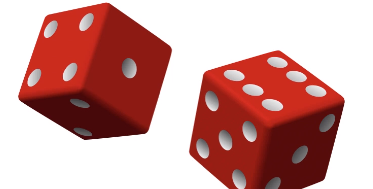
Teoria da Probabilidade

A probabilidade é um número que varia de 0 a 1 e que mede a **chance de ocorrência** de um determinado resultado. Quanto mais próxima de 0 for a probabilidade, menores são as chances de ocorrer o resultado e quanto mais próxima de um for a probabilidade, maiores são as chances. Vários **algoritmos de Machine Learning** tomam como base a teoria da probabilidade.

As probabilidades podem ser expressas de diversas maneiras, inclusive decimais, frações e percentagens. A teoria da probabilidade consiste em utilizar a intuição humana para **estudar fenômenos cotidianos.** Para isso, utilizaremos o princípio básico do aprendizado humano que é a ideia de experimento.

Quando trabalhamos com inteligência artificial alguns modelos de Deep Learning, alguns desses **trabalham com pura probabilidade**. Visão computacional é um exemplo, não temos certeza, mas podemos medir um resultado provável e com base nisso tomar a devidas decisões.

Podemos classificar os experimentos em dois tipos:

Aleatórios (casuais): os experimentos que iremos estudar são os aleatórios, dos quais **não sabemos o resultado** a priori. Usar a probabilidade para **medir a incerteza.** Um experimento é dito aleatório quando **não conseguimos afirmar o resultado** que será obtido antes de realizar o experimento. Quando repetido inúmeras, mesmo aplicando processos semelhantes, possui resultados imprevisíveis. Um experimento é dito equiprovável se todos os possíveis resultados possuem a mesma chance de ocorrer.

Não aleatórios (determinísticos): estes experimentos não aleatórios são totalmente caracterizados a priori, ou seja, são fenômenos em que o **resultado é sabido antes** mesmo em que ele ocorra e desta forma, nada temos a fazer.

Existe uma categoria de algoritmos de Machine Learning chamada de **Algoritmos Probabilísticos,** um deles é o famoso **Naive Bayes**.

Tipos de probabilidade



Probabilidade Clássica: usada quando **sabemos o número de possíveis resultados do evento** de interesse e podemos calcular a probabilidade do evento. Já temos dados à disposição.

Probabilidade Empírica: envolve conduzir um experimento, para se **observar a frequência com que um evento ocorre.** Ainda não sabemos os possíveis resultados e observaremos a frequência.

Probabilidade Subjetiva: quando dados ou experimentos **não estão disponíveis** para calcular a probabilidade. Considerado em experiências passadas

Regras da Probabilidade

1. Se a P(A) = 1, então podemos garantir que evento A ocorrerá.
2. Se P(A) = 0, então podemos garantir que o evento A **NÃO** ocorrerá.
3. A probabilidade de qualquer evento **sempre será entre 0 e 1**.
4. A soma de todas as probabilidades para um evento simples, em um espaço de amostra, será **igual a 1.** Em um dado momento podemos ter diversas variáveis e queremos calcularas probabilidades dos eventos para essas diversas variáveis. A soma dessas probabilidades será igual a 1.
5. O complemento do evento A é definido como todos os resultados em um espaço de amostra, que **não fazem parte do evento A.** Ou seja:

P(A) = 1 – P(A’), onde P(A’) é o complemento do evento A.

Eventos e Espaço Amostral

O primeiro elemento na modelagem de um experimento é o **espaço amostral**, que consiste no conjunto de todos os **possíveis resultados** do experimento. Espaço amostral é o conjunto de possíveis resultados de um experimento ou fenômeno aleatório, representado por **S**.

S = {defeituoso, não defeituoso}  
S = {conceder crédito, não conceder crédito}

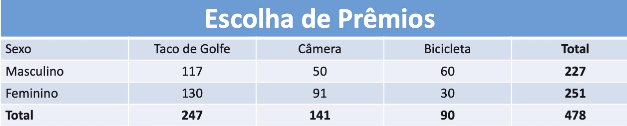
Eventos Complementares: sabemos que um evento **pode ocorrer ou não.** Sendo **P** a probabilidade de que ele ocorra (sucesso) e **Q** a probabilidade de que ele não ocorra (insucesso), para um mesmo evento existe sempre a relação

P + Q = 1 🡪 q = 1 – p

Eventos Independentes: dois eventos são independentes quando a realização ou não realização de um dos eventos **não afeta a probabilidade** **da realização do outro** e vice-versa. Assim, sendo p1 a probabilidade de realização do primeiro evento e p2 a probabilidade do segundo evento, a probabilidade de que tais eventos se realizem simultaneamente é dada por

P = p1 x p2

Probabilidade Conjunta



Se o vencedor for escolhido aleatoriamente por esses clientes, a probabilidade de selecionarmos uma **mulher** é apenas a frequência relativa correspondente. Há 251 mulheres nos dados de um total de 478 dando probabilidade de:

P(mulher) = 251 / 478 = 0,525.

