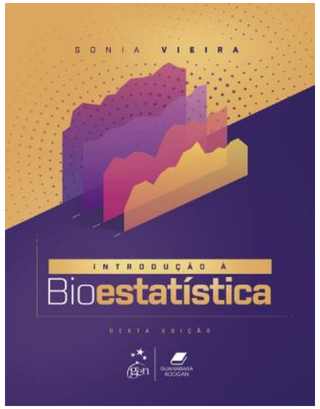
Aula 8 – Introdução aos modelos lineares

Teste *t* de Student

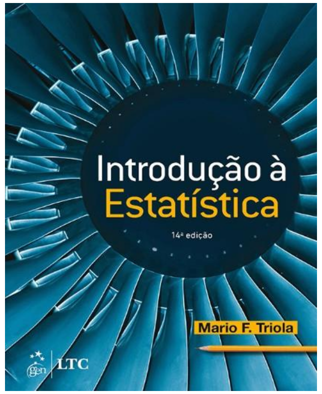
Ao final desta aula, vocês serão capazes de:

- Identificar situações em que o Teste *t* de Student é o teste apropriado
- Diferenciar entre o Teste t para uma amostra, para duas amostras independentes e para dados pareados
- Formular hipóteses nula (H_0) e alternativa (H_1) para diferentes cenários biológicos
- Compreender os pressupostos do teste: normalidade e homogeneidade de variâncias
- Interpretar o resultado de um Teste *t* (valor de *p* e estatística de teste) para tomar uma decisão e redigir uma conclusão contextualizada

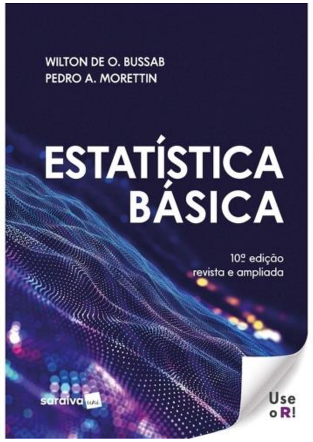
Esta aula está baseada em



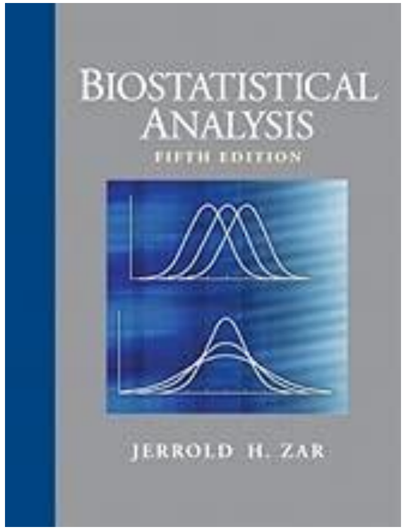
Cap 10 e 11



Cap 9



Cap 12 e 13



Cap 7, 8 e 9

Perguntas comuns

- Um novo fertilizante realmente aumenta o crescimento médio das plantas em comparação com o fertilizante antigo?

- A concentração de um poluente em um rio está, em média, acima do limite de segurança estabelecido por lei?

- Um medicamento para baixar a pressão arterial realmente funciona? Como podemos provar isso comparando a pressão *antes* e *depois* do tratamento?

O que é um teste *t*

Ferramenta estatística que permite comparar médias de uma ou duas populações. Ele avalia se as médias são significativamente diferentes, levando em conta a média, o desvio padrão, e o tamanho da amostra.

Os 4 Passos do Teste de Hipóteses:

1. **Construir as hipóteses** (H0 e H1).
2. **Especificar o nível de significância** (α).
3. **Calcular o valor do teste** (a estatística *t*).
4. **Interpretar o resultado** e tomar uma decisão

Hipótese nula e alternativa(s)

- **Hipótese Nula (H_0)**: A hipótese da "não diferença" ou da "nulidade".
  - Ela sempre contém o sinal de igualdade (=, ≤, ou ≥). Ex: "A média de peso dos filhotes é igual a 25 kg" (H_0:μ=25).
- **Hipótese(s) Alternativa(s) (H_1, H_A):** contradiz a hipótese nula. É geralmente o que o pesquisador quer apoiar. Ex. A média de peso é diferente de 25 Kg (H_1: $\mu \neq 25$)

Nível de significância e erros

Decisão	H_0 é Verdadeira	H_0 é Falsa
Não Rejeitar H_0	Decisão Correta	Erro do Tipo II (β)
Rejeitar H_0	Erro do Tipo I (α)	Decisão Correta (Poder)

Tabela 1. Sugestão de alguns testes estatísticos a empregar de acordo com o tipo de variável observada. Entre parênteses alguns testes não-paramétricos.

Variável Dependente	Variável Independente	Teste
Quantitativa	1 Categórica com 2 níveis	Teste t (teste U)
Quantitativa	1 Categórica com + 2 níveis	ANOVA 1-fator (Kruskall-Wallys)
Quantitativa	2 Categóricas	ANOVA 2-fatores (Friedman[1])
Quantitativa	1 Quantitativa	Regressão simples (correlação Spearman)
Quantitativa	2 ou mais quantitativas	Regressão múltipla
Quantitativa	1 categórica e 1 ou mais quantitativas	ANCOVA
Categórica	1 Categórica	Qui-quadrado[2]; Teste G[2]
Categórica	2 ou mais categóricas	Log-linear[2]

(1) No caso de amostras dependentes, (2) Esses testes eventualmente verificam não a relação de dependência entre variáveis, mas sim a associação entre elas, descaracterizando, portanto a classificação de variáveis dependentes e independentes.

