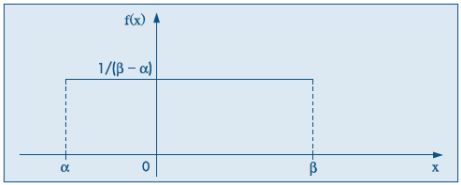
Introdução à Estatística

Modelos Probabilísticos para Variáveis Aleatórias Contínuas

Distribuição de Probabilidade Contínua

- A v.a. $\mathcal{X}$ tem distribuição uniforme no intervalo $[\alpha, \beta]$ se sua função densidade de probabilidade é dada por:

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta-\alpha}, & \text{se } \alpha \leqslant x \leqslant \beta, \\ 0, & \text{caso contrário.} \end{cases}$$

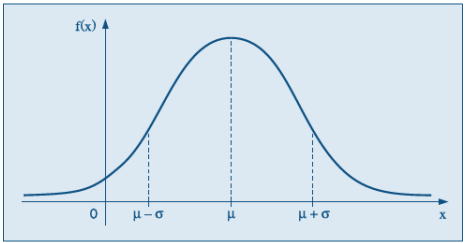


- $E(\mathcal{X}) = \dfrac{\alpha+\beta}{2}$
- $Var(\mathcal{X}) = \dfrac{(\beta-\alpha)^2}{2}$

Distribuição de Probabilidade Normal

- A v.a. $\mathcal{X}$ tem distribuição normal com parâmetros μ e σ^2, $-\infty \leq \mu \leq \infty$, $0 \leq \sigma^2 \leq \infty$ e $-\infty \leq x \leq \infty$ se sua função densidade de probabilidade é dada por:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$



- $E(\mathcal{X}) = \mu$
- $Var(\mathcal{X}) = \sigma^2$
- $\mathcal{X} \sim N(\mu, \sigma^2)$

Distribuição de Probabilidade Normal

- Normal Padrão ($\mu = 0$, $\sigma^2 = 1$)
  - Função Densidade de Probabilidades:

$$\Phi(z) = \frac{1}{2\sqrt{(2\pi)}} e^{-z^2/2}, \ -\infty < z < \infty$$

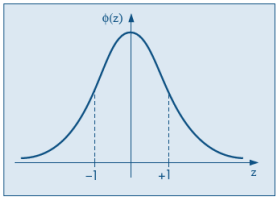


Figure: Função Densidade de Probabilidades para Normal Padrão ($\mathcal{Z} \sim N(0, 1)$)

Dúvidas?