

DigitalHouse >  
Coding School

# DATA SCIENCE

UNIDADE 1  
MÓDULO 2

Introdução à inferência  
estatística

Agosto de 2017

1

**Definir os termos população-amostra;  
parâmetro-estimador**

2

**Compreender o conceito de amostragem e de  
distribuição amostral**

3

**Calcular estimativas pontuais e por intervalos de  
confiança sob amostragem aleatória simples**

# POPULAÇÃO-AMOSTRA



- Nos problemas de diferentes matérias, o comportamento de diversas variáveis definidas em um conjunto de objetos é estudado. O conjunto de objetos será chamado de população.
- Nestes elementos são observadas variáveis, indicadas por  $X_1, X_2, \dots, X_k$ , que são características que mudam de indivíduo para indivíduo.
- Uma forma de simbolizar as “N” unidades da população é:
  - $\{U_i; i = 1 \dots N\} = (U_1, U_2, \dots, U_i, \dots, U_N)$
- A população é definida em relação ao problema de pesquisa a ser abordado.
  - Vamos considerar o conjunto P de eleitores em uma determinada eleição em que 3 candidatos são apresentados, os quais chamamos de 1, 2 e 3. Podemos observar  $X(a)$  como o número do candidato que recebeu o voto de a.

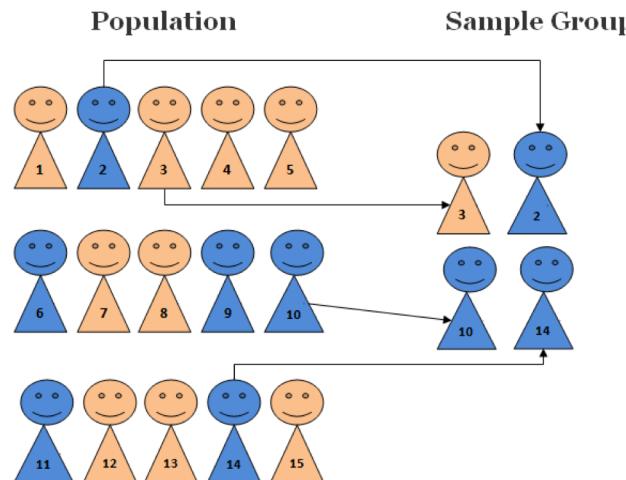
- Em muitos problemas, interessa a distribuição de uma variável aleatória  $X$  que é observada toda vez que a mesma experiência é repetida.
- Nesses casos, cada elemento a ser estudado corresponde ao resultado de uma experiência e, pelo menos teoricamente, você pode repetir a experiência quantas vezes quiser.
- É possível, então, pensar em uma população infinita composta pelas infinitas experiências possíveis de serem realizadas, ainda que esta população não exista realmente.
  - Por exemplo: A experiência consiste em jogar uma moeda e o valor  $V$  recebe 0 ou 1, conforme o resultado seja cara ou coroa.

- **Amostra:** é selecionada da população (definida em relação ao problema de pesquisa). É o subconjunto de unidades selecionadas da população definida.
- Sobre ela recai a realização das observações, medições, etc. As “ $n$ ” unidades ou a amostra selecionada de uma população de “ $N$ ” são simbolizadas:
  - $\{ ui; i = 1 \dots n \} = (u_1, u_2, \dots, ui, \dots u_n)$ .

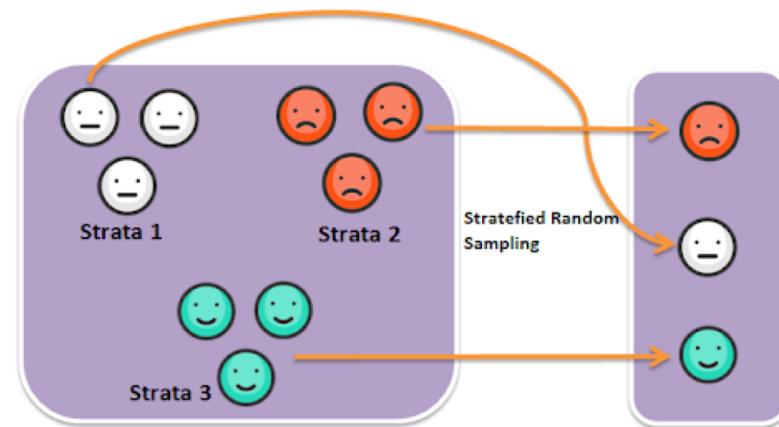
- **Probabilísticas:** Posso calcular a probabilidade de seleção de cada uma das unidades da amostra. => Posso calcular uma medida do erro.

