Planejamento e Análise de Experimentos

6 - Comparações Simples e Pareadas

Michel Bessani
Departamento de Engenharia Elétrica - DEE

Programa de Pós-Graduação em Engenharia Elétrica - PPGEE

UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

Belo Horizonte

Sumário

Introdução

A comparação entre duas populações, utilizando estatísticas calculadas com as amostras aleatórias destas populações, segue os mesmos princípios dos testes de hipóteses para uma única população.

Inferências com duas amostras são rotineiramente utilizadas para comparar o efeito de uma técnica (tratamento) com um grupo de controle: placebo, padrão-ouro, técnica clássica, etc.

As perguntas usualmente envolvem:

- Comparação das médias;
- Comparação da variância;
- Comparação de proporções;
- ...

Comparação Simples

Ilustrando:

Um dos aspectos principais na fabricação de hastes de aço é o corte das barras com um comprimento correto, o qual é o esperado pelos clientes.

Tal processo está sujeito a erros, que resultam em custos adicionais para a padronização e reprocessamento das hastes.

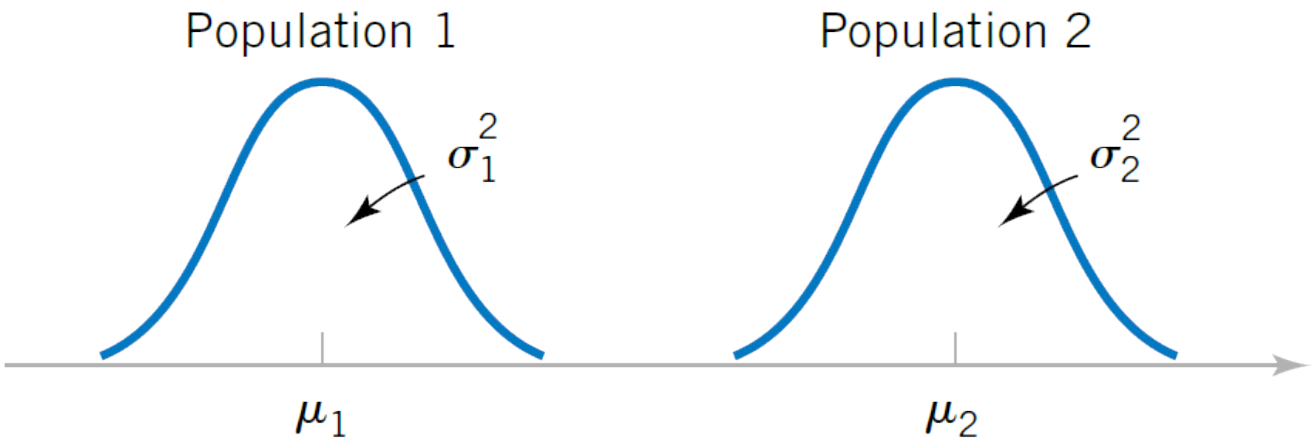
Um engenheiro está interessado em comparar o processo atual de corte das hastes com um novo método que poderia melhorar o desempenho do processo de fabricação[1].

[1] Adaptado de um trabalho final da disciplina desenvolvido por D.F. Carvalho em 2012.

Podemos definir um **modelo** estatístico para os dados que coletaríamos ao fazer um experimento no contexto desse exemplo:

$$y_{ij} = \mu_i + \epsilon_{ij}, \text{ onde: } \begin{cases} i = 1, 2 \\ j = 1, \ldots, n_i \end{cases}$$

Podemos também assumir que os resíduos ϵ_{ij} são iid e seguem uma $\mathcal{N}(0, \sigma_i^2)$, o que implica em:



Desejamos inferir sobre a diferença entre os valores médios da variação entre os dois métodos (atual e novo).

Nesse caso, uma variável de resposta poderia ser o erro absoluto de cada método, e.g., $y_i = |l_i - l_{nominal}|$.