Apostila Paglia & de Marco

Teste *t* para uma amostra

- **Quando usar:** para comparar a média de uma amostra com um valor de referência específico (μ_0)
- **Exemplos biológicos:**
- A OMS preconiza uma taxa de cesáreas de 15%. Um pesquisador coleta dados de uma maternidade para verificar se a taxa média é diferente desse valor de referência.
- Verificar se a quantidade média de flúor em uma marca de dentifrício corresponde ao valor especificado na embalagem

Hipóteses uni- e bicaudais

- Bicaudal

$H_0\colon \mu = \mu_0$ vs. $H_1\colon \mu \neq \mu_0$

- Unicaudal

$H_0\colon \mu \geq \mu_0$vs. $H_1\colon \mu < \mu_0$ (ou $H_0\colon \mu \leq \mu_0$vs. $H_1\colon \mu > \mu_0$)

Cálculo da estatística t

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- Média da amostra ($\bar{x}$)
- Valor de referência (μ_0)
- Desvio padrão (s)
- Tamanho da amostra (n)

Tomada de decisão

Compare o *t* calculado com
o *t* crítico da tabela
(com n−1 graus de liberdade)

Ou, de forma mais moderna,
observe o **valor-p** fornecido
pelo software

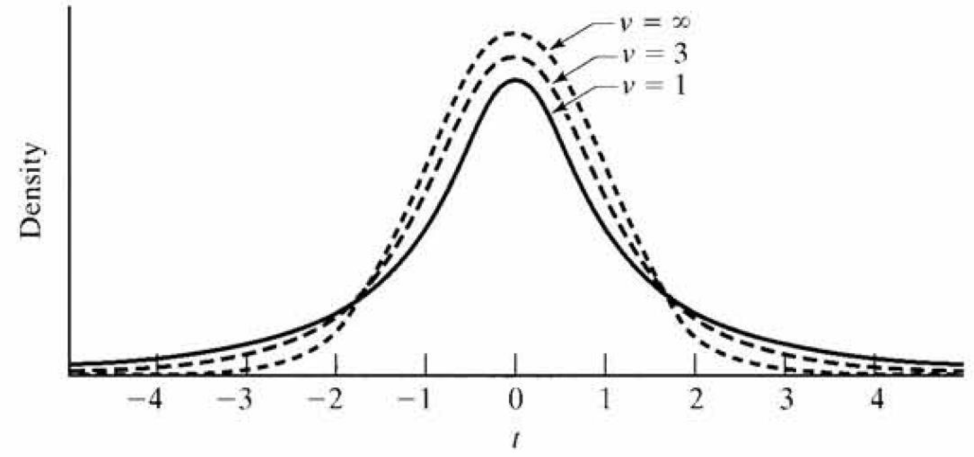


FIGURE 7.1: The *t* distribution for various degrees of freedom, ν. For $\nu = \infty$, the *t* distribution is identical to the normal distribution.

Atividade prática 1 – Pense-pareie-compartilhe (10 min)

- **Cenário:** "Um biólogo sabe que, historicamente, uma espécie de peixe em um lago tem um comprimento médio de 15 cm. Após a introdução de uma nova espécie invasora, ele mede o comprimento de 30 peixes da espécie nativa e suspeita que o tamanho médio diminuiu. Quais são as hipóteses H_0 e H_1 para este estudo?
- **Instruções:**

1. Pensem individualmente por 1 minuto
2. Discutam com um colega ao lado por 2 minutos
3. 2-3 duplas compartilham suas respostas

Teste *t* para duas amostras

- **Amostras Independentes:** Os valores de uma amostra não estão relacionados ou pareados com os da outra. Os grupos são distintos
  - **Exemplos:** Comparar o nível de ansiedade entre um grupo de meninos e um grupo de meninas ; comparar o tempo de alívio da dor entre um grupo que recebeu um anti-inflamatório e *outro* grupo que recebeu laser.
- **Amostras Dependentes (Dados Pareados):** Os valores da amostra são pareados. Cada dado numa amostra está ligado a outro na segunda amostra.
  - **Exemplos:** Medir uma variável nos mesmos indivíduos *antes* e *depois* de uma intervenção (ex: peso em setembro vs. abril) ; comparar uma característica entre mãe e filho, ou em gêmeos.

Teste t para amostras independentes

- **Objetivo:** Comparar as médias de dois grupos completamente separados.
- **Pressuposto Chave:** As variâncias dos dois grupos devem ser aproximadamente iguais (**Homogeneidade de Variâncias**)
  - Como verificar? Usa-se um **Teste F** para comparar as duas variâncias. A hipótese nula do Teste F é que as variâncias são iguais (H_0: $\sigma_1^2 = \sigma_2^2$)

Caso 1: variâncias homogêneas

- Calcula-se a **variância combinada (*pooled variance*, s_p^2)**, que é uma média das duas variâncias de amostra, ponderada pelos graus de liberdade