Alguns tipos:

## Amostragem aleatória simples



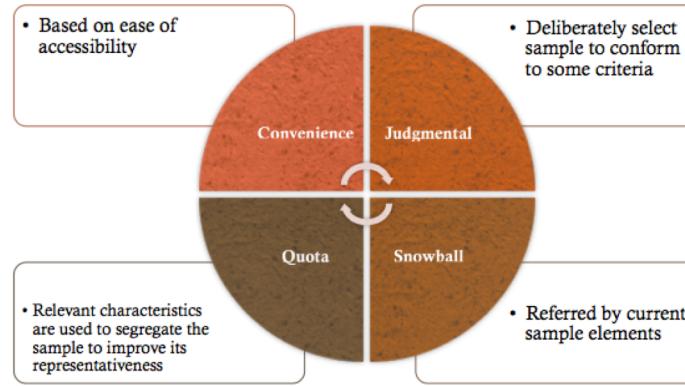
## Amostragem aleatória estratificada



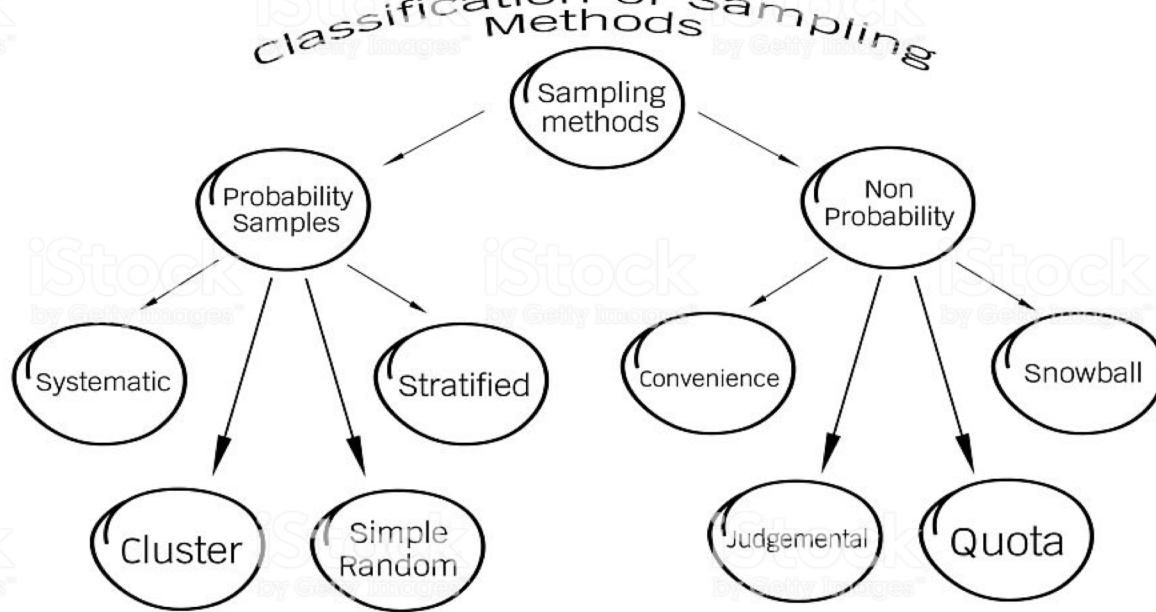
— **Não probabilísticas:** a amostra não probabilística não é um produto de um processo de seleção aleatória. Os indivíduos de uma amostra não probabilística são geralmente selecionados em função de sua acessibilidade ou a critério pessoal e intencional do pesquisador.

- **Amostragem por conveniência;**
- **Amostragem por julgamento.**
- **Amostragem por quota**
- **Amostragem Snowball.**

## Non-Probability Methods



Exemplo Um hospital quer fazer um estudo para testar a eficácia de sua nova vacina contra a gripe que acaba de ser patenteada por um laboratório farmacêutico. **É realizado o estudo de seus pacientes, pois, dessa forma, o hospital tem menos custos financeiros.**



# PARÂMETROS, ESTIMATIVAS E ESTIMADORES



- O objetivo é sempre "estimar" alguma característica da população, que supostamente é fixa (não aleatória).
- Podem ser características simples, como:
  - Uma média, uma proporção, uma variação
- Ou medidas mais complexas, como por exemplo:
  - Os coeficientes de uma regressão ou a associação entre variáveis
- Essa característica é chamada de **parâmetro**.
- Em geral, os parâmetros podem ser considerados relativamente “constantes” (em tempo e espaço).
  - Isso os diferencia dos **estimadores** que veremos a seguir.