As hipóteses estatísticas podem ser definidas como:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{cases} \quad \text{ou} \quad \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$$

Vamos supor que é desejado um nível de significância $\alpha = 0.05$ e que o engenheiro está interessado em detectar qualquer diferença maior do que 15 cm no erro absoluto médio com um poder $(1 - \beta) = 0.8$.

Comparação Simples: Variâncias Iguais

Vamos também assumir que a variância do processo é desconhecida mas pode-se considerar que é similar para os dois métodos.

Uma vez que σ^2 é desconhecida, teremos que estimá-la a partir das observações. Como estamos assumindo que $\sigma_1^2 \approx \sigma_2^2$, podemos utilizar um estimador da variância agrupada:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = wS_1^2 + (1 - w)S_2^2$$

Considerando esse estimador e as premissas assumidas, podemos escrever:

$$T = \frac{(\bar{y}_1 - \bar{y}_1) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{(n_1+n_2-2)}$$

Relembrando as nossas hipóteses:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 < 0 \end{cases}$$

Sob H_0:

$$t_0 = \frac{(\bar{y}_1 - \bar{y}_1) - \cancelto{0}{(\mu_1 - \mu_2)}}{S_p\sqrt{1/n_1 + 1/n_2}} = \frac{(\bar{y}_1 - \bar{y}_1)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{(n_1+n_2-2)}$$

Dessa forma, rejeitaremos H_0 com um nível de confiança de $(1-\alpha)$ se $t_0 \leq t_{(\alpha,n_1+n_2-2)}$.

Vamos relembrar os requisitos especificados pelo engenheiro para o teste de hipóteses:

- Significância $\alpha = 0.05$;
- Poder $(1 - \beta) = 0.80$;
- Tamanho de efeito minimamente relevante $\delta^* = 15$ cm.

A partir dessas especificações, podemos obter o número de observações necessárias.

Para o caso de variâncias aproximadamente iguais[2], o tamanho ótimo das amostras é $n_1 = n_2 = n$, onde:

$$n \cong 2\left(\frac{t_{(\alpha,2n-2)} + t_{(\beta,2n-2)}}{d^*}\right)^2$$

onde, $d^* = \delta^*/\sigma$ é o tamanho de efeito minimamente relevante normalizado, e $t_{(\alpha,2n-2)}$ e $t_{(\beta,2n-2)}$ são os quantils α e β da distribuição $t_{(2n-2)}$.

[2]Paul Mathews', Sample Size Calculations, MMB, 2010.

Apesar da conveniência dessa fórmula, existe um problema: precisamos da estimativa da variância para calcular o tamanho amostral, mas precisamos de uma amostra para estimar a variância.

Podemos proceder de algumas formas:

- Utilizar o conhecimento do processo ou dados históricos para obter uma estimativa inicial da variância;
- Utilizar o tamanho de efeito minimamente relevante normalizado;
- Realizar um estudo piloto para estimar a variância.

Cada uma das abordagens possuem vantagens e desvantagens.

Vamos supor, para o experimento das hastes de aço, que o engenheiro utilizou informações disponíveis nos manuais em conjunto com dados históricos de medidas, para estimar um valor máximo razoável para o desvio padrão $\sigma \approx 15$ cm.

Assumindo a premissa das variâncias iguais, podemos então simplesmente utilizar a fórmula:

$$n \cong 2\left(\frac{t_{(\alpha,2n-2)} + t_{(\beta,2n-2)}}{\delta^*/\sigma}\right)^2$$

Existe ainda um último problema, os valores de $t_{(\alpha,2n-2)}$ e $t_{(\beta,2n-2)}$ também dependem de n, o que torna a equação do tamanho amostral transcendental em n.