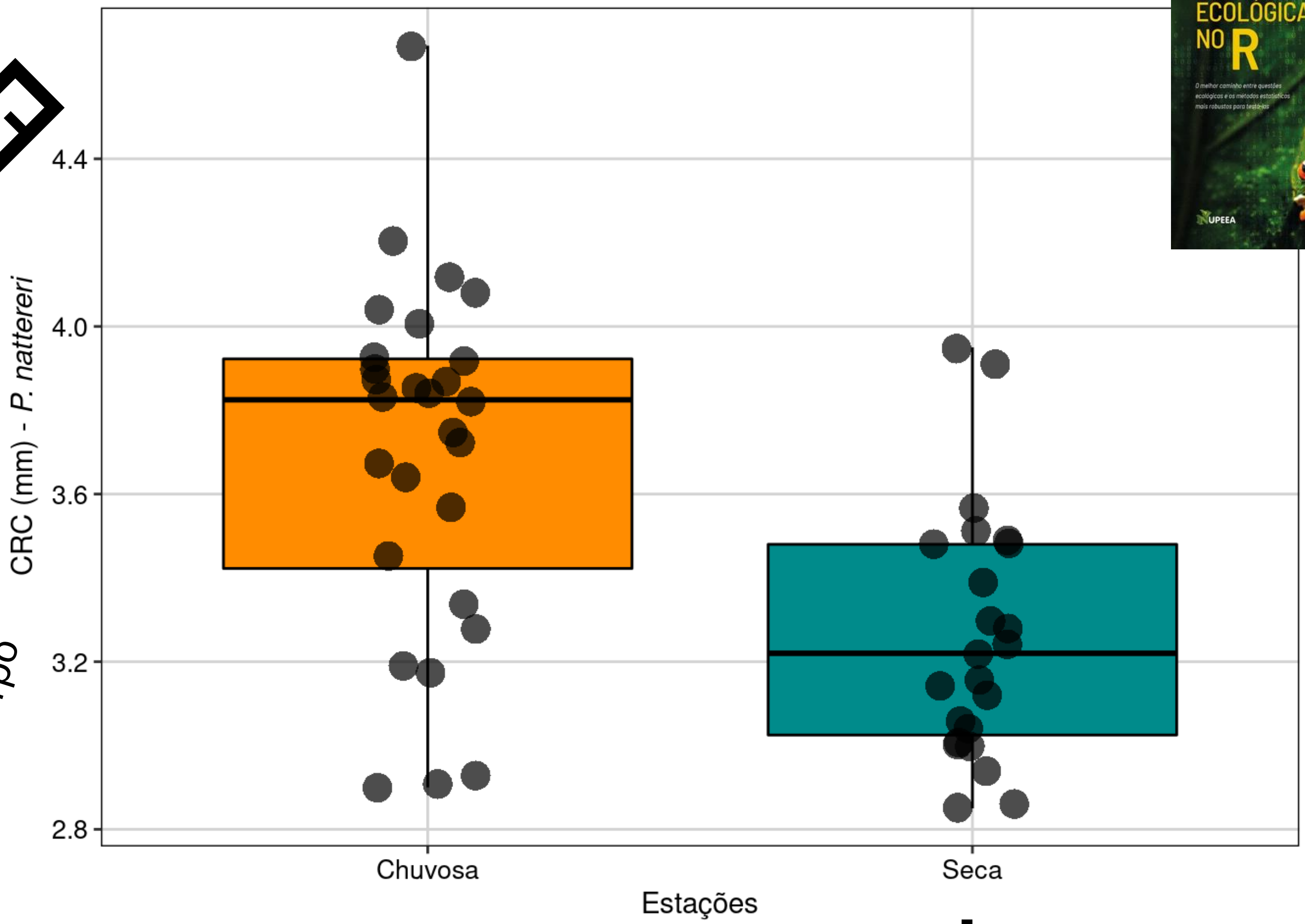
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

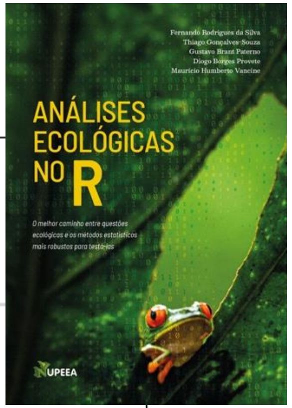
- Graus de liberdade

$$\nu = n_1 + n_2 - 2$$

Exemplo

Tamanho do corpo





Caso 2 – Variâncias heterogêneas

- Se o Teste F indica que as variâncias são diferentes, uma versão modificada do Teste *t* é usada (conhecido como Teste *t* de Welch).
- Não se calcula a variância ponderada

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Graus de liberdade: cálculo feito por programa

Teste t para dados pareados

- **Lógica :** Este teste reduz um problema de duas amostras a um problema de uma amostra. Ele não analisa os dados originais, mas sim a **diferença (d)** dentro de cada par.
- **Procedimento:**

1. Para cada par, calcule a diferença: d = $x_{antes} - x_{depois}$
2. Calcule a média das diferenças (d) e o desvio padrão das diferenças (sd)
3. Execute um Teste t para *uma amostra* nessas diferenças, testando se a média das diferenças é igual a zero.