- Um estimador é uma estatística (isto é, uma função da amostra) usada para estimar um parâmetro desconhecido da população.
- Por exemplo, a **média amostral** é um estimador da **média populacional** que é calculada tomando a média dos dados.
- Qual estimador deve ser usado depende (entre outras coisas) de dois fatores:
  - Do parâmetro a ser estimado.
  - Do desenho amostral (amostragem aleatória simples, amostragem estratificada, etc.).

- Os estimadores são variáveis aleatórias. O valor da média amostral não é conhecido antes de tomar a amostra.
- É possível conhecer quais valores podem ser tomados, inclusive com qual probabilidade (de acordo com o desenho amostral da experiência, o amostrador “escolhe” as probabilidades de observar cada amostra).

- **Estimativa pontual**: Uma estimativa pontual consiste em usar o valor de um estatístico (alguma função dos dados), que chamaremos de estimador, para calcular o valor de um parâmetro desconhecido de uma população. Na estatística clássica, esses parâmetros são considerados fixos (não aleatórios).
  - Por exemplo, quando usamos a média amostral para estimar a média de uma população, ou a proporção de uma amostra para estimar o parâmetro  $p$  de uma distribuição binomial.
- Uma estimativa pontual de algum parâmetro de uma população é um valor único obtido a partir de uma estatística.

- Em geral, há:
  - **Estimativa por intervalos de confiança:** É dado um "intervalo" de valores possíveis. Geralmente, um limite inferior  $L_i$  e um limite superior  $L_s$  são dados de forma que a probabilidade de que o parâmetro esteja entre  $L_i$  e  $L_s$  é conhecida.



# DISTRIBUIÇÕES AMOSTRAIS



- Conceito-chave: **distribuições amostrais** de... estimadores (médias, proporções, coeficientes de regressão, coeficientes de correlação... etc.).
- Vamos distinguir três tipos de distribuições:
  - Distribuição da variável ( $X$ ) na população. Em amostragem, as características ( $X$ ) das unidades são consideradas não aleatórias.
  - Distribuição da variável ( $x$ ) na amostra.
  - Distribuição amostral - de um estimador - de todas as amostras possíveis de tamanho  $n$  fixo de uma população. É induzido pelo esquema de amostragem.

É um **conceito TEÓRICO**: Para, por exemplo, construir a distribuição amostral de médias, seria necessário conhecer todos os indivíduos da população, com todos os valores da variável, e poder extrair **TODAS** as amostras possíveis. Isso é impossível. Para uma população de  $N = 10.000$ , há aproximadamente  $6,5208E+241$  amostras possíveis de tamanho 100 (sem reposição).

# TEOREMA DO LIMITE CENTRAL



- Pense em jogar um dado e anotar a soma dos valores
  - Se o dado for lançado uma vez, o resultado poderá ser qualquer número de 1 a 6 e haverá exatamente 1 maneira de obter cada resultado.
  - Se for lançado 2 vezes, a soma poderá ir de 2 a 12 e a distribuição será...
  - A tabela a seguir mostra qual é a distribuição de cada valor possível da soma para cada quantidade total de lançamentos (os valores possíveis são as colunas, as células mostram quantas maneiras essa soma pode ocorrer). Por exemplo, com 2 lançamentos, há 3 maneiras de que a soma dos dados seja 4 (sair 1 e, em seguida, 3, saírem 2 e 2 ou sair 3 seguido por 1).

Soma

| Lançamento |   | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|------------|---|---|---|---|---|---|---|---|---|---|----|----|----|
|            | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0  | 0  | 0  |
|            | 2 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 5 | 4 | 3  | 2  | 1  |