Podemos resolver utilizando uma primeira estimativa de $t_{k,(2n-2)} \approx z_k$, e depois iterar até achar o menor n que satisfaz:

$$n \geq 2\left(\sigma/\delta^*\right)^2\left(t_{(\alpha,2n-2)} + t_{(\beta,2n-2)}\right)^2$$

Tamanho amostral necessário:

```
> (ss.calc <- power.t.test(delta       = 15,
+                          sd          = 15,
+                          sig.level   = 0.05,
+                          power       = 0.8,
+                          type        = "two.sample",
+                          alternative = "one.sided"))

     Two-sample t test power calculation

              n = 13.09777
          delta = 15
             sd = 15
      sig.level = 0.05
          power = 0.8
    alternative = one.sided

NOTE: n is number in *each* group
```

Podemos realizar o teste t para comparar as médias de duas populações independentes de forma computacional:

```
> y <- read.table("steelrods.txt",
+                 header = TRUE)
> t.test(y$Length.error ~ y$Process,
+        alternative = "less",
+        mu          = 0,
+        var.equal   = TRUE,
+        conf.level  = 0.95)

 Two Sample t-test

data:  y$Length.error by y$Process
t = -14.312, df = 32, p-value = 9.244e-16
alternative hypothesis: true difference in means between group new
and group old is less than 0
95 percent confidence interval:
      -Inf -7.156884
sample estimates:
mean in group new mean in group old
         7.782353         15.900000
```

Precisamos verificar as premissas assumidas para o teste.

Neste caso particular são:

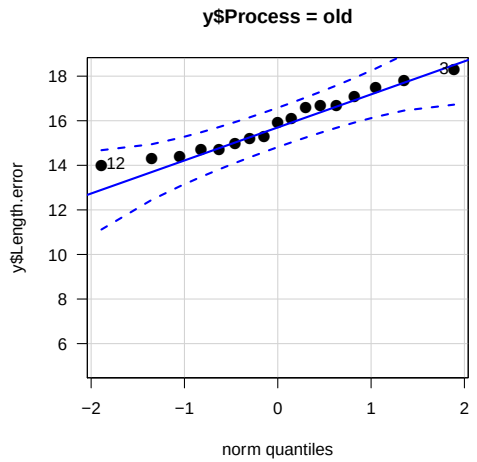
1. Normalidade;
2. Igualdade das variâncias;
3. Independência.

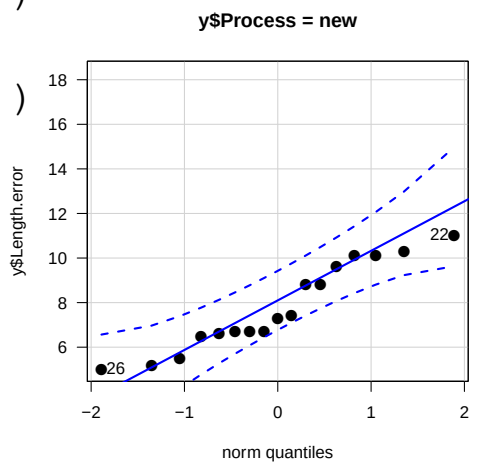
Verificação da Normalidade por Gráfico ou Teste de Hipóteses:

```
> qqPlot(y$Length.error, groups = y$Process,
         cex = 1.5, pch = 16, las = 1,
         layout = c(2, 1))
```

```
> shapiro.test(y$Length.error[y$Process == "new"])
# W = 0.92269, p-value = 0.164
> shapiro.test(y$Length.error[y$Process == "old"])
# W = 0.94971, p-value = 0.4519
```

Lembrete: o teste t é robusto a violações leves ou moderadas na normalidade dos resíduos.