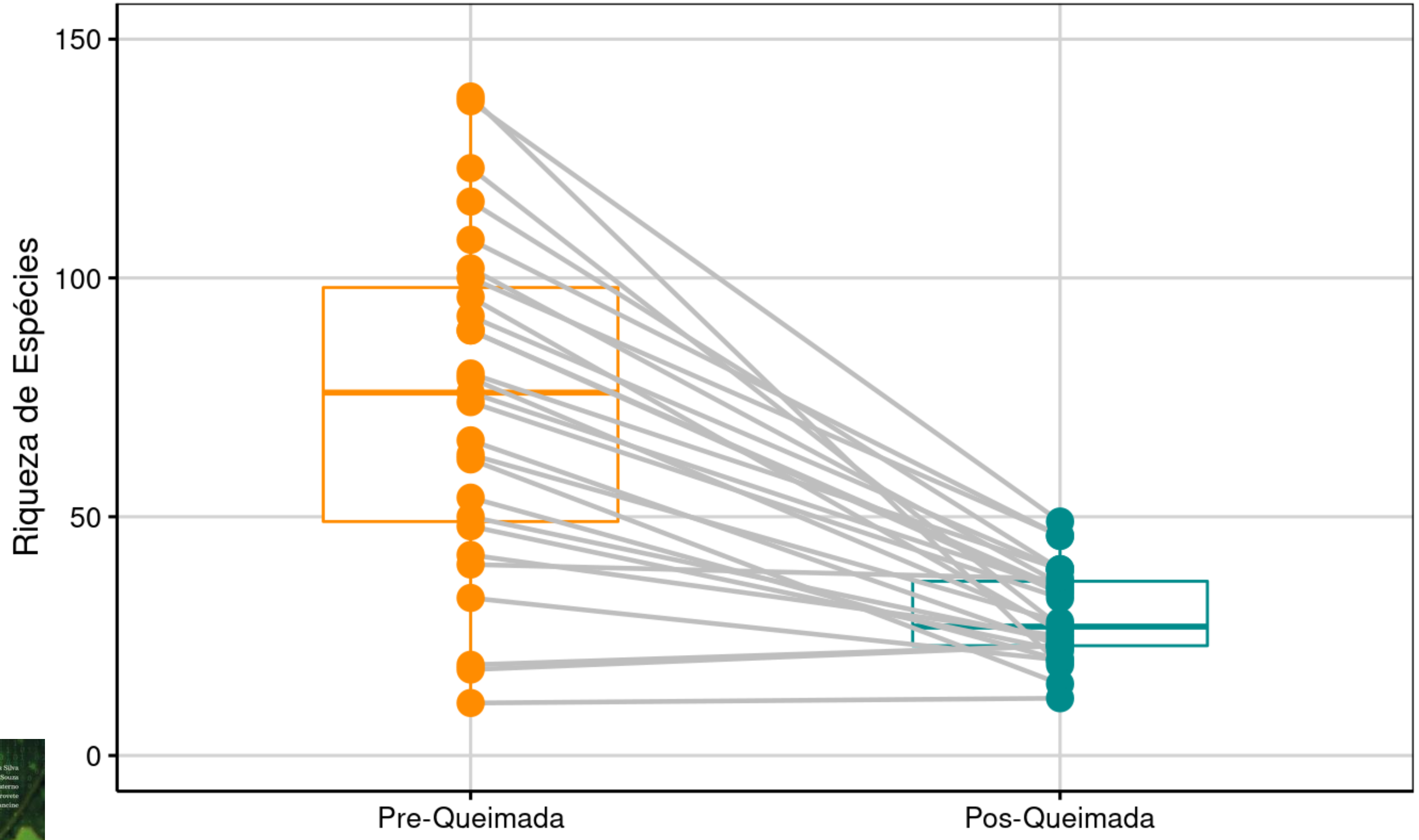
Teste t para dados pareados

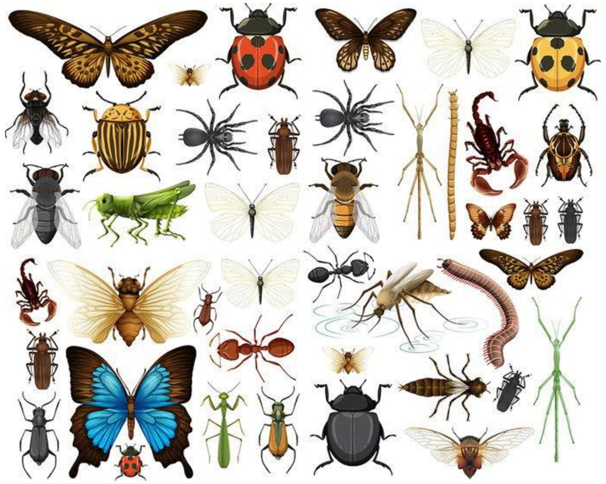
- Hipóteses: H_0: $\mu_d = 0$ vs. H_1: $\mu_d \neq 0$

- Fórmula: $$t = \frac{\bar{d} - 0}{s_d / \sqrt{n}}$$

- Graus de liberdade: $\nu = n - 1$ (n = número de pares)