Soma

Lançamento

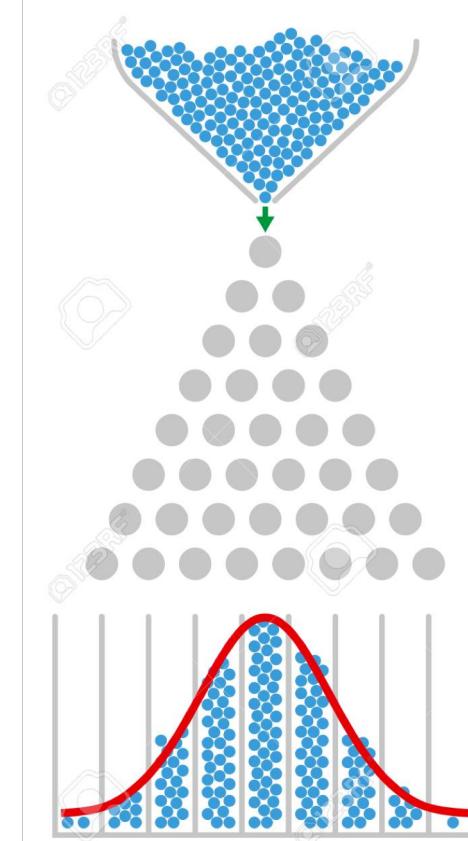
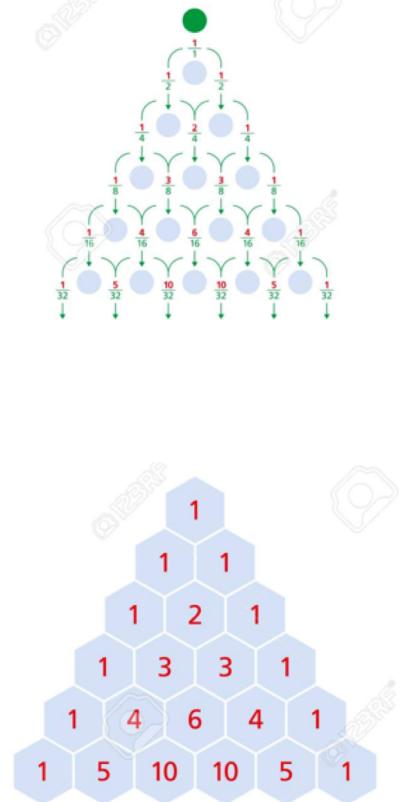
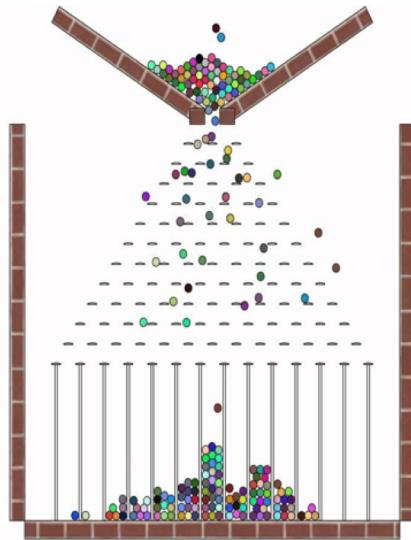
|   | 1 | 2 | 3 | 4 | 5 | 6  | 7  | 8  | 9  | 10 | 11 | 12 |    | 14 | 15 | 16 | 17 | 18 |
|---|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1 | 1 | 1 | 1 | 1 | 1 | 1  | 0  | 0  | 0  | 0  | 0  | 0  |    |    |    |    |    |    |
| 2 | 0 | 1 | 2 | 3 | 4 | 5  | 6  | 5  | 4  | 3  | 2  | 1  |    |    |    |    |    |    |
| 3 | 0 | 0 | 1 | 3 | 6 | 10 | 15 | 21 | 25 | 27 | 27 | 25 | 21 | 15 | 10 | 6  | 3  | 1  |

- Pense em jogar um dado e anotar a soma dos valores...
  - Se lançamos o dado 3 vezes, o intervalo de valores para a soma passa a ser de 3 a 18.

Soma

| Lançamento | 1 | 2 | 3 | 4 | 5 | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 |
|------------|---|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1          | 1 | 1 | 1 | 1 | 1 | 1  | 0  | 0  | 0  | 0  | 0  | 0  |    |    |    |    |    |    |
| 2          | 0 | 1 | 2 | 3 | 4 | 5  | 6  | 5  | 4  | 3  | 2  | 1  |    |    |    |    |    |    |
| 3          | 0 | 0 | 1 | 3 | 6 | 10 | 15 | 21 | 25 | 27 | 27 | 25 | 21 | 15 | 10 | 6  | 3  | 1  |

- Galton board
- Triângulo de Pascal.
- Distribuição Normal.





- A distribuição normal é gerada a partir da aleatoriedade e da imprevisibilidade.
- A ordem emerge do caos.
- O comportamento do indivíduo é imprevisível.
- O comportamento dos arranjos de indivíduos pode ser previsto.

- Teorema central do limite. Conexão com amostragem.
  - O teorema central do limite nos dá uma distribuição para nosso estimador da média populacional dada pela média amostral.
  - Se as observações vierem de uma distribuição normal, a distribuição terá média  $\mu$  e variância  $\sigma^2 / n$  independente de  $n$  ser suficientemente grande ou não.
  - Lembre-se que se estimamos a média amostral e o desvio padrão amostral, então o erro padrão da média amostral é o desvio padrão amostral dividido pela raiz do tamanho amostral.
  - O erro padrão permite medir a dispersão em relação à média amostral, mas ajustando pelo tamanho amostral (podemos comparar a dispersão para estimadores da média que usam diferentes tamanhos de amostra).