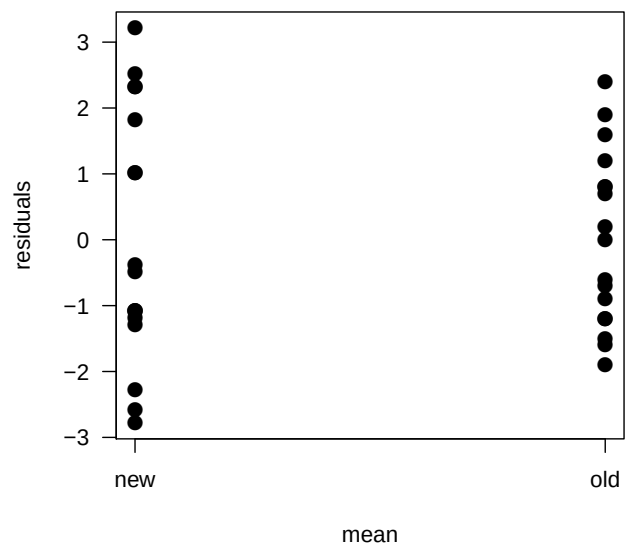




Verificação da Igualdade das Variâncias por Gráfico ou Teste de Hipóteses:

```
> fligner.test(Length.error ~ Process, data = y)
 Fligner-Killeen test of homogeneity of variances
data:  Length.error by Process
Fligner-Killeen:med chi-squared = 2.0785, df = 1, p-value = 0.1494
```

```
> resid <- tapply(X     = y$Length.error, INDEX = y$Process,
                  FUN   = function(x){x - mean(x)})
> stripchart(x = resid,
+            vertical = TRUE,
+            pch = 16,
+            cex = 1.5,
+            las = 1,
+            xlab = "mean",
+            ylab = "residuals")
```



Não existe um teste para verificar a premissa de independência, isso deve ser garantido na etapa de planejamento do experimento.

Pode-se, no máximo, avaliar se existe alguma autocorrelação serial nos resíduos utilizando o teste de Durbin-Watson, que é um teste totalmente dependente da ordenação das observações.

É um teste útil para detectar tendências relacionadas com a ordenação das observações, mas não vai além disso.

Comparação Simples: Variâncias Diferentes

Vamos considerar um caso mais geral em que as variâncias são desconhecidas e não podem ser assumidas iguais.

Para esse caso, geralmente se utiliza o teste **t de Welch**, que é uma modificação no teste t. A estatística de Welch é calculada como:

$$t_0^* = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Sob H_0, t_0^* é aproximadamente distribuída como uma $t_{(v)}$, com:

$$v = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1} + \frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$$

Como o teste t de Welch para duas amostras considera que as variâncias são diferentes, ele costuma ser uma das primeiras escolha, uma vez que ele reduz o número de premissas no teste.

Essa exclusão da premissa resulta em uma redução do poder do teste.

Calcular o tamanho amostral para o caso geral (variâncias diferentes) pode ser feito para o caso balanceado ($n_1 = n_2 = n$) ou para o caso desbalanceado ($n_1 \neq n_2$).

Para o caso desbalanceado, a alocação ótima de observações deve seguir os desvios padrões (caso boas estimativas dos σ_1 e σ_2 estejam disponíveis):

$$\frac{n_1}{n_2} = \frac{\sigma_1}{\sigma_2}$$

Comparações Pareadas

Ilustrando

Um jovem pesquisador desenvolveu um novo algoritmo de otimização (A) para uma família de problemas e quer comparar sua velocidade de convergência com um método que representa o estado-da-arte (B).

O pesquisador implementa os dois métodos e quer determinar se o método proposto possui um desempenho médio melhor para um família de problemas em específico, que pode ser representada por um conjunto de teste padrão (benchmark).

As medidas de desempenho são realizadas em condições homogêneas (mesmo computador, mesma condição operacional, etc.) e o tempo é medido de uma forma que ele não é sensível aos outros processos em execução pelo sistema.

O problema possui algumas questões importantes a serem consideradas:

- Qual a questão de interesse?
- Qual a população para qual esta questão é relevante?
- Quais são as observações independentes desta população?
- Qual o tamanho amostral necessário para o experimento?

A variabilidade devido às diferentes instâncias de teste do problema é uma importante fonte de variabilidade, e esta pode ser controlada.

Uma forma elegante de eliminar a influência desse fator indesejado é realizar o **pareamento** das observações para cada instância de teste:

- As observações são consideradas em pares (A, B) para cada instância de teste;
- O teste de hipóteses é realizado com as diferenças observadas.

Seja y_{Aj} e y_{Bj} os pares de observações do tempo médio para os métodos A e B para a instância j.

A *diferença pareada* das observações é: $d_j = y_{Aj} - y_{Bj}$.