Isso é chamado de **probabilidade marginal,** porque depende apenas dos totais encontrados nas **margens da tabela**. O mesmo método funciona para eventos mais complicados. Qual a probabilidade de escolher uma mulher cujo prêmio preferido é a câmera? Como 91 mulheres nomearam a câmera como preferência, a probabilidade passa a ser:

P (mulher e câmera) = 91 / 478 = 0,190.

Probabilidades como essas são chamadas de **probabilidades conjuntas** porque elas dão a probabilidade de dois eventos ocorrerem juntos

Probabilidade Condicional e Independência

Escrevemos a probabilidade de um cliente selecionado querer uma bicicleta, uma vez que selecionamos uma mulher como:

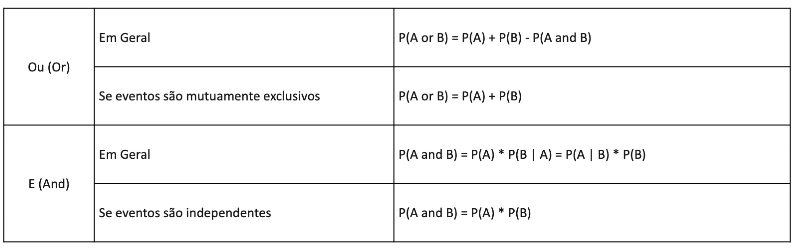
P(bicicleta | mulher) = 30/251 = 0,120

Para os homens, olhamos para a distribuição condicional de prêmios preferidos dado “Homem” mostrado na linha superior da tabela:

Lá, dos 227 homens, 60 disseram que seu prêmio preferido era uma bicicleta. Então:

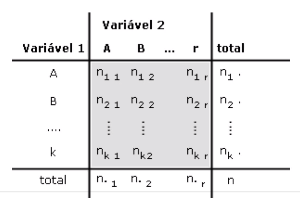
P(bicicleta | homem) = 60/227 = 0,264

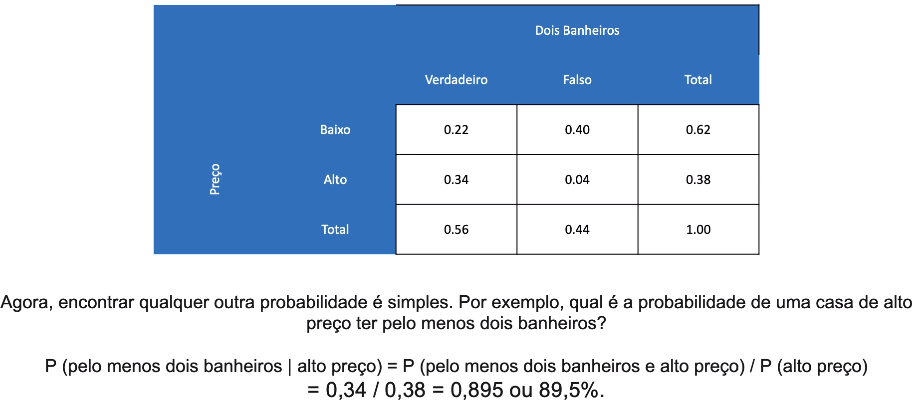
Quando buscamos a probabilidade de um evento de uma distribuição condicional, escrevemos P(B|A) e o pronunciamos “a probabilidade de B dado A.”



As **Tabelas de Contingência** são os meios de organizar as informações correspondentes aos dados classificados segundo **dois critérios.** Muito útil na análise de dados.

Podemos ter os dados das linhas representados por um critério e os dados das colunas representados por outro critério totalmente diferente.





Distribuições de Probabilidade Discreta e Contínua

Em Estatística, uma **Distribuição de Probabilidade** descreve a **chance** que uma variável (discreta ou contínua) pode assumir ao longo de um espaço de valores. Uma distribuição de probabilidade é um modelo matemático que relaciona um certo valor da variável de estudo com a sua probabilidade de ocorrência.