- Se forem obtidas sucessivas amostras de tamanho  $n$  (fixo) de uma população cuja variável em estudo ( $X$ ) é distribuída de forma normal, a **distribuição amostral das médias** dessa distribuição será normal com média igual à média populacional e com desvio padrão igual:

$$\sigma_x = \frac{\sigma}{\sqrt{n}}$$

- Esse resultado não depende do TLC e é exato (não assintótico).

- Lei dos grandes números: Por que faz sentido estimar a média populacional (valor esperado) com a média amostral?
- Lei dos grandes números (Kolmogorov): Vamos supor que temos extrações  $X_1, X_2, \dots, X_n$ . Se as  $X_i$  forem uma sucessão de observações independentes e identicamente distribuídas, de modo que  $E(X_i)$  seja igual a uma constante  $\mu$  finita. Então, a média das  $X_i$  convergirá em probabilidade para  $E(X_i)$ .
- A LGN garante que, à medida que o tamanho amostral aumente, a média amostral se aproxime da média populacional.

# TESTE DE HIPÓTESES



- Apresentação mais geral:
  - O teste de uma hipótese estatística é uma maneira formal de decidir entre duas opções com base em certas observações empíricas. Em particular, para distinguir entre distribuições de probabilidade baseadas em variáveis aleatórias geradas por uma delas.
  - Esta regra formal pode indicar 1 (rejeitar hipótese inicial) ou 0 (não rejeitar).
  - Para testar (estatisticamente) uma hipótese, a dividimos em duas:
    1. Hipótese nula ( $H_0$ ): baseada no conhecimento prévio da população a ser testada.
    2. Hipótese alternativa ( $H_1$  :(é aquela que será tomada como (provavelmente) verdadeira caso os dados da amostra sejam derivados da rejeição da  $H_0$ .

- Apresentação mais geral:
  - A distribuição do estimador (variável aleatória) fornece a probabilidade de observar um certo valor para a estatística de prova (uma função dos dados cuja distribuição é conhecida sob  $H_0$ ).
  - Em seguida, é estabelecido um nível de significância  $\alpha$  que indica a probabilidade de rejeitar  $H_0$  quando  $H_0$  é verdadeira, esta é obtida dada a distribuição conhecida da estatística de prova sob  $H_0$ .
  - É importante notar que a escolha de  $H_0$  não é arbitrária. Na estatística clássica:
    1. O erro tipo I: Rejeitar  $H_0$  quando ela é verdadeira) é considerado mais grave do que;
    2. O erro tipo II: Aceitar ou não rejeitar  $H_0$  quando ela for falsa ou, o que é o mesmo, quando  $H_1$  é verdadeiro. Isso significa que antes de rejeitar  $H_0$  é preciso ter muitas evidências que corroborem que ela seja falsa.
  - Portanto, são usados testes com um determinado nível (probabilidade fixa de erro tipo I) que tentam minimizar a probabilidade de erro tipo 2.

- Apresentação mais geral:
  - Vamos ver outro exemplo sobre a escolha de  $H_0$  e  $H_1$ .
  - Queremos decidir se um paciente tem ou não uma doença, para, em caso afirmativo, fornecer um tratamento adequado. Temos, então, duas hipóteses:
    - (A) O paciente tem a doença.
    - (B) O paciente não tem a doença.
  - Percebemos que é mais grave rejeitar (A) quando ela é verdadeira, do que rejeitar (B) quando ela é verdadeira, pois no primeiro caso o paciente pode ter sua situação piorada, enquanto que no segundo caso ele receberia um tratamento desnecessário cujas consequências não seriam tão graves quanto deixar de tratá-lo caso estivesse doente.