Se modelarmos nossas observações como um processo aditivo:

$$y_{ij} = \underbrace{\mu + \tau_i}_{\mu_j} + \beta_j + \epsilon_{ij}$$

onde μ é a média global, τ_i é o efeito do i-ésimo algoritmo na média global, β_j é o efeito do j-ésimo problema e ϵ_{ij} é o resíduo do modelo.

Então podemos escrever:

$$d_j = \cancelto{0}{(\mu + \beta_j - \mu - \beta_j)} + \tau_A - \tau_B + \epsilon_{Aj} - \epsilon_{Bj} = \mu_D + \epsilon_j$$

Podemos definir nosso teste de hipóteses em relação ao μ_D:

$$\begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases}$$

O qual pode ser tratado como um teste de hipóteses para uma única amostra aleatória coletada da população das diferenças entre os tempos médios de convergência para a classe de problemas investigada.

Nossa estatística de teste é:

$$t_0 = \frac{\overline{D}}{S_D/\sqrt{N}}$$

que, sob H_0 é distribuída de acordo com uma distribuição t com $N-1$ graus de liberdade, e N é o número de instâncias de teste utilizadas no experimento.

Existem outras questões importantes a serem consideradas:

- Neste exemplo, o tamanho de efeito minimamente interessante δ^* deve ser expressado em termos do ganho no tempo médio para a classe do problema e não para instâncias individuais.
- O tamanho amostral mais importante é o número de instâncias e não o número de repetições para cada instância.
- O número de repetições para cada instância terá impacto na incerteza associada a cada observação (o tempo médio de convergência para cada instância), o qual será propagado para a variância residual;
- O pareamento remove os efeitos controláveis de fatores indesejáveis;
- O pareamento é indicado em casos onde existe uma grande correlação entre as observações, e.g., condições experimentais heterogêneas.

Voltando ao exemplo, vamos assumir os seguintes fatos sobre a comparação:

- O conjunto de testes é composto por sete instâncias ($N = 7$);
- O pesquisador está interessado em avaliar diferenças no tempo médio de convergência que sejam maiores do que dez segundos ($\delta^* = 10$) com um poder de pelo menos $(1 - \beta) = 0.8$ e com um nível de significância $\alpha = 0.05$;
- O pesquisador realiza $n = 30$ repetições da execução de cada algoritmo para cada instância e partindo de condições iniciais aleatórias[3].

[3] Não que esse número seja necessariamente bom, mas geralmente é uma alternativa fácil se você não quiser continuar justificando suas escolhas para revisores menos experientes em estatística

Carregar e pré-processar os dados.

```
> data<-read.table("soltimes.csv",
+                  header=T)
> # "Problem" and $Algorithm$ are categorical variable, not a continuo
> data$Problem<-as.factor(data$Problem)
> data$Algorithm<-as.factor(data$Algorithm)

> # Summarize within-problem observations by mean
> aggdata<-aggregate(Time~Problem:Algorithm,
+                    data=data,
+                    FUN=mean)
> summary(aggdata)
 Problem Algorithm      Time
 1:2     A:7       Min.   : 37.63
 2:2     B:7       1st Qu.:109.45
 3:2               Median :178.73
 4:2               Mean   :175.48
 5:2               3rd Qu.:245.25
 6:2               Max.   :296.79
 7:2
```

Realizar a análise.

```
> # Perform paired t-test
> t.test(Time~Algorithm,
+        paired=T,
+        data=aggdata)

 Paired t-test

data:  Time by Algorithm
t = -9.1585, df = 6, p-value = 9.54e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.85862 -12.64118
sample estimates:
mean of the differences
                -17.2499
```

Outra forma de analisar.

```
> difTimes<-with(aggdata, Time[1:7]-Time[8:14])
> t.test(difTimes)

 One Sample t-test

data:  difTimes
t = -9.1585, df = 6, p-value = 9.54e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -21.85862 -12.64118
sample estimates:
mean of x
 -17.2499
```

