Introdução ao Planejamento e Análise Estatística de Experimentos

Prof. Caio Azevedo

- Grande parte do conhecimento científico é desenvolvido através de evidências empíricas.
- Tais evidências, em geral, são oriundas dos resultados de experimentos devidamente planejados e analisados.
- Utilização de planejamento e análise de experimentos: Psicometria, Física, Química, Biologia, Economia entre outras.
- Extrapolação dos resultados para um “universo” maior.

- Psicometria: medir conhecimento de indivíduos, medir predisposição em se adquirir um produto.
- Biologia: comparar o desenvolvimento de diferentes tipos de árvores, comparar a resistências de diferentes tipos de bactérias à temperaturas extremas.
- Física: estudar o impacto de fontes radiotativas em ambientes urbanos ou lugares turísticos.
- Economia: comparar o desempenho de diferentes carteiras de investimento.

Estatística & Experimentação

- Planejamento e análise estatística de experimentos: trata da correta estruturação, implementação e análise de experimentos.
- Permite extrapolar os resultados da amostra para a população (inferência).
- Identificação de características e nuâncias (desconhecidas, não previstas).

Exemplo: comparação de altura de árvores

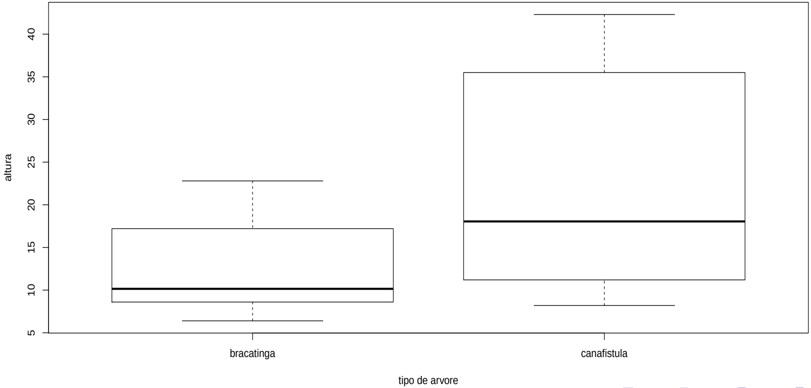
- Um agrônomo realizou um levantamento para estudar o desenvolvimento de duas espécies de árvores: a bracatinga (Mimosa scabrella) e canafístula (Peltophorum dubium).
- Para essa finalidade foram coletadas duas amostras de tamanhos iguais à 30 árvores de cada espécie.
- Considere que o desenvolvimento das árvores se traduz em medir suas alturas.

Observações

- Como responder à pergunta de interesse?
- Dados

Bracatinga						Canafístula					
6,4	7,0	9,0	10,2	16,2	20,1	8,2	10,1	14,1	20,2	25,7	40,1
6,8	8,3	9,1	11,4	16,3	20,3	9,7	10,3	14,2	20,3	30,9	40,2
6,9	8,6	9,3	13,7	17,2	21,4	9,8	11,2	14,4	20,6	35,5	40,5
6,9	8,7	9,9	14,8	18,4	22,8	10,0	13,2	14,8	29,9	38,2	41,8
6,9	8,7	10,1	15,2	20,0	22,8	10,0	13,4	15,9	23,8	40,0	42,3

Boxplot


altura
5
10
15
20
25
30
35
40
bracatinga
canafistula
tipo de arvore

Medidas-resumo

Medida resumo	bracatinga	canafístula
Média	12,78	22,31
Mediana	10,15	18,20
Desvio-padrão	5,48	12,22
CV(%)	42,90	54,78
Mínimo	6,40	8,20
Máximo	22,80	42,30

Considerações

- Lembrando : $X \sim N(\mu, \sigma^2)$ (variável aleatória), x (valor observado).
- Seja Y_{ij} a variável aleatória (va) que representa a altura da j-ésima árvore selecionada do i-ésimo tipo.
- Neste caso, $i = 1, 2$ e $j = 1, 2, ..., 30$ e y_{ij} valores observados (dados na tabela anterior).
- Suponha que $Y_{ij} \sim N(\mu_i, \sigma_i^2)$ (mutuamente independentes) (qual o significado desta suposição?).
- Teste de Kolmogorov para normalidade: bracatinga (pvalor = 0,1274), canafístula (pvalor = 0,1813) (qual o significado deste resultado?).