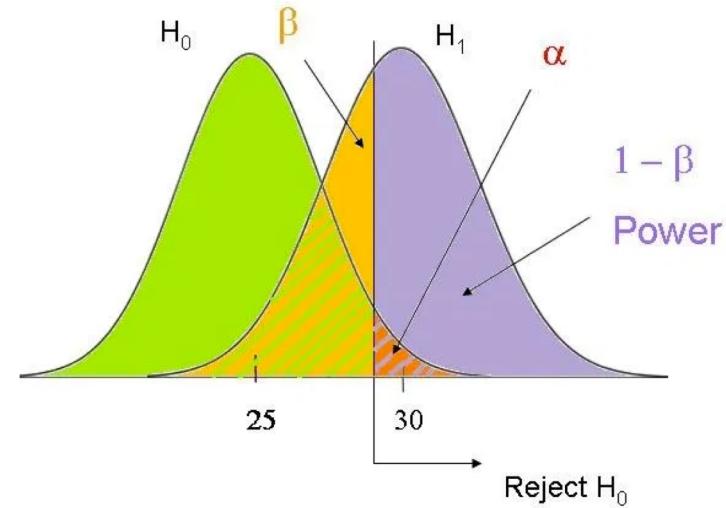
- Apresentação mais geral. Etapas de um teste de hipótese.
  - Formular hipóteses nulas e alternativas.
  - Identificar a estatística de prova apropriada e sua distribuição sob  $H_0$  (supondo que ela seja verdadeira).
  - Determinar o nível de significância (probabilidade de erro ao rejeitar  $H_0$  que pode ser tolerada).

- Apresentação mais geral. Etapas de um teste de hipótese.
  - Estabelecer a regra de decisão (dado o nível de significância, teremos valores críticos que separam a região de não rejeição da região de rejeição).
  - Coletar dados e calcular o valor amostral da estatística de prova.
  - Tomar a decisão estatística.

| Decisão aprovada   | Hipótese $H_0$                           |   |
|--------------------|--|---|
|                    | Verdadeira                               | Falsa                                   |
| Não rejeitar $H_0$ | <i>Decisão correta</i><br>$(1 - \alpha)$ | <i>Erro tipo II</i><br>$(\beta)$        |
|                    | <i>Erro tipo I</i><br>$(\alpha)$         | <i>Decisão correta</i><br>$(1 - \beta)$ |

- Nível máximo de erro que estamos dispostos a tolerar.
- Como estamos trabalhando com amostras, nunca podemos ter 100% de certeza de termos tomado a decisão correta ao rejeitar ou não uma  $H_0$ . Por quê?

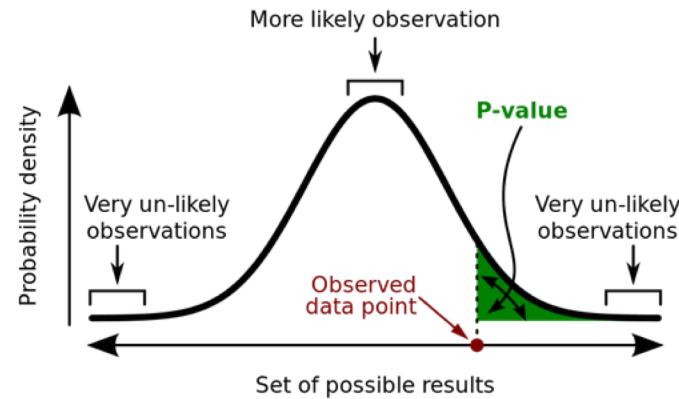
- Os coeficientes  $\alpha$  e  $\beta$  variam de forma inversa.
- O coeficiente  $\alpha$  é a probabilidade de erro do tipo I em qualquer teste de hipótese - rejeitando incorretamente a hipótese nula.
- O coeficiente  $\beta$  é a probabilidade de erro do tipo II em qualquer teste de hipótese - falha incorreta em rejeitar a hipótese nula.



1. Formulação de  $H_0$  e  $H_1$ .
2. Seleção de uma prova/teste adequado.
3. Determinação de um nível de erro/significância (probabilidade de rejeitar  $H_0$  ou  $\alpha$ ).
4. Supor que  $H_0$  é verdadeira e assumir uma determinada distribuição amostral “centralizada” na hipótese nula.
5. Estabelecer a regra de decisão (dado o nível de significância, teremos valores críticos que separam a região de não rejeição da região de rejeição).
6. Calcular a estatística de prova sob a  $H_0$  e sob os valores observados.
7. Comparar ambos.
8. “Decidir”.

- **P-value:**

- O valor-p é o menor valor de significância para o qual rejeitamos  $H_0$  para uma dada amostra (o que gera um certo valor observado da estatística de prova).
- É importante notar que o cálculo do valor-p para uma amostra depende tanto da distribuição da estatística sob  $H_0$  e quanto da regra de decisão escolhida.
- Por exemplo, se minha regra de decisão determinar rejeição quando a estatística de prova for maior ou igual a um valor  $C$ , o valor-p será a probabilidade de que uma variável aleatória que segue a mesma distribuição da estatística de prova sob  $H_0$  seja maior ou igual ao valor empírico obtido para essa estatística na amostra dada.



