Statistics: An Introduction using R – M. J. Crawley

Síntese por: Leandro M. Gimenez

# Capítulo 1 – Fundamentals (Fundamentos)

Para a realização de qualquer análise estatística é necessário conhecer quais os tipos de variáveis estão envolvidas. O primeiro passo é a definição sobre qual é a variável resposta, ou seja aquela do eixo y e qual a ou as variável (is) que explicam este comportamento, ou seja as do eixo x.

Se a variável que se deseja entender ou explicar for contínua, categórica, se é uma contagem ou proporção, então o método estatístico adequado irá variar.

**Tudo varia**

A estatística pode ser compreendida como uma ferramenta para isolar a variabilidade natural daquela causada por um fator cujo comportamento se deseja compreender. Quando se consegue esta separação diz-se que existe uma diferença estatística significativa. È importante ter em mente que nem sempre a falta de diferença estatística implica em um resultado ruim ou mesmo que não haja diferença.

**Significância**

A palavra significância pode ter diferentes sentidos, dependendo de quem a emprega. No caso da estatística, a significância se refere à incerteza de afirmar que um determinado fenômeno não ocorreu devido o acaso. Admite-se um nível de incerteza nesta definição, o qual pode variar, mas que geralmente é de 5% de repetibilidade.

**Boa e má hipóteses**

A escolha de hipóteses para comprovar se um determinado evento ocorre ao acaso ou não, deve ser cuidadosa. Boas hipóteses são aquelas que podem ser irrefutavelmente rejeitadas.

*Hipótese nula* – è aquela que afirma que nada está acontecendo, ou melhor, que não há efeito do fator estudado para a resposta observada.

Rejeita-se a hipótese nula quando os dados demonstram que a falta de efeito ocorre muito raramente.

**p-valores**

Valor que reflete a probabilidade que um determinado resultado ou que um resultado mais extremo do que aquele observado possa ter ocorrido por acaso, caso a hipótese nula fosse verdadeira.

Sintetizando, o p-valor é uma da credibilidade da hipótese nula. Por exemplo, p <0,001, implica que muito raramente o efeito observado se deve ao acaso. Por outro lado p = 0,23, indica que não há evidência de que a hipótese nula não seja verdadeira.

**Interpretação dos modelos estatísticos**

Há dois tipos de enganos na interpretação dos modelos estatísticos. O primeiro é o de rejeitar a hipótese nula quando ela é verdadeira ->erro tipo I**. Na realidade não há efeito mas o procedimento utilizado erroneamente infere que há.**

Quando se aceita a hipótese nula e ela é falsa tem-se o erro tipo II. **Na realidade existe efeito, porém o procedimento utilizado erroneamente infere que não há.**

**Modelagem estatística**

O objetivo da modelagem estatística é o de encontrar um modelo matemático que melhor se ajuste aos dados. Deve-se evitar o erro comum de ajustar os dados ao modelo, os dados são como são, deve se assumir que refletem algo que realmente ocorreu.

O melhor modelo é aquele que produz a menor variação inesperada nos dados originais e cujos parâmetros todos sejam estatisticamente significativos.

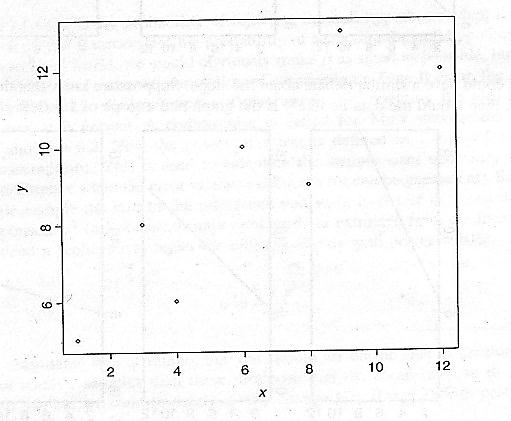
A concepção do modelo está a cargo da pessoa que deseja avaliar os fatores responsáveis por uma determinada resposta. Para tal é fundamental um bom conhecimento das relações existentes entre os fatores variados e entre eles e a resposta estudada. Desta maneira se pode produzir um modelo que seja simples e que consiga descrever adequadamente boa parte da variação encontrada nos dados. Podem existir vários modelos que se ajustam mais ou menos aos dados, parte da análise estatística está justamente na escolha daquele que é o mais adequado.