O conjunto de todos os possíveis resultados de um experimento aleatório é denominado **espaço amostral**. O resultado de um experimento de probabilidade geralmente é uma contagem ou uma medida. Quando isso ocorre, o resultado é chamado de **variável aleatória** contínua ou discreta. Uma variável aleatória **discreta**, é aquela que assume valores em um conjunto enumerável, não podendo assumir valores decimais ou não inteiros. Número de filhos, funcionários, veículos... Uma variável aleatória **continua** pode assumir diversos valores num intervalo de valores reais. Renda, peso, altura, temperatura.

v

A **Distribuição Binomial** é utilizada para descrever cenário em que os resultados de uma variável aleatória podem ser agrupados em duas categorias mutuamente exclusivas de forma que não haja dúvida na classificação do resultado da variável. O resultado só pode assumir dois valores.

No geral, as duas categorias de uma distribuição binomial são classificadas como **sucesso e falha**. Estes termos são meramente descritivos, falando de resultados do experimento, ou seja, resultado 0 ou 1, positivo ou não positivo, de acordo com o objetivo do estudo não há uma conotação com relação ao resultado do experimento, mas sim puramente uma descrição

Sucesso  **P**  = 1 – q

Os parâmetros da Distribuição Binomial são **n** e **p**. A **média de uma Distribuição Binomial,** representa a média de longo prazo de sucessos esperados, com base no **número de observações.**

A **variância** de uma Distribuição Binomial, representa a variação que existe no número de sucessos (**p)** sobre um número de (**n)** observações.

A **Distribuição Poisson** é utilizada para descrever cenários onde existe a probabilidade de ocorrência do evento em um **intervalo contínuo.** Esse intervalo contínuo geralmente é dado por tempo ou área. Número de atendentes atendidos por hora, número de acidentes por mês, número de defeitos de uma instalação. A Distribuição Poisson é caracterizada pelo parâmetro único chamado **lambda**, que representa a taxa média de ocorrência por unidade de medida.

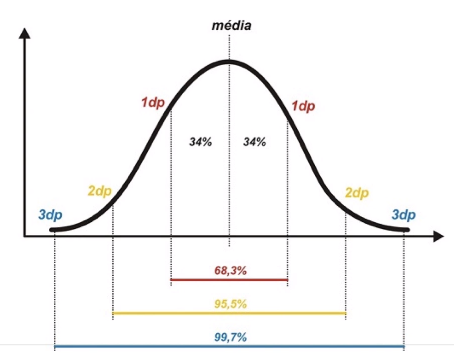
Um dos pontos chave das Distribuições binomial e Poisson é que os **eventos são independentes** um dos outros.

Cada amostra de cada experimento e um **conjunto novo de dados.** Desta forma**,** a **probabilidade de sucesso** ou número de ocorrências se mantém **constante.**

A **Distribuição Hipergeométrica** é uma distribuição de probabilidade discreta que descreve o número de **sucessos numa frequência de (n)** extrações de uma população finita, ou seja, sem reposição.

Quando a amostragem é sem substituição, a probabilidade de **sucesso muda** durante o processo de amostragem, isso viola os requisitos para uma distribuição de **probabilidade binominal.** Nesse caso, o ideal é usar a **Distribuição Hipergeométrica.**

A **Distribuição Normal** é uma das mais importantes para quem trabalha com **Machine Learning.** É a distribuição de probabilidade mais utilizada e mais importante, pois permite modelar uma infinidade de fenômenos naturais. Estudos do comportamento humano, processos industriais e muito mais. Além de facilitar o uso de **aproximações para o cálculo de probabilidades** de muitas variáveis aleatórias.



Empregada em série de análises, inclusive no pré-processamento dos dados antes de treinar o modelo de Machine Learning. A Distribuição Normal é útil quando os dados têm a estar próximos ao **centro da distribuição** (**próximos da média)** e quando **valores extremos (outliers)** são muito raros. Muito comumente usada em diversas **estatísticas inferenciais.**

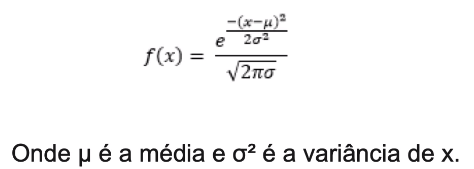
Essa distribuição estatística é uma função que define uma curva e a **área sob essa curva** **determina** a probabilidade de ocorrer o evento por ela correlacionado. É uma curva simétrica em torno do seu ponto médio

Há um **Modelo Normal** diferente para cada combinação de **m** e **s**, mas se padronizarmos nossos dados primeiro, criando **z-scores** e subtraindo da média pra fazer a média igual a 0 e dividindo pelo desvio padrão para fazer o desvio padrão igual a 1, então precisaremos apenas do modelo com média 0 e desvio padrão 1, Chamamos isso de Modelo Normal Padrão ou Distribuição Normal Padrão (**Standard Normal Distribution).**