- P-value:

- Se p-value < nível de significância, então devemos rejeitar  $H_0$ . Forma alternativa de fixar a regra de decisão. Não é necessário buscar valores críticos em uma tabela.
- O procedimento do teste de hipóteses é um argumento por contradição para mostrar que a manutenção da hipótese nula (que sempre é assumida como verdadeira) leva a conclusões absurdas e, portanto, a  $H_0$  deve ser rejeitada.
- O problema é determinar se a diferença entre os dados observados e o que é esperado deles sob  $H_0$  (supondo que  $H_0$  seja verdadeira) pode ser atribuído à variação aleatória (porque trabalhamos com uma amostra).
- Um p-value pequeno implica em fortes evidências para rejeitar a hipótese nula  $H_0$ .

- P-value:

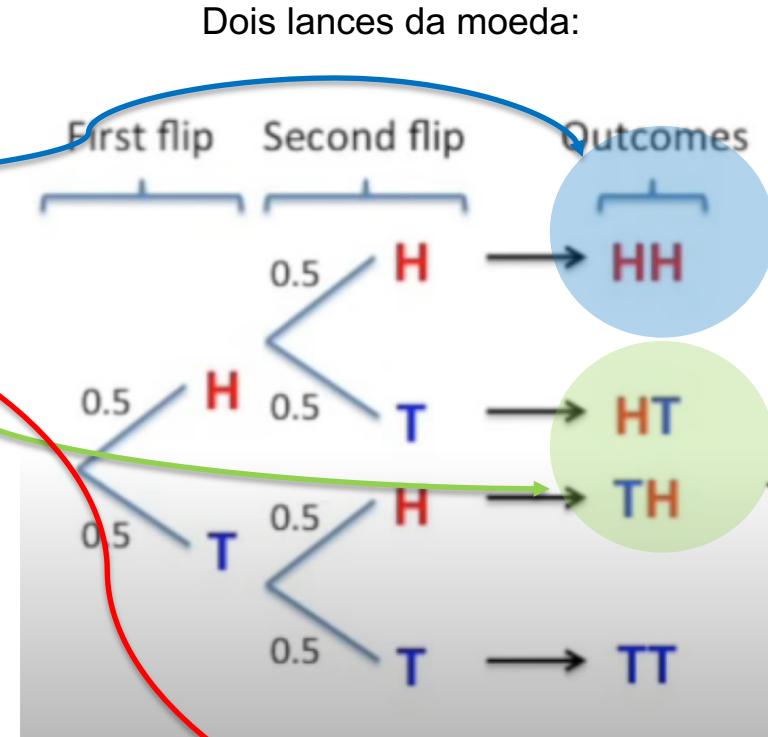
- Probabilidade de que a aleatoriedade gerou os dados ou algo do tipo ou mais raro.

- $p-value = 0.25 + 0.25 + 0.0 = 0.5.$

- $P(HH) = 0.25$

- Limite de signifiância tradicional  
 $p-value = 0.05.$

- É mais provável encontrar valores de  $HT$  ou  $TH$ , que de  $HH$  ou  $TT$



Nulo.

- **P-value:**

- Vamos supor que, dado um conjunto de dados, a estatística de prova seja calculado e o p-value resulte em 0,001. Vamos começar supondo que  $H_0$  é verdadeira e imaginar outros pesquisadores repetindo o experimento em condições idênticas.
- Esse valor do p-value diz que se  $H_0$  for verdadeira, apenas 1 em cada 1.000 pesquisadores poderá obter um valor da estatística tão extremo quanto o obtido.
- Note que o p-value não é a probabilidade de que  $H_0$  seja verdadeira.
- Independentemente da quantidade de repetições da experiência,  $H_0$  é sempre verdadeira ou falsa.

- Intimamente relacionado com a distribuição de probabilidades (amostral) de uma estatística sob a suposição de que  $H_0$  é verdadeira.
  - “A probabilidade de obter o valor observado ou mais extremos da estatística de prova se a hipótese nula for verdadeira”
- O p-value dá a probabilidade de obter evidências contra  $H_0$  (aleatoriamente), assumindo que  $H_0$  é verdadeira. Quanto menor for o p-value, mais evidências contra  $H_0$  temos, sempre assumindo que  $H_0$  é verdadeira.