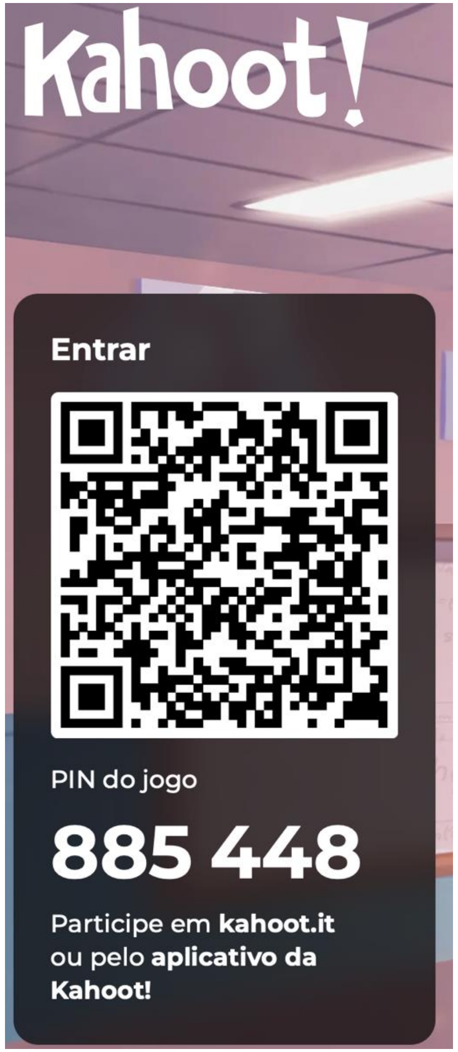
Exemplo



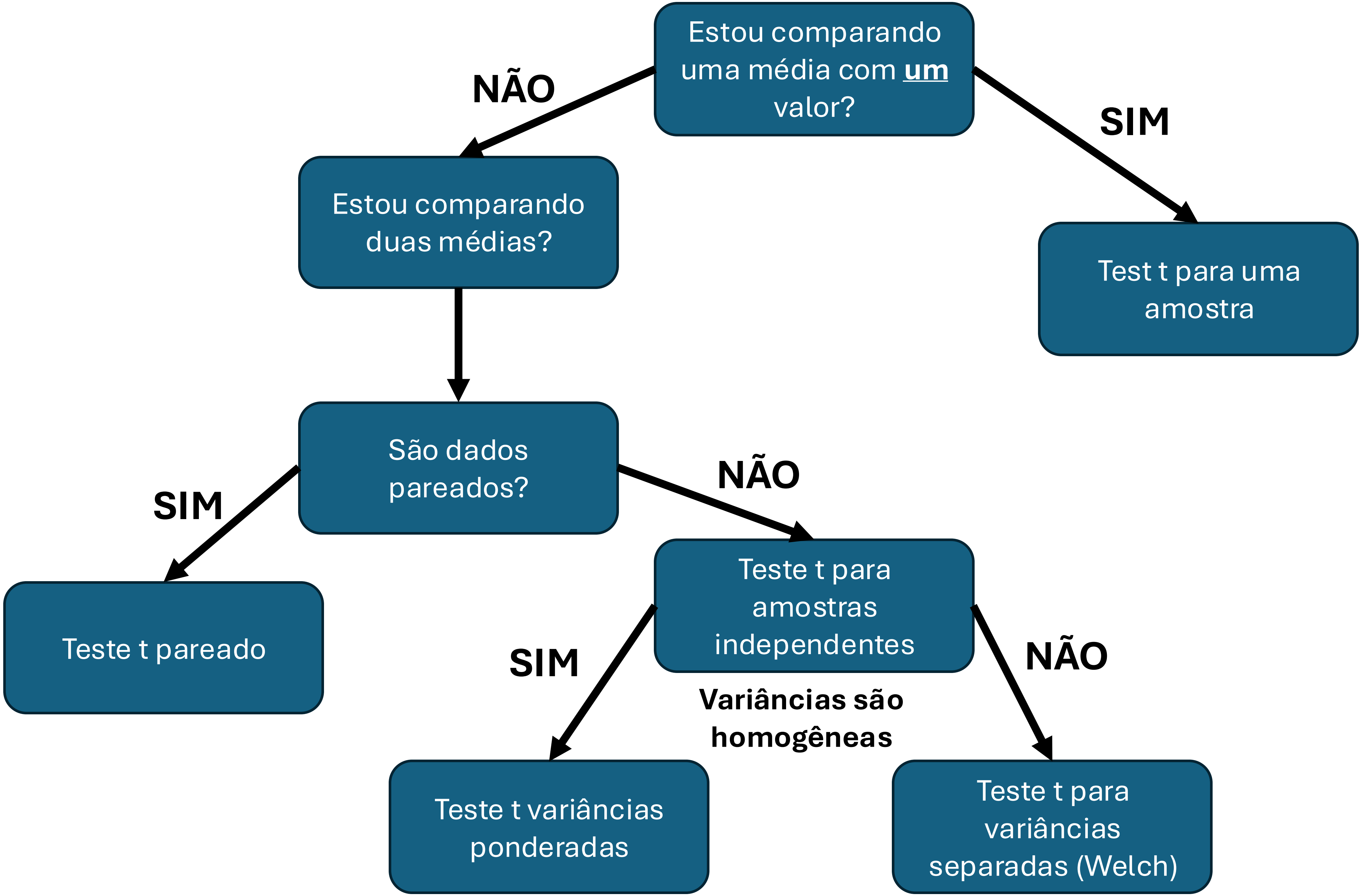


Riqueza de artrópodes de 27 áreas comparadas antes e depois de queimada

Atividade prática 2 – Qual teste usar? (5 min)



https://play.kahoot.it/v2/lobby?quizId=095cb8f0-fc9f-437a-9449-105b02de0750


Estou comparando uma média com **um** valor?
NÃO
SIM
Estou comparando duas médias?
Test t para uma amostra
São dados pareados?
SIM
NÃO
Teste t pareado
Teste t para amostras independentes
SIM
NÃO
Variâncias são homogêneas
Teste t variâncias ponderadas
Teste t para variâncias separadas (Welch)

Como interpretar o resultado do teste

- A conclusão nunca “prova” que H_1 é verdadeira (Lembrem-se do Popper). Usamos frases como: “Há evidência estatística suficiente para rejeitar a hipótese nula...” ou “Não há evidência estatística para concluir que as médias são diferentes...”.
- A conclusão deve sempre se referir ao contexto do problema original. Exemplo: “Não foi encontrada evidência de que os oito ratos machos (...) tinham peso médio diferente do especificado (70 g)”

E se os pressupostos não forem atendidos?

- **Normalidade:** O Teste *t* é considerado **robusto** a desvios da normalidade, especialmente com amostras maiores (n>30) ou de tamanhos iguais. No entanto, se os dados forem muito assimétricos ou contiverem outliers, os resultados podem não ser confiáveis.
- **Possível solução**: testes não paramétricos ou transformações nos dados

- Até aqui aprendemos três versões do Teste t de Student: para uma amostra, para duas amostras independentes e para dados pareados. Cada um parecia ter sua própria fórmula e seu próprio procedimento.
- Isso pode parecer uma 'coleção de testes', uma receita de bolo para cada situação.
- Mas e se eu dissesse que todos eles são, na verdade, **a mesma ideia estatística fundamental** disfarçada de maneiras diferentes?


Médias amostrais são estimadores não enviesados de médias populacionais
Posso comparar médias de duas populações
Existem vários tipos de teste t
Diferentes teste t são tipos particulares de um modelo linear unificado

- Agora, vamos revisitar o Teste t, mas sob uma nova ótica: a de **modelagem linear**.
- Vamos ver como todos os testes t são, na essência, modelos lineares simples.
- Entender isso não só unifica o conhecimento, mas abre as portas para quase todas as técnicas estatísticas que vocês usarão na carreira, como a ANOVA.