Uma das coisas que fundamentalmente teremos que fazer como Cientista de Dados é provavelmente **padronizar os dados,** colocando assim os dados com a **média 0** e **desvio-padrão igual a 1**. Isso caracteriza os dados em uma **distribuição normal.** Muitos algoritmos de Machine Learning quando foram construídos eles consideraram que os dados apresentavam distribuição normal

Se o **histograma não for em forma de sino**, as pontuações (**scores)** **z** não serão bem modeladas pelo modelo normal. E a padronização não ajudará por **não alterar a forma da distribuição.** Portanto **verificar o histograma dos dados** antes de aplicar o **modelo normal.**

Como determinar se a Distribuição é Normal?

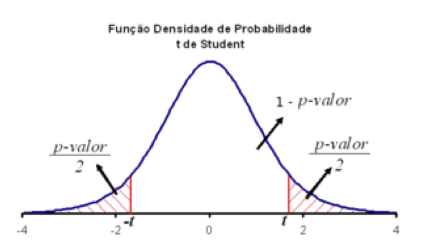


A notação N(u,s²) é usada para representar tal distribuição normal. Para calcularmos então a probabilidade de um resultado, bastar integrar a função f(x) em relação a x, com os limites de integração representando a faixa de valores que se quer obter a probabilidade, Vale notar que a integrar da função densidade de probabilidade normal não possui solução analítica. Sendo asism, seu cálculo deve ser realizado através de um método numérico.

A **Distribuição Contínua** é usada para descrever os dados quando todos os valores têm a mesma chance de ocorrer. A probabilidade de se gerar qualquer ponto em um intervalo contido no espaço amostral é **proporcional ao tamanho do intervalo**, visto que na distribuição uniforme a função f(x) é igual para qualquer valor de x no intervalo considerado. A razão para construir uma distribuição de probabilidade é para encontrar a **probabilidade de ocorrência** de qualquer valor para uma variável aleatória. **Encontrar uma função matemática.**

Outra maneira de se dizer “distribuição uniforme” seria “um número finito de resultados com chances iguais de acontecer”.

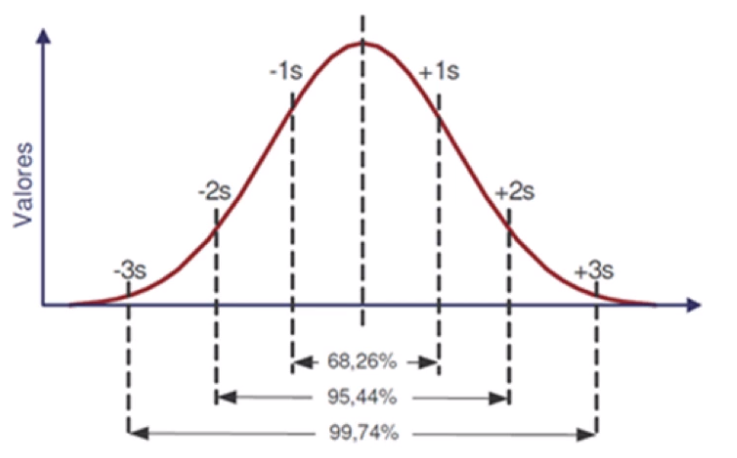
A **Distribuição t de Student** é uma das principais distribuições de probabilidade, com inúmeras aplicações em **inferência estatística.** A distribuição é simétrica em torno da média, com caudas mais longas, podendo com isso gerar **valores mais extremos do que uma distribuição normal.**