Verossimilhança

- Sob as suposições $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$
  $L(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}(\sigma_1^2)^{n_1/2}(\sigma_2^2)^{n_2/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{30}\frac{(y_{ij}-\mu_i)^2}{\sigma_i^2}\right]$
- Estatística suficiente (o modelo acima pertence à família exponencial multiparamétrica): $\mathbf{T} = (\overline{Y}_1, \overline{Y}_2, S_1^2, S_2^2)$, em que $\overline{Y}_i = \frac{1}{30}\sum_{j=1}^{30} Y_{ij}, i = 1,2$ e $S_i^2 = \frac{1}{30-1}\sum_{j=1}^{30}\left(Y_{ij} - \overline{Y}_i\right)^2, i = 1,2$ (exercício).
- Estimadores de máxima verossimilhança $\widehat{\mu}_i = \overline{Y}_i$, $\widehat{\sigma}_i^2 = \frac{30-1}{30} S_i^2$, $i = 1, 2$.
- Neste caso, é possível construir intervalos de confiança exatos (exercício).

Teste para igualdade das variâncias

- Hipóteses: $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_1 : \sigma_1^2 \neq \sigma_2^2$.
- Estatísticas $F = \frac{S_1^2}{S_2^2} \sim F_{(n_1-1,n_2-1)}$, sob H_0, em que n_i é o tamanho da amostra do grupo $i = 1, 2$, lembrando que $\frac{(n_i-1)S_i^2}{\sigma_i^2} \sim \chi^2_{(n_i-1)}$.
- Valor observado $f = \frac{s_1^2}{s_2^2} = \frac{30{,}037}{149{,}409} = 0, 20$, $pvalor = \min\left(P(F > f|H_0), P(F < f|H_0)\right)$, $F \sim F_{(n_1-1,n_2-1)}$. Neste caso, $pvalor < 0, 0001$.

Teste para igualdade das médias

- Hipóteses: $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$.
- Estatísticas $T = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$ que tem uma distribuição **aproximada**, sob H_0, t_ν, $\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$.
- Valor observado $t = -3,90$, $pvalor = 2P(T > |t||H_0) = 0,0004$, $T \sim t_\nu$.
- $\widetilde{\mu}_1 = 12,78(1,00)$, $IC(\mu_1, 95\%) = [10,81; 14,74]$ e $\widetilde{\mu}_2 = 22,31(2,23)$, $IC(\mu_1, 95\%) = [17,94; 26,68]$.

Outras formas/métodos de comparação

- Comparar as próprias distribuições.
- Testes não paramétricos para igualdade de médias (testes de Kruskal-Wallis e de Mann-Whitney).
- Comparar as medianas ou as modas ao invés das médias.
- Assumir normalidade e igualdade de variâncias e testar a igualdade de médias, através do estudo da variabilidade dos dados. Se a hipótese de igualdade for aceita implica que as distribuições são as mesmas.

Análise de Variância $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$

- Assume-se iguais variâncias para os grupos.
- Soma de quadrados total (defina $\overline{y} = \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}$, $n = n_1 + n_2$) (provar a expressão abaixo)

  $SQT = \sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 =$

  $$\underbrace{\sum_{i=1}^{2} n_i (\overline{y}_i - \overline{y})^2}_{\text{entre os tipos de árvore}} + \underbrace{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2}_{\text{dentro de cada tipo de árvore}} =$$

  $SQF + SQR$
- Se SQF (soma de quadrados dos fatores) for **significativamente** maior do que SQR (soma dos quadrados dos resíduos), conclui-se que as médias populacionais são diferentes.

Generalizando para k grupos

- Neste caso, as hipóteses de interesse são $H_0 : \mu_1 = \mu_2 = ...\mu_k$ vs H_1 : há pelo menos uma diferença.
- Soma de quadrados total $\overline{y} = \frac{1}{n}\sum_{i=1}^{K}\sum_{j=1}^{n_i} y_{ij}, n = n_1 + n_2$ (provar a expressão abaixo)

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 =$$
$$\underbrace{\sum_{i=1}^{k} n_i (\overline{y}_i - \overline{y})^2}_{\text{entre os grupos}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2}_{\text{dentro de cada grupo}} = SQF + SQR$$

- Se SQF for significativamente maior do que SQR, conclui-se que as médias populacionais são diferentes.

