ME623 - Planejamento e Pesquisa

Parte 03

Tatiana Benaglia - 2024S1

Passo-a-Passo em Delineamento de Experimentos:

Relembrando:

1. Definir os objetivos do experimento. Especificar o que será medido.
2. Identificar as fontes de variação:
   i. Fatores e os seus níveis,
   ii. Unidades experimentais,
   iii. Blocos e covariáveis.
3. Aleatoriamente, alocar unidades experimentais aos tratamentos.
4. Especificar um modelo.
5. Rodar um experimento piloto.
6. Calcular o número de replicações e/ou observações necessárias.

Experimentos Completamente Aleatorizados

Na aula anterior vimos como comparar dois tratamentos através do:

- Teste t para amostras independentes
- ANOVA com um único fator de dois níveis

Muitos experimentos desse tipo envolvem mais de dois níveis (tratamentos) do fator.

Iremos então generalizar o caso de um fator A com dois tratamentos para o caso de a tratamentos.

Esses experimentos são conhecidos como **Experimentos Completamente Aleatorizados**.

Exemplo: Algodão

Antes de definir formalmente o modelo, vamos pensar no seguinte exemplo.

- Uma engenheira quer investigar a resistência de uma nova fibra sintética usada para fazer camisetas.
- Ela sabe que a porcentagem de algodão na composição da fibra afeta a resistência.
- Será quer aumentar a porcentagem de algodão aumentará a resistência da fibra?
- A porcentagem de algodão deve ser entre 10 e 40% para que o produto final tenha outras características de qualidade desejáveis (como poder aplicar uma estampa)
- Interesse: testar 5 níveis do percentual de algodão (15%, 20%, 25%, 30% e 35%) e, para cada nível, repetir o experimento 5 vezes.

Perguntas

1) Quanto(s) e qual(is) fator(es) temos neste experimento?
2) Quantos níveis temos em cada fator? Quais são?
3) Quantas replicações serão feitas e quantas UEs são necessárias?

Exemplo: Algodão

A tabela abaixo apresenta o resistência medida em lb/in^2 nas unidades experimentais.

Percentual de Algodão (%)	Observações 1	2	3	4	5	Total	Média
15	7	7	15	11	9	49	9.8
20	12	17	12	18	18	77	15.4
25	14	18	18	19	19	88	17.6
30	19	25	22	19	23	108	21.6
35	7	10	11	15	11	54	10.8
						376	15.04

Análise Gráfica

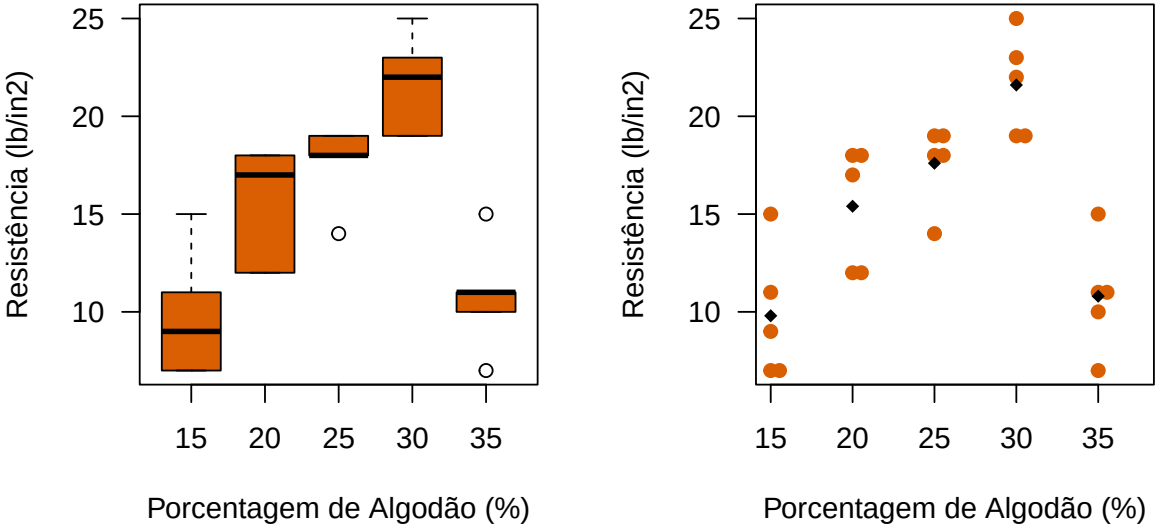

Figure 1: Boxplot (esquerda) e dotplot (direita) da resistência para cada percentual de algodão

A Análise de Variância

- Queremos testar se existe diferença entre as resistências média para todos os $a = 5$ níveis do fator A.
- E por que não aplicar o teste t para todos os pares de médias?