**Máxima verossimilhança (maximum likelihood)**

É o que se busca para definir qual o melhor modelo. As técnicas utilizadas para chegar ao modelo devem levar a estimadores que reduzam a variância e que sejam livres de ruído.

O exemplo na sequencia é para demonstrar o conceito de máxima verossimilhança:

Em um gráfico cartesiano, sejam os pontos definidos pelos pares de x e y de acordo como segue, x(1,3,4,6,8,9,12) e y(5,8,6,10,9,13,12). Então teremos o gráfico abaixo.



No pacote R, podemos criar os objetos x e y e depois plotar em um gráfico através dos comandos abaixo:

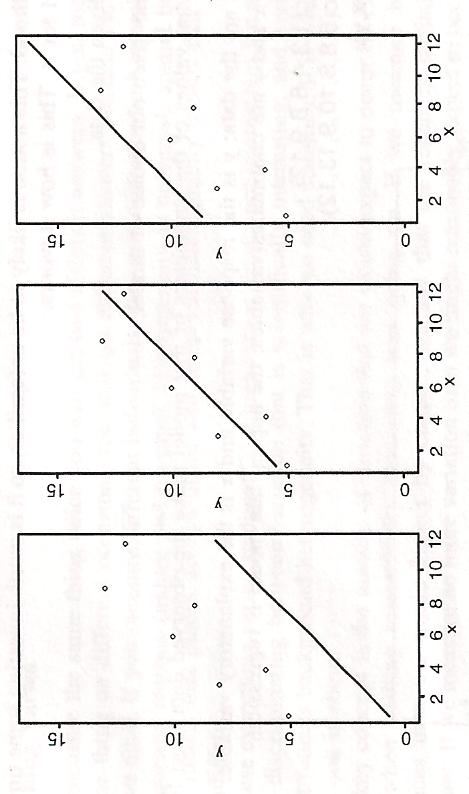
**x<-c(1,3,4,6,8,9,12)**; *Este comando cria um conjunto de valores e o denomina como objeto x.*

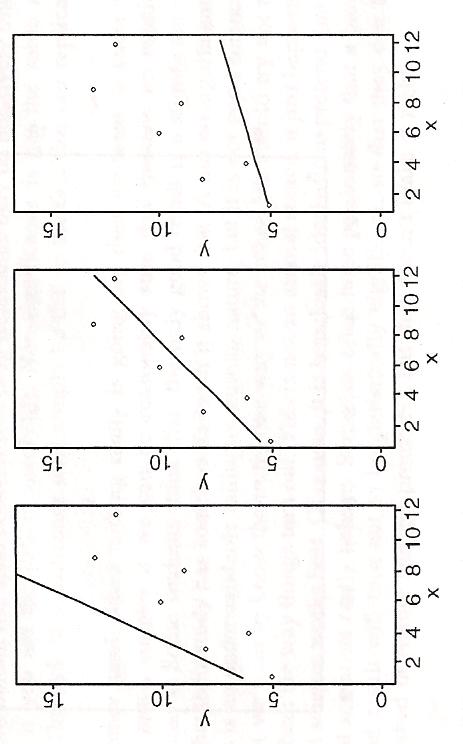
**y<-c(5,8,6,10,9,13,12)**; *Este comando cria um conjunto de valores e o denomina como objeto y.*

***plot(x,y)****; Este comando cria um gráfico cartesiano de x,y*

Como ambas variáveis resposta (y) e causadora (x), são contínuas, o melhor modelo a se empregar é o de regressão. Há uma série de modelos de regressão, porém ao se observar ao gráfico nota-se que o modelo mais simples, ou seja, pode ser utilizado: y = a + bx.