Estatística do teste

- Defina $QMF = SQF/(2-1)$ e $QMR = SQR/(60-2))$ e $F = QMF/QMR$
- Sob H_0, temos que $F \sim F_{(2-1,60-2))}$.
- Para K grupos, temos $QMF = SQF/(k-1)$ e $QMR = SQR/(n-k)$ e, sob H_0, $F \sim F_{(k-1,n-k)}$
- Note que (sob H_0) $SQF/\sigma^2 \sim \chi^2_{(k-1)}$, $SQR/\sigma^2 \sim \chi^2_{(n-k)}$ e $SQF \perp SQR$.
- Logo $F = \frac{SQF/[(k-1)\sigma^2]}{SQR/[(n-k)\sigma^2]} = QMF/QMR \sim F_{(k-1),(n-k)}$

Justificativa para o resultado anterior

- Sob $H_0(\mu_1 = \mu_2 = ... = \mu_k = \mu)$ temos que $Y_{ij} \sim N(\mu, \sigma^2)$. Por simplicidade, assuma que $\mu = 0$
- Observação: melhor maneira de se provar os resultados é utilizando resultados de formas quadráticas associadas à vetores aleatórios.
- Defina $W = SQT/\sigma^2$, $V = SQF/\sigma^2$ e $Z = SQR/\sigma^2$.
- Note que $\sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2 = \sum_{i=1}^{k} \left(Y_{ij} - \overline{Y}\right)^2 + n\overline{Y}^2$.
- Assim $Q = W + S$, onde $Q = \frac{1}{\sigma^2} \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}^2$ e $S = \frac{n}{\sigma^2} \overline{Y}^2$

Cont.

- Pode-se provar que W e S são independentes e, assim $M_Q(t) = M_W(t)M_S(t)$.
- Além disso, $Q \sim \chi^2_n$ e $S \sim \chi^2_1$.
- Logo $M_W(t) = \frac{M_Q(t)}{M_S(t)} = \frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}} = (1-2t)^{-(n-1)/2}$.
- Assim $W \sim \chi^2_{(n-1)}$.
- Analogamente pode-se provar que $V \sim \chi^2_{(k-1)}, Z \sim \chi^2_{(n-k)}$ e $V \perp Z$ (exercício).
- Veja livro Mood, Graybill and Boes, páginas 243 e 244.

Voltando ao exemplo

- Neste caso $f = \frac{1363{,}3/1}{5203{,}9/58} = \frac{1363{,}27}{89{,}72} = 15,19$, $pvalor = 0,0003$.
- Estamos assumindo que $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (embora, aparentemente, esta suposição não seja verdadeira).
- Neste caso, um estimador apropriado para σ^2 é $S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$, que resulta em $s^2 = 89,72$.
- Assim, $\widetilde{\mu}_1 = 12,78(1,73)$, $IC(\mu_1, 95\%) = [9,39; 16,17]$ e $\widetilde{\mu}_2 = 22,31(1,73)$, $IC(\mu_1, 95\%) = [18,91; 25,70]$.

Aspectos sobre planejamento e análise estatística de experimentos

- Objetivo: perguntas de interesse a serem respondidas, em geral, são traduzidas por hipóteses de interesse (não estatísticas). Exemplo: as árvores da espécie bracatinga apresentam maior desenvolvimento do que as árvores da espécie canafístula.
- Experimento: conjunto de procedimentos que seguem um protocolo a fim de leventar informações para alcançar os objetivos. Exemplo: selecionar 30 árvores de cada espécie, em uma determinada região, segundo um plano de amostragem aleatória simples sem reposição.

Cont.

- Variável resposta ou resposta: característica de principal interesse no experimento. Exemplo: altura de cada árvore.
- Variável explicativa ou covariável: característica de interesse (secundário) que auxilia no estudo do comportamento da variável resposta. Em geral, será chamada de fator ou tratamento. Exemplo: tipo de árvore.
- Fator de confundimento: característica que pode influenciar nos resultados do experimento e que, em princípio, não é considerado como fator de interesse. Exemplo: tipo de solo.

Cont.

- Unidade experimental. Elemento de interesse no experimento. Exemplo: árvore
- Planejamento amostral: mecanismo de seleção das unidades experimentais. Exemplo: selecionar 30 árvores de cada tipo, segundo um plano AASs (amostragem aleatória simples sem reposição).
- Planejamento experimental: estrutura que indica como os fatores interagem entre si e como as unidades experimentais são alocadas às combinações dos fatores. Exemplo: serão selecionadas 15 árvores de cada espécie para serem submetidas à determinado inseticida (as outras 15 não serão submetidas).

Aspectos relevantes

- O teste proposto anteriormente fornece indívio se existe ou não pelo menos uma média populacional diferentes das demais.
- Como idenditificar quais médias são diferentes?
- Se tivermos mais de um fator de interesse (e quisermos estudar como os fatores interagem entre si em termos da resposta).?
- Se quisermos ter estimativas mais precisas dos parâmetros de interesse (eg, média) ?
- Como forncer estimativas para as possíveis médias comuns entre os tratamentos?
- Como avaliar as suposições consideradas?