  Tendo 5 tratamentos, temos um total de 10 pares possíveis de média.

$$P(\text{Não rejeitar } H_0 | H_0 \text{ é verdadeira}) = (1 - 0.05)^{10} = 0.60.$$

  Portanto,

$$P(\text{Erro Tipo I}) = P(\text{Rejeitar } H_0 | H_0 \text{ é verdadeira}) = 1 - 0.60 = 0.40 >> 0.05$$

- O procedimento apropriado para testar a igualdade de várias médias simultaneamente é conhecido como **Análise de Variância** ou ANOVA.

Representação e Notação

Suponha que temos um fator A com a níveis (tratamentos) e n replicações para cada tratamento.

A resposta observada de cada aplicação do tratamento em uma unidade experimental é denotada pela variável aleatória y.

A tabela abaixo mostra a apresentação típica dos dados em experimentos com um fator.

Tratamento	**Observações**				**Totais**	**Média**
1	y_{11}	y_{12}	$\cdots$	y_{1n}	$y_{1\cdot}$	$\bar{y}_{1\cdot}$
2	y_{21}	y_{22}	$\cdots$	y_{2n}	$y_{2\cdot}$	$\bar{y}_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	
a	y_{a1}	y_{a2}	$\cdots$	y_{an}	$y_{a\cdot}$	$\bar{y}_{a\cdot}$
					$y_{\cdot\cdot}$	$\bar{y}_{\cdot\cdot}$

Notação

- y_{ij} é a j-ésima observação no i-ésimo tratamento.
- $y_{i\cdot}$ é a soma das observações no i-ésimo tratamento.
- $\bar{y}_{i\cdot}$ é a média das observações para o i-ésimo tratamento.
- $\bar{y}_{\cdot\cdot}$ é a média de todas as observações.

Simbolicamente,

$$y_{i\cdot} = \sum_{j=1}^{n} y_{ij} \qquad \text{e} \qquad \bar{y}_{i\cdot} = \frac{y_{i\cdot}}{n}, \quad \text{para } i = 1, \ldots, a.$$

Além disso,

$$y_{\cdot\cdot} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij} \qquad \text{e} \qquad \bar{y}_{\cdot\cdot} = \frac{y_{\cdot\cdot}}{an} = \frac{y_{\cdot\cdot}}{N} = \frac{\bar{y}_{1\cdot} + \bar{y}_{2\cdot} + \cdots + \bar{y}_{a\cdot}}{a},$$

em que $N = an$ é o número total de observações (tamanho da amostra).

Modelo de Análise de Variância de Um Fator

Suponha que temos experimento completamente aleatorizado de um fator, cujos tratamentos são aplicados aleatoriamente nas UEs com igual número de replicações para cada tratamento.

O modelo estatístico de análise de variância com um fator (*One-Way ANOVA*) postula a seguinte relação para os dados:

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, a; \quad j = 1, 2, \ldots, n;$$

em que:

- μ_i é a média das respostas para o i-ésimo tratamento. Note que μ_i é um número e não uma v.a.;
- ε_{ij} é o erro experimental. Assumimos que, se tudo estiver controlado, os erros são identicamente distribuídos com $\mathbb{E}(\varepsilon_{ij}) = 0$, $\text{Var}(\varepsilon_{ij}) = \sigma^2$ para todo i, j, e $\mathbb{E}(\varepsilon_{ij}\varepsilon_{i'j'}) = 0$ se $i' \neq i$ ou $j' \neq j$ (não-correlacionados).