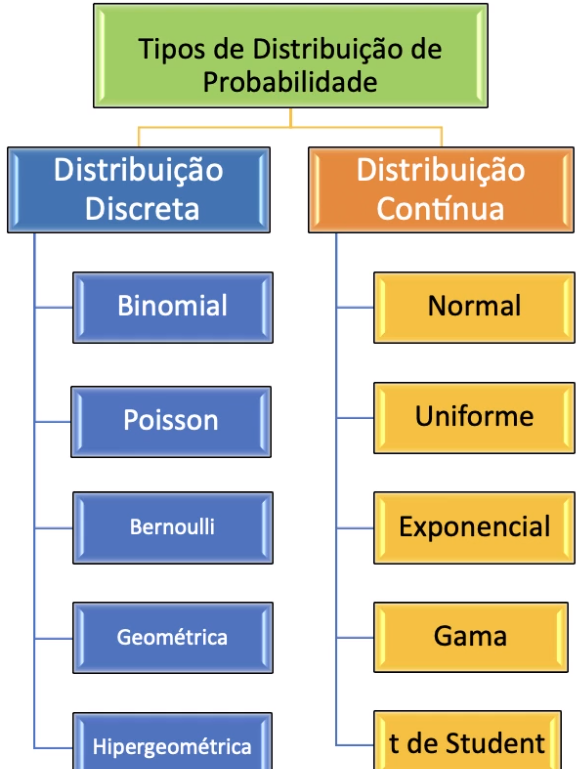


Na caracterização das **distribuições de probabilidade** é de grande importância a utilização de medidas que indiquem aspectos relevantes da distribuição, como medidas de posição (média, mediana e moda), medidas de dispersão (variância e desvio-padrão) e medidas de assimetria e curtose.

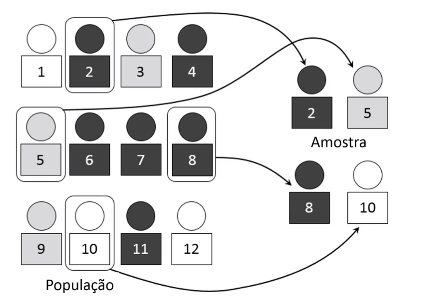
Teorema do Limite Central

Em uma **Distribuição Normal** de dados, simétrica, nós podemos esperar que 68%, 95% e 99.7% dos valores estarão em 1, 2, e 3 desvios-padrão acima e abaixo da média. Com a **Distribuição Normal podemos realizar uma série de inferências.** A área abaixo da curva Normal representa 100% de probabilidade associada a uma variável.





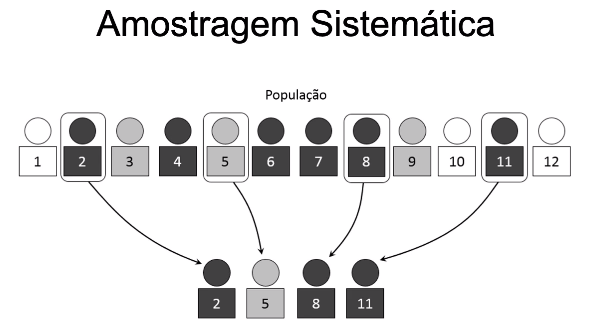
Amostragem Probabilística ou Aleatória: neste tipo de amostragem as amostras são obtidas de forma aleatória, ou seja, a probabilidade de cada elemento da população fazer parte da amostra é igual. E todas as amostras selecionadas são igualmente prováveis. Neste método apenas escolhemos os componentes de forma aleatória. Qualquer unidade da população tem a mesma probabilidade.



* Amostragem Aleatória Simples
* Amostragem Aleatória Simples sem reposição da população depois do sorteio
* Amostragem Aleatória Simples com reposição da população depois do sorteio

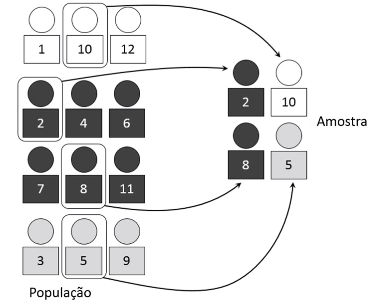
Técnica de Bootstraping auxilia na escolha de qual tipo de amostragem optar para o uso de Machine Learning.

Amostragem Sistemática: elementos da população estão ordenados e são retirados periodicamente. Como numa linha de produção para testar a qualidade. Correndo o risco de não detectar algum problema na população.



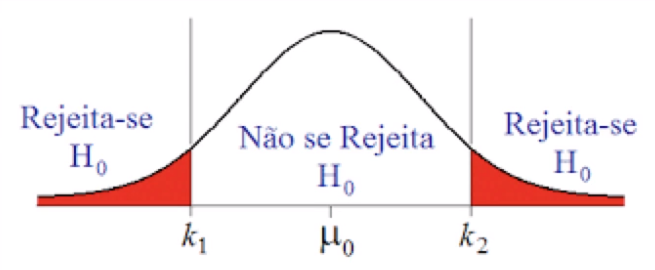
a

Amostragem Estratificada: uma população heterogênea é estratificada e subdividida em populações ou substratos homogêneos. Para cada estrato uma amostra é retirada.