Equação de um modelo linear passo-a-passo

Variável resposta = modelo + erro

Equação de um modelo linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

Equação de um modelo linear passo-a-passo

$$\boxed{y_i} = \boxed{\alpha + \beta x_i} + \boxed{\varepsilon}$$

Equação de um modelo linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

Variável resposta

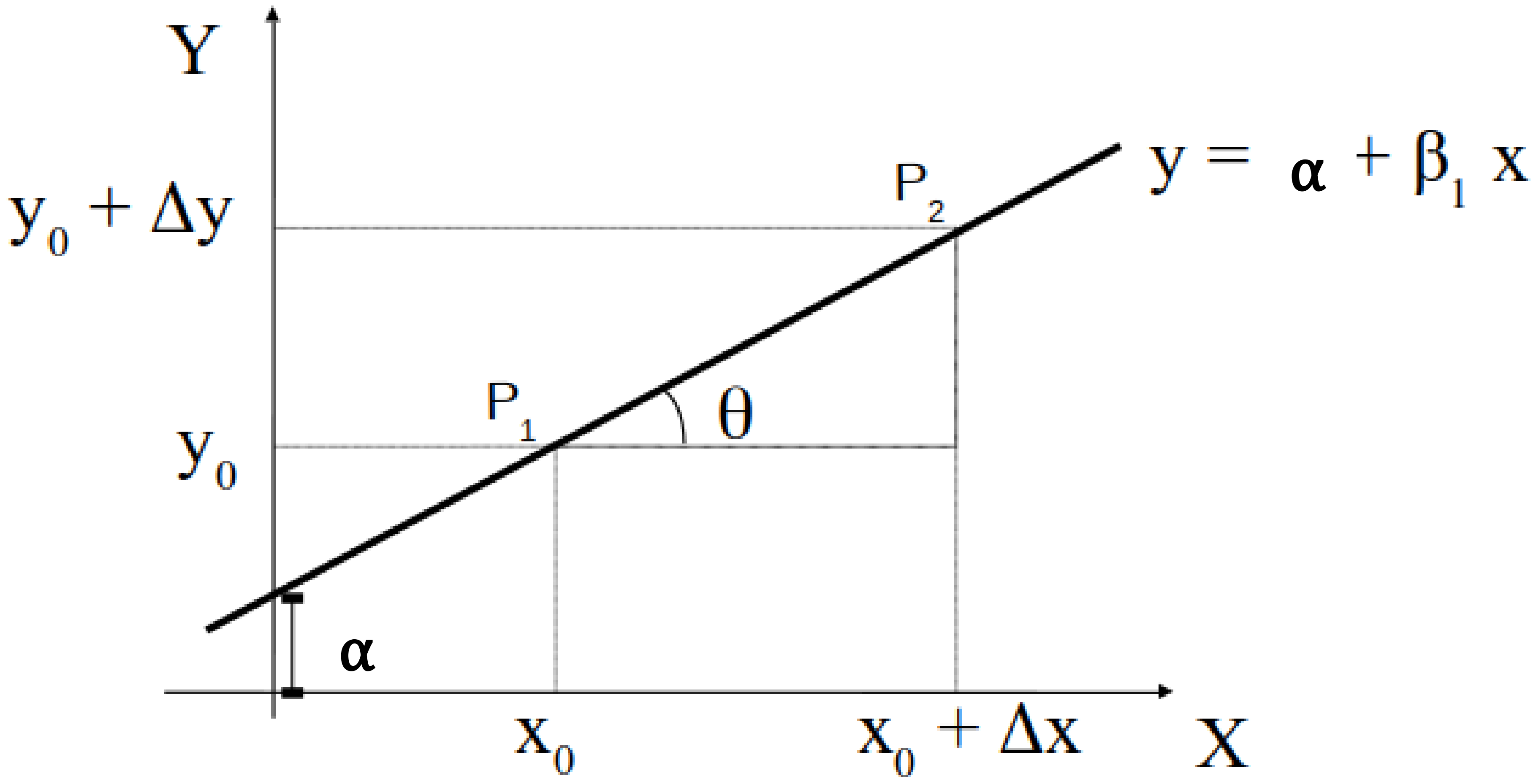
Variável preditora

Equação de um modelo linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

intercepto

Inclinação (slope)



https://www.lampada.uerj.br/arquivosdb/_book/regressão-linear.html

Exemplo

Altura dos filhos = α + β*altura dos pais + resíduos

Pressão arterial = α + β*grupo de tratamento + resíduos

Parâmetros do modelo

- Intercepto = valor médio da variável resposta quando a variável preditora é zero

- Inclinação (slope, coeficiente angular) = inclinação da reta predita pelo modelo. Representa o efeito de X sobre Y. Mede o quanto a variável resposta (Y) muda em unidades de β.
  - Pode-se calcular o IC95% para o slope
- Resíduo (erro): Representa a variabilidade biológica natural e tudo mais que não conseguimos explicar com nosso modelo.

Preditores categóricos

- Até agora, parece uma equação de reta da 8ª série. Mas o grande poder do modelo linear é que a variável preditora, X, **não precisa ser numérica**
- Podemos usar uma variável categórica, como ‘grupo tratado’. E a maneira de fazer isso é com algo chamado **variáveis Dummy**
  - Se temos dois grupos (ex: Controle e Tratamento), criamos uma variável X que assume o valor **0** para um grupo (o grupo de referência, ex: Controle) e **1** para o outro (ex: Tratamento).

Variáveis *dummy* como “tradutores”

- Mas e se o nosso preditor não for um número e sim uma categoria? Por exemplo:
  - **Grupo:** Tratamento vs. Controle
  - **Sexo:** Macho vs. Fêmea
  - **Local:** Área Preservada vs. Área Poluída
- Não podemos colocar a palavra “Tratamento” na equação e multiplicá-la por β. Precisamos de uma forma de “traduzir” essas categorias para uma linguagem numérica que o modelo linear entenda.
- A variável dummy é um “tradutor” que converte categorias em números (0 e 1), permitindo que usemos preditores categóricos em modelos matemáticos como o modelo linear.

Solução para o problema

- A **variável *dummy*** (também chamada de variável fictícia, indicadora ou binária) é a solução para este problema. Ela é uma variável numérica que representa uma informação categórica.
- **Como funciona?** A forma mais comum é a **codificação 0/1**:
- **Escolha um “Grupo de Referência”:** Um dos seus grupos será a base de comparação. A escolha é arbitrária, mas afeta a interpretação dos resultados. Geralmente, o grupo "Controle" ou o grupo mais comum é escolhido como referência.
- **Atribua os Códigos:**
  - O grupo de referência recebe o código **0**.
  - O outro grupo recebe o código **1**.
- Essa simples codificação transforma nossa variável categórica em um número que pode ser usado na equação do modelo linear. Ela atua como um “interruptor”: está “desligada” (0) para o grupo de referência e “ligada” (1) para o outro grupo.