Dito isto, vamos olhar uma versão um pouco mais complicada, mas sem a qual não conseguiremos generalizar o modelo para *Two-Way ANOVA* e coisas mais complicadas.

Modelo de Análise de Variância de Um Fator

Suponha que temos experimento completamente aleatorizado de um fator, cujos tratamentos são aplicados aleatoriamente nas UEs com igual número de replicações para cada tratamento.

Vamos considerar este modelo:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, 2, ..., a; \quad j = 1, 2, ..., n;$$

em que:

- μ é a média global das respostas;
- τ_i é o efeito do i-ésimo tratamento;
- ε_{ij} é o erro experimental. Assumimos que, se tudo estiver controlado, os erros são identicamente distribuídos com $\mathbb{E}(\varepsilon_{ij}) = 0$, $\text{Var}(\varepsilon_{ij}) = \sigma^2$ para todo i, j, e $\mathbb{E}(\varepsilon_{ij}\varepsilon_{i'j'}) = 0$ se $i' \neq i$ ou $j' \neq j$ (não-correlacionados).

Este modelo tem um problema de identificabilidade. De forma bem geral, note que temos $a + 1$ parâmetros para estimar a médias. Então, precisaremos de alguma restrição.

Identificabilidade

Vamos verificar o problema de identificabilidade de uma maneira mais formal.

Do seu curso de Inferência: um modelo probabilístico $\mathbb{P}_\theta$ é chamado **identificável** se

$$\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2} \Rightarrow \theta_1 = \theta_2,$$

ou olhando a contra-positiva,

$$\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}.$$

Isto significa que dois pontos diferentes no espaço paramétrico do modelo implicam que a probabilidade de um evento (por exemplo, a verossimilhança) será diferente.

De volta à ANOVA

A formulação que

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, 2, \dots, a; \quad j = 1, 2, \dots, n;$$

não é identificável.

Defina $\theta = (\mu, \tau_1, \dots, \tau_a, \sigma^2)$ os parâmetros do modelo.

Defina $\theta^* = (\mu - \gamma, \tau_1 + \gamma, \dots, \tau_a + \gamma, \sigma^2)$ para um $\gamma \in \mathbb{R}$ qualquer.

Então quando usamos como verossimilhança $\mathbb{P}_\theta$, temos que

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, 2, \dots, a; \quad j = 1, 2, \dots, n;$$

mas quando usamos como verossimilhança $\mathbb{P}_{\theta^*}$, temos que

$$\begin{aligned} y_{ij} &= (\mu - \gamma) + (\tau_i + \gamma) + \varepsilon_{ij} \\ &= \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, 2, \dots, a; \quad j = 1, 2, \dots, n. \end{aligned}$$

Isto é, a distribuição de y_{ij} será a mesma.

Uma restrição que garante identificabilidade

Existem diversas maneiras de introduzir uma **restrição** no espaço paramétrico que introduza identificabilidade.

No caso, a que usamos com mais frequência é a imposição (no caso em que o número de replicações é o mesmo) que

$$\sum_{i=1}^{a} \tau_i = 0.$$

Essa restrição tem algumas propriedades que vamos mostrar em aulas futuras:

- O estimador de mínimos quadrados $\widehat{\mu}$ sob esta restrição coincide com a média dos dados.
- Os efeitos estimados $\widehat{\tau}_i$ também têm uma interpretação clara: são efeitos diferenciais, com relação à média global, do i-ésimo nível do tratamento na resposta.

Contudo, nem todo *software* usa essa restrição.

Prova que a restrição garante identificabilidade

Suponha que para θ_1 e θ_2 vale a condição $\sum_{i=1}^{a} \tau_i^{(k)} = 0$, para $k = 1, 2$. Então como a distribuição de ε_{ij} não depende de μ e τ_i, temos que

$$\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2} \Rightarrow \begin{cases} \mu^{(1)} + \tau_1^{(1)} = \mu^{(2)} + \tau_1^{(2)}, \\ \vdots \\ \mu^{(1)} + \tau_a^{(1)} = \mu^{(2)} + \tau_a^{(2)}, \\ \sigma^2_{(1)} = \sigma^2_{(2)}. \end{cases}$$