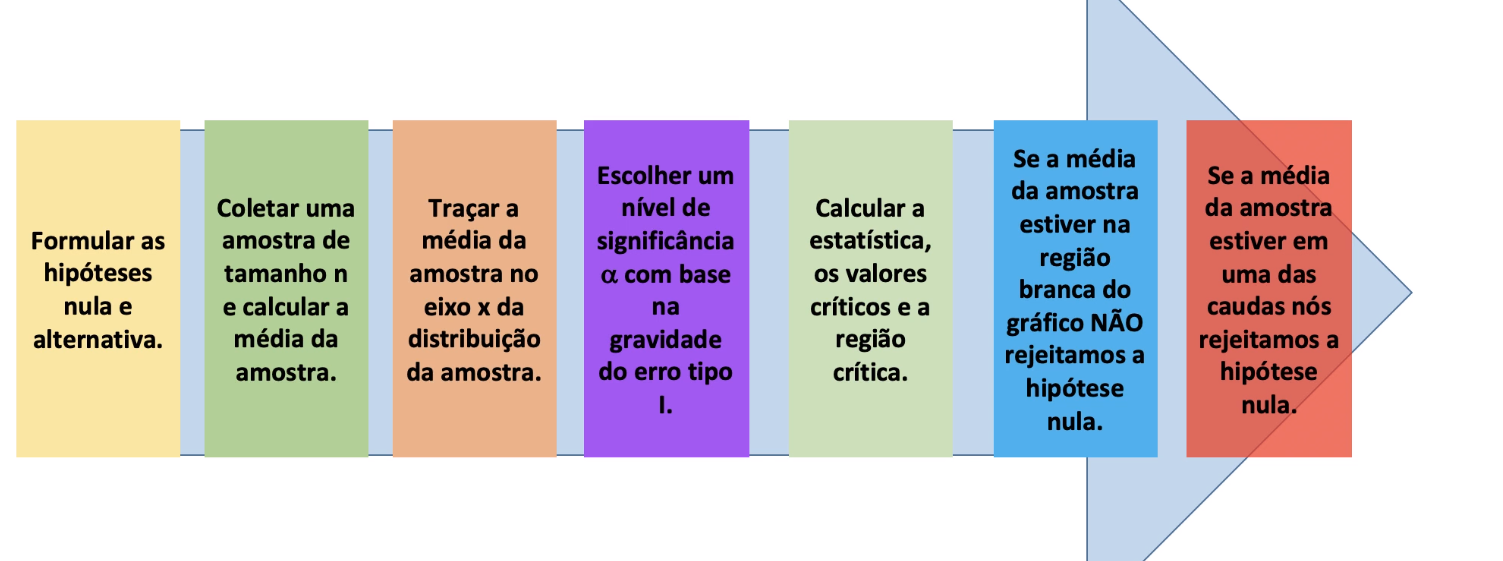
Testes de Hipótese: Uma hipótese estatística é uma suposição sobre um determinado **parâmetro** da população, como média, desvio-padrão, coeficiente de correlação etc. Um teste de hipótese é um procedimento para decisão sobre a veracidade ou falsidade de uma determinada hipótese. Para que uma hipótese seja validada ou rejeitada com certeza, seria necessário examinar toda população.

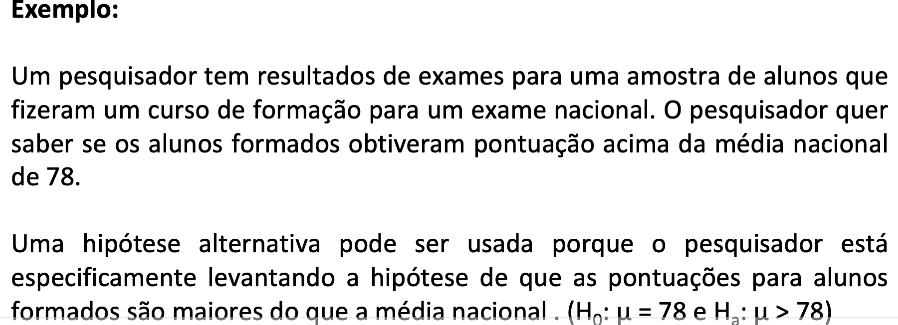
Um **Teste de Hipótese Estatística** é um procedimento de decisão que nos possibilita decidir entre H° (hipótese nula) ou Ha (hipótese alternativa), com base nas informações contidas na amostra.