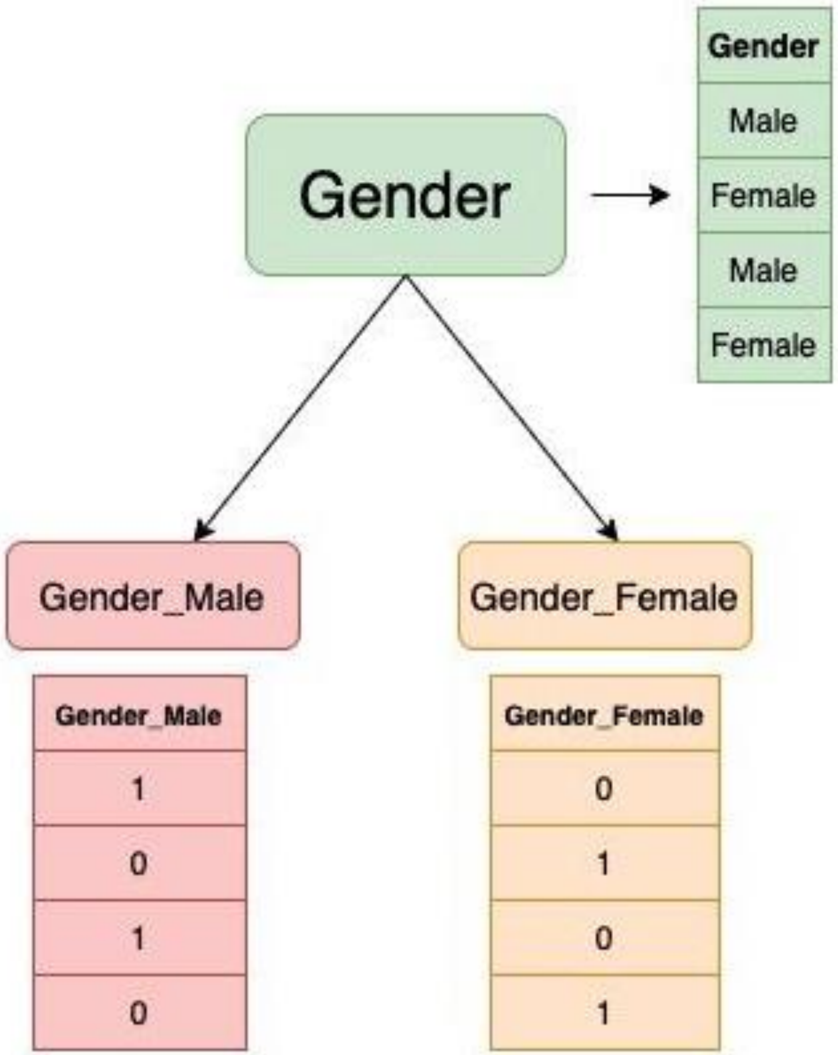
Codificando um fator como *dummy*

- Vamos usar um cenário biológico: um experimento para testar o efeito de um suplemento na dieta de ratos.
- **Variável Categórica Original:** Grupo (com os níveis "Controle" e "Suplemento")
- **Variável Resposta (Y):** Peso_Final (em gramas)

Rato	Grupo	Peso_Final (g)
1	Controle	205
2	Controle	215
3	Suplemento	230
4	Suplemento	240
5	Controle	210

Rato	Grupo	Peso_Final (g)	Suplemento_dummy
1	Controle	205	0
2	Controle	215	0
3	Suplemento	230	1
4	Suplemento	240	1
5	Controle	210	0

Dummy para um fator com 2 níveis



A mágica na interpretação do modelo

- Agora, vamos ver o que acontece quando colocamos essa variável dummy na nossa equação:

$$Peso\_Final_i = \alpha + \beta * (Suplemento\_dummy_i)$$

- **Para um rato do grupo Controle (Suplemento_dummy=0):** A equação se torna:

$$Peso\_Final_{Controle} = \alpha + \beta \cdot (0) = \alpha$$

- **Interpretação:** O coeficiente α (o intercepto) é a **média de peso do grupo de referência** (o grupo Controle)
- **Para um rato do grupo Suplemento (Suplemento_dummy=1):** A equação se torna:

$$Peso\_Final_{Suplemento} = \alpha + \beta \cdot (1) = \alpha + \beta$$

- **Interpretação:** A média de peso do grupo Suplemento é a média do grupo Controle (α) **mais** o valor de β.

Ligando os pontos

- Isso significa que o coeficiente β representa exatamente a **diferença entre a média do grupo “1” e a média do grupo “0”**

$$\beta = Média_{Suplemento} - Média_{Controle}$$

- Portanto, fazer um **Teste t para amostras independentes** para testar se as médias dos dois grupos são iguais (H_0:$\mu_{Suplemento} = \mu_{Controle}$) é matematicamente idêntico a fazer um **modelo linear** e testar se o coeficiente da variável *dummy* é igual a zero (H_0: $\beta = 0$)

E se tivermos mais de 2 grupos?

- A beleza dessa abordagem é que ela se expande facilmente. Se você tiver *k* grupos, precisará de *k-1* variáveis *dummy*
- **Exemplo com 3 grupos:** “Controle”, “Tratamento A”, “Tratamento B”
- **Referência:** “Controle”
- **Variáveis *Dummy* (precisamos de 3 - 1 = 2):**
  - Trat_A_dummy: É 1 se o grupo for "Tratamento A", 0 se for o contrário.
  - Trat_B_dummy: É 1 se o grupo for "Tratamento B", 0 se for o contrário.

Dummy para um fator com 3 níveis

Grupo	Trat_A_dummy	Trat_B_dummy
Controle	0	0
Tratamento A	1	0
Tratamento B	0	1

E se tivermos mais de 2 grupos?

- O modelo seria

$$y = \alpha + \beta_1 * \left(Trat_{A_{dummy}}\right) + \beta_2 \left(Trat_{B_{dummy}}\right)$$

- α seria a média do grupo Controle.
- β_1 seria a diferença entre a média do Tratamento A e o Controle.
- β_2 seria a diferença entre a média do Tratamento B e o Controle.
- Isso é, fundamentalmente, como a **Análise de Variância (ANOVA)** funciona!

Construindo o modelo passo-a-passo

- Queremos comparar o teor de sódio em duas marcas de sopa. Nossa variável resposta (Y) é “teor de sódio”. Nossa variável preditora (X) é a “marca da sopa”

Modelo

Marca de Sopa	Codificação
Knor (Referência)	0
Vono	1

- $Sódio = \alpha + \beta * Marca + \varepsilon_i$
- Interpretando os Coeficientes:
- Para a Knor (X=0): $\boldsymbol{Média}(\boldsymbol{Sódio}_{\boldsymbol{Knor}}) = \boldsymbol{\alpha} + \boldsymbol{\beta} * (\mathbf{0}) = \boldsymbol{\alpha}$
  - *Conclusão: é simplesmente a média do teor de sódio do grupo de referência (Marca Knor)!*
- Para a Vono (X=1): $\boldsymbol{Média}(\boldsymbol{Sódio}_{\boldsymbol{Vono}}) = \boldsymbol{\alpha} + \boldsymbol{\beta} * (\mathbf{1}) = \boldsymbol{\alpha} + \boldsymbol{\beta}$
  - *Conclusão: A média da Vono é a média da Knor mais o valor de β .*
- Se $\boldsymbol{Média}(\boldsymbol{Sódio}_{\boldsymbol{Vono}}) = \boldsymbol{Média}(\boldsymbol{Sódio}_{\boldsymbol{Knor}}) + \boldsymbol{\beta}$, então $\beta = \boldsymbol{Média}(\boldsymbol{Sódio}_{\boldsymbol{Vono}}) - \boldsymbol{Média}(\boldsymbol{Sódio}_{\boldsymbol{Knor}})$
- β é exatamente a diferença entre as médias dos grupos!