# INTERVALOS DE CONFIANÇA



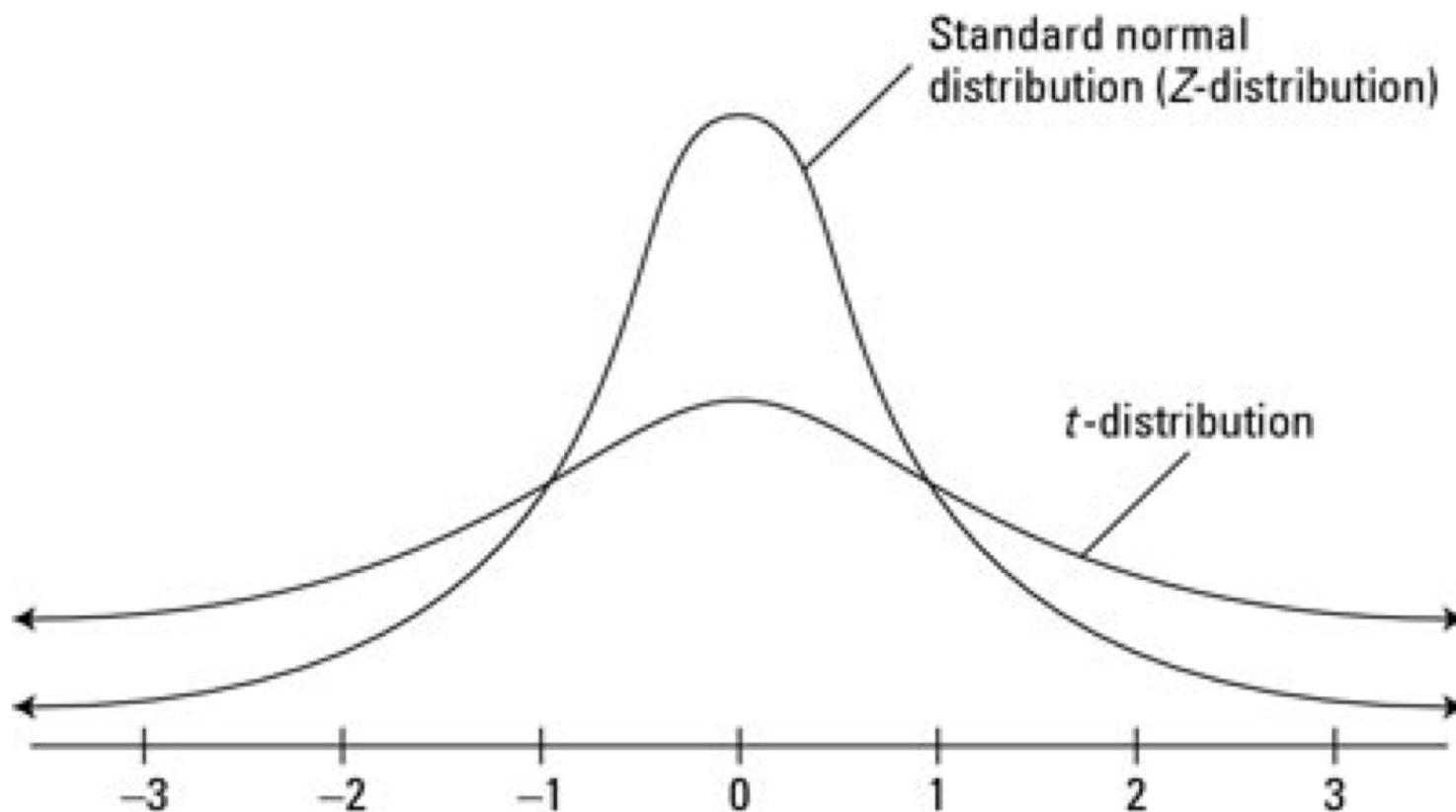
- **Média da amostra:** estimativa pontual => é o valor da estatística nessa amostra
- **Estimativa por intervalos:** a ideia é fornecer um intervalo de valores possíveis para o parâmetro com um valor de probabilidade associado.
- **Como conseguir isso?** Começando por poder obter uma estatística cuja distribuição seja conhecida e não dependa de parâmetros desconhecidos.

O z-score indica o quanto distante, em desvios padrão, uma medida se encontra da média:

$$z = \frac{X - \mu}{\sigma}$$

O intervalo de confiança faz uso do z-score para descrever a quantidade de incerteza associada à estimativa de amostra, de um parâmetro populacional.

- Observações.
  - E se a variância populacional não for conhecida? É mais uma fonte de incerteza que devemos incorporar, se na derivação anterior o desvio padrão da população para a amostragem for alterado, a distribuição da variável normalizada se tornará um T-Student.
  - A distribuição T-Student é simétrica.



- Observações.
  - O intervalo de confiança é sempre uma função da amostra, que é aleatória. Então, as bordas do intervalo também são variáveis aleatórias! Nesse caso, a borda do intervalo depende da média da amostra (que muda a cada amostra).