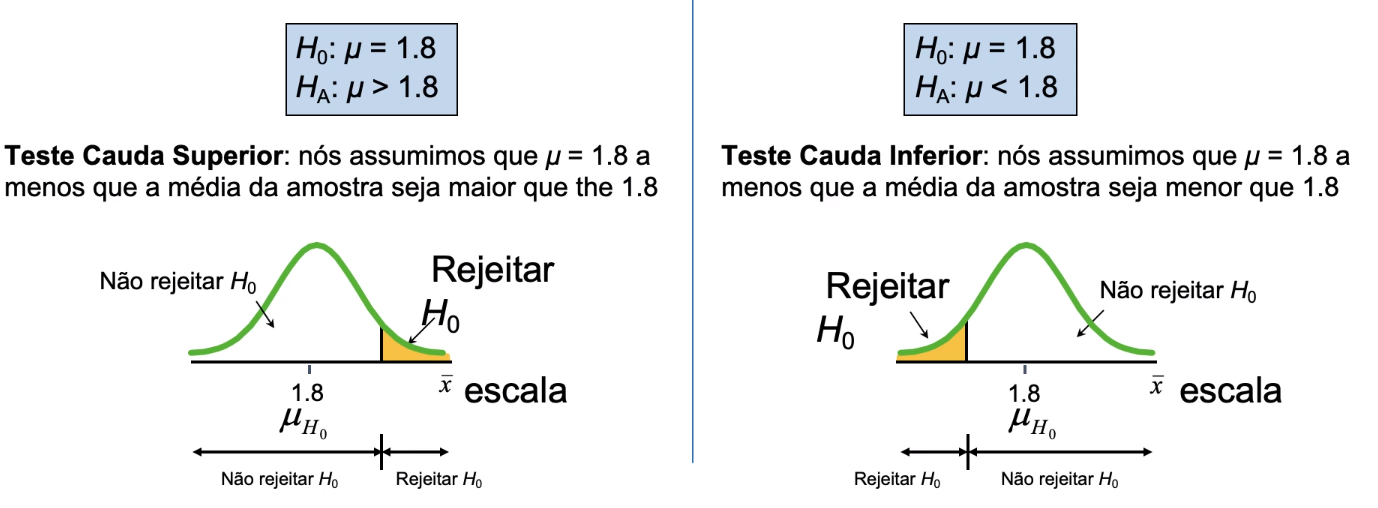
A Hipótese nula afirma que um parâmetro da população (como a média, desvio padrão, e assim por diante) é igual a um valor hipotético. A Hipótese nula é, muitas vezes, uma **alegação inicial** baseada em análises anteriores ou conhecimentos especializados.

A hipótese alternativa afirma que um parâmetro da população é menor, maior ou diferente do valor hipotético na hipótese nula. A hipótese alternativa é aquela que **acredita que pode ser verdadeira** ou **espera provar ser verdadeira.**