Conectando com o teste de hipótese

- A hipótese nula do Teste t para amostras independentes é $H_0{:}\mu_A = \mu_B$
- Isso é o mesmo que dizer que a diferença entre as médias é zero: $H_0{:}\mu_B - \mu_A = 0$
- Vimos que β é a diferença entre as médias. Portanto, a hipótese nula do Teste t é **exatamente a mesma** que testar se o coeficiente β é igual a zero no modelo linear: $\mathbf{H_0{:}\ \beta = 0}$
- A estatística *t* e o valor-p que você obtém ao testar se $\beta = 0$ num programa são idênticos aos que você obtém com um Teste t para amostras independentes

Atividade – Desafio do coeficiente

- **Cenário:** Pesquisadores testaram um novo fármaco para inibir o crescimento de tumores em ratos. Um grupo recebeu o fármaco (Tratado) e outro, um placebo (Controle). Eles rodaram um modelo linear e obtiveram a seguinte saída de computador:

```
Variável Resposta: Tamanho_Tumor (mm³)
--------------------------------------------------
Coeficiente    | Estimativa | Erro Padrão | t-valor | p-valor
--------------------------------------------------
(Intercepto)   |    52.5    |     3.1     | 16.9    | < 0.001
GrupoTratado   |   -15.3    |     4.5     | -3.4    |  0.003
--------------------------------------------------
(Grupo de Referência: Controle)
```

Atividade – Desafio do coeficiente

1. Qual é o tamanho médio do tumor no grupo Controle?
2. Qual é a *diferença* no tamanho médio do tumor entre o grupo Tratado e o Controle?
3. Qual era a hipótese nula que o Teste t estaria avaliando? E qual a hipótese nula correspondente no modelo linear?
4. Com um nível de significância de α=0.05, o fármaco teve um efeito estatisticamente significante? Por quê?

Atividade – Desafio do coeficiente

1. Qual é o tamanho médio do tumor no grupo Controle?
   a) *52.5 mm^3, o valor do intercepto α*
2. Qual é a *diferença* no tamanho médio do tumor entre o grupo Tratado e o Controle?
   a) *-15.3 mm^3, o valor de β. O sinal negativo indica que o grupo tratado teve um tumor menor*
3. Qual era a hipótese nula que o Teste t estaria avaliando?
   a) $H_0: \mu_{Controle} = \mu_{Tratado}$
4. E qual a hipótese nula correspondente no modelo linear?
   a) $H_0: \beta = 0$
5. Com um nível de significância de α=0.05, o fármaco teve um efeito estatisticamente significante? Por quê?

Test t pareado: o modelo mais simples

- **Lógica:** Como vimos, o Teste t pareado funciona calculando a **diferença** (d_i) para cada par. Isso cria uma *nova e única variável*.
- **O Modelo:** Estamos simplesmente testando se a média dessa nova variável é zero. $d_i = \beta + \varepsilon_i$
- **A Hipótese:** A hipótese nula H_0:μ_d= 0 é idêntica a testar se o intercepto desse modelo é zero: **H_0: α = 0**.
- **Conexão:** Um Teste t pareado é apenas um modelo linear com um único parâmetro (o intercepto), aplicado aos dados de diferença

Teste t para uma amostra

- **Lógica:** Queremos testar se a média da nossa amostra é igual a um valor de referência, μ_0.
- **O Modelo:** Em vez de modelar Y, vamos modelar *a diferença para o valor de referência*: $Y_i^{'} = Y_i - \mu_0$.
- Se a média original $(E[Y_i])$ for igual a μ_0, então a média da nossa nova variável $(E[Y_i])$ = 0
- **Equação do Modelo:** $(Y_i - \mu_0) = \alpha + \varepsilon_i$
- **A Hipótese:** A hipótese nula H_0:μ=μ_0 se torna um teste para saber se o intercepto (que é a média da variável transformada) é zero: **H_0: α = 0**

Por que pensar em modelos?

Tipos de teste *t*	**Pergunta**	**Modelo linear**	**Hipótese nula**
Uma amostra	A média da população é μ_0?	$(Y - \mu_0) \sim \alpha$	$H_0: \alpha = 0$
Pareado	A média das diferenças é 0?	$Y_{diferença} \sim \alpha$	$H_0: \alpha = 0$
Amostras independentes	As médias dos dois grupos são iguais?	$Y \sim \alpha + \beta X_{grupo}$	$H_0: \beta = 0$

Vantagens da abordagem de modelagem

- **Escalabilidade:** E se tivermos 3 grupos ou mais? Não precisamos de um novo teste "do zero". Apenas adicionamos mais preditores *dummy* ao nosso modelo linear. Isso é a **ANOVA**.
- **Flexibilidade:** E se quisermos comparar dois grupos, mas também controlar o efeito de uma variável contínua (como a idade do paciente)? Apenas adicionamos outra variável preditora ao modelo. Isso é a **ANCOVA**.
  - $Y \sim \alpha + \beta_1 X_{grupo} + \beta_2 X_{idade}$
- **Compreensão aprofunda:** Em vez de decorar uma lista de testes, você aprende um único arcabouço para formular e responder perguntas científicas

O Teste t, a ANOVA e a regressão não são ferramentas diferentes; são variações de um mesmo tema: os modelos lineares.
Ao entender essa conexão, vocês estão se preparando não apenas para usar a estatística, mas para pensar como estatístico(a)!