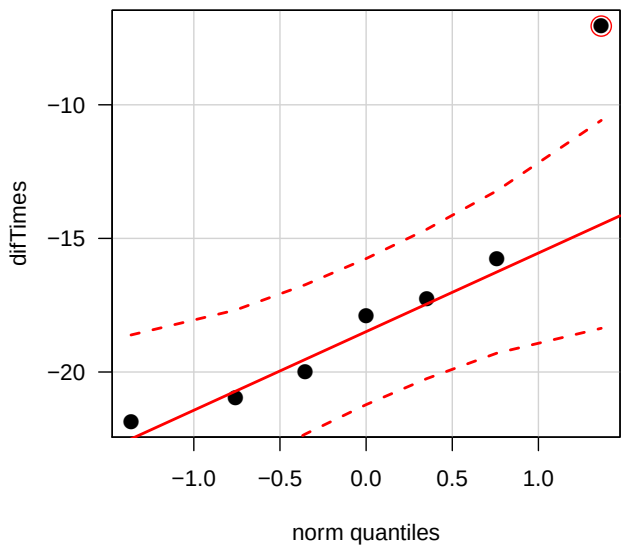
Verificar as premissas

```
> shapiro.test(difTimes)

 Shapiro-Wilk normality test

data:  difTimes
W = 0.83866, p-value = 0.09655

# Redo test without outlier
> indx<-which(difTimes==max(difTimes))
> t.test(difTimes[-indx])$p.value
[1] 6.179743e-06
> t.test(difTimes[-indx])$conf.int
[1] -21.41856 -16.48037
attr(,"conf.level")
[1] 0.95
```



Comparações Pareadas: Tamanho Amostral

Comparações pareadas podem requerer tamanhos amostrais menores para um poder do teste equivalente ao experimento não pareado quando a variabilidade é relativamente alta.

Mais especificamente, se a variabilidade entre os níveis é dada por σ_ϵ e a variabilidade entre as unidades é σ_μ, nos temos então, para um N grande o suficiente (e.g., $N \geq 10$),

$$\frac{N_{\text{não pareado}}}{N_{\text{pareado}}} \approx \sqrt{2\left[\left(\frac{\sigma_u}{\sigma_\epsilon}\right)^2 + 1\right]}$$

Como seriam os resultados se não considerássemos os efeitos das instâncias, i.e., testar como comparação simples?

```
> t.test(Time~Algorithm,data=aggdata)

 Welch Two Sample t-test

data:  Time by Algorithm
t = -0.36088, df = 11.993, p-value = 0.7245
alternative hypothesis: true difference in means between group A and
group B is not equal to 0
95 percent confidence interval:
 -121.40320   86.90341
sample estimates:
mean in group A mean in group B
       166.8527        184.1026
```

Comparações Múltiplas

E se quisermos comparar mais de duas amostras?

- Em muitos casos se faz necessário comparar três ou mais amostras (múltiplas amostras).
- Uma abordagem óbvia para este tipo de problema é uso de testes de comparação de duas amostras, realizados nos pares de amostras. Entretanto, a realização dessas múltiplas comparações demandam:
  - Um **procedimento de correção de significância**, por exemplo a correção de Bonferroni;
  - ou o uso de **testes específicos para comparações múltiplas**, por exemplo o teste de Tukey.
- Essa situação é muito comum após realizarmos uma teste de Análise de Variância (Teste ANOVA);

Esses aspectos serão explorados na aula de Teste ANOVA.

Referências

- D.C. Montgomery, G.C. Runger (2010), *Applied Statistics and Probability for Engineers*, John Wiley & Sons;
- D.C. Montgomery (2005), *Design and Analysis of Experiments*, John Wiley & Sons;
- F. Campelo, F. Takahashi, *Sample size estimation for power and accuracy in the experimental comparison of algorithms*. J. Heuristics, 2018 - `https://doi.org/10.1007/s10732-018-9396-7`;
- Paul Mathews' (2010), *Sample Size Calculations*, MMB;
- Felipe Campelo (2018), *Lecture Notes on Design and Analysis of Experiments*. Online: `http://git.io/v3Kh8` Version 2.12; Creative Commons BY-NC-SA 4.0.