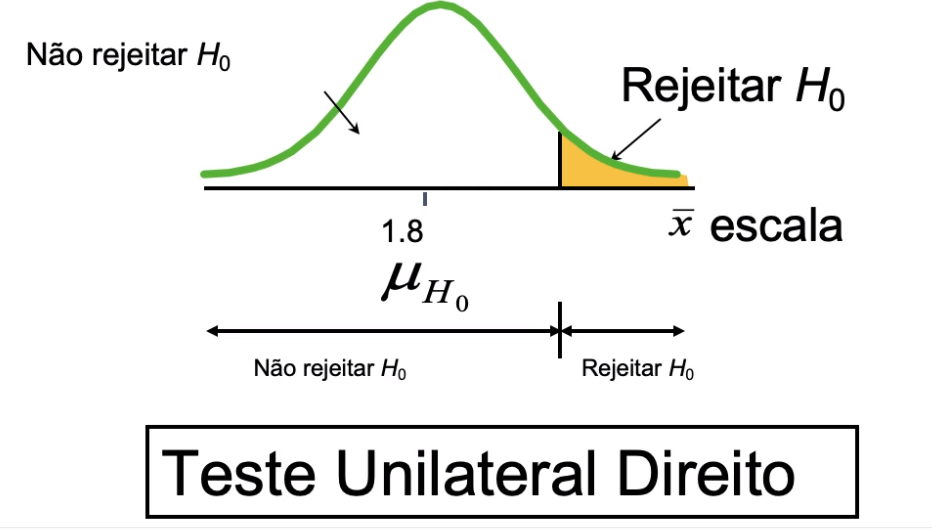




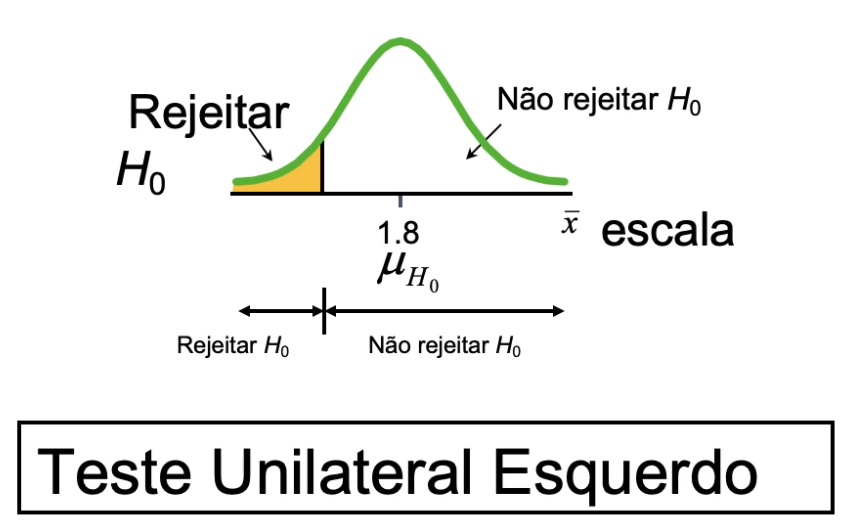
Teste de Hipótese Unilateral



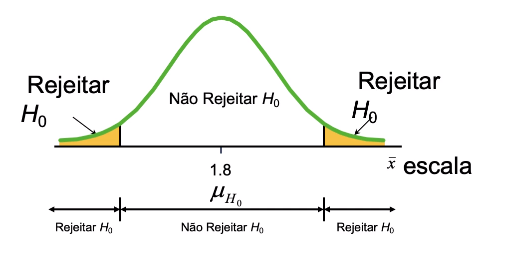
Se a média estiver dentro da região branca do gráfico, não rejeitamos a hipótese nula, caso contrário, a rejeitamos.

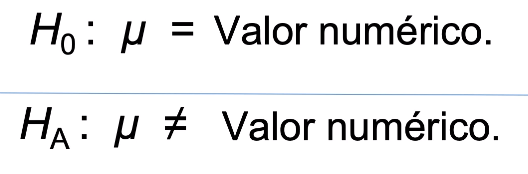


Se a média estiver dentro da região branca do gráfico, não rejeitamos a hipótese nula, caso contrário, a rejeitamos.



Como a hipótese nula é expressa como # ela pode ser maior ou menor que, por isso o teste é **bilateral.**

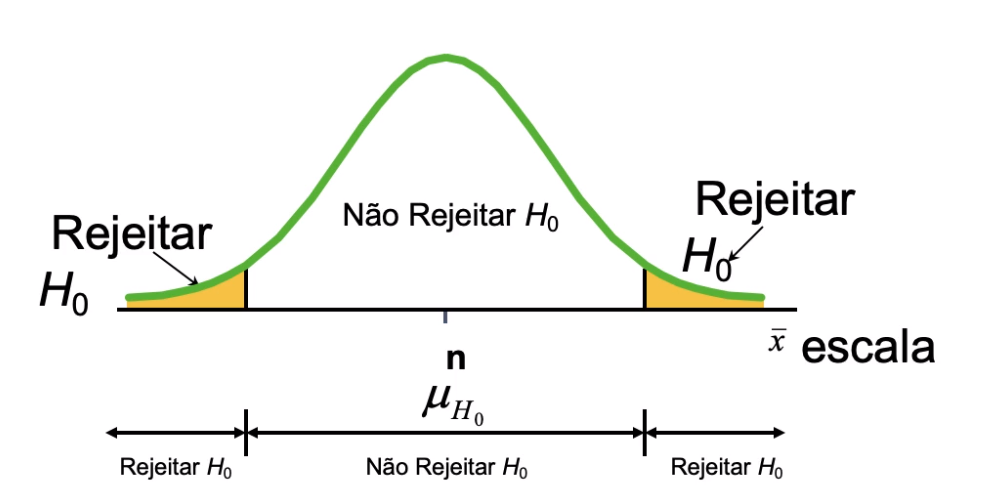




Para Testar Ho, é preciso definir uma regra de decisão com o objetivo de estabelecer uma zona de rejeição da hipótese, ou seja, definir um nível de significância, **alfa,** sendo os mais consensuais os **alfas** 0.10, 0.05 e 0.01. Basicamente o que estamos definindo é a nossa **margem de erro**, e definimos isso através do **nível de significância.**



Se o valor do parâmetro da população, defendido pelo Ho, cair na **zona de rejeição**, então esse valor é muito **pouco provável** **de ser o valor verdadeiro** da população e o Ho será rejeitado em favor de Ha. Pode acontecer, que apesar de rejeitada com base em dados de uma amostra, a Ho de fato seja verdadeira. Nesse caso, estaríamos cometendo um **erro de decisão. Erro Tipo 1.**



Quando o valor defendido pelo Ho sair da zona de rejeição, então consideramos que não há evidência para rejeitar Ho em prejuízo a Ha. **Erro Tipo 2.**