Considerando os parâmetros desse modelo, a= intercepto e b=declividade da reta, a escolha entre as várias possibilidades de ajuste da reta será aquela que permitir a redução da variância ao máximo, obtendo então a máxima verossimilhança. Nos gráficos abaixo são demonstrados os efeitos de se variar os parâmetros a, na primeira série e b, na segunda.





O melhor ajuste é naturalmente aquele dos gráficos do centro. Este ajuste geralmente é realizado automaticamente e simultaneamente para os dois parâmetros, pelos vários softwares estatísticos disponíveis.

**Planejamento do experimento**

Conceitos fundamentais: repetição e aleatorização

* Repetição: o objetivo é aumentar a confiabilidade
* Aleatorização: o objetivo é reduzir o ruído

Há outras questões também primordiais para que uma análise estatística possa ser feita de modo adequado:

**O princípio da parcimônia**

Dado um conjunto de explicações igualmente boas para um fenômeno, então a explicação correta é aquela mais simples.

* Modelos devem ter o menor número de parâmetros possível,
* Modelos lineares devem ser preferidos em relação aos não-lineares,
* Experimentos fundamentados em poucos pressupostos são preferidos em relação àqueles fundamentados em vários,
* Modelos devem ser reduzidos enquanto forem aceitáveis,
* Explicações simples devem ser preferidas em relação às complexas,
* A observação, a teoria e a experimentação devem andas juntas para a resolução efetiva de problemas científicos.

“Um modelo deve ser o mais simples possível, mas não simplista”.

Cuidados: não simplificar demais para “não jogar fora o bebê junto com a água do banho”.

**Repetição**

A repetição é fundamental para aumentar a confiabilidade das estimativas de um determinado parâmetro. Em um experimento há muitas coisas que podem mudar, ainda que estejamos exercendo controle. A repetição ajuda a quantificar a variabilidade dentro de um mesmo tratamento. Para ser considerada repetição há alguns preceitos:

* Deve ser independente,
* Não deve fazer parte de uma série histórica,
* Não devem estar todas muito próximas,
* Deve ter uma escala espacial adequada.

**Qual o número de repetições?**

Não existe uma regra única, mas é importante ter em mente que geralmente fazemos menos que o necessário. Uma regra bastante geral é: abaixo de 30 repetições a amostragem é pequena, acima de 30 a amostragem é grande.

A definição do número de amostras depende de qual é a variância do parâmetro observado, como muitas vezes não conhecemos esta variância, a sugestão é que se realize um estudo piloto.

Estudo piloto: permite obter um indicativo da variância entre unidades antes da aplicação dos tratamentos e também da magnitude aproximada das respostas devidas ao tratamento.

Uma observação muito importante é que às vezes é melhor reduzir a complexidade e se concentrar em menos parâmetros, fazendo assim mais repetições, do que chegar ao término com um resultado que seja ambíguo.

**Poder do teste**

Reflete a probabilidade de rejeitar a hipótese nula quando ela é falsa, ou seja, de definir com certeza que não há efeito.

Está associada ao erro do tipo II, ou seja, aceitar a hipótese nula quando ela é falsa, a representação é dada com base em probabilidade de β. Desejamos sempre minimizar ao máximo β, porém à medida que o fazemos aumentamos a probabilidade do erro do tipo I: negar a hipótese nula quando ela é verdadeira.

O poder de um teste é dado por 1- β. Geralmente o valor utilizado para β é 0,2 e dessa forma o poder do teste é geralmente 0,8.

Este valor do poder do teste é utilizado para calcular o tamanho da amostra necessária para detectar uma desejada diferença quando a variância do erro é conhecida ou pode ser estimada. Utiliza-se a equação abaixo:

Onde:

n = número de repetições

s2 = variância da resposta

= diferença que se deseja detectar

Por exemplo, se quiser detectar 5% de diferença em produtividade de soja, cuja média é 3500 kg.ha-1 , e o desvio padrão for de 150 kg.ha-1. Então, a diferença desejada será de 175 kg/ha-1, a variância será de 22500 (1502) e o número de repetições será de 6, como pode ser visto abaixo

*= 180000/30625=5,9*

Para o calculo no sistema R existem distintas funções, entre elas **power.t.test**, que pode ser utilizada no caso considerado bastando entrar os parâmetros conforme segue abaixo:

power.t.test(type=”one sample”, power=0.8, sd=150,delta=2)

Além desta existe outras funções como **power.anova.test** e **power.prop.test**

**Aleatorização**

É fundamental para obter resultados livres de tendências. Deve ser realizado da maneira mais impessoal possível, mediante sorteio.