Mas como $\sum_{i=1}^{a} \tau_i^{(k)} = 0$, para $k = 1, 2$. eu posso somar as primeiras a linhas e concluir que $\mu^{(1)} = \mu^{(2)}$. Consequentemente $\tau_i^{(1)} = \tau_i^{(2)}$ para todo i, e

$$\mathbb{P}_{\theta_1} = \mathbb{P}_{\theta_2} \Rightarrow \theta_1 = \theta_2. \blacksquare$$

Modelo de Análise de Variância de Um Fator

Suponha que temos experimento completamente aleatorizado de um fator, cujos tratamentos são aplicados aleatoriamente nas UEs com igual número de replicações para cada tratamento.

O modelo de análise de variância é:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, a; \quad j = 1, 2, \ldots, n;$$

sujeita à restrição que

$$\sum_{i=1}^{a} \tau_i = 0,$$

em que:

- μ é a média global das respostas;
- τ_i é o efeito do i-ésimo tratamento;
- ε_{ij} é o erro experimental. Assumimos que, se tudo estiver controlado, os erros são identicamente distribuídos com $\mathbb{E}(\varepsilon_{ij}) = 0$, $\text{Var}(\varepsilon_{ij}) = \sigma^2$ para todo i, j, e $\mathbb{E}(\varepsilon_{ij}\varepsilon_{i'j'}) = 0$ se $i' \neq i$ ou $j' \neq j$ (não-correlacionados).

Balanceamento

O cenário em que o número de replicações em cada tratamento é igual será chamado de **balanceado**.

Mostraremos no futuro que este é o caso ideal. Sempre que possível, deve-se realizar o experimento balanceado. Agora, não sendo possível, precisamos usar um modelo não-balanceado (*One-way ANOVA* não-balanceada).

Seja:

- Um fator A com a níveis (tratamentos).
- Cada tratamento i tem n_i replicações independentes, $i = 1, \ldots, a$.
- Ao todo temos $N = \sum_{i=1}^{a} n_i$ observações no experimento.

Modelo de Análise de Variância Não-Balanceada

Suponha que temos um tratamento completamente aleatorizado aplicado nas u.e. com um número n_i de replicações para cada nível do tratamento.

O modelo de análise de variância é:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, 2, \ldots, a; \quad j = 1, 2, \ldots, n_i;$$

sujeita à restrição que

$$\sum_{i=1}^{a} n_i \tau_i = 0,$$

em que:

- μ é a média global das respostas;
- τ_i é o efeito do i-ésimo tratamento;
- ε_{ij} é o erro experimental. Assumimos que, se tudo estiver controlado, os erros são identicamente distribuídos com $\mathbb{E}(\varepsilon_{ij}) = 0$, $\text{Var}(\varepsilon_{ij}) = \sigma^2$ para todo i, j, e $\mathbb{E}(\varepsilon_{ij}\varepsilon_{i'j'}) = 0$ se $i' \neq i$ ou $j' \neq j$ (não-correlacionados).

Algumas Considerações

- A restrição $\sum_{i=1}^{a} n_i \tau_i = 0$ garante que $\hat{\mu}$ coincida com $\bar{y}_{..}$, a média global.
- Note que no caso não balanceado, a média global é uma média ponderada:

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij} = \frac{n_1 \bar{y}_{1.} + n_2 \bar{y}_{2.} + \cdots + n_a \bar{y}_{a.}}{\sum_{i=1}^{a} n_i}.$$

Falamos aqui vagamente sobre os estimadores $\hat{\mu}$ e $\hat{\tau}_i$, porque precisamos construir um procedimento antes, que evita falsas descobertas.

Nas próximas aulas veremos:

- O procedimento de ANOVA para testar a hipótese $\tau_1 = \cdots = \tau_a = 0$, e como estimar a variância $\hat{\sigma}^2$.

Leitura

- Montgomery, DC. Design and Analysis of Experiments. Capítulo 3.
- Dean, A. e Voss, D. - Design and Analysis of Experiments. Capítulo 3.

Agradecimentos
Parte deste material foi criado pelo prof. Guilherme Ludwig - IMECC/UNICAMP