Cada unidade amostral deve ter exatamente a mesma probabilidade de ser escolhida ao acaso.

Muitos pensam que estão aleatorizando, porém é comum cometer erros. Um profundo entendimento sobre o fenômeno em estudo é necessário para realizar uma boa aleatorização.

**O que é um inferência forte**

Inferência = chegar a uma opinião ou decidir que algo é verdadeiro, baseado em informações disponíveis.

* Quanto melhor a qualidade da informação, mais forte é a inferência, para tal há dois requisitos.
* A hipótese deve ser muito bem formulada
* Os testes escolhidos devem ser adequados
* Não fazer experimentos para “ver o que acontece”, esses experimentos são fracos pois podem haver muitas explicações plausíveis para o que está sendo observado.

Quanto mais bem estruturado o experimento e a análise mais forte será o poder de inferência sobre o que realmente aconteceu.

**O que é uma inferência fraca**

Em alguns estudos onde não se pode adotar os preceitos de repetição, aleatorização e não há como fazer bons controles, só é possível obter inferências fracas: as conclusões obtidas nestes estudos não podem ser considerada como resultado do teste de hipóteses.

**Por quanto tempo conduzir o experimento? Deve-se repetir?**

Dependendo do tipo de fenômeno estudado é necessário repetir o experimento para obter resultados confiáveis. Por exemplo na área médica ou na das ciências ambientais, ou biológicas. Há naturalmente pressões econômicas que tendem a reduzir o experimento e a sua duração. O que se observa em um experimento pontual pode não ser o mesmo que se observa se este mesmo experimento fosse realizado novamente, a dinâmica pode ser um fator importante par o que se está avaliando. Na agricultura há toda a questão do clima, por exemplo.

Para evitar problemas a recomendação é a de definir com antecedência a duração do experimento.

**O que é uma pseudo-repetição?**

Ocorre quando se detecta, na análise de dados, um maior número de graus de liberdade do que havia na realidade, havendo dois tipos básicos

* Pseudo-repetições temporais – repetição de mensurações em um mesmo indivíduo ao longo do tempo,
* Pseudo-repetições espaciais – repetição de mensurações em uma mesma vizinhança.

Quando se faz pesudo-repetições está se abrindo mão do pressuposto da independência dos erros. Se este for um problema detectado então sugere-se:

* Fazer a média e desenvolver os testes estatísticos nas médias
* Fazer análises para cada período
* Utilizar séries temporais adequadas, ou modelos com efeitos mistos.

**Condições iniciais para realizar o experimento**

Como podemos saber se algo mudou sem que se saiba como começou?

Embora esteja implícito no experimento, que as condições iniciais são as mesmas para todos os indivíduos, isto nem sempre ocorre.

Para evitar problemas, ou para poder detecta-los e talvez corrigir, sugere-se que se façam mensurações no período inicial do mesmo. Com estes dados realiza-se teste estatístico para assegurar que as condições não eram distintas.

Além disso a mensuração das condições iniciais pode auxiliar na análise final. Utilizando a análise de covariância pode-se melhorar a resolução da análise final.

**Dados ortogonais e não ortogonais – o problema dos experimentos observacionais**

Em experimentos planejados não há valores faltando, portanto todas as combinações de tratamentos estão igualmente representadas.

No caso de experimentos observacionais, não há controle do número de indivíduos para os quais se têm as informações. A ocorrência de tratamentos com falta de mensurações são comuns e diz-se que não há ortogonalidade.

(17/03/2011).